Cell Research (2014) 24:1181-1200.

www.nature.com/cr

ORIGINAL ARTICLE

Systematically labeling developmental stage-specific genes for the study of pancreatic β-cell differentiation from human embryonic stem cells

Haisong Liu[1, *], Huan Yang[1, *], Dicong Zhu[1, *], Xin Sui[1], Juan Li[2], Zhen Liang[1], Lei Xu[3], Zeyu Chen[4], Anzhi Yao[4], Long Zhang[5], Xi Zhang[5], Xing Yi[4], Meng Liu[1], Shiqing Xu[6], Wenjian Zhang[6], Hua Lin[7], Lan Xie[8], Jinning Lou[6], Yong Zhang[5], Jianzhong Xi[2], Hongkui Deng[1, 4, 9]

[1]Shenzhen Stem Cell Engineering Laboratory, Key Laboratory of Chemical Genomics, Peking University Shenzhen Graduate School, Shenzhen, Guangdong 518055, China; [2]Department of Biomedical Engineering, College of Engineering, Peking University, Beijing 100871, China; [3]Department of Hematopoietic Stem Cell Transplantation, 307 Hospital, Academy of Military Medicine Sciences, Beijing 100071, China; [4]The MOE Key Laboratory of Cell Proliferation and Differentiation, College of Life Sciences, Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China; [5]BGI-Shenzhen, Shenzhen, Guangdong 518083, China; [6]Institute of Clinical Medical Sciences, China-Japan Friendship Hospital, Beijing 100029, China; [7]Department of Gynecology and Obstetrics, China-Japan Friendship Hospital, Beijing 100029, China; [8]Department of Gynecology and Obstetrics, Beijing Renhe Hospital, Beijing 102600, China; [9]Peking University Stem Cell Research Center, Department of Cell Biology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China

The applications of human pluripotent stem cell (hPSC)-derived cells in regenerative medicine has encountered a long-standing challenge: how can we efficiently obtain mature cell types from hPSCs? Attempts to address this problem are hindered by the complexity of controlling cell fate commitment and the lack of sufficient developmental knowledge for guiding hPSC differentiation. Here, we developed a systematic strategy to study hPSC differentiation by labeling sequential developmental genes to encompass the major developmental stages, using the directed differentiation of pancreatic β cells from hPSCs as a model. We therefore generated a large panel of pancreas-specific mono- and dual-reporter cell lines. With this unique platform, we visualized the kinetics of the entire differentiation process in real time for the first time by monitoring the expression dynamics of the reporter genes, identified desired cell populations at each differentiation stage and demonstrated the ability to isolate these cell populations for further characterization. We further revealed the expression profiles of isolated NGN3-eGFP$^+$ cells by RNA sequencing and identified sushi domain-containing 2 (SUSD2) as a novel surface protein that enriches for pancreatic endocrine progenitors and early endocrine cells both in human embryonic stem cells (hESC)-derived pancreatic cells and in the developing human pancreas. Moreover, we captured a series of cell fate transition events in real time, identified multiple cell subpopulations and unveiled their distinct gene expression profiles, among heterogeneous progenitors for the first time using our dual reporter hESC lines. The exploration of this platform and our new findings will pave the way to obtain mature β cells *in vitro*.

***Keywords*:** gene labeling; pancreatic β cell; directed differentiation; embryonic stem cell; SUSD2
Cell Research (2014) **24**:1181-1200. doi:10.1038/cr.2014.118; published online 5 September 2014

Introduction

Human pluripotent stem cells (hPSCs), including human embryonic stem cells (hESCs) and induced pluripotent stem cells, can be efficiently induced into various cell types by mimicking key developmental events in

*These three authors contributed equally to this work.
Correspondence: Hongkui Deng[a], Jianzhong Xi[b]
[a]E-mail: hongkui_deng@pku.edu.cn
[b]E-mail: xi@coe.pku.edu.cn
Received 8 July 2014; revised 12 July 2014; accepted 14 July 2014; published online 5 September 2014

a stepwise manner and thereby hold great potential for studying developmental biology, disease modeling and cell-replacement therapy [1]. Despite tremendous efforts made in the past decade, no fully matured cell types have actually been obtained *in vitro* [2], which greatly hinders the further application of these hPSC-derived cells. To obtain mature cell types *in vitro* from hPSCs, two major problems still need to be solved.

First, because the best current protocols for hPSC differentiation are typically implemented in a stepwise fashion, deviations in the induction process at each stage (and especially the early stages) will accumulate and can be dramatically amplified stepwisely [3]. Previous studies on the directed differentiation of hPSCs into pancreatic β cells have shown that the yield of INSULIN (INS)-producing cells at later stages was sensitive to both the intensity and timing of TGF-β signaling at the first two stages [4]. Moreover, the cells generated at each step are in fact heterogeneous. Undesired cell-cell interactions and factors secreted by the unwanted cells will mask the desired signals and misdirect the differentiation process [5]. Thus, to optimize the induction conditions and thereby pave the way for accurate control of the entire stepwise differentiation process, the establishment of strategies that allow both monitoring the kinetics of the entire differentiation process and purifying desired cell populations for characterization and culture at each step could be of a great help [3].

Second, the establishment of current protocols for hPSC differentiation largely relies on developmental knowledge that was mostly extrapolated from studies of nonhuman experimental models, especially mice [2]. In the case of the pancreas, the different requirements for key regulators like *GATA6* during pancreatic development in the two species shows that differences in developmental mechanisms may exist [6, 7]. Moreover, even in mouse models, there are some knowledge gaps in the developmental process of pancreatic β cells, resulting in a lack of sufficient developmental clues for directing differentiation. It is therefore essentially necessary to develop appropriate tools that permit the isolation of *in vitro*-derived cell populations at each stage for careful evaluation and thereby shed light on the less understood aspects of development [8].

Collectively, to address the questions above, a systematic strategy to monitor, purify and analyze hPSC-derived intermediate cells at each step would be highly desirable. Prior studies showed that genetic labeling, especially for lineage-specific transcriptional factors, could be used in the study of development and differentiation in hPSCs [9-11]. However, partly because of the low efficiency of traditional gene targeting techniques, such studies mostly focused on narrow windows of the differentiation process and cannot provide a systematic solution for the study of the entire continuous process [12]. The breakthroughs made in gene targeting strategies in recent years make it possible to conduct the systematic gene labeling of certain cell lineages in hPSCs with high efficiency [13-16].

HPSC-derived pancreatic β cells hold great potential for cell-replacement therapy in type I diabetes, but no strategies have been reported that can efficiently generate mature β cells *in vitro* [17]. We first demonstrated that pancreatic β cells could be generated from mouse ESCs by recapturing sequential endogenous developmental signals in a stepwise fashion [18]. Thereafter, significant progress has been made in generating pancreatic cell populations from hPSCs, and a variety of stepwise protocols have been established [19-26]. The hPSC-derived pancreatic progenitors can differentiate into mature β cells after transplantation in mice, but these *in vitro*-derived β cells still function poorly. This defect mainly results from the lack of a systematic approach for the study of multistep differentiation processes and the lack of the developmental knowledge, which is necessary to direct or facilitate *in vitro* differentiation.

In this study, we systematically labeled sequential genes in pancreatic development covering the major developmental stages of pancreatic β cells from hESCs for the first time, with the aid of transcription activator-like effector nuclease (TALEN). We therefore generated a large panel of reporter cell lines; several of these are dual reporter cell lines that were constructed on the basis of the *NGN3-eGFP* cell line due to the key role of *NGN3* in the commitment of pancreatic endocrine progenitors [27]. With this unique platform, we successfully visualized the kinetics of the entire differentiation process in real time and made it possible to recognize and therefore isolate intermediate cell populations at each differentiation stage for further characterization. Using the dual-reporter hESC lines, we captured the process of cell fate transition and unveiled distinct gene expression profiles of multiple cell subpopulations. We further profiled hPSC-derived NGN3-eGFP$^+$ cells by RNA-sequencing and identified sushi domain-containing 2 (SUSD2) as a novel surface protein that enriches for pancreatic endocrine progenitors and early endocrine cells both in hESC-derived pancreatic cells and in the developing human pancreas. An exploration of this platform and these new findings will pave the way for obtaining mature β cells *in vitro* from hPSCs.

Results

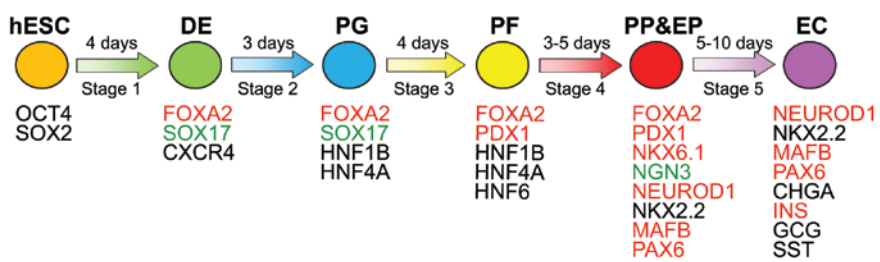
Generating a panel of pancreatic-specific hESC reporter lines by labeling sequential developmental genes of pan-

creatic β cells

To develop a systematic platform for the study of hPSC differentiation into β cells, we attempted to label a large panel of sequential developmental genes to cover the major developmental stages of pancreatic β cells. So far, the major developmental process of β cells has been divided into several sequential stages, which are represented by the expression of key genes: (1) definitive endoderm (e.g., *FOXA2* and *SOX17*), (2) primitive gut tube (e.g., *HNF1B*, *HNF4A* and *FOXA2*), (3) posterior foregut (e.g., *PDX1* and *HNF6*), (4) pancreatic progenitors (e.g., *PDX1*, *NKX6.1* and *PTF1A*), (5) endocrine progenitors (e.g., *NGN3*, *NKX2.2*, *NEUROD1*, *MAFB*, *PAX4* and *PAX6*) and (6) β cells (e.g., *INS*, *PDX1* and *MAFA*) [28]. Currently, the stepwise protocols for the generation of β cells from hPSCs have been established by recapturing developmental stages (Figure 1A) [19-23, 25, 26]. Therefore, we chose *FOXA2*, *SOX17*, *PDX1*, *NKX6.1*, *NGN3*, *NEUROD1*, *MAFB*, *PAX6* and *INS* as our target genes for labeling (Figure 1A).

The *NEUROG3* (*NGN3*) gene was selected as our first target because of its important role in endocrine commitment. *In vitro*-derived pancreatic progenitors can mature into functional β cells after transplantation in mouse [25, 29], but they cannot generate functional β cells *in vitro* [19, 30], implying that the subsequent endocrine progenitor stages should be prioritized. Initially, we employed a bacterial artificial chromosome (BAC) recombineering-based strategy for gene targeting. The stop codon of a BAC harboring the *NGN3* locus was replaced with a *2A-eGFP-loxp-CAG-neo-loxp* cassette by recombination (*CAG*, *chicken β-actin promoter with CMV enhancer*; *2A*, *foot-and-mouth disease virus 2A segment*). The *2A* sequence can mediate the separation of the adjacent genes via translational skipping, which allows the target gene to be traced by fluorescence while preserving gene function [31]. H1 cells were electroporated with the linearized BAC-based targeting vector and treated with G418 and ganciclovir for 2 weeks. 192 Drug-resistant colonies were screened by polymerase chain reaction (PCR), yielding 20 positive colonies (Supplementary information, Figure S1B). The derived cell lines were further confirmed by Southern blotting (Supplementary information, Figure S1C). Subsequently, the antibiotic resistance cassette of one cell line was removed by CRE recombinase (Supplementary information, Figure S1D).

A



B

Reporter cell line name	Gene targeting details
NGN3-eGFP	*NGN3-eGFP*
SOX17-eGFP	*SOX17-eGFP*
FOXA2-DR	*FOXA2-Tdtm*/*NGN3-eGFP*
PDX1-DR	*PDX1-Tdtm*/*NGN3-eGFP*
NKX6.1-DR	*NKX6.1-Tdtm*/*NGN3-eGFP*
NEUROD1-DR	*NEUROD1-Tdtm*/*NGN3-eGFP*
MAFB-DR	*MAFB-Tdtm*/*NGN3-eGFP*
PAX6-DR	*PAX6-Tdtm*/*NGN3-eGFP*
INS-DR	*INS-Tdtm*/*NGN3-eGFP*

Figure 1 The generation of stage-specific hESC reporter lines by labeling sequential developmental genes of β cells. **(A)** Schematic of a five-stage pancreatic endocrine differentiation protocol from hESCs represented by important marker genes. The genes labeled with *eGFP* are in green and those labeled with *Tdtm* are in red; unlabeled genes are in black. **(B)** The reporter cell lines constructed in this study. Dual-reporter cell lines were named with the second targeted gene locus followed by "DR"; for example, the "FOXA2-DR" cell line stands for the "*FOXA2-Tdtm*/*NGN3-eGFP* dual-reporter" cell line. Abbreviations: DE (definitive endoderm); PG (primitive gut tube); PF (posterior foregut); PP (pancreatic progenitor); EP (endocrine progenitor); EC (endocrine cells); *INS* (*INSULIN*); *GCG* (*GLUCAGON*); *Tdtm* (*TdTomato*).

Although we tried to tag other pancreatic genes using a similar approach, we failed to obtain the desired reporter cell lines.

TALENs were then introduced to our gene targeting strategy to improve the homologous recombination efficiency (Supplementary information, Figure S1A). The DNA sequences of TALENs that were specific for the selected gene loci were synthesized using a high-throughput integrated chip method [32]. The ability of these nucleases to carry out efficient genomic editing was confirmed by a T7 endonuclease assay and sequencing (Supplementary information, Figure S1A, upper right). To create gene targeting vectors, the stop codon of each gene was replaced with a *2A-Tdtomato (Tdtm)-loxp-CAG-neo-loxp* cassette in the corresponding BAC, and the whole cassette (with the homology arms) was retrieved by a small vector. A herpes simplex virus thymidine kinase (*HSV-TK*) cassette was also included in the targeting vector to select against random integration (Supplementary information, Figure S1A, upper left). Target hESCs were nucleofected with linearized targeting vector and the corresponding TALENs, and the resulting colonies were characterized as described above. In this way, we successfully targeted *Tdtm* into seven loci: *FOXA2*, *PDX1*, *NKX6.1*, *MAFB*, *PAX6*, *NEUROD1* and *INS* (Figure 1A). Because *NGN3-eGFP* cells were used as target cells, we generated seven dual reporter cell lines (Figure 1B). We also constructed a *SOX17-eGFP* reporter cell line based on wild-type H1 using a similar strategy. The targeting efficiency varied between the different gene loci, ranging from 3% to 63% (Supplementary information, Figure S1B). The correct integration of the reporter genes was confirmed by Southern blotting (Supplementary information, Figure S1C), and normal karyotypes of the derived cells lines were demonstrated by G-band staining (Supplementary information, Figure S1E).

In summary, by combining high-throughput TALEN synthesis techniques and BAC recombineering, we successfully tagged a panel of sequential developmental genes covering the principal stages of pancreatic β-cell development, and thereby obtained a pool of mono- and dual reporter cell lines.

Expression dynamics and fidelity of individual gene reporters

To evaluate the developmental regulation of our reporter cell lines, we converted them into pancreatic endocrine cells using a stepwise protocol (Figure 1A and Supplementary information, Data S1) that was modified from previous reports [21, 23, 25, 29]. Before induction, none of the reporter cell lines expressed Tdtm or eGFP (data not shown). After induction, the fidelity of the reporters for each gene locus was independently evaluated from stage 1 to stage 4 or 5. To facilitate the description of these effects, the reporter protein expressed from a certain targeted gene locus was named "targeted gene name-reporter protein name". For instance, "FOXA2-Tdtm" stands for the Tdtm protein expressed from the *FOXA2* locus.

The expression dynamics of each reporter gene was monitored and evaluated by flow cytometric analysis. *FOXA2* expression was first detected in the anterior primitive streak in intraembryonic tissues, followed by definitive endoderm and endoderm-derived organs such as the pancreas [28, 33]. As expected, Tdtm could be detected after FOXA2-DR cells were induced for 24 h (Supplementary information, Figure S2B). The percentage of $Tdtm^+$ cells increased and was maintained at peak levels during stages 3 and 4 (Figure 2A and Supplementary information, Figure S2A).

SOX17 was also required for endoderm specification, but its expression was initiated later than that of *FOXA2* in the early embryonic development of mouse [34]. Accordingly, the expression of eGFP was first observed after *SOX17-eGFP* cells were induced for 2-3 days (data not shown). Similar to a previous report [10], the number of $eGFP^+$ cells peaked at stages 1-2 and declined thereafter (Figure 2A and Supplementary information, Figure S2A).

PDX1 was first observed in the posterior foregut domain of mouse embryos. PDX1 is a key marker both for pancreatic progenitors and β cells [28, 33]. Accordingly, $Tdtm^+$ cells were first detected in PDX1-DR cell cultures at stage 3 (Figure 2A and Supplementary information, Figure S2A). The intensity of both Tdtm and PDX1 increased significantly at stage 4 (Figure 2A and data not shown).

NKX6.1 is also a key marker both for pancreatic progenitors and β cells [28, 33]. Consistent with previous reports [29], the expression of Tdtm was first detected in NKX6.1-DR cell cultures at stage 4, days 2-3 (Figure 2A and Supplementary information, Figure S2A).

NGN3 is the master gene for endocrine progenitor specification [28, 33]. The expression of eGFP appeared at the end of stage 3 in the *NGN3-eGFP* cell cultures (Figure 2A), peaked at stage 4 (15% - 40%; Figure 2A and Supplementary information, Figure S2A) and declined thereafter (data not shown).

NEUROD1 is a downstream target of *NGN3*. Its expression was restricted to endocrine cells, especially β cells in later stage pancreas [28, 33]. As expected, we first detected NEUROD1-Tdtm expression at the beginning of stage 4 (Figure 2A), just after NGN3-eGFP expression. The number of $Tdtm^+$ cells increased during

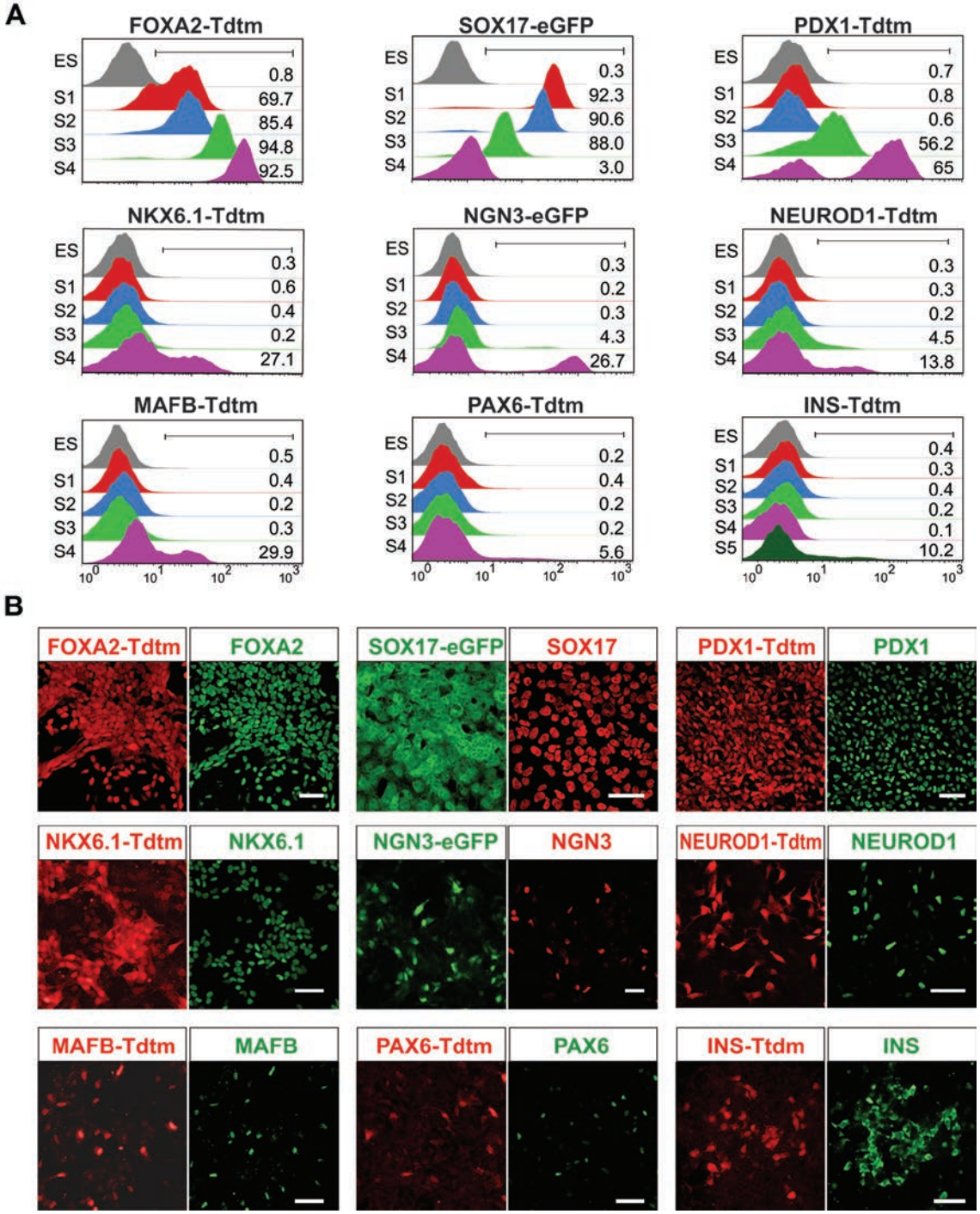


Figure 2 Expression dynamics and fidelity of individual reporter genes. **(A)** Flow cytometric analysis of the expression dynamics of individual reporter genes during hESC differentiation. The cell cultures were analyzed by flow cytometry at the end of each stage. The expression of FOXA2-Tdtm, SOX17-eGFP, PDX1-Tdtm, NKX6.1-Tdtm, NGN3-eGFP, NEUROD1-Tdtm, MAFB-Tdtm and PAX6-Tdtm was measured for stages (S) 1-4; whereas *INS-Tdtm* for S1-S5. The numbers in the figure indicate percentage of reporter-expressing cells at the end of specific stages. **(B)** Confocal imaging after co-staining of reporter gene expression with endogenous gene expression. The expressions of FOXA2-Tdtm and SOX17-eGFP were analyzed at S1, day (D) 4; PDX1-Tdtm at S3, D4; NGN3-eGFP was analyzed at S4, D1.5; NKX6.1-Tdtm and NEUROD1-Tdtm were analyzed at S4, D3; MAFB-Tdtm was analyzed at S4, D4; PAX6-Tdtm was analyzed at S4, D5, and INS-Tdtm at S5, D5. The endogenous protein staining is pseudocolored, and the *eGFP* expression of the dual-reporter cell lines is not shown. The superimposed images are shown in Supplementary information, Figure S2D. Scale bar, 50 μm. Abbreviations: *INS* (*INSULIN*); *Tdtm* (*TdTomato*).

stage 4 (Figure 2A and Supplementary information, Figure S2A).

MAFB and *PAX6* are required for the correct development of endocrine precursor cells in mouse [28, 33]. In our study, MAFB-Tdtm expression was initiated at stage 4 days 3-4 (Figure 2A and Supplementary information, Figure S2A); while the expression of PAX6-Tdtm was detected relatively later (Figure 2A; Supplementary information, Figure S2A). However, the expression of PAX6-Tdtm was relatively weak and the PAX6-Tdtm$^+$ cells could not be clearly recognized as a separate cell fraction by flow cytometry unless magnified by an RFP antibody (Supplementary information, Figure S2D).

INS is the key hormone for regulating the blood glucose levels and is produced by pancreatic β cells. In accordance with previous reports [19, 23, 25], INS-Tdtm$^+$ cells were observed at stage 5 (Figure 2A and Supplementary information, Figure S2A). Dispersed INS-Tdtm$^+$ cells aggregated into clusters after prolonged cultivation (Supplementary information, Figure S2C).

The fidelity of the reporter genes was evaluated by immunostaining (Figure 2B and Supplementary information, Figure S2E). As shown in Supplementary information, Table S1, the expression of most of the reporter genes was highly co-localized with that of the targeted genes in the time window evaluated. The agreement between the eGFP protein and NGN3 protein varied according to the day (Supplementary information, Figure S2F). Approximately 80% of the eGFP$^+$ cells expressed NGN3 and 82% of the NGN3$^+$ cells expressed eGFP$^+$ at stage 4, day 1. The proportion of NGN3$^+$ cells among the eGFP$^+$ cells gradually decreased thereafter, which is consistent with previous observations in *NGN3-eGFP* mice that eGFP has a much longer half-life (> 24 h) than NGN3. Similarly, some PDX1-Tdtm$^+$/PDX1^{low} cells were also observed at stage 4 (as discussed later).

Taken together, these results indicate that the expression of most of these reporter genes mimics that of the endogenous genes with high fidelity. More importantly, we visualized the kinetics of the entire differentiation process in real time for the first time by tracking the expression dynamics of the reporter genes.

Individual pancreatic reporters marked and thereby permitted the isolation of specific cell populations during differentiation of hESCs into β cells

Because the labeled genes cover the major stages of pancreatic β-cell development, each reporter for one gene locus could mark specific cell populations resembling the corresponding cell populations at specific developmental stages *in vivo*. To validate this possibility, all the reporter cell lines were induced for evaluation as described above.

FOXA2-Tdtm$^+$ cells expressed SOX17, and SOX17-eGFP$^+$ cells expressed FOXA2 at the end of stage 1 (Figure 3A). Both FOXA2-Tdtm$^+$ cells and SOX17-eGFP$^+$ cells lacked OCT4 expression (Supplementary information, Figure S3A). HNF1B and HNF4A, markers of the primitive gut tube, were highly co-expressed in both FOXA2-Tdtm$^+$ cells and SOX17-eGFP$^+$ cells at stage 2 but not stage 1 (Figure 3B, Supplementary information, Figure S3A and data not shown). Taken together, these results indicate that both FOXA2-Tdtm and SOX17-eGFP can mark the definitive endoderm cells at the end of stage 1 and the primitive gut tube cells at the end of stage 2.

Tdtm$^+$ cells were detected in PDX1-DR cell cultures at the end of stage 3. These Tdtm$^+$ cells also expressed FOXA2, HNF1B and HNF6, indicating a posterior foregut endoderm identity when combined with PDX1 (Figure 3C and Supplementary information, Figure S3A). Thus, the PDX1-Tdtm can be used to label posterior foregut endoderm cells at the end of stage 3.

A variety of pancreas-associated gene expression was initiated at stage 4, including *NKX6.1*, *NGN3*, *NEUROD1*, and *MAFB*, whose expression could be indicated by Tdtm or eGFP (Figure 1A).

At stage 4 day 3, immunostaining of NKX6.1-DR cell cultures indicated that most of the NKX6.1-Tdtm$^+$ cells (> 95%) expressed the pancreatic progenitor makers PDX1, SOX9, HNF1B and HNF6, but not the endocrine cell markers CHGA and hormones (Figure 3D and Supplementary information, Figure S3A). A few Tdtm$^+$ cells also expressed NKX2.2 (Supplementary information, Figure S3A), a pan-endocrine lineage marker, similar to that observed *in vivo* [35]. These results suggest that most of the NKX6.1-Tdtm$^+$ cells can be regarded as pancreatic progenitor cells at stage 4.

Approximately 80% of the NGN3-eGFP$^+$ cells expressed NGN3 protein at stage 4 day 1, and most of them also expressed NKX2.2 and NEUROD1 (Figure 3D; Supplementary information, Figures S2F and S3B), demonstrating that most NGN3-eGFP$^+$ cells consisted of endocrine progenitor-like cells at that time. The proportion of NGN3$^+$ cells among the NGN3-eGFP$^+$ cells gradually declined thereafter (Supplementary information, Figure S2F). However, nearly 100% of the NGN3-eGFP$^+$ cells expressed NKX2.2 at stage 4 day 4, and most of these cells also expressed CHGA and exhibited weak PDX1 staining (Supplementary information, Figure S3B and S3C), implying that these NGN3-eGFP$^+$/NGN3$^-$ cells are newborn endocrine cells. Thus, these results indicate that the NGN3-eGFP$^+$ cells represent a mixture of endocrine progenitor cells and their derivatives in hESC

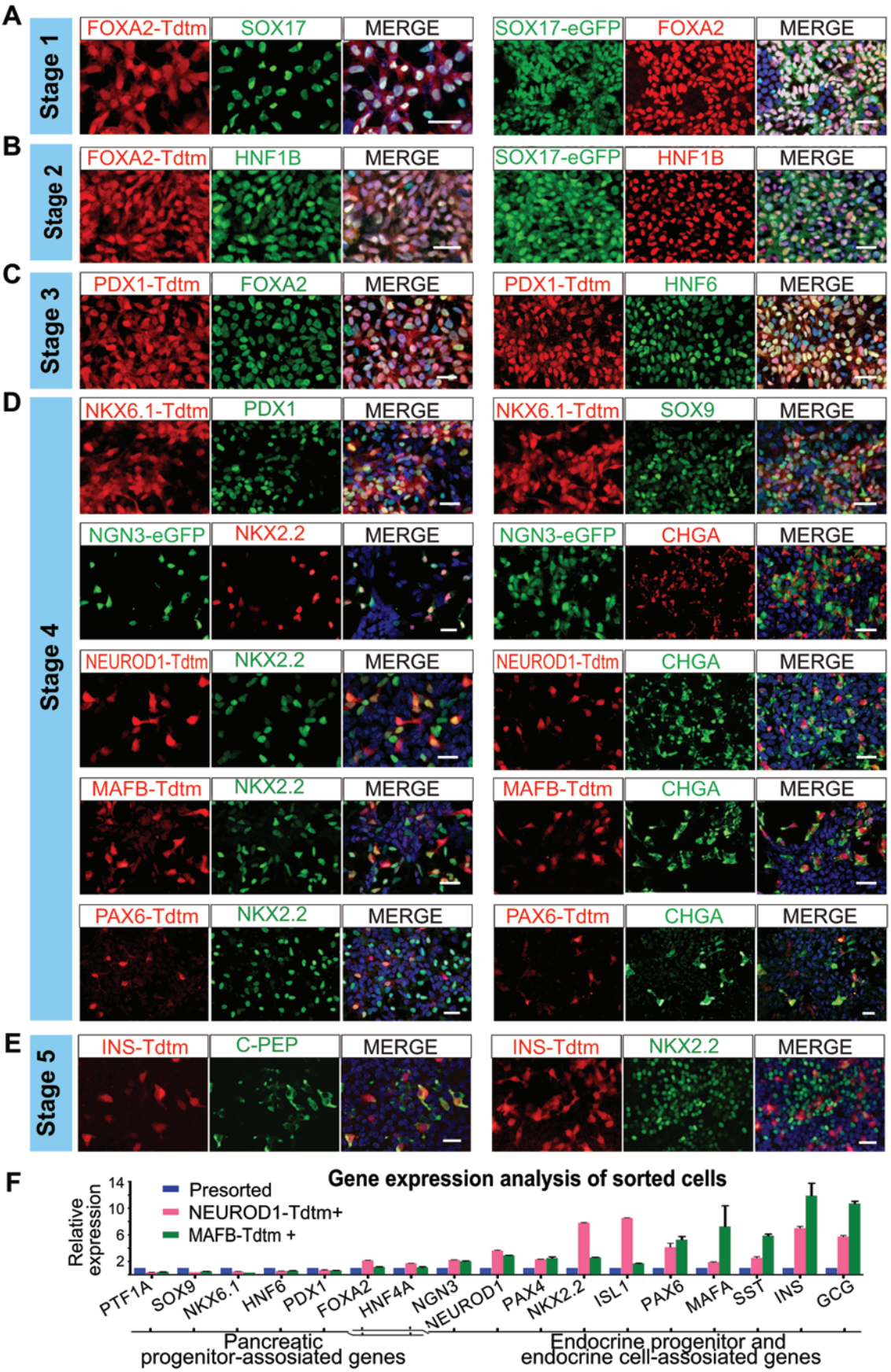

A
Stage 1
FOXA2-Tdtm
SOX17
MERGE
SOX17-eGFP
FOXA2
MERGE
B
Stage 2
FOXA2-Tdtm
HNF1B
MERGE
SOX17-eGFP
HNF1B
MERGE
C
Stage 3
PDX1-Tdtm
FOXA2
MERGE
PDX1-Tdtm
HNF6
MERGE
D
Stage 4
NKX6.1-Tdtm
PDX1
MERGE
NKX6.1-Tdtm
SOX9
MERGE
NGN3-eGFP
NKX2.2
MERGE
NGN3-eGFP
CHGA
MERGE
NEUROD1-Tdtm
NKX2.2
MERGE
NEUROD1-Tdtm
CHGA
MERGE
MAFB-Tdtm
NKX2.2
MERGE
MAFB-Tdtm
CHGA
MERGE
PAX6-Tdtm
NKX2.2
MERGE
PAX6-Tdtm
CHGA
MERGE
E
Stage 5
INS-Tdtm
C-PEP
MERGE
INS-Tdtm
NKX2.2
MERGE
F
Gene expression analysis of sorted cells
Relative expression
14
10
6
2
Presorted
NEUROD1-Tdtm+
MAFB-Tdtm +
PTF1A
SOX9
NKX6.1
HNF6
PDX1
FOXA2
HNF4A
NGN3
NEUROD1
PAX4
NKX2.2
ISL1
PAX6
MAFA
SST
INS
GCG
Pancreatic progenitor-assosiated genes
Endocrine progenitor and endocrine cell-assosiated genes

Figure 3 Different gene reporters mark distinct cell populations at the different differentiation stages of hESCs. **(A)** At the end of stage (S)1, immunostaining analysis demonstrated that the FOXA2-Tdtm$^+$ cells expressed the definitive endoderm marker SOX17, and the SOX17-eGFP$^+$ cells expressed the definitive endoderm marker FOXA2. **(B)** At the end of S2, the FOXA2-Tdtm$^+$ cells and SOX17-eGFP$^+$ cells expressed the primitive gut-tube cell marker HNF1B. **(C)** At the end of S3, the PDX1-Tdtm$^+$ cells expressed the posterior foregut endoderm markers HNF6 and FOXA2. **(D)** At the end of S4, the NKX6.1-Tdtm$^+$ cells expressed the pancreatic progenitor markers PDX1 and SOX9; the NGN3-eGFP$^+$ cells express the endocrine-associated genes NKX2.2 and CHGA. The NEUROD1-Tdtm$^+$ cells, MAFB-Tdtm$^+$ cells and PAX6-Tdtm$^+$ cells represented different populations of endocrine precursors; most of them expressed NKX2.2, but accounted for different percentages of NKX2.2$^+$ cells. They also expressed CHGA. **(E)** At the end of S5, the INS-Tdtm$^+$ cells marked the INS-producing cells; these cells expressed the β-cell markers C-PEPTIDE and NKX2.2. Nuclear staining with DAPI (blue) is shown in the merged images. **(F)** Gene expression analysis of sorted cells at the end of stage 4 showed that MAFB-Tdtm$^+$ cells expressed higher level of late-stage endocrine-associated genes (e.g., MAFA, PAX6, INS, *etc.*) than NEUROD1-Tdtm$^+$ cells. Imaging was performed using confocal microscopy. Scale bar, 25 μm. Abbreviations: INS (INSULIN); CHGA (CHROMOGRANIN A); Tdtm (TdTomato); C-PEP (C-PEPTIDE).

cultures at stage 4.

As mentioned above, NEUROD1, MAFB and PAX6 are three transcriptional factors that function in different time windows and contexts in endocrine precursors [28]. Similar to the *in vivo* expression sequence, NEUROD1-Tdtm was detected first, followed by MAFB-Tdtm and PAX6-Tdtm. At stage 4 day 5, immunostaining showed that NEUROD1-Tdtm$^+$, MAFB-Tdtm$^+$ and PAX6-Tdtm$^+$ cells accounted for ~55%, 78% and 23% of the NKX2.2-expressing cells, respectively (Figure 3D). These cells also expressed CHGA and lacked high PDX1 expression (Supplementary information, Figure S3A). Thus, NEUROD1-Tdtm, MAFB-Tdtm and PAX6-Tdtm can identify different endocrine-associated cell populations in hESC cultures.

Substantial INS-Tdtm$^+$ cells were detected at stage 5. These cells were positive for both INS and C-PEPTIDE (Figure 3E), implying that the INS proteins were synthesized *de novo*. These newly emerged INS-Tdtm$^+$ cells expressed NKX2.2 but little or no PDX1 (Figure 3E and Supplementary information, Figure S3A and data not shown), suggesting that they were immature. Thus, INS-Tdtm can mark INS-producing cells in hESC cultures.

The above results indicate that these reporters mark different cell populations representing distinct developmental stages, potentially allowing for the isolation of specific cell populations for characterization of molecular properties and differentiation potency. To validate this possibility, we took the *MAFB-Tdtm* and *NEUROD1-Tdtm* reporters as examples and performed quantitative PCR (qPCR) to assess gene expression after fluorescence-activated cell sorting (FACS)-based purification. Purified NEUROD1-Tdtm$^+$ and MAFB-Tdtm$^+$ cells both showed enriched endocrine-associated gene expression and lower levels of pancreatic progenitor-associated genes. Nevertheless, MAFB-Tdtm$^+$ cells expressed higher levels of later stage endocrine-associated genes, such as MAFA and hormone genes, suggesting that these different reporters might be used to purify endocrine precursors of different types and/or different developmental stages (Figure 3F). To characterize the differentiation potency of isolated cell populations, we took the *NGN3-eGFP* reporter as an example and performed further culture of FACS-purified NGN3-eGFP$^+$ cells and NGN3-eGFP$^-$ cells at stage 4. INS$^+$ cells were enriched in NGN3-eGFP$^+$ cell-derived cultures but not NGN3-eGFP$^-$ cell-derived cultures, demonstrating that the *in vitro*-derived INS-producing cells arose from NGN3-expressing cells (Supplementary information, Figure S3D). Thus, the reporter cell lines can be used to purify distinct cell populations at different stages for further characterization.

Collectively, during hESC differentiation into pancreatic β cells, these reporters label cell populations that resemble the corresponding cell populations at different developmental stages *in vivo*, thus allowing the isolation of these cell populations for further characterization.

Dual reporter cell lines are versatile tools for capturing cell fate transition and identifying cell subpopulations

Because our dual reporter systems were constructed by labeling paired genes expressed at neighboring developmental stages, we investigated whether they could be used as a tracing tool to capture cell fate transitions. Dual-reporter cells were induced as described above and examined by flow cytometry.

As shown in differentiating FOXA2-DR cell cultures (Figure 4A and 4B), the expression of Tdtm preceded that of eGFP. EGFP$^+$ cells first emerged as eGFP$^+$/Tdtm$^+$ cells at stage 4, implying that NGN3-eGFP$^+$ cells were expressing FOXA2. This finding is consistent with recent observations that FOXA2 is co-expressed with NGN3 in the developing mouse pancreas [36]. These data also demonstrate that FOXA2-Tdtm$^+$ cells have at least two

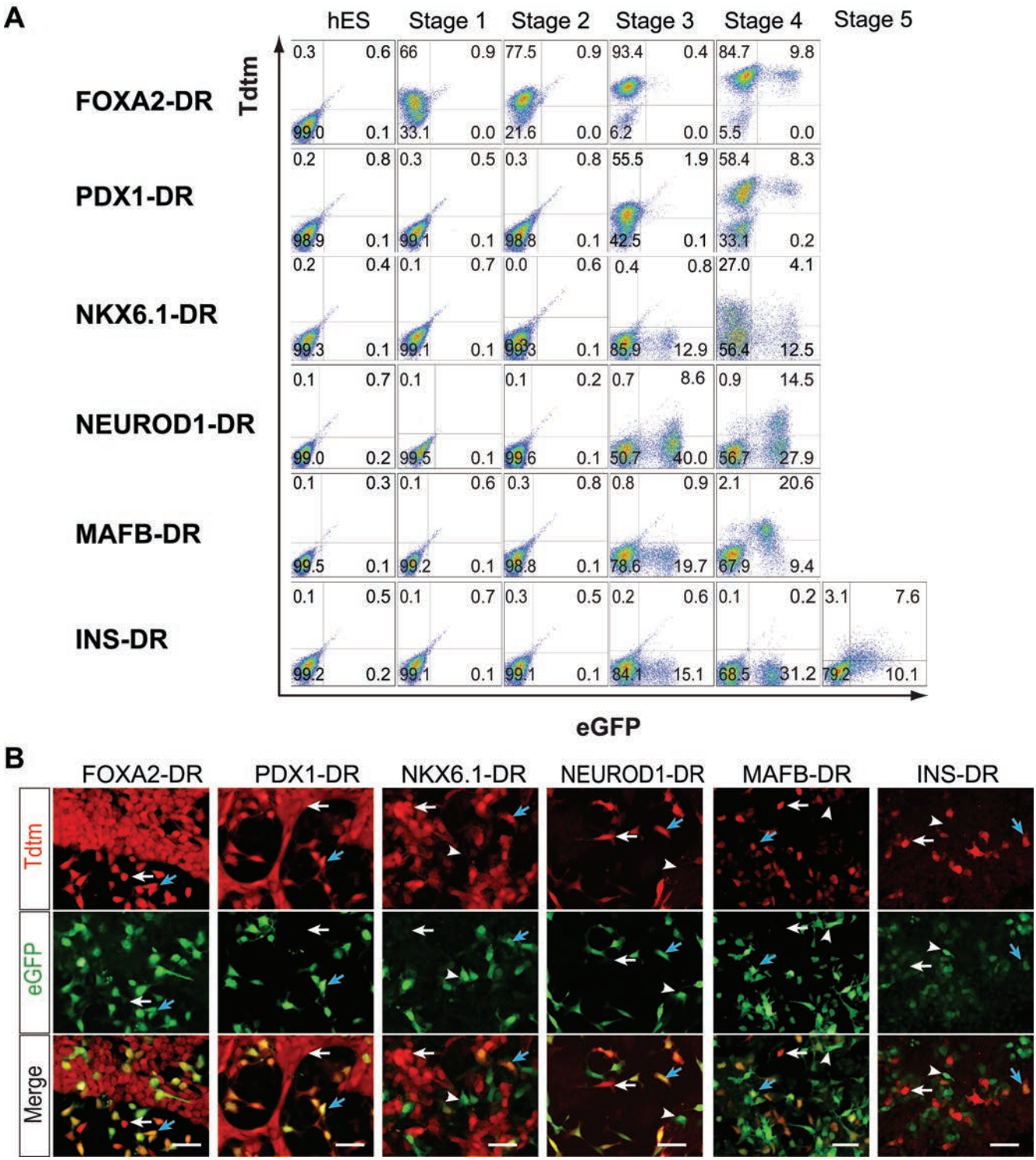


Figure 4 Dual-reporter cell lines identify cell subpopulations. **(A)** The expression dynamics of two reporter genes (*eGFP* and *Tdtm*) in the dual-reporter cell lines, showing the existence of cell subpopulations and gene expression transitions between pairs of targeted genes. **(B)** Distinct cell populations express different combinations of fluorescent proteins in the dual-reporter cell line-derived cultures. The left five cell cultures were from the end of stage 4, and last culture is from stage 5. White arrows, arrow heads and blue arrows indicate $eGFP^-/Tdtm^+$, $eGFP^+/Tdtm^-$ and $eGFP^+/Tdtm^+$ cells, respectively. Scale bar, 25 μm. Abbreviations: INS (INSULIN); Tdtm (TdTomato).

subpopulations defined by the two reporters at this stage.

A similar phenomenon was observed in PDX1-DR cell cultures at stage 4, implying that NGN3$^+$ cells were also derived from PDX1$^+$ cells (Figure 4A and 4B). In accordance with the observation that PDX1 is downregulated in endocrine progenitors [37, 38], the NGN3-eGFP$^+$ cells expressed relatively low level of PDX1 (Supplementary information, Figure S3C). The minor inconsistency in the expression intensity between PDX1 and PDX1-Tdtm may have resulted from the fact that Tdtm had a slightly longer half-life than PDX1, as observed with the *NGN3-eGFP* reporter.

In NKX6.1-DR cell cultures at stage 4 day 3, two subpopulations of NGN3-eGFP$^+$ cells were identified by additionally examining Tdtm (Figure 4A and 4B). Because most NGN3-GFP$^+$ cells also express NKX2.2 at this stage (Supplementary information, Figure S3B), most NGN3-GFP$^+$/NKX6.1-Tdtm$^+$ cells would be considered NKX2.2$^+$/NKX6.1$^+$ cells, which are potential endocrine progenitors or precursors specific for β cells. Moreover, because eGFP$^+$ cells made their first appearance at the end of stage 3 (before the onset of Tdtm$^+$ cells (Figure 4A)), a portion of NGN3-eGFP$^+$ cells should come from NKX6.1$^-$ cells.

NEUROD1 and *MAFB* are expressed by intermediate endocrine precursors at different developmental stages *in vivo* [28]. In our study, NEUROD1-Tdtm$^+$ cells made their first appearance as Tdtm$^+$/eGFP$^+$ cells at the end of stage 3, while MAFB-Tdtm$^+$ cells first appeared as Tdtm$^+$/eGFP$^+$ cells at the end of stage 4 (Figure 4A, 4B and data not shown). These data imply that the NEUROD1$^+$ cells and MAFB$^+$ cells are derived from NGN3$^+$ cells in different time windows.

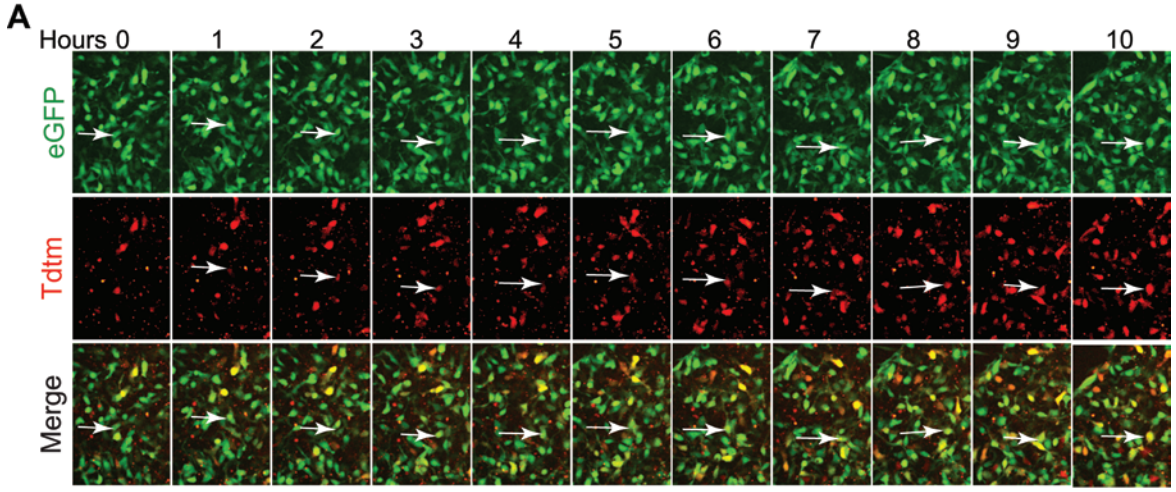
To induce *INS* expression, INS-DR cells were induced through stage 5, when both the intensity and the percentage of NGN3-eGFP$^+$ cells declined (Figure 4A and 4B). However, > 75% of the INS-Tdtm$^+$ cells expressed weak eGFP, reflecting their origin from NGN3$^+$ cells. The minority INS-Tdtm$^+$/NGN3-eGFP$^-$ cells thus represented early-emerged INS-producing cells and showed downregulated NGN3-eGFP.

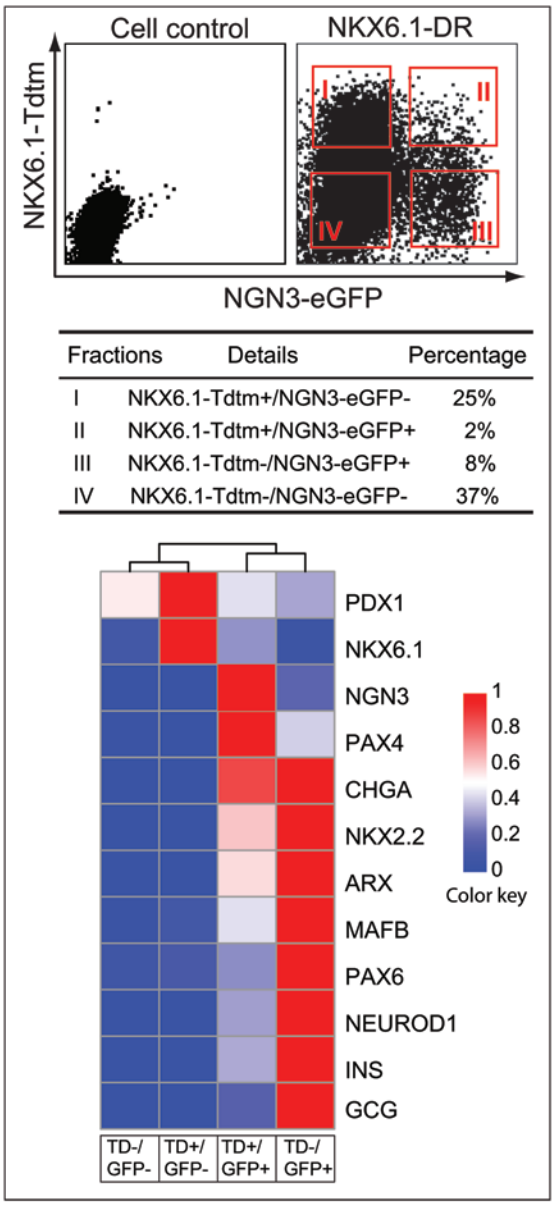
To further confirm the above findings, we used the NEUROD1-DR cell line as an example and performed real-time imaging during differentiation. As shown in Figure 5A and Supplementary information, Movie S1, a portion of the NGN3-eGFP$^+$ cells gradually turned on the expression of Tdtm within 10 h at stage 4 day 2. Thus, this direct evidence implies that NEUROD1$^+$ cells are derived from NGN3$^+$ cells.

The dual reporter cell lines can be used to recognize distinct cellular subpopulations, potentially allowing for the isolation of these subpopulations for characterization of molecular properties. To demonstrate the feasibility of this option, we used the NKX6.1-DR and INS-DR cell lines as examples and performed qPCR to evaluate the gene expression patterns of subpopulations isolated by FACS.

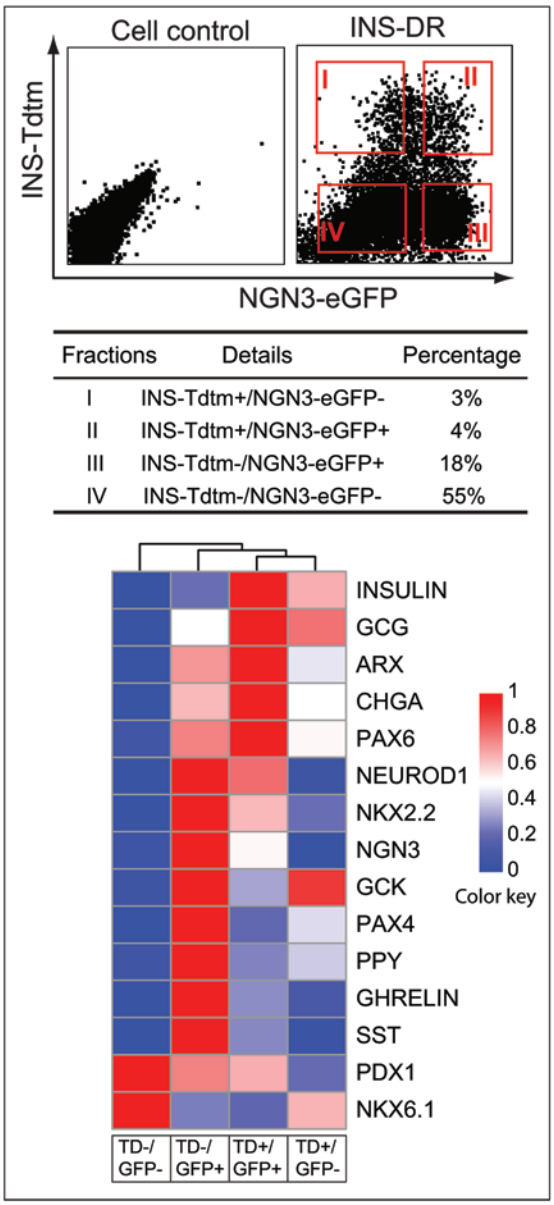
There were four main cell subpopulations in the NKX6.1-DR cell-derived cultures at the end of stage 4; each of these subpopulations had a distinct gene expression profile (Figure 5B). The NKX6.1-Tdtm$^+$/NGN3-eGFP$^-$(N6$^+$/GP$^-$) cells expressed the highest levels of *PDX1* and *NKX6.1* but low level of endocrine-related genes, implying a pancreatic progenitor identity; while the NKX6.1-Tdtm$^-$/NGN3-eGFP$^-$ (N6$^-$/GP$^-$) cells expressed low levels of all pancreatic-associated genes, implying enrichment of non-pancreatic cells. Most of the endocrine-associated genes, particularly those associated with later stage (such as *PAX6*, *MAFB*, *INS* and *GCG*), were enriched in NKX6.1-Tdtm$^-$/NGN3-eGFP$^+$ (N6$^-$/GP$^+$) cells, implying that endocrine-like cells are highly represented in this cell subpopulation. The NKX6.1-Tdtm$^+$/NGN3-eGFP$^+$ (N6$^+$/GP$^+$) cells expressed the highest level of *NGN3* and moderate levels of most other endocrine genes, implying enrichment of endocrine progenitor-like cells in this subpopulation. During pancreatic development, the fate choice between β cells and α cells within the endocrine progenitors is primarily determined by the mutual repression of PAX4 (which favors β-cell identity) and ARX (which favors α cell identity) [39]. It is interesting that the N6$^+$/GP$^+$ cells also expressed highest levels of *PAX4* but not *ARX*, relative to the other cell subpopulations. Thus, these N6$^+$/GP$^+$ cells might be enriched for β-cell-specific endocrine progenitors. As these N6$^+$/GP$^+$ cells also expressed both NKX6.1 and NKX2.2 (Figures 2B, 5B and Supplementary information, Figures S2E, S3A and S3B), our results are consistent with the long-standing assumption that NKX6.1$^+$/NKX2.2$^+$ cells are β-cell-specific endocrine progenitors or precursors [35].

The INS-DR cell-derived cultures also included four cell subpopulations with distinct gene expression profiles by the end of stage 5 (Figure 5C). As expected, the INS-Tdtm$^-$/NGN3-eGFP$^-$ (IN$^-$/GP$^-$) cells showed the highest levels of pancreatic progenitor genes *PDX1* and *NKX6.1*. Interestingly, the INS-Tdtm$^+$/NGN3-eGFP$^+$(IN$^+$/GP$^+$) cells expressed the highest level of *INS*, *GCG* and the α cell determinant *ARX*, but very low levels of *PAX4*. This result could partly explain why those *in vitro*-derived INS/GCG co-expressing cells adopted an α but not β-cell identity [40]. Endocrine progenitor-associated genes, and other hormone genes including *SOMATOSTATIN* (*SST*), *GHRELIN* and *PANCREATIC POLYPEPTIDE (PPY)*, were all enriched in the INS-Tdtm$^-$/NGN3-eGFP$^+$ (IN$^-$/





Fractions	Details	Percentage
I	NKX6.1-Tdtm+/NGN3-eGFP-	25%
II	NKX6.1-Tdtm+/NGN3-eGFP+	2%
III	NKX6.1-Tdtm-/NGN3-eGFP+	8%
IV	NKX6.1-Tdtm-/NGN3-eGFP-	37%



Fractions	Details	Percentage
I	INS-Tdtm+/NGN3-eGFP-	3%
II	INS-Tdtm+/NGN3-eGFP+	4%
III	INS-Tdtm-/NGN3-eGFP+	18%
IV	INS-Tdtm-/NGN3-eGFP-	55%

Figure 5 Capture of cell-fate transitions and characterization of cell subpopulations using dual-reporter cell lines. **(A)** Real-time tracing of NEUROD1-DR cell cultures at stage 4 day 2 implies that NEUROD1-Tdtm$^+$ cells were derived from NGN3-eGFP$^+$ cells. The arrows indicate newborn Tdtm$^+$ cells. **(B)** Q-PCR analysis of gene expression in isolated cell subpopulations from NKX6.1-DR cell cultures at the end of stage 4. **(C)** Q-PCR analysis of gene expression in isolated cell subpopulations from INS-DR cell cultures at the end of stage 5. Gene expression values in **B** and **C** were first normalized to GAPDH, and then the highest expression value of each gene among the corresponding group of cell subpopulations was set to 1. The color scale indicates the normalized expression values. Abbreviations: INS (INSULIN); Tdtm (TdTomato); Td (TdTomato); GFP (eGFP).

GP$^+$) subpopulation. Compared with IN$^+$/GP$^+$ cells, the INS-Tdtm$^+$/NGN3-eGFP$^-$ (IN$^+$/GP$^-$) cells, which represent early-emerged INS-producing cells, exhibited lower expression of most endocrine-associated genes, except *GCK*, *PAX4* and *NKX6.1*.

Collectively, using the dual-reporter cell lines, we captured a series of cell fate transition events in real time, identified multiple cell subpopulations and unveiled their distinct gene expression profiles, among heterogeneous hESC-derived cultures for the first time. These results demonstrate that our dual reporter cell lines are unique tools for the study of cell fate transitions and cell subpopulations, which cannot be achieved by traditional methods.

Expression profiling of NGN3-eGFP$^+$ cells led to the identification of a novel surface protein, SUSD2, for enriching pancreatic endocrine progenitor cells and early endocrine cells

NGN3-eGFP can be used to mark endocrine progenitor cells and their progeny in differentiating hESCs, suggesting that a profiling analysis of NGN3-eGFP$^+$ cells could lead to the discovery of novel participating genes. Six batches of NGN3-eGFP$^+$ and NGN3-eGFP$^-$ cells purified by FACS from PDX1-DR cell cultures at stage 4 were pooled together, respectively. RNA-sequencing-based analysis of the two cell fractions identified a total of 3 976 differentially expressed genes ($P < 0.05$; fold > 2), of which 72% are enriched in NGN3-GFP$^+$ cells (Figure 6A and Supplementary information, Table S2). As expected, all the genes expressed in pancreatic progenitor cells are enriched in NGN3$^-$eGFP$^-$ cells, and all the endocrine-associated genes are enriched in NGN3-eGFP$^+$ cells (Figure 6A). Similar results were obtained when comparison was performed between the gene expression profiles of NGN3-GFP$^+$ cells and an available data set of hESC-derived CD142-enriched pancreatic endoderm cells (Supplementary information, Figure S4A) [24]. Among the 3 976 genes, 195 have potential transcriptional factor activity and 453 are potential noncoding RNA genes (Figure 6B and Supplementary information, Table S2), most of which have not been well investigated in pancreatic development. The gene ontology term 'integral to membrane' (GO: 0016021) identified 1 003 differentially expressed genes that encoded ion channels, adhesion molecules and transporters (Supplementary information, Figure S4B and Table S2).

We further filtered the above transmembrane proteins with strict criteria and identified three proteins that are specifically expressed in certain subsets of endocrine-associated cell populations by immunostaining human fetal pancreas sections with corresponding antibodies. Two of the three proteins were PLEXIN A2 (PLXNA2) and HISTAMINE H1 RECEPTOR (HRH1), both of which were specifically expressed by GCG-producing cells, but not INS-producing cells or other hormone-positive cells in the 18-week human pancreas (Figure 6C, Supplementary information, S4D and data not shown). This result implies that signaling mediated by PLXNA2/NEUROPILIN receptor complexes and the histamine-HRH1 axis acts in the developing α cells. As these two proteins did not exhibit classic cell membrane localization, we focused on the third protein, SUSD2, for further characterization.

We first assessed the expression of SUSD2 in *NGN3-eGFP* reporter cell line-derived cultures. SUSD2 was first detected at stage 4 day 1 (Figure 6D and 6E). SUSD2 was highly co-localized with NGN3-eGFP throughout stage 4, except in a limited number of NGN3-eGFP^{low} cells on days 0-1 (Figure 6D). This finding indicated that SUSD2 might mark NGN3$^+$ cells and their progeny, similar to NGN3-eGFP. As expected, 80%, 93% and 88% of the SUSD2$^+$ cells also expressed NGN3, NKX2.2 and NEUROD1 at stage 4, day 1, respectively (Figure 6D). Given that *NKX2.2* and *NEUROD1* are downstream targets of *NGN3*, these findings imply that most of these SUSD2$^+$ cells are NGN3$^+$/NKX2.2$^+$/NEUROD1$^+$ cells, characteristic of endocrine progenitor. At stage 4, day 0, the NGN3$^{+/low}$ cells did not express SUSD2 or NEUROD1 but did weakly express NKX2.2 and NGN3-eGFP (Figure 6D), suggesting an identity of early endocrine progenitor cells. The proportion of NGN3$^-$ cells in SUSD2$^+$ cells increased gradually after stage 4, day 1. At stage 4, day 4, most of the SUSD2$^+$ cells expressed NKX2.2 and CHGA but little

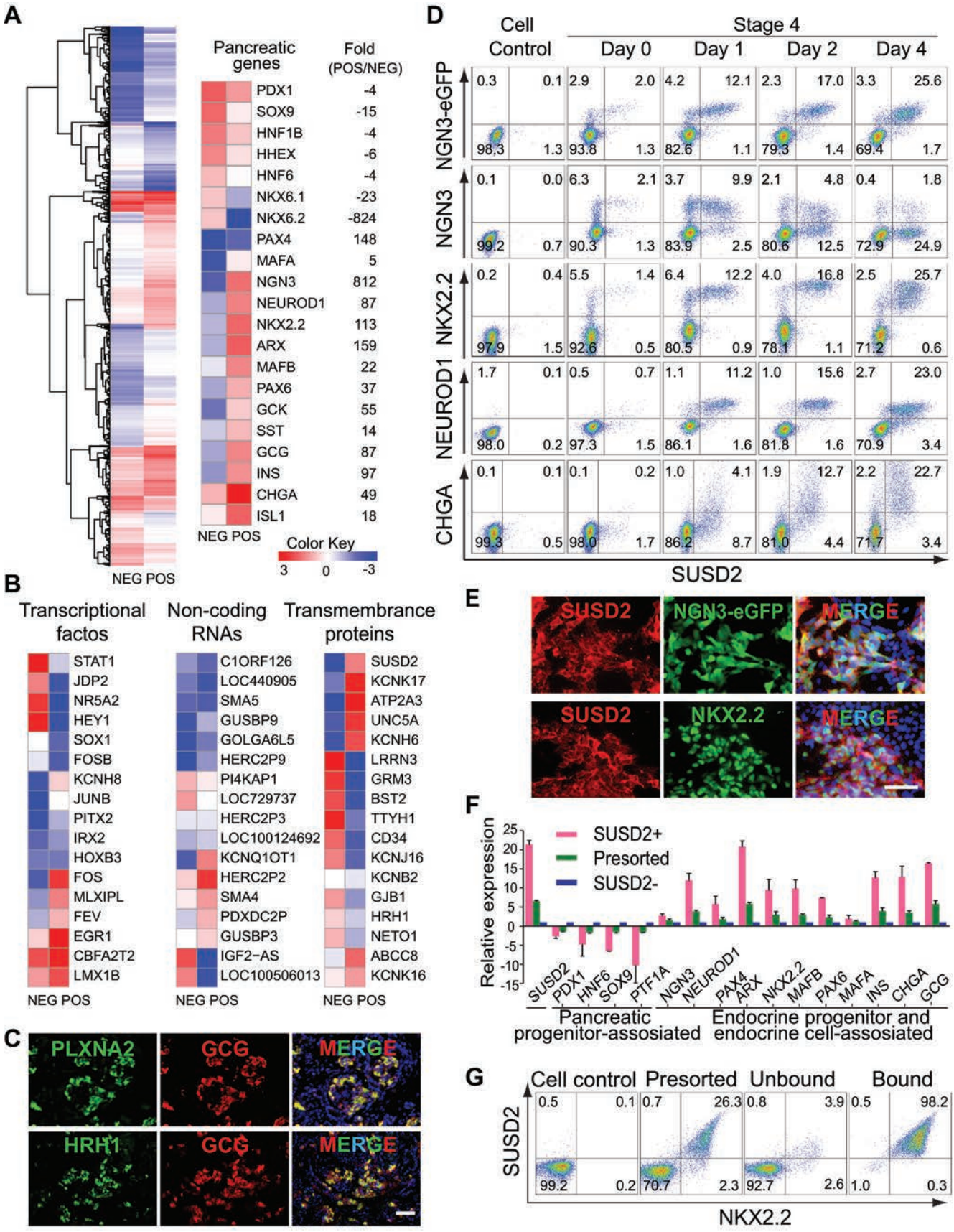

A
Pancreatic genes
Fold (POS/NEG)
PDX1 -4
SOX9 -15
HNF1B -4
HHEX -6
HNF6 -4
NKX6.1 -23
NKX6.2 -824
PAX4 148
MAFA 5
NGN3 812
NEUROD1 87
NKX2.2 113
ARX 159
MAFB 22
PAX6 37
GCK 55
SST 14
GCG 87
INS 97
CHGA 49
ISL1 18
NEG POS
Color Key
3 0 -3
NEG POS
B
Transcriptional factos
Non-coding RNAs
Transmembrance proteins
STAT1 JDP2 NR5A2 HEY1 SOX1 FOSB KCNH8 JUNB PITX2 IRX2 HOXB3 FOS MLXIPL FEV EGR1 CBFA2T2 LMX1B
C1ORF126 LOC440905 SMA5 GUSBP9 GOLGA6L5 HERC2P9 PI4KAP1 LOC729737 HERC2P3 LOC100124692 KCNQ1OT1 HERC2P2 SMA4 PDXDC2P GUSBP3 IGF2-AS LOC100506013
SUSD2 KCNK17 ATP2A3 UNC5A KCNH6 LRRN3 GRM3 BST2 TTYH1 CD34 KCNJ16 KCNB2 GJB1 HRH1 NETO1 ABCC8 KCNK16
NEG POS NEG POS NEG POS
C
PLXNA2 GCG MERGE
HRH1 GCG MERGE
D
Cell Control
Stage 4
Day 0 Day 1 Day 2 Day 4
NGN3-eGFP
NGN3
NKX2.2
NEUROD1
CHGA
SUSD2
E
SUSD2 NGN3-eGFP MERGE
SUSD2 NKX2.2 MERGE
F
Relative expression
SUSD2+
Presorted
SUSD2-
SUSD2 PDX1 HNF6 SOX9 PTF1A NGN3 NEUROD1 PAX4 ARX NKX2.2 MAFB PAX6 MAFA INS CHGA GCG
Pancreatic progenitor-assosiated
Endocrine progenitor and endocrine cell-assosiated
G
Cell control Presorted Unbound Bound
SUSD2
NKX2.2

Figure 6 Gene profiling of NGN3-eGFP$^+$ cells identified SUSD2 as a novel surface marker for endocrine progenitors and early endocrine cells. **(A)** Hierarchical clustering of 3 976 genes that are differentially expressed in NGN3-eGFP$^+$ and NGN3-eGFP$^-$ cells (left). The pancreatic progenitor-associated genes are enriched in NGN3-eGFP$^-$ cells; the pancreatic endocrine-associated genes are enriched in NGN3-eGFP$^+$ cells (right). The color scale indicates the normalized expression values. **(B)** Representative differentially expressed potential transcriptional factors, noncoding RNAs and transmembrane proteins. **(C)** Section staining demonstrated that both PLXNA2 and HRH1 are specifically expressed in GCG-producing cells, but not INS-producing cells in the 18-week human pancreas. **(D)** The gene expression dynamics of *NGN3-eGFP* cells during stage 4, showing the co-expression of SUSD2 and eGFP, and the co-expression of SUSD2 and endocrine-associated proteins, varied over time and by gene. **(E)** Immunostaining of NGN3-eGFP cell cultures at stage 4 day 4 showed that a high co-localization existed between SUSD2 and NGN3-eGFP and between SUSD2 and NKX2.2. Nuclear staining with DAPI (blue) is shown in the merged images. **(F)** FACS-based gene expression analysis further confirmed the enrichment of endocrine-associated genes in SUSD2$^+$ cells and the enrichment of pancreatic progenitor-associated genes in SUSD2$^-$ cells (mean ± SEM, n = 3). **(G)** Flow cytometry analysis showed that NKX2.2$^+$ cells were highly enriched in the bound fraction from the MACS experiments using an anti-SUSD2 antibody. Cells at the end of stage 1 were used as cell control. Abbreviations: POS (positive); NEG (negative); *INS* (*INSULIN*); *GCG* (*GLUCAGON*); *CHGA* (*CHROMOGRANIN A*); *SST* (*SOMATOSTATIN*); *PLXNA2* (*PLEXIN A2*); *HRH1* (*HISTAMINE H1 RECEPTOR*).

or no PDX1 or NGN3 (Figure 6D and Supplementary information, S4F), implying that the NGN3$^-$/SUSD2$^+$ cells resembled early endocrine cells. Accordingly, an almost equal number of CHGA$^+$ cells and NGN3$^-$/SUSD2$^+$ cells were detected throughout this stage (Figure 6D). Analysis of the gene expression dynamics of the entire hESC differentiation process further confirmed that the expression of *SUSD2* tightly followed that of endocrine progenitor-associated genes (Supplementary information, Figure S4G). The expression of *SUSD2* declined after an extended culture period, slightly later than the down-regulation of *NGN3-eGFP* (Supplementary information, Figure S4G and data not shown). Further analysis of available RNA-sequencing data set showed that high *SUSD2* expression is detected in NGN3-eGFP$^+$ cells but not NGN3-eGFP$^-$ cells, hESCs, CD142-enriched pancreatic endoderm cells or human islet cells (Supplementary information, Figure S4C) [24, 41]. Thus, *SUSD2* is a specific surface marker for endocrine progenitor cells and early endocrine precursor cells in hESC-derived cultures at stage 4.

To explore the utility of SUSD2 as a surface marker for enriching specific cell populations, we conducted both FACS and magnetic-activated cell sorting (MACS) experiments in *NGN3-eGFP* cell-derived cultures. RT-qPCR analyses of FACS-purified cells at stage 4, day 4 showed that the expression of several endocrine-associated genes was enriched in SUSD2$^+$ cells, while pancreatic progenitor markers were enriched in the SUSD2$^-$ population (Figure 6F). *NKX2.2* is a pan-endocrine marker that is expressed in pancreatic progenitors, endocrine precursors and endocrine cells. The MACS experiments showed a high enrichment of NKX2.2$^+$/NGN3-eGFP$^+$ (97.1% ± 2.1%, n = 5) cells in the bound fractions, and only a small population of NKX2.2^{low} cells (6.1% ± 2.5%, n = 5) or NGN3-eGFP^{low} cells (5.1% ± 2.2%, n = 5) remained in the unbound fractions (Figure 6G). The further differentiation of MACS-purified cells for 5 days in hormone-inducing conditions showed that the cultures from the unbound fractions expressed a high level of PDX1 and sporadically NKX2.2 and INS, and formed compact epithelial-like colonies. The cultures from the bound fractions expressed PDX1 weakly but NKX2.2 and INS robustly, and did not form large colonies (Supplementary information, Figure S4H and data not shown). Moreover, SUSD2 exhibited a comparable ability to enrich NKX2.2$^+$ cells in wild-type H1 and H9 cell cultures (Supplementary information, Figure S4E).

To further characterize the developmental potential of SUSD2-enriched cells, we transplanted these cells under the kidney capsule of SCID-beige mice. Section immunostaining of 19-week engraftments demonstrated that SUSD2-enriched cells can differentiate into all five types of hormone-producing cells (n = 6; Supplementary information, Figure S5A-S5D). In particular, the derived β cells expressed high levels of PDX1 and NKX6.1 proteins, markers of mature β cells (Supplementary information, Figure S5A and data not shown). As predicted, given that the pancreatic progenitor cells and non-pancreatic cells were enriched in the SUSD2-negatively enriched cells, these engraftments derived from SUSD2-negatively enriched cells also comprised hormone-producing cells as well as many non-hormone-producing cells (Supplementary information, Figure S5A). Collectively, these results imply that SUSD2 can mark and permit the efficient isolation of cells resembling pancreatic endocrine progenitor cells and early endocrine cells in hESC-derived pancreatic-associated cell cultures.

To test whether *SUSD2* expression is conserved *in vivo*, we evaluated the expression of SUSD2 in week 11 and 18 human fetal pancreas and the adult human pancreas by section immunostaining. In the week 11,

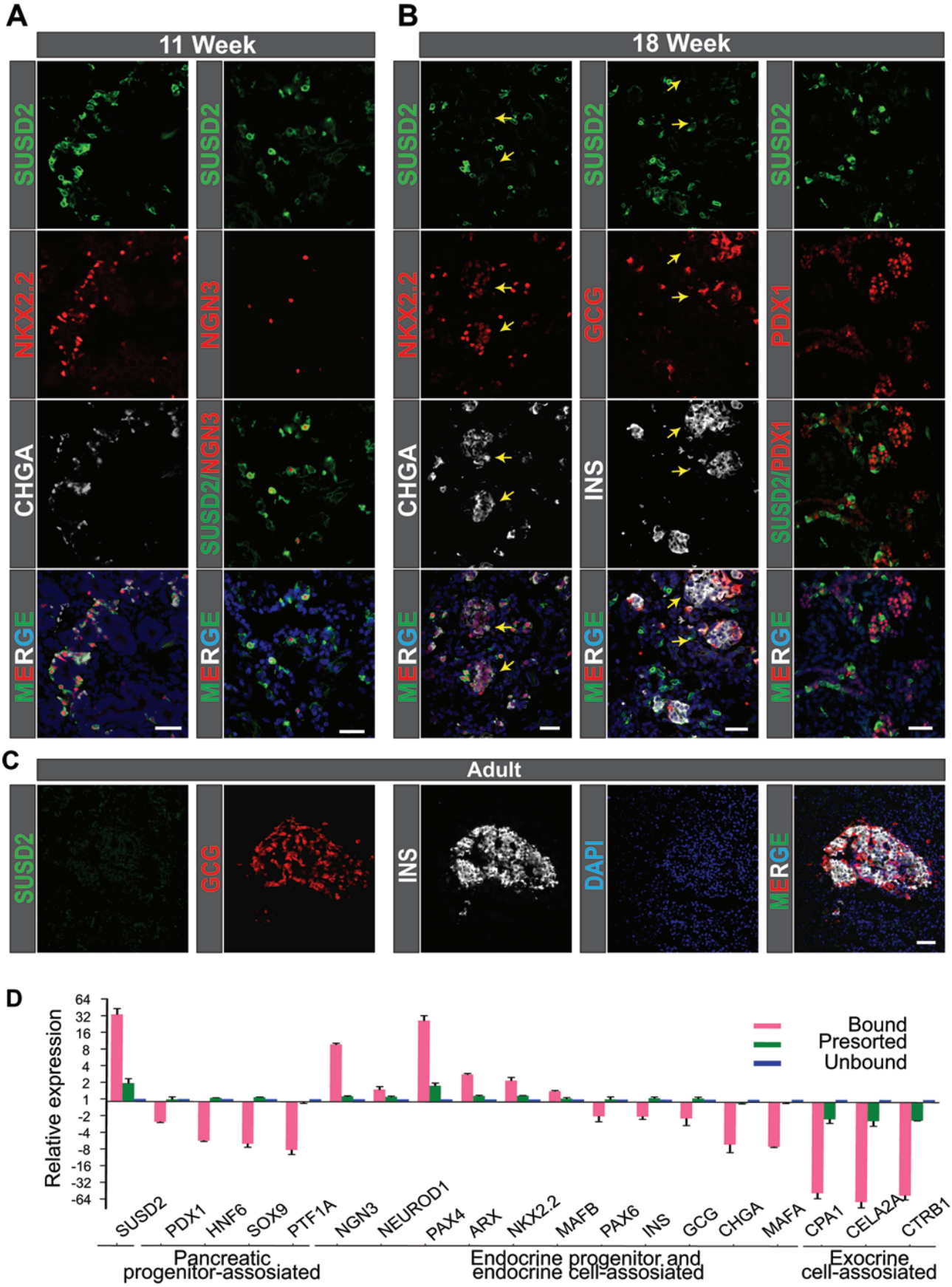

A
11 Week
SUSD2
NKX2.2
CHGA
MERGE
SUSD2
NGN3
SUSD2/NGN3
MERGE
B
18 Week
SUSD2
NKX2.2
CHGA
MERGE
SUSD2
GCG
INS
MERGE
SUSD2
PDX1
SUSD2/PDX1
MERGE
C
Adult
SUSD2
GCG
INS
DAPI
MERGE
D
Relative expression
64
32
16
8
4
2
1
-2
-4
-8
-16
-32
-64
Bound
Presorted
Unbound
SUSD2
PDX1
HNF6
SOX9
PTF1A
NGN3
NEUROD1
PAX4
ARX
NKX2.2
MAFB
PAX6
INS
GCG
CHGA
MAFA
CPA1
CELA2A
CTRB1
Pancreatic progenitor-assosiated
Endocrine progenitor and endocrine cell-assosiated
Exocrine cell-assosiated

Figure 7 SUSD2 marks endocrine progenitors and early endocrine cells in the developing human pancreas. **(A)** In the 11-week pancreas, SUSD2 was highly co-localized with NKX2.2, and ~87% of NGN3$^+$ cells also expressed SUSD2. **(B)** In the 18-week human pancreas, SUSD2 expression was restricted to dispersed NKX2.2^{high} cells and excluded from islet structures, indicated by clusters of CHGA$^+$/NKX2.2^{low} cells (left) or by clusters of hormone-positive cells (middle); SUSD2$^+$ cells expressed little PDX1(right). **(C)** In the adult pancreas, the expression of SUSD2 decreased to a very low level in pancreatic endocrine-associated cells. **(D)** RT-qPCR analysis of cells sorted from week 18 to 23 human fetal pancreas samples by MACS using an anti-SUSD2 antibody demonstrated that endocrine progenitor- and early endocrine cell-associated genes were enriched in the bound fractions, whereas pancreatic progenitor-, late-stage endocrine cell- and exocrine-associated genes were enriched in the unbound fractions (mean ± SEM, n = 3). Nuclear staining with DAPI (blue) is shown in the merged images. Scale bar, 50 μm. Imaging was performed using confocal microscopy. Abbreviations: *INS* (*INSULIN*); *GCG* (*GLUCAGON*); *CHGA* (*CHROMOGRANIN A*).

pancreas SUSD2 was highly co-localized with NKX2.2 (~95% NKX2.2$^+$ cells expressed a high level of SUSD2; ~94% SUSD2$^+$ cells also expressed NKX2.2; Figure 7A). Low-level expression of *CHGA* was also detected in the majority of SUSD2$^+$/NKX2.2$^+$ cells. At this stage of pancreatic development, INS-producing and GCG-producing cells were sparsely detected, both of which were slightly co-localized with SUSD2$^+$ cells. In the week 18 fetal pancreas, cells expressing high level of SUSD2 were mainly restricted to the NKX2.2^{high} population, most of which were dispersed in the pancreas as single cells (Figure 7B). Most of the SUSD2$^-$/NKX2.2^{low} cells expressed a high level of CHGA and formed small clusters or large islet structures. Further immunostaining confirmed that SUSD2$^+$ cells were excluded from the newly formed islet-like structures consisting of INS-producing and GCG-producing cells. Only a relatively low level of SUSD2 expression was observed in some hormone-producing single cells and in some small hormone-positive clusters (Figure 7B and Supplementary information, S6A). In the adult pancreas and islet, SUSD2 expression decreased to a very low level (Figure 7C). Similar to the *in vitro* observations, approximately 87% of the NGN3$^+$ cells co-expressed SUSD2 both in the 11-week and 18-week pancreas (Figure 7A and Supplementary information, S6A), while SUSD2$^+$ cells expressed little PDX1 or HNF1B (Figure 7B and Supplementary information, S6A). Similar to NKX2.2^{high} cells in the week 18 pancreas, most of the NEUROD1$^+$ cells also expressed high level of SUSD2 (Supplementary information, Figure S6A). Thus, *SUSD2* expression is generally conserved *in vivo*. In the developing human pancreas, high SUSD2 expression mainly occurs in endocrine progenitors and early endocrine cells that have not been incorporated into islet structures.

We further assessed whether SUSD2 could be used as a surface marker to isolate *in vivo* endocrine progenitors and their newborn progeny. We conducted MACS experiments to separate cells in the week 18-23 pancreas using an anti-SUSD2 antibody. RT-qPCR analyses demonstrated that pancreatic progenitor-associated and exocrine-associated genes were enriched in the unbound fraction (Figure 7D). Endocrine progenitor-associated genes (such as *NGN3*, *PAX4* and *NKX2.2*) were enriched in the bound fraction; while late-stage endocrine cell-associated genes (such as *MAFA*, *GCG* and *INS*) were enriched in the unbound fraction (Figure 7D). This is consistent with our observations by section staining that SUSD2 can mark endocrine progenitors and newborn endocrine cells, but its expression is excluded from the new-formed islets, which are composed of late-stage endocrine cells and expressed much higher levels of late-stage endocrine-associated genes. Still, the *in vivo* SUSD2-enriched cells expressed higher levels of hormones and other late-stage endocrine genes than the *in vitro*-enriched cells (Supplementary information, Figure S6B). Thus, these results imply that SUSD2 can mark and permit the enrichment of pancreatic endocrine progenitor cells and early endocrine cells in the developing human pancreas.

In summary, SUSD2 can specifically mark and permit the enrichment of cells representing pancreatic endocrine progenitor cells and early endocrine cells both in hESC-derived pancreatic-associated cell cultures and in the developing human pancreas.

Discussion

In this study, for the first time, we systematically labeled sequential developmental genes covering the major developmental stages of pancreatic β cells from hESCs. Our studies demonstrate that the systematic labeling of key developmental genes and the application of dual reporters provide a comprehensive approach to more precisely investigate pancreatic β-cell differentiation from hESCs.

We demonstrated the value of this strategy for unveiling developmental clues at different differentiation stages and identified a novel surface protein, SUSD2, that can be used for the study of the cell fate commitment mechanism associated with endocrine progenitors and early

endocrine cells. Because NGN3$^+$ endocrine progenitor cells are thought to be unipotent [42], multiple types of endocrine progenitors or precursors may coexist. In accordance with this assumption, we found that a portion of hESC-derived SUSD2-enriched cells can differentiate into mature β cells after transplantation in the mouse. A detailed comparative analysis of the gene expression profiles and developmental potentials of SUSD2$^+$ cells isolated *in vivo* and *in vitro* will shed further light on the specification mechanism for each cell population at this critical stage, particularly if these comparisons are performed at the single-cell level. We also found that a limited portion of the hormone-producing cells in the engraftment derived from SUSD2-enriched cells were actually β cells. Considering that SUSD2 can enrich for most endocrine progenitors and early endocrine cells in hPSC cultures, this phenomenon implies that the current protocol for endocrine lineage differentiation from hPSC-derived pancreatic progenitors is not efficient for β-cell generation. As the endocrine progenitor stage comes first during the differentiation of pancreatic progenitors into β cells, our "SUSD2-enrichment-based transplantation" assay may represent a practical approach that can be used to screen for conditions that induce hPSC-derived pancreatic progenitors into β-cell-specific endocrine progenitors or precursors. Moreover, in the engraftment derived from SUSD2-negatively enriched cells, many non-hormone-producing cells and duct-like structures were observed. This result implies that SUSD2 could be used to eliminate non-pancreatic cells, thus providing a safer method for cell replacement therapy for diabetes once conditions for inducing β-cell-specific endocrine progenitors are identified. In addition, SUSD2 is a migration-associated protein that is involved in the epithelial-to-mesenchymal transition (EMT) process during tumor invasion [43]. A similar EMT process also occurs during the specification of endocrine cells [44]. Our data imply that the expression of SUSD2 is restricted to endocrine progenitors and early stage endocrine precursor cells, both of which display a mesenchymal phenotype. Therefore, SUSD2 may also be involved in the EMT process that occurs during endocrine cell specification.

We also show that our dual reporter system offers a unique tool for the study of pancreatic β-cell differentiation from hESCs. The dual reporter cell lines helped us to capture the fate transition events and identify intermediate subpopulations for characterization in a non-invasive, real-time manner. Different cell subpopulations may represent cells at different developmental stages and/or with different developmental potentials. Our studies revealed that a portion of SUSD2-enriched cells are β-cell-specific endocrine progenitors or precursors. As most NGN3-eGFP$^+$ cells at the end of stage 4 also express SUSD2, NKX6.1-Tdtm$^+$/NGN3-eGFP$^+$ cells should represent a portion of the SUSD2-enriched cells. We found that *NGN3* and *PAX4* were highly enriched in NKX6.1-Tdtm$^+$/NGN3-eGFP$^+$ cells. It would therefore be interesting to ascertain in the future whether the NKX6.1-Tdtm$^+$/NGN3-eGFP$^+$ cells are the cell subpopulation that is specific for β-cell identity.

The reporter hESC lines generated in this study are also valuable tools for investigating hPSC differentiation into other cell lineages, particularly neurons. Each of the genes targeted in this study is expressed in at least one other lineage (e.g., *SOX17* in blood) [45]. More importantly, all of the genes labeled in this study are expressed during the development of the neural system, and most of them are critical for the development of specific neuronal cell types [46]. Thus, our reporter cell lines should also be very useful for studying the differentiation of neurons and other cell lineages in hPSC cultures.

Collectively, we have developed a valuable platform for the study of β-cell differentiation from hESCs. Further applications of this platform and our new findings will help us to unveil the overall molecular mechanism underlying the sequential commitment process and to identify conditions for obtaining mature human β cells *in vitro*. The directed differentiation of other cell lineages faces problems similar to those found in pancreatic β cells. Thus, the systematic generation of stage-specific reporter cell lines represents a practical general solution for guiding the differentiation and studying the *in vitro* development of various cell lineages in hPSCs.

Materials and Methods

Generation of targeting vectors and TALENs

To replace the stop codon of *NGN3*, the heat-induced DY380 cells containing a BAC were electroporated with the *2A-GFP-loxp-CAG-neo-loxp* cassette flanked with 1-kb homology arms. To shorten the right arm of the BAC-based targeting vector to 6.5 kb, the heat-induced DY380 cells containing the modified BAC were electroporated with the *HSV-TK-Amp^r* cassette flanked with 1-kb homology arms, leaving the left arm at a length of 134.5 kb.

To replace the stop codon for the other genes, the L-arabinose-induced *E. coli* cells containing BACs and plasmid PSC101-BAD-gbaA were electroporated with the PCR products of the *2A-Tdtm-loxp-CAG-neo-loxp* or *2A-eGFP-loxp-CAG-neo-loxp* cassettes flanked with 50-bp homology arms. For retrieval, the L-arabinose-induced *E. coli* cells containing the modified BAC and plasmid PSC101-BAD-gbaA were electroporated with the PCR product of plasmid PL253SNK (derived from PL253) flanked with 50-bp homology arms. All the BACs used in this paper were ordered from Children's Hospital Oakland Research Institute (CHORI) as listed in Supplementary information, Data S1.

TALENs specific for the selected gene loci were synthesized

using a high-throughput integrated chip method as previously described [32]. The DNA sequences recognized by the respective pairs of TALENs are listed in Supplementary information, Data S1. To measure the cutting efficiency, the TALENs were transfected into 293T cells and analyzed by T7 endonuclease (NEB) assays and sequencing. Primers used for gene targeting are provided in Supplementary information, Table S5.

Gene targeting in hES cells

To target the *NGN3* locus, a linearized BAC-based targeting vector was electroporated into H1 cells using an electroporator (Bio-Rad). The *CAG-neo* cassette was deleted from *NGN3-eGFP* cell lines by the transient expression of CRE recombinase. To target the other genes, the *NGN3-eGFP* cell line with the deletion of the *CAG-neo* cassette was used as targeting cell line; to target *SOX17*, wild-type H1 cells were used. The linearized targeting vector and TALENs were nucleofected into hES cells using the 4D-Nucleofector System (Lonza). Three days later, drugs were added into the culture medium for 2 weeks; G418 (Gibco) was added at a concentration of 50 μg/ml for the first week and 100 μg/ml for the second week, and ganciclovir (Sigma) was added at a concentration of 0.2 μM. Up to 192 colonies (96 colonies in most cases) for each gene were picked and expanded in 96-well plates. To reduce the labor required, we treated the colonies in one column as a group for the preparation of genomic DNA (Tiangen) and PCR screening (Takara, DR044A). When one group was positive by PCR, the colonies in this group would be screened one by one. The PCR-positive colonies were expanded for Southern blot and differentiation analysis.

RNA-sequencing sample preparation and analysis

Two micrograms of total RNA was extracted for cDNA library construction. The cDNA libraries were constructed with the Illumina Paired-End DNA Sample Prep Kit according to the manufacturer's protocol. Sequencing was performed on an Illumina Hiseq2000 platform (BGI). For each sample, the sequence reads were aligned to the transcriptome using Tophat software. All raw data and processed data have been submitted to Gene Expression Omnibus (accession number: GSE54879).

hESC culture and differentiation

hESCs, including the reporter cell lines, were maintained as previously reported [23]. A five-stage differentiation protocol was used for directed differentiation, including definitive endoderm (which was treated with a high concentration of Activin-A for 4 days), primitive gut tube (which was treated with KGF and SB525334 for 3 days), posterior foregut (which was treated with RA, Noggin and SANT-1 for 4 days), pancreatic and endocrine progenitors (which were treated with Noggin and TPB for 3-5 days) and hormone-producing cells (which were treated with human LIF and Alk5 inhibitor II for more than 5 days).

Flow cytometry and cell sorting

Single-cell suspensions from cell cultures were acquired by trypsin-EDTA digestion (Life Technologies). Single-cell suspensions from fetal pancreatic tissues were acquired according to the protocol described in Supplementary information, Data S1. For the flow cytometry analysis of living cells, single-cell suspensions were directly analyzed using a BD FACSCalibur flow cytometry system. For surface marker and intracellular flow cytometry analyses, single-cell suspensions were stained with antibodies and analyzed using a BD FACSCalibur. The data were analyzed by FlowJo software. FACS was performed on a BD FACS Aria IIu, and MACS experiments were performed with reagents from Miltenyi Biotec according to the manufacturer's instructions.

Immunofluorescence and imaging

The immunostaining of cells/sections was performed according to the protocols described in the Extended Materials and Methods (Supplementary information, Data S1). The antibodies used here are listed in Supplementary information, Table S3. Real-time cell-tracing was performed using a High-Speed Live Cell Confocal Imaging System (Nikon).

PCR

Regular PCR was performed using 2× EasyTaq PCR SuperMix (TransGene). Genomic PCR was performed using LA Taq (TaKaRa). RT-qPCR was performed with SYBR Green PCR Master Mix (Life Technologies), and gene expression was normalized to GAPDH. The detailed methods are described in Supplementary information, Data S1 and the primers used are listed in Supplementary information, Table S4.

Human pancreatic tissue acquisition

Human fetal/adult pancreatic specimens were obtained from the China-Japan Friendship Hospital. Informed consent was obtained for tissue donations, and approval for the study was granted by institutional review boards.

Detailed experimental procedures are provided in Supplementary information, Data S1.

Acknowledgments

We thank Chenyan Wang, Yang Zhao, Yan Shi and Jun Xu for critical reading of the manuscript and for discussion in the preparation of this manuscript. We also thank Xiaobao Wang for conducting molecular cloning; Weichao Du, Bingqing Xie and Zijian Li for cell culture; Yinan Liu for the RT-qPCR analysis; Li Su, Shiliang Ma, Liying Du and National Center for Protein Sciences at Peking University (including but not limited to Zhonglin Fu; Zailing Bai and Jingshu Wang) for FACS; Hongxia Lv and Xiaochen Li for confocal microscopy; and Junhua Zou for G-banding analysis. This work was supported by the National Basic Research Program of China (973 program; 2012CB966401), the Key New Drug Creation and Manufacturing Program (2011ZX09102-010-03), the Ministry of Science and Technology (2013DFG30680), the National Natural Science Foundation of China (30830061), the Ministry of Education of China (111 project), and National Science and Technology Major Projects Supporting Program from Shenzhen (GJHS20120820102148947).

References

1 Cohen DE, Melton D. Turning straw into gold: directing cell fate for regenerative medicine. *Nat Rev Genet* 2011; **12**:243-252.

2 Murry CE, Keller G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. *Cell* 2008; **132**:661-680.

3 Nostro MC, Keller G. Generation of β cells from human pluripotent stem cells: potential for regenerative medicine. *Semin Cell Dev Biol* 2012; **23**:701-710.
4 Nostro MC, Sarangi F, Ogawa S, *et al.* Stage-specific signaling through TGFβ family members and WNT regulates patterning and pancreatic specification of human pluripotent stem cells. *Development* 2011; **138**:861-871.
5 Nazareth EJ, Ostblom JE, Lucker PB, *et al.* High-throughput fingerprinting of human pluripotent stem cell fate responses and lineage bias. *Nat Methods* 2013; **10**:1225-1231.
6 Rodriguez-Segui S, Akerman I, Ferrer J. GATA believe it: new essential regulators of pancreas development. *J Clin Invest* 2012; **122**:3469-3471.
7 Pan FC, Wright C. Pancreas organogenesis: from bud to plexus to gland. *Dev Dyn* 2011; **240**:530-565.
8 McKnight KD, Wang P, Kim SK. Deconstructing pancreas development to reconstruct human islets from pluripotent stem cells. *Cell Stem Cell* 2010; **6**:300-308.
9 Bu L, Jiang X, Martin-Puig S, *et al*. Human ISL1 heart progenitors generate diverse multipotent cardiovascular cell lineages. *Nature* 2009; **460**:113-117.
10 Wang P, Rodriguez RT, Wang J, Ghodasara A, Kim SK. Targeting SOX17 in human embryonic stem cells creates unique strategies for isolating and analyzing developing endoderm. *Cell Stem Cell* 2011; **8**:335-346.
11 Micallef SJ, Li X, Schiesser JV, *et al.* INS(GFP/w) human embryonic stem cells facilitate isolation of *in vitro* derived insulin-producing cells. *Diabetologia* 2012; **55**:694-706.
12 Giudice A, Trounson A. Genetic modification of human embryonic stem cells for derivation of target cells. *Cell Stem Cell* 2008; **2**:422-433.
13 Kim Y, Kweon J, Kim A, *et al.* A library of TAL effector nucleases spanning the human genome. *Nat Biotechnol* 2013; **31**:251-258.
14 Wang H, Hu YC, Markoulaki S, *et al*. TALEN-mediated editing of the mouse Y chromosome. *Nat Biotechnol* 2013; **31**:530-532.
15 Shalem O, Sanjana NE, Hartenian E, *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* 2014; **343**:84-87.
16 Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 2014; **343**:80-84.
17 Pagliuca FW, Melton DA. How to make a functional β-cell. *Development* 2013; **140**:2472-2483.
18 Shi Y, Hou L, Tang F, *et al.* Inducing embryonic stem cells to differentiate into pancreatic β cells by a novel three-step approach with activin A and all-trans retinoic acid. *Stem Cells* 2005; **23**:656-662.
19 D'Amour KA, Bang AG, Eliazer S, *et al*. Production of pancreatic hormone-expressing endocrine cells from human embryonic stem cells. *Nat Biotechnol* 2006; **24**:1392-1401.
20 Jiang J, Au M, Lu K, *et al.* Generation of insulin-producing islet-like clusters from human embryonic stem cells. *Stem Cells* 2007; **25**:1940-1953.
21 Jiang W, Shi Y, Zhao D, *et al*. *In vitro* derivation of functional insulin-producing cells from human embryonic stem cells. *Cell Res* 2007; **17**:333-344.
22 Chen S, Borowiak M, Fox JL, *et al*. A small molecule that directs differentiation of human ESCs into the pancreatic lineage. *Nat Chem Biol* 2009; **5**:258-265.
23 Zhang D, Jiang W, Liu M, *et al.* Highly efficient differentiation of human ES cells and iPS cells into mature pancreatic insulin-producing cells. *Cell Res* 2009; **19**:429-438.
24 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.
25 Rezania A, Bruin JE, Riedel MJ, *et al.* Maturation of human embryonic stem cell-derived pancreatic progenitors into functional islets capable of treating pre-existing diabetes in mice. *Diabetes* 2012; **61**:2016-2029.
26 Zhu FF, Zhang PB, Zhang DH, *et al.* Generation of pancreatic insulin-producing cells from rhesus monkey induced pluripotent stem cells. *Diabetologia* 2011; **54**:2325-2336.
27 Rukstalis JM, Habener JF. Neurogenin3: a master regulator of pancreatic islet differentiation and regeneration. *Islets* 2009; **1**:177-184.
28 Oliver-Krasinski JM, Stoffers DA. On the origin of the β cell. *Genes Dev* 2008; **22**:1998-2021.
29 Kroon E, Martinson LA, Kadoya K, *et al.* Pancreatic endoderm derived from human embryonic stem cells generates glucose-responsive insulin-secreting cells *in vivo*. *Nat Biotechnol* 2008; **26**:443-452.
30 Bruin JE, Erener S, Vela J, *et al.* Characterization of polyhormonal insulin-producing cells derived *in vitro* from human embryonic stem cells. *Stem Cell Res* 2014; **12**:194-208.
31 Hasegawa K, Cowan AB, Nakatsuji N, Suemori H. Efficient multicistronic expression of a transgene in human embryonic stem cells. *Stem Cells* 2007; **25**:1707-1712.
32 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.
33 Gittes GK. Developmental biology of the pancreas: a comprehensive review. *Dev Biol* 2009; **326**:4-35.
34 Zorn AM, Wells JM. Vertebrate endoderm development and organ formation. *Annu Rev Cell Dev Biol* 2009; **25**:221-251.
35 Sander M, Sussel L, Conners J, *et al.* Homeobox gene Nkx6.1 lies downstream of Nkx2.2 in the major pathway of β-cell formation in the pancreas. *Development* 2000; **127**:5533-5540.
36 Ejarque M, Cervantes S, Pujadas G, *et al.* Neurogenin3 cooperates with Foxa2 to autoactivate its own expression. *J Biol Chem 2013*; **288**:11705-11717.
37 Schwitzgebel VM, Scheel DW, Conners JR, *et al.* Expression of neurogenin3 reveals an islet cell precursor population in the pancreas. *Development* 2000; **127**:3533-3542.
38 Wilson ME, Scheel D, German MS. Gene expression cascades in pancreatic development. *Mech Dev* 2003; **120**:65-80.
39 Collombat P, Mansouri A, Hecksher-Sorensen J, *et al.* Opposing actions of Arx and Pax4 in endocrine pancreas development. *Genes Dev* 2003; **17**:2591-2603.
40 Rezania A, Riedel MJ, Wideman RD, *et al.* Production of functional glucagon-secreting α-cells from human embryonic stem cells. *Diabetes* 2011; **60**:239-247.
41 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.

42 Desgraz R, Herrera PL. Pancreatic neurogenin 3-expressing cells are unipotent islet precursors. *Development* 2009; **136**:3567-3574.
43 Watson AP, Evans RL, Egland KA. Multiple functions of sushi domain containing 2 (SUSD2) in breast tumorigenesis. *Mol Cancer Res* 2013; **11**:74-85.
44 Gouzi M, Kim YH, Katsumoto K, Johansson K, Grapin-Botton A. Neurogenin3 initiates stepwise delamination of differentiating endocrine cells during pancreas development. *Dev Dyn* 2011; **240**:589-604.
45 He S, Kim I, Lim MS, Morrison SJ. Sox17 expression confers self-renewal potential and fetal stem cell characteristics upon adult hematopoietic progenitors. *Genes Dev* 2011; **25**:1613-1627.
46 Arntfield ME, van der Kooy D. β-Cell evolution: How the pancreas borrowed from the brain: The shared toolbox of genes expressed by neural and pancreatic endocrine cells may reflect their evolutionary relationship. *Bioessays* 2011; **33**:582-587.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)