Stem Cell Reports
Article

ISSCR

OPEN ACCESS

Identification and Single-Cell Functional Characterization of an Endodermally Biased Pluripotent Substate in Human Embryonic Stem Cells

Thomas F. Allison,[1,*] Andrew J.H. Smith,[3,5] Konstantinos Anastassiadis,[4] Jackie Sloane-Stanley,[5] Veronica Biga,[2,7] Dylan Stavish,[1] James Hackland,[1] Shan Sabri,[6] Justin Langerman,[6] Mark Jones,[1] Kathrin Plath,[6] Daniel Coca,[2] Ivana Barbaric,[1] Paul Gokhale,[1] and Peter W. Andrews[1]

[1]Centre for Stem Cell Biology, Department of Biomedical Science, University of Sheffield, Sheffield S10 2TN, UK
[2]Signal Processing and Complex Systems Group, Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S10 2TN, UK
[3]MRC Centre for Regenerative Medicine, Institute for Stem Cell Research, School of Biological Sciences, University of Edinburgh, Edinburgh EH16 4UU, UK
[4]Stem Cell Engineering, Biotechnology Center, Technische Universitaet Dresden, 01307 Dresden, Germany
[5]MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK
[6]UCLA School of Medicine, Department of Biological Chemistry, University of California Los Angeles, Los Angeles, CA, USA
[7]School of Medicine, Faculty of Biology and Health, University of Manchester, Manchester M13 9PT, UK
*Correspondence: tomallison24@gmail.com
https://doi.org/10.1016/j.stemcr.2018.04.015

SUMMARY

Human embryonic stem cells (hESCs) display substantial heterogeneity in gene expression, implying the existence of discrete substates within the stem cell compartment. To determine whether these substates impact fate decisions of hESCs we used a GFP reporter line to investigate the properties of fractions of putative undifferentiated cells defined by their differential expression of the endoderm transcription factor, GATA6, together with the hESC surface marker, SSEA3. By single-cell cloning, we confirmed that substates characterized by expression of *GATA6* and SSEA3 include pluripotent stem cells capable of long-term self-renewal. When clonal stem cell colonies were formed from *GATA6*-positive and *GATA6*-negative cells, more of those derived from *GATA6*-positive cells contained spontaneously differentiated endoderm cells than similar colonies derived from the *GATA6*-negative cells. We characterized these discrete cellular states using single-cell transcriptomic analysis, identifying a potential role for SOX17 in the establishment of the endoderm-biased stem cell state.

INTRODUCTION

Human embryonic stem cells (hESCs) offer opportunities for a wide range of applications in human health care, provided that effective methods are developed for controlling their differentiation. A central problem for stem cell biology, whether for pluripotent stem cells from the early embryo, or multipotent stem cells from later tissues, is to establish how such cells make fate decisions between self-renewal or differentiation and then how they choose between alternative pathways of differentiation (Murry and Keller, 2008). In part, the decision any individual stem cell makes depends upon external cues, and many studies focus on the response of stem cells to particular signals, whether diffusible cytokines, the extracellular matrix or cell:cell interactions (Semrau and van Oudenaarden, 2015). However, as cell characterization has become more refined and single-cell analyses have become feasible, many studies have also highlighted the heterogeneity of stem cell populations, making it possible to cluster cells into different subsets (Hough et al., 2009, 2014). This raises the question of whether this heterogeneity is "noise" with no relevance to fate decisions, or whether the different subsets of stem cells respond differently to external cues so that their ultimate fate depends on a combination of extrinsic and intrinsic factors. By definition, stem cells assigned to different subsets must all be capable of self-renewal and the same range of differentiation, but it is possible that the different subsets correspond to different, interconvertible substates in which the stem cells exhibit distinct properties (Arias and Brickman, 2011; Draper et al., 2002; Enver et al., 2005, 2009).

Among hematopoietic stem cells, heterogeneity in the patterns of gene expression at the single-cell level has been used to suggest the existence of multi-lineage priming, whereby subsets of stem cells activate components of different lineage-related regulatory genes prior to commitment to differentiate (Hu et al., 1997; Huang et al., 2007). Further, different subsets of a myeloid progenitor cell separated by differential surface markers appeared to have different propensities for monocyte and erythroid differentiation, although both were capable of self-renewal (Chang et al., 2008). However, in another study based on single-cell analyses (Pina et al., 2012), that conclusion was questioned since the apparent lineage-biased subsets could themselves be further subdivided into self-renewing and lineage-committed cells, emphasizing the need for clonal analyses to confirm the co-existence of self-renewal capacity and lineage bias in a single cell. In the pluripotent context, interconvertible subsets of mouse embryonic stem cells have been identified using reporters for stem cell-associated transcription factors such as

NANOG (Chambers et al., 2007), STELLA (Hayashi et al., 2008), or REX1 (Toyooka et al., 2008), or lineage-associated transcription factors such as HEX (Canham et al., 2010), and shown to exhibit different functional properties.

We previously identified a transitory state of hESCs, marked by lack of the surface marker SSEA3, with an apparently greater tendency to differentiate (Enver et al., 2005), while Laslett et al. (2007) reported a gradation in expression of the surface markers CD9 and GCTM2 as hESCs transited from an undifferentiated to differentiated state (Laslett et al., 2007). However, although these observations indicate substates with a greater or lesser tendency to differentiate, it is unclear whether substates can be identified with different biases with respect to the lineages they follow after differentiation. Previously, we inferred the existence of such lineage-biased substates in the pluripotent human embryonal carcinoma cell line NTERA2, but could not specifically identify the biased cells prior to differentiation (Tonge et al., 2010).

In a recent study of gene expression in individual hESCs, we observed that among cells expressing characteristic features of undifferentiated cells, notably the surface antigen SSEA3, and the transcription factors OCT4 and NANOG, some also expressed genes typically associated with endoderm differentiation, such as *GATA6* (Gokhale et al., 2015). To test whether these cells are functional, self-renewing stem cells, we have produced and analyzed an hESC line, Shef4, carrying a GFP reporter knocked into the *GATA6* locus by gene targeting, as a tool to interrogate whether functionally biased substates exist within the over-arching pluripotent stem cell state. We have found that the undifferentiated cells can not only interconvert between substates that do and do not express *GATA6*, but also that in the *GATA6*-expressing substate they have a higher probability of endoderm differentiation.

RESULTS

A *GATA6-GFP* Reporter Cell Line Reveals Orders of hESC Heterogeneity

To investigate the dynamics of *GATA6* expression in live hESCs, we generated a Shef4 hESC line (Aflatoonian et al., 2010) with an GFP reporter knockin into one allele of the *GATA6* locus by Zinc Finger Nuclease-mediated homologous recombination. The GFP reporter knockin into the translational initiation codon of the *GATA6* locus was designed to express GFP under the control of the endogenous *GATA6* promoter (Figure S1A). Shef4 clones with gene targeted integrations by homologous recombination were identified, and one heterozygous knockin clone (S4G6 4/F-9) was confirmed to contain a single insertion of the GFP reporter at the *GATA6* locus with no additional integrations (Figure S1B). This clone was further genetically modified to delete the neomycin resistance gene selection cassette by recombinase-mediated excision (Supplemental Experimental Procedures), and a resulting clone (S4G6 A3) was generated with the expected DNA rearrangement (Figure S1B) and a normal XY karyotype (Figure S1C). To validate the fidelity of the reporter line, we differentiated both the parental Shef4 cells and the reporter cell line S4G6 A3 toward endoderm. As expected, the Shef4 cells showed increased GATA6 protein, but no GFP expression, whereas the reporter line showed an increase in GFP expression and GATA6 protein in a correlative manner as anticipated for the above knockin strategy (Figure S1D). To assess whether the knockin of the GFP cassette into the *GATA6* locus altered endodermal differentiation capacity, we performed qPCR for genes characteristic of endoderm/primitive streak. Gene expression levels were found to be similar between the parental Shef4 cells and the GFP knockin line, confirming the differentiation capacity of the reporter line (Figure S1E). Additionally, we investigated whether the insertion of GFP into the *GATA6* locus altered the *GATA6* RNA level in the hESC state. We found by performing qPCR a slightly reduced level of *GATA6* expression in the reporter knockin line relative to the Shef4 parental cells qualitatively consistent with the expectation that the reporter integration should result in premature termination of *GATA6* transcription (Figure S1F).

Having validated our reporter line, we subsequently used expression of GFP as a measure of the *GATA6* transcriptional state, which we refer to throughout the manuscript as *GATA6*. By flow cytometry, we observed that the reporter line grown in KO/SR (Knockout DMEM and 20% Knockout Serum Replacement) on mouse embryo fibroblast (MEF) feeders, contained a subset of 2%–10% cells expressing *GATA6* (Figure 1A). We also found varying degrees of *GATA6* expression denoted by "low" and "high." To determine whether GFP expression correlated with GATA6 protein expression in self-renewing conditions, we stained the reporter line in self-renewal conditions with a GATA6 antibody and found that as GFP intensity increased, the levels of GATA6 protein also increased (Figure S2A). To begin characterizing *GATA6* expressing cells, we first tested whether they expressed SSEA3, a sensitive cell surface marker that we have used extensively to identify undifferentiated hESCs (Andrews et al., 1982; Enver et al., 2005; Gokhale et al., 2015). We found a new level of cellular heterogeneity and the appearance of distinct populations of hESCs in culture. The most apparent population expressed high levels of SSEA3 with no *GATA6* expression (3+/6−), with smaller populations expressing high *GATA6* levels with no SSEA3 (3−/6+), and no SSEA3 or *GATA6* (3−/6−). Notably, we saw co-expressing populations consisting of high SSEA3 with low *GATA6* (3+/6L) and high SSEA3 with high *GATA6* (3+/6H) expression (Figure 1B). To determine

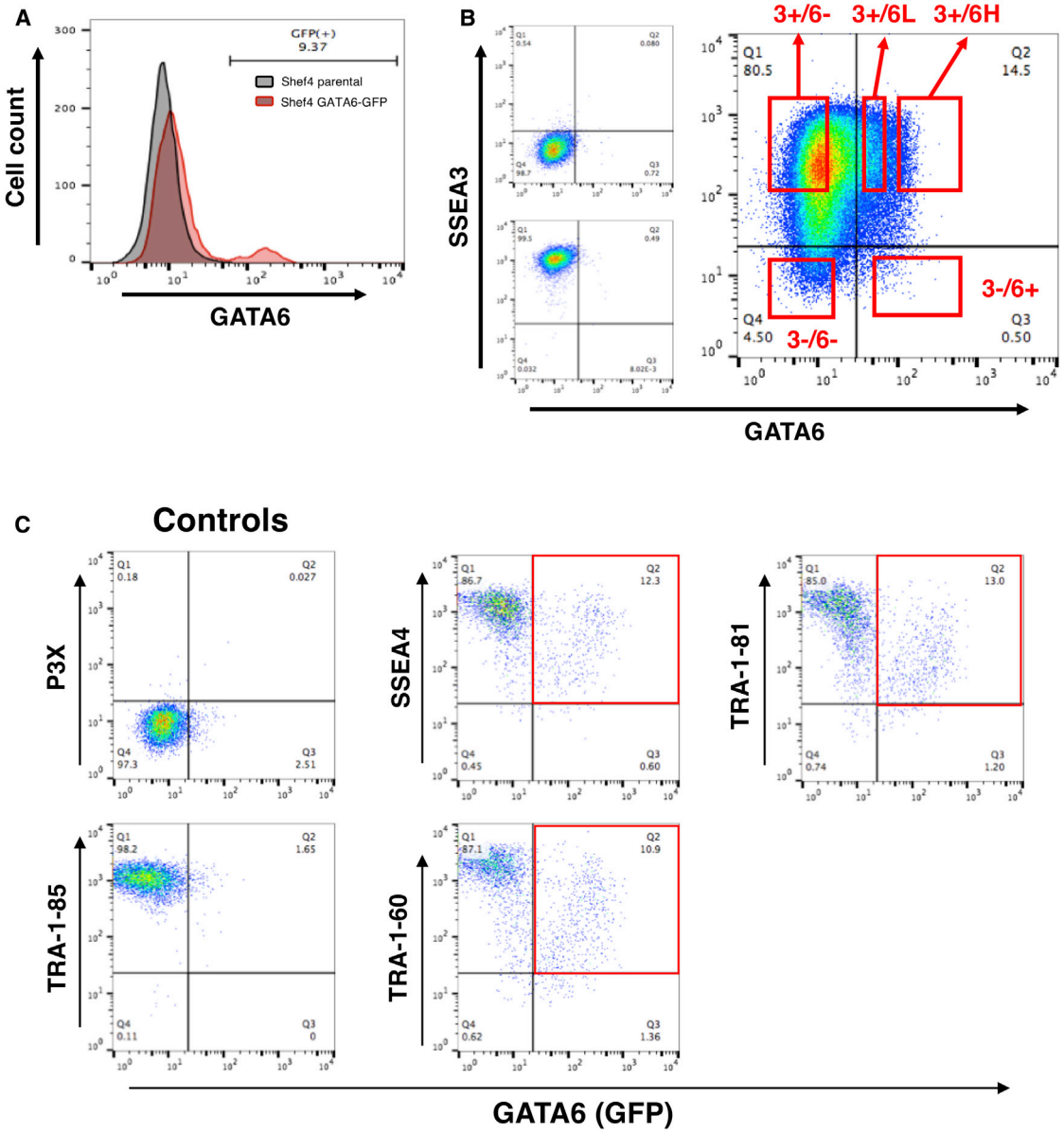


Figure 1. *GATA6* Is Expressed in a Small Subset of hESCs

(A) Representative FACS plot of the Shef4 *GATA6*-GFP reporter line S4G6 A3 cultured in KO/SR and MEF conditions. Black peak represents the unmodified parental Shef4 control line, and red, the Shef4 *GATA6*-GFP reporter line.

(B) Representative FACS plot of SSEA3 vs *GATA6* expression. Left panels show gating controls P3X (above) and TRA-1-85 (below) on the Shef4 parental line. Right panel shows the identification of distinct cell populations: SSEA3 high, *GATA6* negative (3+/6−); SSEA3 high, *GATA6* low (3+/6L); SSEA3 high, *GATA6* high (3+/6H); SSEA3 negative *GATA6* high (3−/6+), of the *GATA6* reporter line.

(C) Representative FACS plots of additional stem cell surface markers, SSEA3, TRA-1-81 or SSEA4 vs *GATA6* expression with the same controls as (B).

whether this co-expression was a feature of just SSEA3, we also examined three other stem cell-associated surface antigens, SSEA4, TRA-1-60, and TRA-1-81 (Adewumi et al., 2007). Similar to SSEA3, these three antigens showed co-expression with *GATA6* (Figure 1C). These results suggest that hESCs exist within substates demarcated by the expression of stem cell surface markers and GATA6, a transcription factor usually associated with endoderm differentiation. This then raised the question of whether GATA6 confers a bias when these cells differentiate.

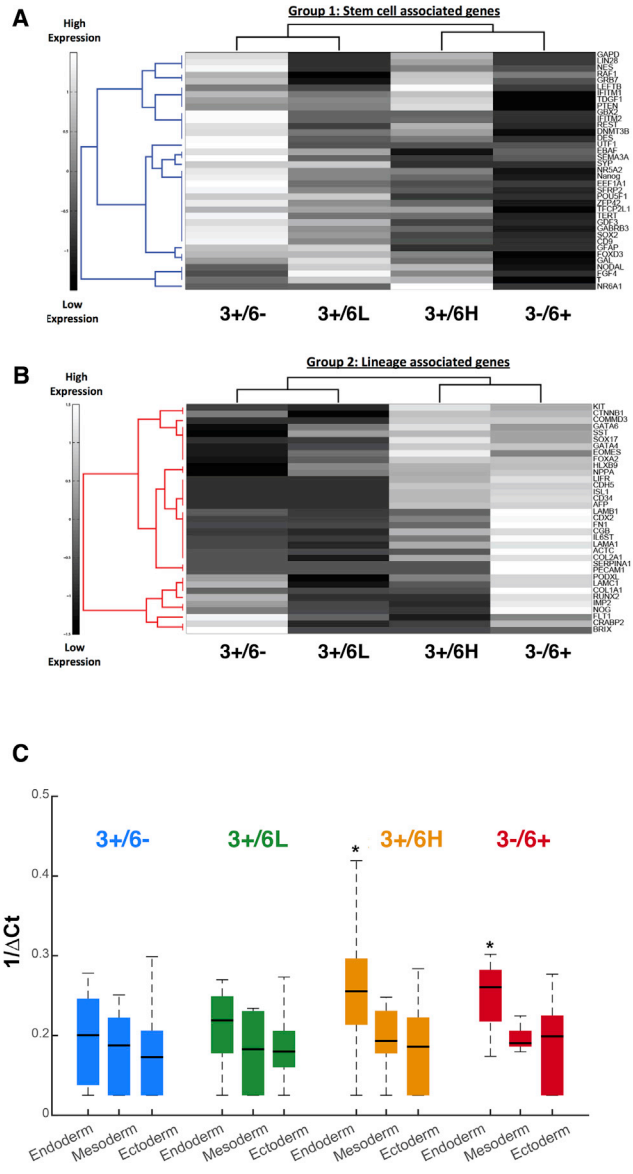


Figure 2. Gene Expression Profiles of Fractions 3+/6−, 3+/6L, 3+/6H, and 3−/6+

(A and B) qPCR using an Applied Biosystems pluripotency TaqMan array on each cell fraction. Hierarchical clustering using Spearman's rank correlation showed strong segregation of genes into two groups: stem cell-associated (group 1) (A) and lineage-associated genes (group 2) (B) with respective gene names. Colormap indicates level of expression of 1/Δ-CT values standardized by row. (C) Boxplot analysis of average gene expression of lineage-specific genes in each cell fraction grouped by specific germ layer. *Kruskal-Wallis statistical test results are indicated for p values <0.05.

GATA6-Expressing Cells Have Gene Expression Patterns Indicative of Early Endoderm Differentiation

To better understand the gene expression differences between the cellular substates we identified, we performed qPCR on the four cell fractions (Figure 1B) using the TaqMan Low Density Pluripotency Array (Adewumi et al., 2007). Hierarchical analysis revealed two major clusters: one cluster, mostly comprising stem cell-related genes, was expressed in the 3+/6− cells, and downregulated in the 3−/6+ subset, whereas a second cluster, mostly comprising various differentiation-related genes, showed the opposite pattern. The 3+/6L and 3+/6H subsets showed intermediate patterns of expression, which could be interpreted to represent intermediate stages in a progression from the 3+/6− state to the 3−/6+ state (Figures 2A and 2B). The changes in expression of a few genes, e.g., *LIN28*, *GRB7*, *NR6A*, and *T*, did not fit this simple progressive view, but most likely this reflects the complexities and persistent heterogeneity of the cell subsets (Figures 2A and S2B). When genes associated *a priori* with endoderm, mesoderm, and ectoderm differentiation were grouped (Adewumi et al., 2007), we found no overall difference between the subsets with respect to mesoderm and ectoderm-related genes, but there was a significant increase in expression of genes associated with endoderm in the 3+/6H and 3−/6+ subsets (Figure 2C). Therefore, *GATA6* expression appeared to be correlated with a reduction in stem cell-associated genes and was coincident with an increase in, specifically, endodermal gene expression.

A Subset of *GATA6*-Expressing Cells Maintain Pluripotency

Whereas the gene expression patterns of the *GATA6*-expressing subsets suggest progressive endoderm differentiation, the continued expression of the stem cell surface antigen, SSEA3, as well as transcription factors, such as OCT4 and SOX2, is consistent with the retention of an undifferentiated hESC phenotype. To test this, we carried out high-content clonogenic assays to test the self-renewal capacity of single cells from the four cell subsets. Cells from each population (3+/6−, 3+/6L, 3+/6H, and 3−/6+) were isolated by fluorescence activated cell sorting (FACS) and seeded at a clonal density (500 cells/cm^2) (Blauwkamp et al., 2012). After 4 days, resulting colonies were immunolabeled for expression of OCT4 or SOX2 and the number and characteristics of the colonies were analyzed using a high-content microscopy platform. Colonies were generated from each substate, including the 3−/6+ subset, though with different efficiencies. The cloning efficiencies of the 3+/6− and 3+/6L subsets were similar at around 6%, whereas the cloning efficiency of the 3+/6H cells was lower at about 2.5% and that of the 3−/6+ cells substantially lower at 0.2% (Figure 3A). We also performed the same experiments on the *GATA6* reporter line S4G6 4/F-9 and found the same trend in cloning efficiencies (Figure S3A), demonstrating no effects resulting from the presence of the selection marker. We next looked at the distribution

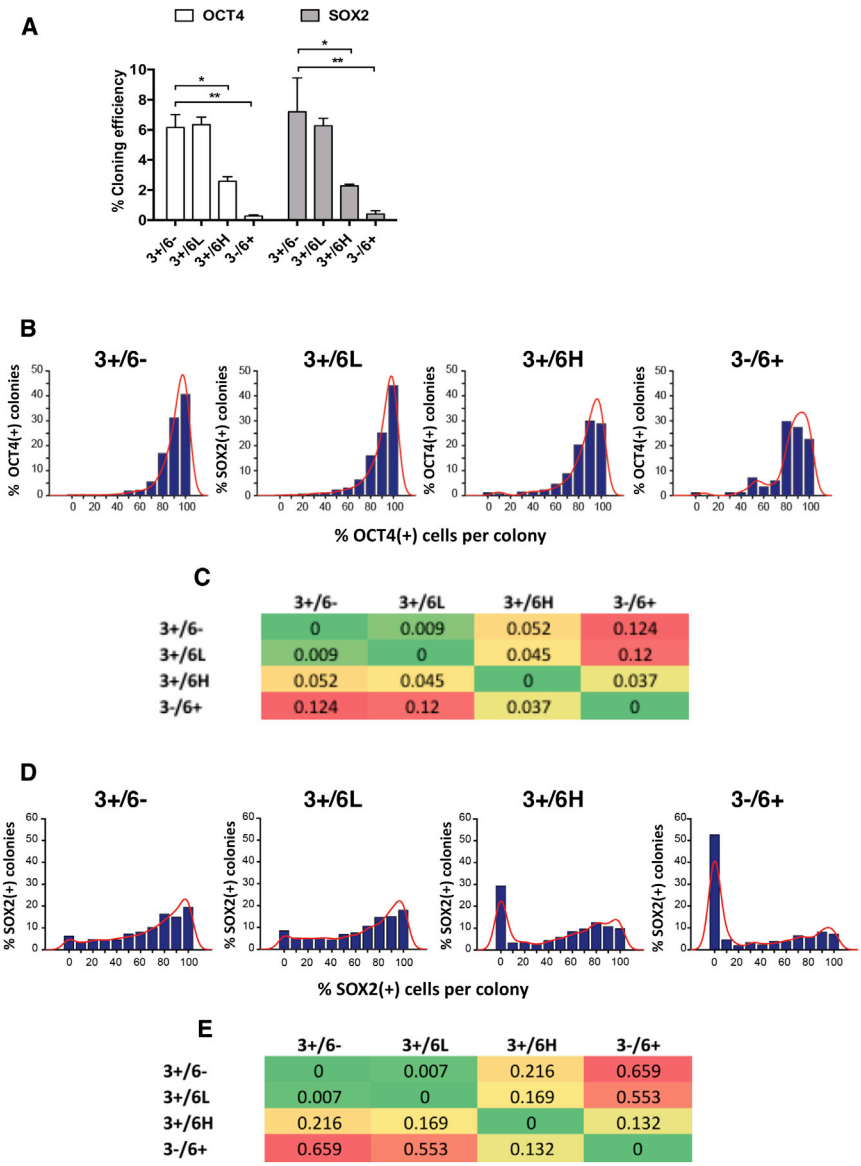


Figure 3. High *GATA6* Expression Results in a Reduced Cloning Efficiency

(A) Percentage cloning efficiency of each cell fraction (3+/6−, 3+/6L, 3+/6H, and 3−/6+) using OCT4 (left) and SOX2 (right) as markers for the stem cell state. Sorted fractions were plated as single cells at clonogenic density in KO/SR and MEF conditions. Cloning efficiency was calculated by dividing the number of OCT4-positive (left) or SOX2-positive (right) colonies by starting seed density. Error bars represent SD of three biological experiments. Student's t test was used to determine significance (OCT4 graph: 3+/6− to 3+/6H *p = 0.0017, 3+/6− to 3−/6+ **p = 0.0001, SOX2 graph; 3+/6− to 3−/6H *p = 0.0019, 3+/6− to 3−/6+ *p = 0.0002).

(B) Proportion of OCT4-positive (OCT4[+]) cells in OCT4-positive colonies derived from single cells from fractions 3+/6−, 3+/6L, 3+/6H, and 3−/6+. Positive colonies include one or more OCT4(+) cells. Counts are shown as bar plots (blue) with superimposed estimated nonparametric distribution (red).

(C) Kullback-Leibler symmetric divergence between OCT4-associated distributions shown in (B). This measure increases with reduced similarity between distributions; zero indicates identical distributions.

(D) Proportion of SOX2(+) (SOX2-positive) cells in SOX2-positive colonies derived from single cells from fractions 3+/6−, 3+/6L, 3+/6H, and 3−/6+. Positive colonies include one or more SOX2(+) cells. Counts are shown as bar plots (blue) with superimposed estimated nonparametric distribution (red).

(E) Kullback-Leibler symmetric divergence between SOX2-associated distributions shown in (D).

of OCT4 expression within colonies from each of the four cell subsets. For each subset, most cells in each colony expressed OCT4, although there was a noticeable downward shift in the proportion of OCT4(+) cells per colony from the 3+/6− and 3+/6L subsets to the 3+/6H and 3−/6+ subsets, as quantified using the Kullback-Leibler divergence analysis (Figures 3B and 3C). A similar pattern was observed with SOX2 expression, although in all cases there was a broader distribution of SOX2 expression and a significant number of colonies, especially from the 3+/6H and 3−/6+ subsets, contained only SOX2-negative cells, likely due to the absence of SOX2 expression in endoderm differentiation (Adachi et al., 2010) (Figures 3D and 3E). Thus, from these functional studies, we found that the 3+/6H and 3−/6+ subsets had a reduced cloning efficiency, implying a greater tendency to differentiate. Nevertheless, a proportion of cells within these subsets retained the ability to remain within the stem cell compartment and self-renew irrespective of their high *GATA6* expression.

As a more robust assay for confirming that *GATA6*-expressing stem cells were indeed *bona fide* stem cells, we sorted single cells from each subset (3+/6−, 3+/6L, and 3+/6H) into individual wells of a 96-well plate to generate clonal lines. From this, we obtained respectively 43, 76, and 49 clones from 288, 960, and 1,920 cells deposited, equivalent to cloning efficiencies of 15%, 8%, and 3% (Figure S3B). We did not include the 3−/6+ fraction in this part of the study due to its very low cloning efficiency. To check

the accuracy of FACS sorting, we used exactly the same conditions to sort mixtures of Chinese hamster ovary (CHO) cells, stably transfected to constitutively express GFP or Tomato fluorescent protein, alongside the sorting for the stem cell fractions (Figure S3C). Using this CHO assay, we detected a misclassification rate of only 1 in every 166 cells sorted (0.6%) (Figure S3D). Based on this rate, as well as the fact that CHO cells have a much higher cloning efficiency than hESCs, thereby over-representing misclassification, we concluded that it was highly unlikely that any clones from the *GATA6*-positive fractions arose from misclassified *GATA6*-negative cells.

All of the clones obtained from each subset grew with a characteristic morphology consistent with that of undifferentiated stem cells (Figure S3E). To confirm this phenotype, six clones were picked from each subset and passaged for a minimum of eight passages with no loss of stem cell morphology. Between passages 5 and 8, two representative clones from each fraction were analyzed by flow cytometry and qPCR for stem cell attributes. Irrespective of the subset of origin, all clones showed similar patterns of SSEA3, TRA-1-81, and SSEA4 expression to that of the unsorted stem cell line (Figure 4A), and expressed similar levels of core stem cell transcription factors OCT4, NANOG, SOX2, and REX1 (Figure 4B). Additionally, gene expression for germ layer differentiation within all subclones was low and comparable to the unsorted line (Figure 4C). To ensure that the clones from each fraction were pluripotent, two representative clones from each were induced to differentiate through a defined, neutral embryoid body differentiation protocol (Ng et al., 2008). Each clone, irrespective of the starting cell, showed strong upregulation of genes associated with mesoderm and ectoderm, demonstrating pluripotency (Figure 4D). Thus, clonal lines generated from hESCs expressing *GATA6* at low and high levels were *bona fide* pluripotent stem cells. Finally, the clones, irrespective of their original *GATA6* status, were able to reconstitute entirely the original culture heterogeneity, so that they were indistinguishable from the starting population after five passages, demonstrating that the *GATA6*-positive substate within the stem cell compartment is interconvertible (Figures 4E and S4).

GATA6-Expressing hESCs Are Biased Toward Endoderm Differentiation

To test whether the hESC subsets expressing *GATA6* exhibit a bias in their propensity to differentiate toward endodermal derivatives, cells from each fraction were isolated by FACS and allowed to differentiate using a defined spin-embryoid body (EB) system without the addition of exogenous proteins or small molecules to direct differentiation. The resulting EBs exhibited structural organization consisting of an inner, middle, and outer mass of cells but there were marked differences in the morphology depending upon the subset of cells from which they were derived. EBs from the 3+/6− and the 3+/6L subsets were similar with a dense, compacted morphology and clear borders. By contrast, the EBs from the 3+/6H and 3−/6+ subsets were much more cystic and showed less structural organization (Figure S5A).

Next, we performed qPCR on day 10 EBs from each subset. Compared with EBs of the 3+/6− subset, EBs from all of the *GATA6* expressing subsets, including 3+/6L cells, showed a marked upregulation of endoderm (*GATA4*, *GATA6*, *AFP*, *SOX17*, *FOXA2*, *SOX7*) and mesoderm-associated genes (*CD4*, *PECAM*, *KDR*, and *DESMIN*). Exceptions were reduced levels of *GATA4*, *SOX7*, and *CD34* in the EBs from the 3−/6+ subsets, potentially due to these cells being further along in differentiation, past the point of normal developmental expression of these genes. By contrast, genes associated with ectodermal differentiation (*SOX2*, *PAX6*, *TH*, and *SOX1*) were markedly downregulated in EBs from the *GATA6*-expressing subsets, with a notable gradation from 3+/6L to 3+/6H and the 3−/6+ derived EBs (Figure 5A).

These results indicate that, on a population basis, the *GATA6* expressing subsets show a strong bias toward endoderm and mesoderm differentiation, at the expense of ectoderm differentiation. Together with the data that these subsets also contain long-term self-renewing undifferentiated stem cells, the results are consistent with the conclusion that, within the stem cell compartment, undifferentiated hESCs can transit reversibly between *GATA6*-positive and *GATA6*-negative substates, but while in these substates they exhibit a differential bias in the pathways of differentiation they are likely to follow. However, the possibility that the *GATA6*-positive subsets contain both undifferentiated, unbiased stem cells together with cells already committed to an endodermal fate, cannot be excluded and may account for the differentiation bias. To address this, we carried out a high-content clonogenic assay to assess the differentiation propensity of individual hESCs under conditions that did permit limited spontaneous differentiation.

Cells from the 3+/6−, 3+/6L, 3+/6H, and 3−/6+ subsets were isolated by FACS and seeded at a clonogenic density of 500 cells/cm^2 into self-renewing conditions (Barbaric et al., 2014; Blauwkamp et al., 2012). The resulting colonies were dual stained for expression of OCT4, as an indicator of undifferentiated stem cells, and an early endodermal marker, SOX17 or GATA4. Four emerging colony types with respect to SOX17 were apparent, and classified as OCT4(+)/SOX17(−), OCT4(−)/SOX17(+), OCT4(+)/SOX17(+) and OCT4(−)/SOX17(−) (Figure 5B). A similar set with respect to GATA4 expression was also identified (not shown).

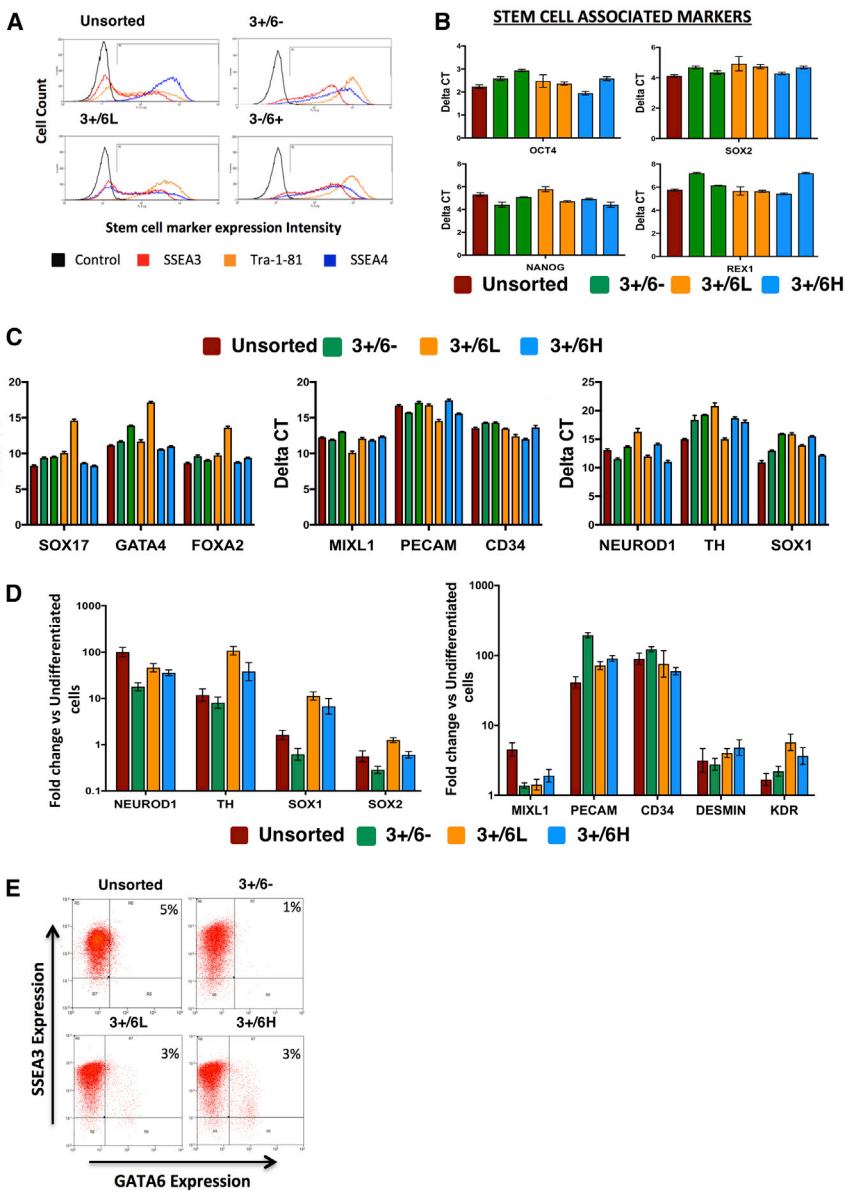


Figure 4. Stable, Long-Term Self-Renewing hESC Subclones Can Be Derived from *GATA6*-Expressing Cells

(A) Flow cytometric analysis of subclones derived from the 3+/6−, 3+/6L, and 3+/6H fractions. Unsorted represents the unsorted cells of the reporter line. P3X was used as a negative control, and markers SSEA3 (red), TRA-1-81 (orange), and SSEA4 (blue) were used to identify stem cells. FACS plots show one clone from each fraction, which is representative of four clones analyzed from each fraction.

(B) qPCR analysis of two subclones from each fraction for core stem cell transcription factors, shown as Delta-CT normalized to beta-actin; error bars are the SD from three technical repeats. Red bar represents the reporter line, and individual 3+/6−, 3+/6L, and 3+/6H subclones are shown by green, orange, and blue bars, respectively.

(C) qPCR for lineage-specific markers of each germ layer in unsorted (red) and two subclones from each fraction. Bar color as in (B), showing Delta-CT normalized to beta-actin with three technical repeats.

(D) qPCR of day 10 EBs from unsorted (red), and subclones from 3+/6− (green), 3+/6L (orange), and 3+/6H (blue) fractions for genes specifying mesoderm (left panel) and ectoderm (right panel) to demonstrate pluripotency of the lines. Data shown as fold change against undifferentiated cells from the same starting population. Error bars are SD of three technical repeats.

(E) Flow cytometric analysis of reporter line (top left) and 3+/6− (top right), 3+/6L (bottom left), and 3+/6H (bottom right) subclones for SSEA3 versus *GATA6* expression 5–8 passages after initial single-cell seeding. Gates were set using P3X and Shef4 parental line as SSEA3 and GFP negative controls respectively. Plot shows one clone representative of four clones analyzed from each fraction.

OCT4(+)/SOX17(−) colonies or OCT4(+)/GATA4(−) colonies predominated among those derived from 3+/6− and 3+/6L cells compared with fewer such undifferentiated colonies from the 3+/6H and 3−/6+ subsets, as we previously observed. On the other hand, considerably more colonies that contained SOX17 or GATA4 expressing cells were found among those originating from 3+/6H or 3−/6+ cells, consistent with the population differentiation data. Importantly, however, among these fractions was also a higher proportion of SOX17 or GATA4-expressing colonies that also contained OCT4-expressing cells, particularly in the colonies derived from 3+/6H cells (Figures 5C and 5D). We also repeated this experiment on the SG4 4/F-9 reporter clone and found a similar trend (Figure S5B). By looking at the distribution of SOX17 or GATA4 in OCT4-positive colonies from each subset, we also found that there was a small yet distinct increase in the proportion of SOX17(+) or GATA4(+) cells per OCT4-positive colony within the 3+/6H and 3−/6+ biased fractions (Figures 5E, 5F, S5C, and S5D). Taken together, these results indicate that the 3+/6H and even the 3−/6+ subsets contain individual undifferentiated stem cells that exhibit an endoderm differentiation bias.

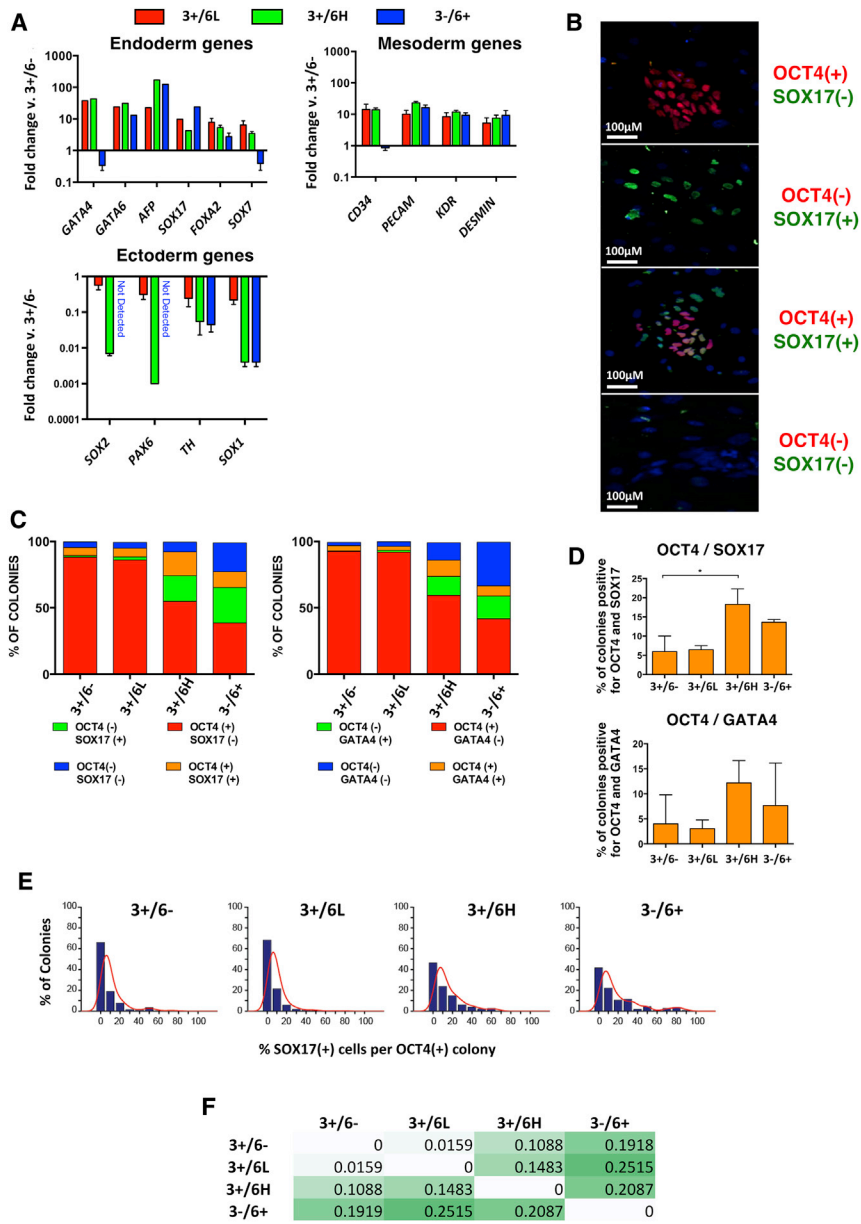


Figure 5. High *GATA6* Expression Results in Endoderm Differentiation Bias at Population and Single-Cell Level

(A) qPCR of differentiating cells from the 3+/6L (red), 3+/6H (green), 3-/6+ (blue) fractions in a non-directed EB differentiation assay, shown as fold change against differentiating cells from 3+/6− fraction, for genes expressed in endoderm (top left), mesoderm (top right), and ectoderm (bottom left). Beta-actin was the normalizing gene. Error bars represent three biological replicates.

(B) Representative images of colonies derived from 3+/6−, 3+/6L, 3+/6H, and 3−/6+ fractions. Images were taken at ×10 magnification on an InCell Analyzer 2000 and automated quantitative analysis performed using developer toolbox software. The same algorithms were used for each technical and biological repeat and the process was automated to eliminate human bias.

(C) Quantification of colony types from 3+/6−, 3+/6L, 3+/6H, and 3−/6+ fractions showing the percentage of colonies per fraction with colony phenotype shown in (B) from three biological repeats.

(D) Percentage of colonies containing OCT4 and SOX17 (top graph) or OCT4 and GATA4 (bottom graph) positive cells only. Significance was calculated using t test of three biological replicates and stars represent degree of significance (*$p < 0.05$). Numbers for each fraction: 3+/6− = 83, 3+/6L = 103, 3+/6H = 122, 3−/6+ = 66.

(E) Histogram showing the distribution of SOX17(+) cells in OCT4-positive colonies resulting from single cells from 3+/6−, 3+/6L, 3+/6H, and 3−/6+ fractions. Positive colonies include at least two OCT4(+) cells. Counts are shown as a bar plot (blue) with superimposed estimated nonparametric distribution (red).

(F) Kullback-Leibler symmetric divergence between SOX17-associated distributions in OCT4-positive colonies. This measure increases with reduced similarity between distributions; zero indicates identical distributions.

Single-Cell Transcriptomic Analysis of Endodermally Biased hESCs

Having established at the single-cell level that a distinct endoderm-biased substate exists within the stem cell compartment, and with evidence that these four cell fractions represent discrete developmental stages (Figures 2A and 2B), we performed single-cell RNA sequencing, using the Drop-seq methodology (Macosko et al., 2015) on each of the four cell fractions to gain a mechanistic understanding of the populations of cells comprising each fraction. Using tSNE (t-distributed stochastic neighbor embedding) analysis, we defined 13 distinct cell clusters comprising 3,500 cells from all four cell fractions (Figure 6A). We mapped clusters back to cell fraction of origin, and found that clusters were generally fraction specific, so that 3+/6− were confined to clusters 1 and 2, 3+/6L to clusters 1, 5 and 6, 3+/6H to clusters 8, 11, 12, 13, and 3−/6+ to clusters 7, 9, and 10 (Figure 6A). Nevertheless, we saw some

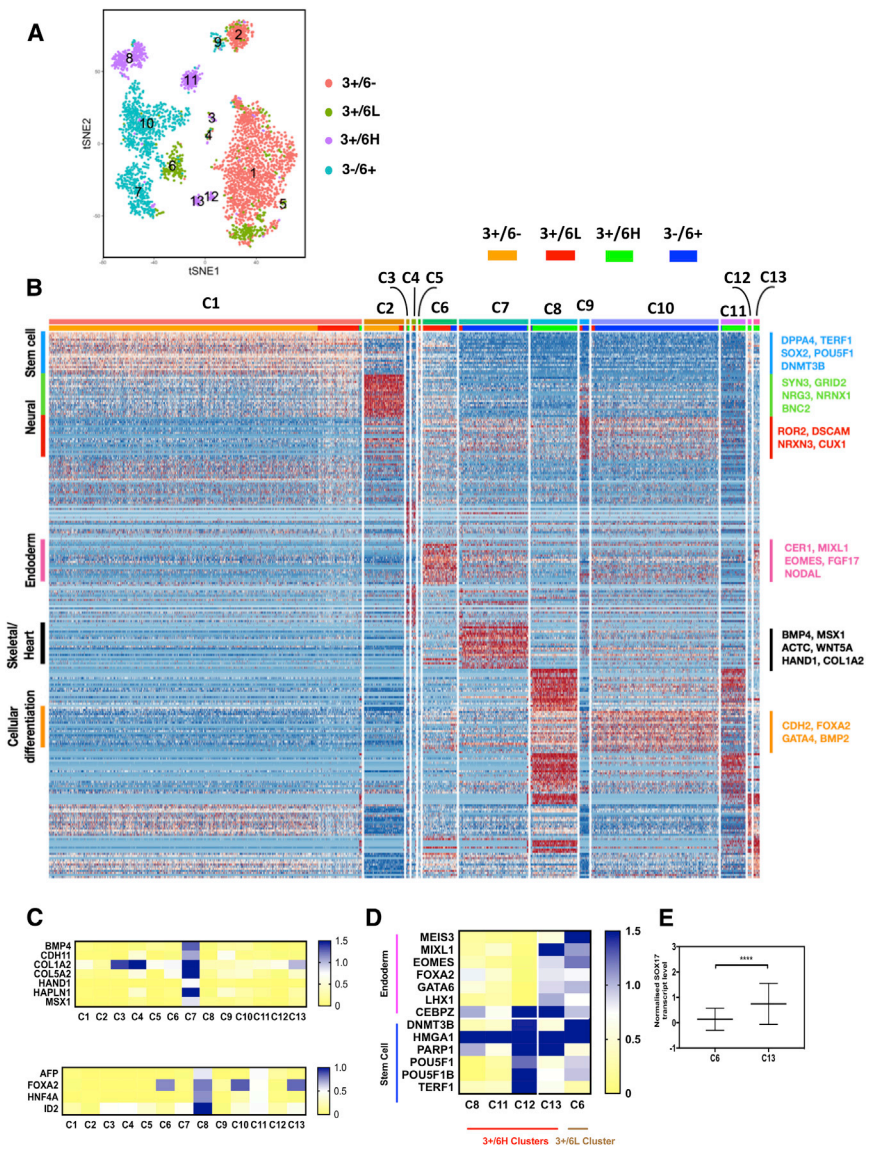


Figure 6. Single-Cell Transcriptomic Analysis of Endodermally Biased hESCs

(A) tSNE analysis of the four cellular subsets representing 13 putative clusters separated according to gene expression per single cells. Single cells are represented by individual dots and are colored according to cell fraction library. Numbers represent cluster number assigned arbitrarily.

(B) Heatmap of the top 30 most differentially expressed genes between the 13 individual putative clusters, as described in Figure 6A. Color scheme is based on *Z* score distribution from −2 (blue) to +2 (red). Right margin color bars represent gene sets specific to each cluster. Left margin color bars represent top Gene Ontology terms of the top 30 most differentially expressed genes for each cluster.

(C) Heatmap of the average expression of genes typically associated with later developmental processes including heart/skeletal (top) and hepatic (lower) lineages across the 13 clusters. Color scheme is based on the averaged normalized expression of each gene from no expression (yellow) to expression (blue).

(D) Heatmap of the average expression of genes associated with both the stem cells and early differentiating cells across all the individual 3+/6H clusters and the 3+/6L cluster. Color scheme is based on the averaged normalized expression of each gene from no expression (yellow) to expression (blue).

(E) Scatter dot plot to show the mean, upper, and lower limit expression of SOX17 between cluster 6 of the 3+/6L and cluster 13 of the 3+/6H fractions. Student's t test was used to determine statistical significance of ****$p > 0.0001$.

overlap of cell fractions within single clusters, particularly for the 3+/6L fraction in clusters 1, 2, 7, 9, and 10, and 3−/6+ in clusters 6 and 8. To ensure these observations were not due to FACS sorting misclassification, we looked at the expression of *GATA6* across the tSNE space and found that *GATA6* was only expressed in clusters composed of GFP(+) sorted cells (Figure S6A). Further, other endoderm-specific genes were only present in GFP(+) sorted cells and strongly correlated with *GATA6* expression (Figure S6B). Thus, the single-cell data showed further heterogeneity within sorted cell fractions as evidenced by the generation of multiple clusters per fraction, and it was apparent that some cells within a fraction showed more transcriptomic similarities to cells of other fractions.

As an unbiased approach to investigate which cell types were being generated in the cell fractions, we performed cluster-specific binomial differential gene expression analysis. We found that the 3+/6− fraction in cluster 1 showed the highest level of stem cell-associated gene expression. Interestingly, we found that cluster 2, although derived from the 3+/6− fraction, showed strong upregulation of neural associated genes including *SYN3*, *GRID2*, and *NRG3*. Cells of the 3+/6L fraction, which showed similar self-renewal behavior to cells of 3+/6− were split between cluster 1 and cluster 6, whereby both clusters showed high stem cell gene expression. The 3+/6L cells within cluster 6, however, also expressed high levels of early endoderm-associated genes, including *EOMES*, *FGF17*, *NODAL*,

and *LEFTY1*, which may account for the observed endoderm differentiation bias within our neutral EB differentiation assay. The 3+/6H fraction, except clusters 12 and 13, and the 3−/6+ fractions consisted of cells with low stem cell expression yet high expression of genes involved in cellular differentiation, gastrulation, and endoderm (clusters 7–11), consistent with their general lack of ability for self-renewal (Figure 6B). Additionally, it was apparent that *GATA6* expression correlated with multiple lineages, including mesoderm (cluster 7), and definitive endoderm (clusters 8 and 11) differentiation, although we found no strong evidence for primitive endoderm by *SOX7* expression (Figure S6C). Further, cells within the 3+/6H and 3−/6+ fractions generated clusters that showed higher expression of more mature endoderm-associated genes (*AFP*, *FOXA2*, *ID2*, and *HNF4A*; cluster 8) and mesoderm-associated genes (*MSX1*, *HAND1*, *CDH11*, and *ALPK2*; cluster 7) (Figure 6C), confirming our previous observations that these cell fractions represent a later developmental time point than the 3+/6L and 3+/6− fractions. Thus, these data enabled us to capture discrete subpopulations of cells progressing along a developmental trajectory that correlates with the increased expression of *GATA6* and the subsequent loss of SSEA3. We next sought to identify which cluster may represent the endoderm-biased stem cells of the 3+/6H fraction. Of all the clusters composed of 3+/6H cells, only cluster 13 had robust and significant co-expression of both endoderm and stem cell genes (Figure 6D). Further, cells within cluster 13 also showed co-expression of *OCT4*, *SOX2*, and *GATA6*, indicative of mesendodermally biased cells (Nazareth et al., 2013) (Figure S6E). This cluster, however, was not unique in the sense that cluster 6 of the 3+/6L fraction also showed strong co-expression of genes for these opposing lineages, but did not show functional bias toward endoderm differentiation in our single-cell assay. To investigate what specific genes may be driving this unique biased state of the 3+/6H cells at single-cell level, we performed pairwise differential expression analysis between clusters 6 and 13, and then filtered results for transcription factors. We identified one transcription factor gene, *SOX17*, as significantly more highly expressed in the 3+/6H fraction compared with the 3+/6L (Figure 6E). Therefore, it appears that the retention of expression of stem cell genes is imperative to remain within the stem cell compartment, and *SOX17* may be a main driving force for cells to enter an endoderm-biased substate.

DISCUSSION

Using a *GATA6*-GFP reporter line, we have corroborated our previous observations (Gokhale et al., 2015) and confirmed in live cells that *GATA6* is heterogeneously expressed in a subset of cells alongside the surface stem cell marker SSEA3. GATA6 is a key lineage-associated transcription factor implicated in specifying the endoderm lineage during the segregation of the inner cell mass and extra-embryonic lineages in the blastocyst; later during gastrulation, it is expressed in cells of the lateral plate mesoderm (Koutsourakis et al., 1999). On the other hand, SSEA3 is associated with a cell surface globoseries glycolipid expressed by undifferentiated hESCs (Andrews et al., 1982; Kannagi et al., 1983). Compared with other surface markers of these cells, SSEA3 is lost most quickly upon differentiation (Draper et al., 2002; Enver et al., 2005; Fenderson et al., 1987).

Our results demonstrate that undifferentiated hESCs can transiently express a lineage regulatory transcription factor, GATA6, while retaining the capacity for long-term self-renewal. Further, these undifferentiated stem cells can oscillate between a *GATA6*-positive and *GATA6*-negative expression state. Also, on a population basis, when differentiation was induced by EB formation, the *GATA6*-positive cells showed a greater propensity to differentiate toward endoderm-related lineages, than do the *GATA6*-negative cells, which appear to exhibit a greater propensity for ectodermal differentiation. Further, qPCR analysis of these subsets demonstrated that the increased expression of *GATA6* correlated with the increased expression of genes involved in early gastrulation and differentiation. More specifically, genes associated with endoderm but not mesoderm or ectoderm were upregulated, suggesting directional activation of an endodermal program. This pattern of gene upregulation is consistent with the role of GATA6 in the early specification of extra-embryonic endoderm and definitive endoderm during mouse gastrulation (Chazaud et al., 2006; Koutsourakis et al., 1999; Plusa et al., 2008), as well as the expression of *GATA6* in hESC-derived definitive endoderm (McLean et al., 2007). The subsets revealed a clear hierarchy of cells in culture such that the 3+/6− and 3−/6+ fractions showed quite opposite gene expression patterns, with the 3+/6− subset representing a more pristine stem cell state and the 3−/6+, a more differentiated state, with the 3+/6L and 3+/6H subsets in between. The cloning efficiency of these subsets similarly reduced progressively from the 3+/6− subset through the 3+/6L and 3+/6H subsets and was lowest in the 3−/6+ subset implying a corresponding reduction in the proportions of clonogenic stem cells in each subset.

The reduced cloning efficiency and increased propensity for endoderm differentiation of the *GATA6*-positive subsets could be explained by a lineage bias in self-renewing stem cells that co-express pluripotent associated and lineage-associated genes, with a corresponding reduction in cloning efficiency, or it could reflect the presence of two further subsets within each of the 3+/6L and 3+/6H subsets, one

self-renewing but not lineage biased and one not self-renewing but committed progenitor cells, as reported by Pina et al. (2012) for hematopoietic stem cells. These possibilities are not mutually exclusive. Unfortunately, given the low plating efficiency of hESCs, it is not possible to conclude directly from population-level data whether this population bias reflects a differentiation bias at the level of individual self-renewing stem cells.

However, using OCT4 or SOX2 as surrogate markers of self-renewing stem cells, in addition to SSEA3, we were able to show that single-cell-derived colonies that we classified as arising from self-renewing stem cells contained more spontaneously differentiated cells of the endoderm pathway, marked either by SOX17 or GATA4, when derived from the 3+/6L or 3+/6H subsets, than when derived from the 3+/6− subset. Further, many of the cells within each colony expressed these stem cell markers implying continued expression through at least four to five cell divisions. We conclude that the colonies classified as OCT4-positive or SOX2-positive were derived from self-renewing undifferentiated embryonic stem cells, and that not only can self-renewing stem cells express the lineage regulator transcription factor GATA6 but also that its expression does increase the probability of those stem cells following an endoderm route when they commit to differentiation.

By single-cell RNA sequencing we are able to identify single cells co-expressing both stem cell- and endoderm-specific genes, beyond that of SSEA3 and *GATA6* alone. Using tSNE analysis, we found that almost all 3+/6L and a small proportion of 3+/6H cells retained the expression of key stem cell-associated genes, likely representing cells within the stem cell compartment and in line with our functional data. In particular, we found that the co-expression of *OCT4* and *SOX2* was strongly retained within self-renewing associated clusters but lost in all other clusters, implicating an important role for OCT4 and SOX2 in the ability for endoderm gene expressing cells to remain within the stem cell compartment. This is supported by the established role of OCT4 and SOX2 as master regulators of the stem cell state (Buitrago and Roop, 2007; Huangfu et al., 2008; Takahashi et al., 2007). Thus, the status of OCT4/SOX2 expression may dictate cellular residence inside or outside of the stem cell compartment. We also found that the 3+/6− fraction showed heterogeneity, consistent with previous functional reports (Tonge et al., 2011). Interestingly, we found a subset of cells, approximately 11% of the 3+/6− fraction, with neural gene expression profiles. We also found co-expression of stem cell genes alongside mesodermal-associated genes, so one could imagine a system that contains pluripotent stem cells biased toward each primary germ layer. To support this hypothesis, however, further work is required to elucidate whether these cells also identify functional lineage-biased substates within the stem cell compartment.

Hierarchies of human pluripotent stem cells based on the co-expression levels of the surface stem cell markers GCTM-2 and CD9 have also shown the existence of lineage marker expression in stem cell populations, albeit with little functional relevance (Hough et al., 2009, 2014). It was suggested that cultures of these cells contain metastable self-renewing cells in a continuum with intermediate pluripotent states that eventually become primed for lineage specification. Our results are similarly consistent with a continuum in which the self-renewing capacity of the stem cells diminishes as they progressively acquire lineage-associated features while retaining the ability to revert to a more pristine, less lineage-associated, state. Evidently, heterogeneity has functional relevance to the behavior of hESCs. With substantial evidence for functional substates within the stem cell compartment, a deeper understanding of the mechanisms that govern and stabilize these substates would offer a new level of control for the efficient and uniform differentiation of hESCs, and so facilitate the development of applications such as in regenerative medicine.

EXPERIMENTAL PROCEDURES

Cell Culture

The Shef4 hESC line (Aflatoonian et al., 2010) and its derivatives were cultured on mitomycin C inactivated mouse embryonic fibroblasts in Knockout DMEM with 20% Knockout serum replacement as previously described (Draper et al., 2002) or in feeder-free conditions using E8 medium and vitronectin (Life Technologies). Embryoid bodies were produced and grown in the serum-free, defined medium, APEL (Stem Cell Technologies), as described by Ng et al. (2008). See Supplemental Experimental Procedures for more details.

Generation of *GATA6*-GFP Reporter hESCs

GATA6 reporter Shef4 hESCs were generated using a standard gene targeting replacement vector designed to insert an GFP reporter cassette by homologous recombination into exon 2 of the human *GATA6* locus at the position of the ATG translational initiation codon. See Supplemental Experimental Procedures for more details.

Immunoassays, Flow Cytometry, and Cell Sorting

For details including a list of antibodies, see Supplemental Experimental Procedures.

Gene Expression Analysis

Quantitative real-time PCR was performed on the QuantStudio 12K Flex Real-Time PCR system (Invitrogen) using TaqMan universal master mix (Invitrogen) in conjunction with the Roche universal probe library system (Roche). Drop-seq analysis was carried out as described in Macosko et al. (2015). For full details including a list

of qPCR primers, and analytical methods, see Supplemental Experimental Procedures.

Statistical Analysis

For full details of statistical tests including clustering, boxplot, Kullback-Leibler divergence analysis, and tSNE analysis of single-cell RNA-sequencing data, see Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession number for the RNA-sequencing data reported in this paper is GEO: GSE113168.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and six figures and can be found with this article online at https://doi.org/10.1016/j.stemcr.2018.04.015.

AUTHOR CONTRIBUTIONS

T.A. devised experimental plans and performed the bulk of the experiments and interpretation of the results within this manuscript. A.J.H.S., J.S.S., and K.A. generated the targeting construct and subsequently the *GATA6* reporter line. V.B. and S.S. provided the bioinformatics analyses. J.L. provided detailed instructions to establish the single-cell drop-seq technique. D.S. and J.H. performed qPCRs to satisfy reviewer comments. M.J. performed FACS sorting and analysis. K.P., D.C., I.B., P.G., and P.W.A. are principal investigators who helped devise and interpret all experiments and results.

ACKNOWLEDGMENTS

T.A. was a recipient of a BBSRC PhD studentship. This work was supported in part by grants from the European Community's Sixth and Seventh Framework Programs (LSHG-CT-2006-018739 and FP7/2007-2013 agreement no. 602423) and the MRC through the UK Regenerative Medicine Platform (grant no. MR/L012537/1).

Received: September 9, 2015
Revised: April 17, 2018
Accepted: April 17, 2018
Published: May 17, 2018

REFERENCES

Adachi, K., Suemori, H., Yasuda, S.Y., Nakatsuji, N., and Kawase, E. (2010). Role of SOX2 in maintaining pluripotency of human embryonic stem cells. Genes Cells *15*, 455–470.

Adewumi, O., Aflatoonian, B., Ahrlund-Richter, L., Amit, M., Andrews, P.W., Beighton, G., Bello, P.A., Benvenisty, N., Berry, L.S., Bevan, S., et al. (2007). Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. Nat. Biotechnol. *25*, 803–816.

Aflatoonian, B., Ruban, L., Shamsuddin, S., Baker, D., Andrews, P., and Moore, H. (2010). Generation of Sheffield (Shef) human embryonic stem cell lines using a microdrop culture system. In Vitro Cell. Dev. Biol. Anim. *46*, 236–241.

Andrews, P.W., Goodfellow, P.N., Shevinsky, L.H., Bronson, D.L., and Knowles, B.B. (1982). Cell-surface antigens of a clonal human embryonal carcinoma cell line: morphological and antigenic differentiation in culture. Int. J. Cancer *29*, 523–531.

Arias, A.M., and Brickman, J.M. (2011). Gene expression heterogeneities in embryonic stem cell populations: origin and function. Curr. Opin. Cell Biol. *23*, 650–656.

Barbaric, I., Biga, V., Gokhale, P.J., Jones, M., Stavish, D., Glen, A., Coca, D., and Andrews, P.W. (2014). Time-lapse analysis of human embryonic stem cells reveals multiple bottlenecks restricting colony formation and their relief upon culture adaptation. Stem Cell Reports *3*, 142–155.

Blauwkamp, T.A., Nigam, S., Ardehali, R., Weissman, I.L., and Nusse, R. (2012). Endogenous Wnt signalling in human embryonic stem cells generates an equilibrium of distinct lineage-specified progenitors. Nat. Commun. *3*, 1070.

Buitrago, W., and Roop, D.R. (2007). Oct-4: the almighty POUripotent regulator? J. Invest. Dermatol. *127*, 260–262.

Canham, M.A., Sharov, A.A., Ko, M.S., and Brickman, J.M. (2010). Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. PLoS Biol. *8*, e1000379.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature *450*, 1230–1234.

Chang, H.H., Hemberg, M., Barahona, M., Ingber, D.E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. Nature *453*, 544–547.

Chazaud, C., Yamanaka, Y., Pawson, T., and Rossant, J. (2006). Early lineage segregation between epiblast and primitive endoderm in mouse blastocysts through the Grb2-MAPK pathway. Dev. Cell *10*, 615–624.

Draper, J.S., Pigott, C., Thomson, J.A., and Andrews, P.W. (2002). Surface antigens of human embryonic stem cells: changes upon differentiation in culture. J. Anat. *200*, 249–258.

Enver, T., Pera, M., Peterson, C., and Andrews, P.W. (2009). Stem cell states, fates, and the rules of attraction. Cell Stem Cell *4*, 387–397.

Enver, T., Soneji, S., Joshi, C., Brown, J., Iborra, F., Orntoft, T., Thykjaer, T., Maltby, E., Smith, K., Dawud, R.A., et al. (2005). Cellular differentiation hierarchies in normal and culture-adapted human embryonic stem cells. Hum. Mol. Genet. *14*, 3129–3140.

Fenderson, B.A., Andrews, P.W., Nudelman, E., Clausen, H., and Hakomori, S.I. (1987). Glycolipid core structure switching from globo-to lacto-and ganglio-series during retinoic acid-induced differentiation of TERA-2-derived human embryonal carcinoma cells. Dev. Biol. *122*, 21–34.

Gokhale, P.J., Au-Young, J.K., Dadi, S., Keys, D.N., Harrison, N.J., Jones, M., Soneji, S., Enver, T., Sherlock, J.K., and Andrews, P.W. (2015). Culture adaptation alters transcriptional hierarchies among single human embryonic stem cells reflecting altered patterns of differentiation. PLoS One *10*, e0123467.

Hayashi, K., de Sousa Lopes, S.M.C., Tang, F., and Surani, M.A. (2008). Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. Cell Stem Cell *3*, 391–401.

Hough, S.R., Laslett, A.L., Grimmond, S.B., Kolle, G., and Pera, M.F. (2009). A continuum of cell states spans pluripotency and lineage commitment in human embryonic stem cells. PLoS One *4*, e7708.

Hough, S.R., Thornton, M., Mason, E., Mar, J.C., Wells, C.A., and Pera, M.F. (2014). Single-cell gene expression profiles define self-renewing, pluripotent, and lineage primed states of human pluripotent stem cells. Stem Cell Reports *2*, 881–895.

Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C., and Enver, T. (1997). Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev. *11*, 774–785.

Huang, S., Guo, Y.-P., May, G., and Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. Dev. Biol. *305*, 695–713.

Huangfu, D., Osafune, K., Maehr, R., Guo, W., Eijkelenboom, A., Chen, S., Muhlestein, W., and Melton, D.A. (2008). Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. Nat. Biotechnol. *26*, 1269–1275.

Kannagi, R., Cochran, N.A., Ishigami, F., Hakomori, S.-i., Andrews, P., Knowles, B.B., and Solter, D. (1983). Stage-specific embryonic antigens (SSEA-3 and-4) are epitopes of a unique globo-series ganglioside isolated from human teratocarcinoma cells. EMBO J. *2*, 2355.

Koutsourakis, M., Langeveld, A., Patient, R., Beddington, R., and Grosveld, F. (1999). The transcription factor GATA6 is essential for early extraembryonic development. Development *126*, 723–732.

Laslett, A.L., Grimmond, S., Gardiner, B., Stamp, L., Lin, A., Hawes, S.M., Wormald, S., Nikolic-Paterson, D., Haylock, D., and Pera, M.F. (2007). Transcriptional analysis of early lineage commitment in human embryonic stem cells. BMC Dev. Biol. *7*, 12.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

McLean, A.B., D'Amour, K.A., Jones, K.L., Krishnamoorthy, M., Kulik, M.J., Reynolds, D.M., Sheppard, A.M., Liu, H., Xu, Y., Baetge, E.E., et al. (2007). Activin a efficiently specifies definitive endoderm from human embryonic stem cells only when phosphatidylinositol 3-kinase signaling is suppressed. Stem Cells *25*, 29–38.

Murry, C.E., and Keller, G. (2008). Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. Cell *132*, 661–680.

Nazareth, E.J., Ostblom, J.E., Lücker, P.B., Shukla, S., Alvarez, M.M., Oh, S.K., Yin, T., and Zandstra, P.W. (2013). High-throughput fingerprinting of human pluripotent stem cell fate responses and lineage bias. Nat. Methods *10*, 1225–1231.

Ng, E.S., Davis, R., Stanley, E.G., and Elefanty, A.G. (2008). A protocol describing the use of a recombinant protein-based, animal product-free medium (APEL) for human embryonic stem cell differentiation as spin embryoid bodies. Nat. Protoc. *3*, 768–776.

Pina, C., Fugazza, C., Tipping, A.J., Brown, J., Soneji, S., Teles, J., Peterson, C., and Enver, T. (2012). Inferring rules of lineage commitment in haematopoiesis. Nat. Cell Biol. *14*, 287–294.

Plusa, B., Piliszek, A., Frankenberg, S., Artus, J., and Hadjantonakis, A.-K. (2008). Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. Development *135*, 3081–3091.

Semrau, S., and van Oudenaarden, A. (2015). Studying lineage decision-making in vitro: emerging concepts and novel tools. Annu. Rev. Cell Dev. Biol. *31*, 317–345.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

Tonge, P.D., Olariu, V., Coca, D., Kadirkamanathan, V., Burrell, K.E., Billings, S.A., and Andrews, P.W. (2010). Prepatterning in the stem cell compartment. PLoS One *5*, e10901.

Tonge, P.D., Shigeta, M., Schroeder, T., and Andrews, P.W. (2011). Functionally defined substates within the human embryonic stem cell compartment. Stem Cell Res. *7*, 145–153.

Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K., and Niwa, H. (2008). Identification and characterization of subpopulations in undifferentiated ES cell culture. Development *135*, 909–918.

Stem Cell Reports, Volume *10*

Supplemental Information

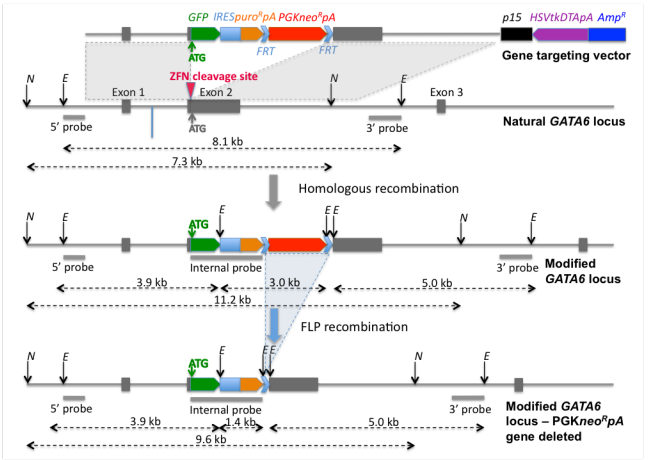
Identification and Single-Cell Functional Characterization of an Endodermally Biased Pluripotent Substate in Human Embryonic Stem Cells

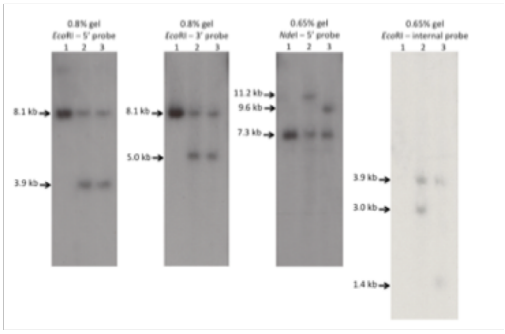
Thomas F. Allison, Andrew J.H. Smith, Konstantinos Anastassiadis, Jackie Sloane-Stanley, Veronica Biga, Dylan Stavish, James Hackland, Shan Sabri, Justin Langerman, Mark Jones, Kathrin Plath, Daniel Coca, Ivana Barbaric, Paul Gokhale, and Peter W. Andrews

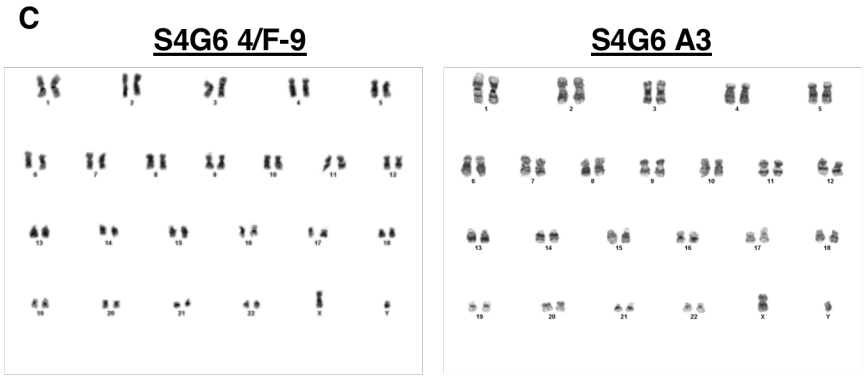
Identification and single cell functional characterization of an endodermally-biased pluripotent sub-state in human embryonic stem cells

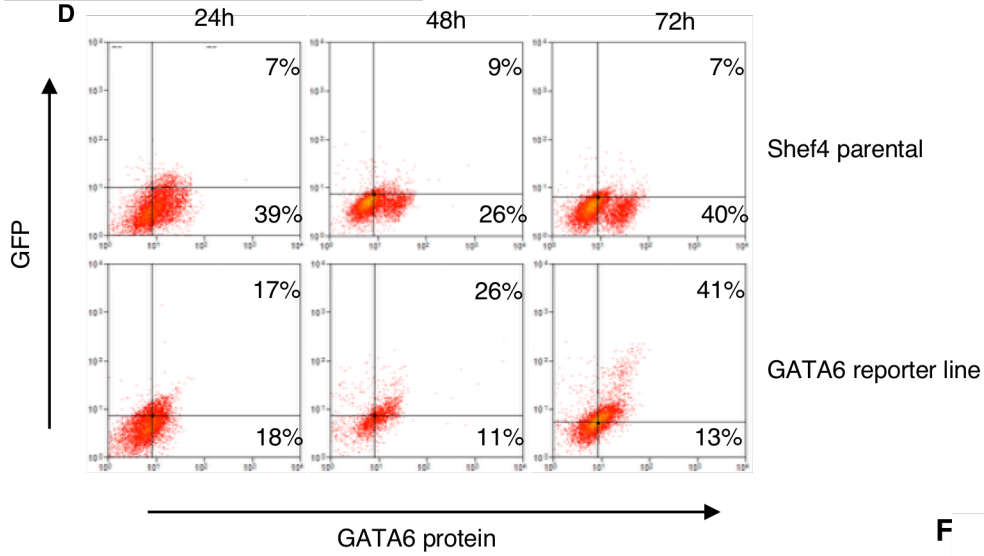
Supplemental Figures

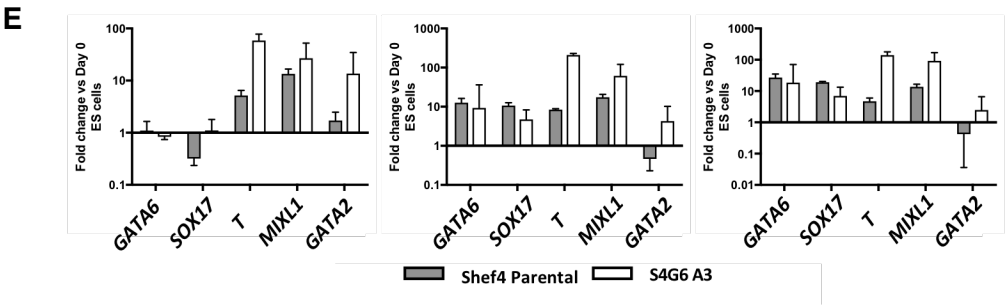
Figure S1


A
GFP IRESpuro^R pA PGKneo^R pA
p15 HSVtkDTApA Amp^R
FRT
FRT
ATG
Gene targeting vector
ZFN cleavage site
N
E
Exon 1
Exon 2
N
E
Exon 3
5′ probe
ATG
3′ probe
Natural GATA6 locus
8.1 kb
7.3 kb
Homologous recombination
ATG
E
E E
N
E
5′ probe
Internal probe
3′ probe
Modified GATA6 locus
3.9 kb
3.0 kb
5.0 kb
11.2 kb
FLP recombination
ATG
E
E E
N
E
5′ probe
Internal probe
3′ probe
Modified GATA6 locus – PGKneo^R pA gene deleted
3.9 kb
1.4 kb
5.0 kb
9.6 kb


B
0.8% gel
EcoRI – 5′ probe
1 2 3
8.1 kb
3.9 kb
0.8% gel
EcoRI – 3′ probe
1 2 3
8.1 kb
5.0 kb
0.65% gel
NdeI – 5′ probe
1 2 3
11.2 kb
9.6 kb
7.3 kb
0.65% gel
EcoRI – internal probe
1 2 3
3.9 kb
3.0 kb
1.4 kb


C
S4G6 4/F-9
S4G6 A3


D
24h
48h
72h
7%
9%
7%
39%
26%
40%
17%
26%
41%
18%
11%
13%
GFP
Shef4 parental
GATA6 reporter line
GATA6 protein


E
Fold change vs Day 0 ES cells
GATA6
SOX17
T
MIXL1
GATA2
Shef4 Parental
S4G6 A3

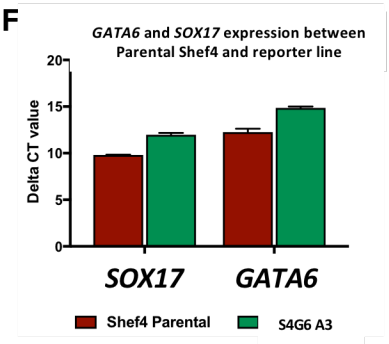
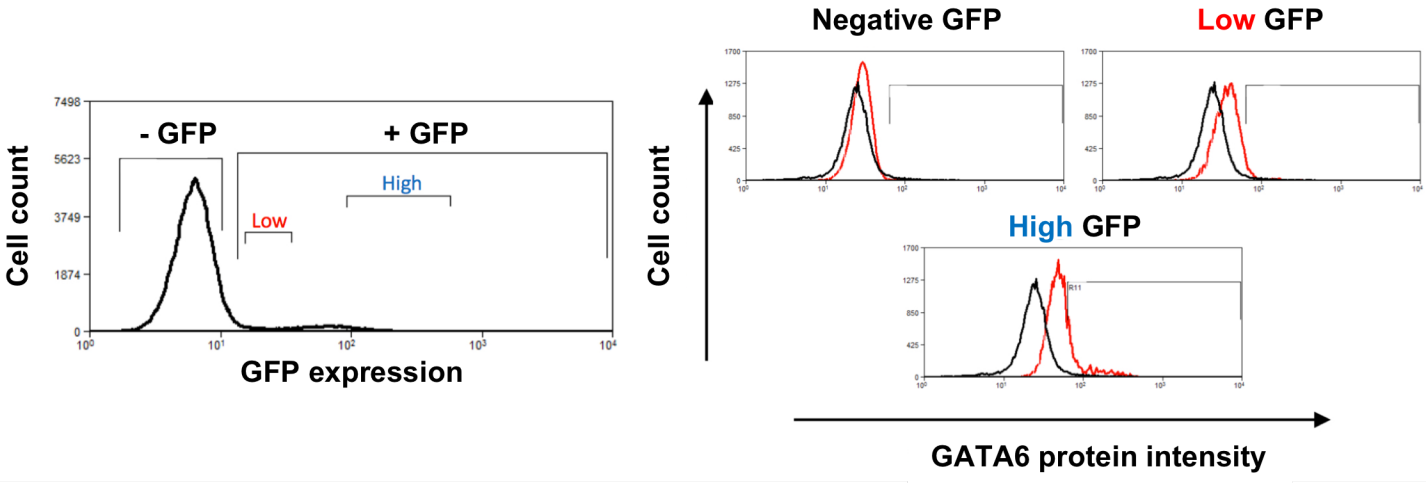

F
GATA6 and SOX17 expression between Parental Shef4 and reporter line
Delta CT value
SOX17
GATA6
Shef4 Parental
S4G6 A3

Figure S1

(A) Diagram of the linearized gene targeting vector, and the stages of genetic modification of the *GATA6* locus. Homology regions between the gene targeting vector and the *GATA6* locus within which homologous recombination can occur are indicated by the grey shaded regions. The *GATA6* gene targeting vector is designed for positive-negative selection with an internal constitutively expressed neomycin resistance-positive selection cassette (*PGKneo^{R}pA)* and, on the terminus of the right homology arm, a constitutively expressed Diphtheria Toxin A (DTA) chain negative selection cassette (*HSVtkDTApA*). The vector contains the enhanced Green Fluorescent Protein (GFP) reporter gene inserted at the position of the ATG translational initiation codon in the sequence of *GATA6* exon 2. Integration by homologous recombination is designed to result in expression of the enhanced GFP reporter and the puromycin resistance (*puro^{R}pA*) gene (linked via an Internal Ribosome Entry Site [IRES] sequence) under transcriptional control from the endogenous *GATA6* locus. The polyadenylation addition signal sequence (pA) at the 3' end of the puromycin resistance gene is predicted to result in premature termination of the *GATA6* transcript from the gene targeted allele. Subsequently the *PGKneo^{R}pA* selection cassette was deleted via the flanking *FRT* recombination sites by FLP site-specific recombination (indicated by the grey shaded triangle). Positions of *Eco*RI (*E*) and *Nde*I (*N*) restriction enzyme sites are shown by vertical arrows with predicted restriction fragments shown by horizontal dashed lines with arrowheads and sizes in kilobases (kb). Positions of the flanking 5' and 3' probes, and the internal probe, used for Southern blot analysis are shown as grey bars. **(B)** Southern blot analysis to confirm correct genetic modification of the *GATA6* locus. In all Southern blots: genomic DNA from the starting Shef4 cells before transfection - track 1; heterozygous *GATA6* gene targeted clone S4G6 4/F-9 with the *PGKneoRpA* selection marker present - track 2; derivative heterozygous *GATA6* gene targeted clone S4G6 A3 with the *PGKneoRpA* selection marker deleted by FLP recombination - track 3. Genomic DNAs were cut with *Eco*RI and *Nde*I and fractionated on 0.8% and 0.65% gels as indicated. Filters were hybridized with 5' and 3' flanking probes, and with the internal vector probe, as indicated. Sizes of bands were determined using a λ *Hind*III marker track (not shown). Restriction fragment sizes are consistent with predicted sizes shown in Supplementary Figure 1A. The internal probe only detects the predicted restriction fragments and importantly no additional restriction fragment sizes are detected proving that there is no random integration of the vector and the only reporter insertion is at the *GATA6* locus. **(C)** karyotype analysis of the hESC S4G6 4/F-9 reporter line and the derivative S4G6 A3 reporter line used within this study, demonstrating in both a normal 46XY karyotype. 30 metaphases were scored. **(D)** Time-course differentiation for the parental Shef4 and *GATA6* reporter (S4G6 A3) hESC lines. The parental and *GATA6* reporter Shef4 hESC lines were differentiated using an endodermal differentiation protocol (D'Amour et al., 2005) and FACS analyzed for GFP (Y axis) vs GATA6 protein (X axis) every 24 hours over a 72 hour period. **(E)** qPCR analysis for early primitive streak markers on a time-course differentiation for the parental Shef4 (grey bars) and the Shef4 S4G6 A3 (unfilled clear bars) hESC lines. Fold changes were calculated against respective hESC pluripotent lines, and β-actin was used as a normalizing gene. **(F)** Comparison of *GATA6* and *SOX17* expression levels between the parental Shef4 and S4G6 A3 reporter hESC lines.

Figure S2

A



Gated sample	% of total population	% of cells expressing GATA6 protein
-GFP	91%	0.9%
Low GFP	2%	5.5%
High GFP	1.5%	23.2%

B

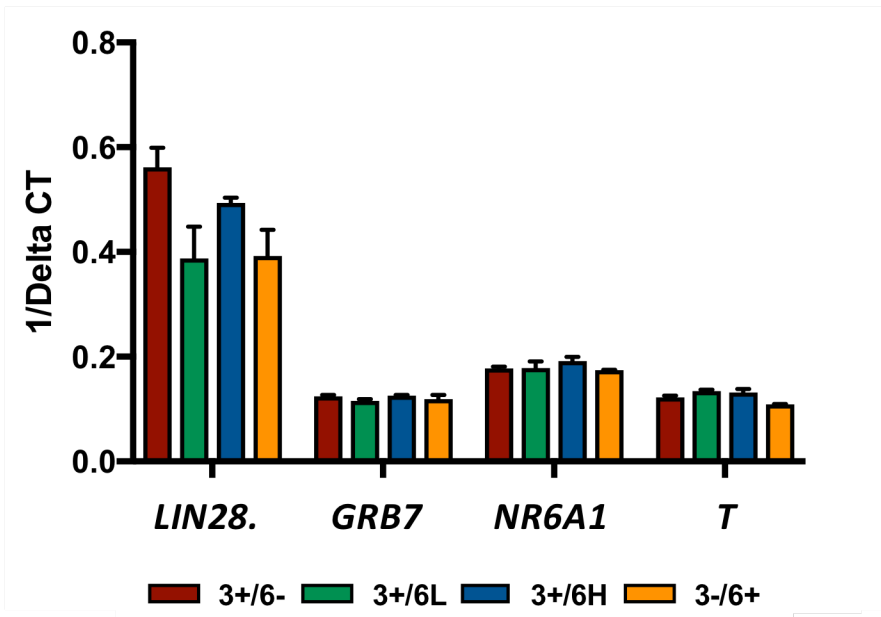
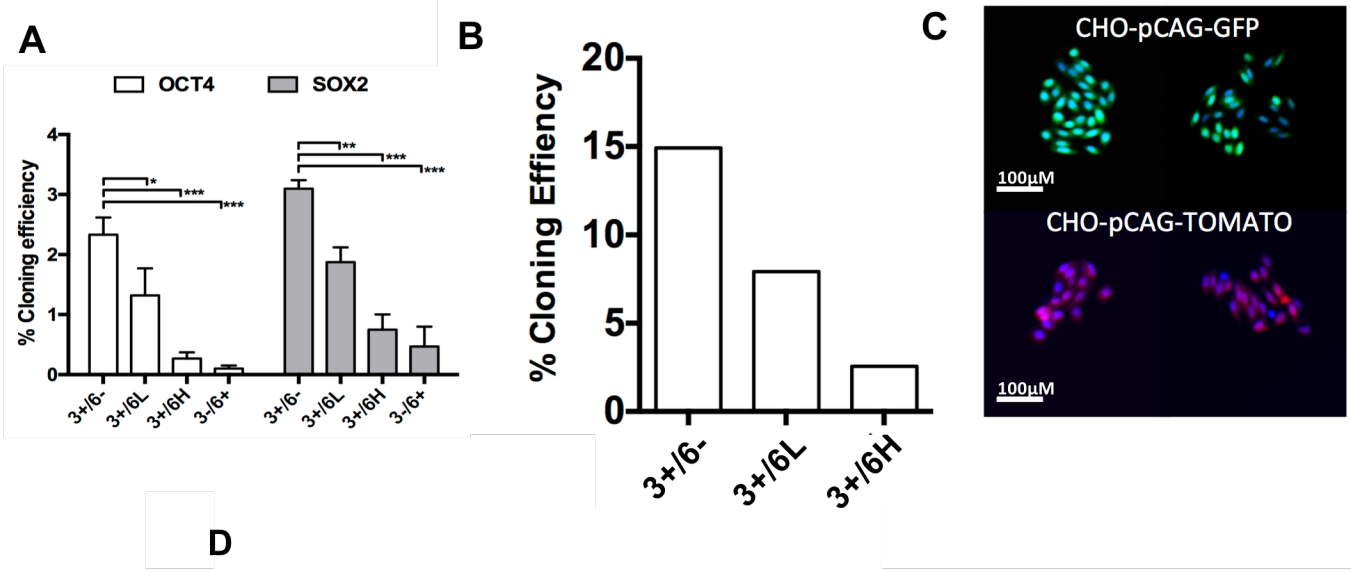


Figure S2

(A) The left panel shows a representative FACS plot of GFP (*GATA6* mRNA) expression within standard MEF/KOSR stem cell cultures, demonstrating the existence of low and high *GATA6* expressing cells. The right panel shows corresponding immunostaining for GATA6 protein levels within these subsets of low or high GFP (*GATA6* mRNA) expressing cells, demonstrating increased levels of GATA6 protein as GFP (*GATA6* mRNA) increases. Black lines show the isotype control, and red lines show GATA6 protein expression. **(B)** qPCR expression levels of genes that did not show significant changes between the four cell fractions (3+/6-, 3+/6L, 3+/6H and 3-/6+). Data is shown as 1/Delta CT values using β-actin as the normalizing gene, error bars representative of three technical repeats.

Figure S3



D

Fraction	Number of cells sorted	Possible error clones at 0.6%	Number of clones observed
3+/6-	288	2	43
3+/6L	960	6	76
3+/6H	1920	12	49

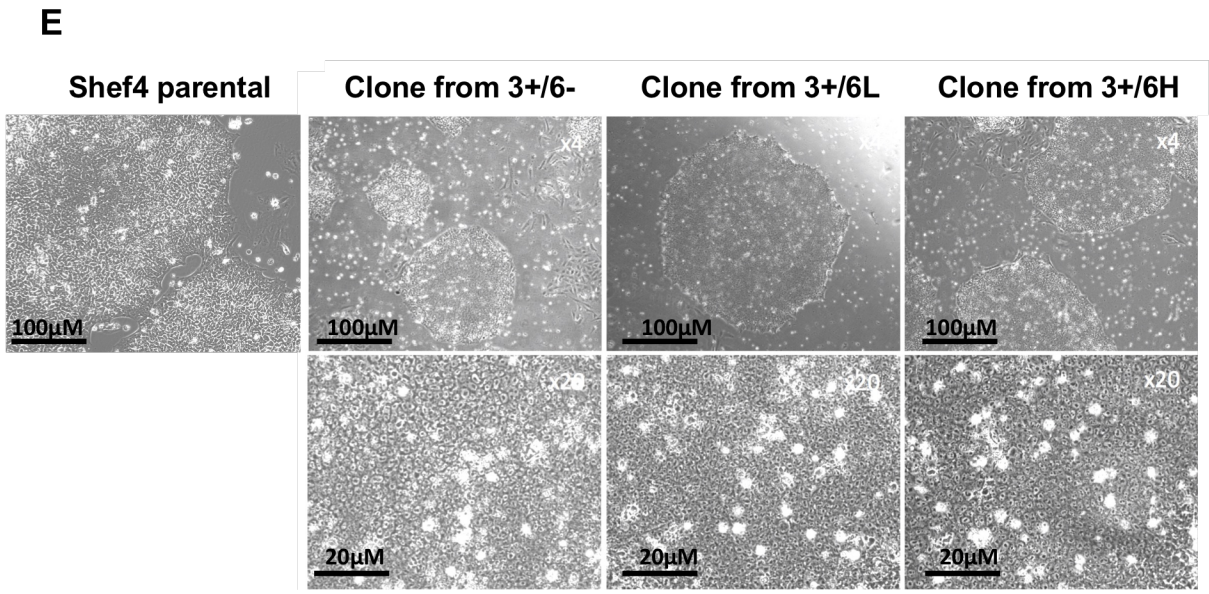


Figure S3

(A) Cloning efficiency of the sorted 3+/6-, 3+/6L, 3+/6H and 3-/6+ cell fractions using the reporter cell line S4G6 4/F-9. White bars represent colonies stained for OCT4, and grey bars, SOX2. The t-test was used to assess significance and is representative of three biological repeats, with p values of 0.0017 for 3+/6- to 3+/6L, 0.0001 for 3+/6- to 3+/6H and 0.0002 for 3+/6- to 3-/6+.**(B)** Cloning efficiency of the 3+/6-, 3+/6L and 3+/6H cell fractions after single cell/well FACS sorting. **(C)** Representative images of CHO-GFP (left) or CHO-Tomato (right) expressing colonies after single cell deposition used to determine FACS sorting misclassification rates. Single CHO cells were maintained for 4 days to form colonies then fixed to determine the proportion of misclassification, quantified in **(D). (E)** Representative images of resulting clones from single cell/well deposition showing typical hESC morphology. Shef4 parental line is shown as the control.

Figure S4

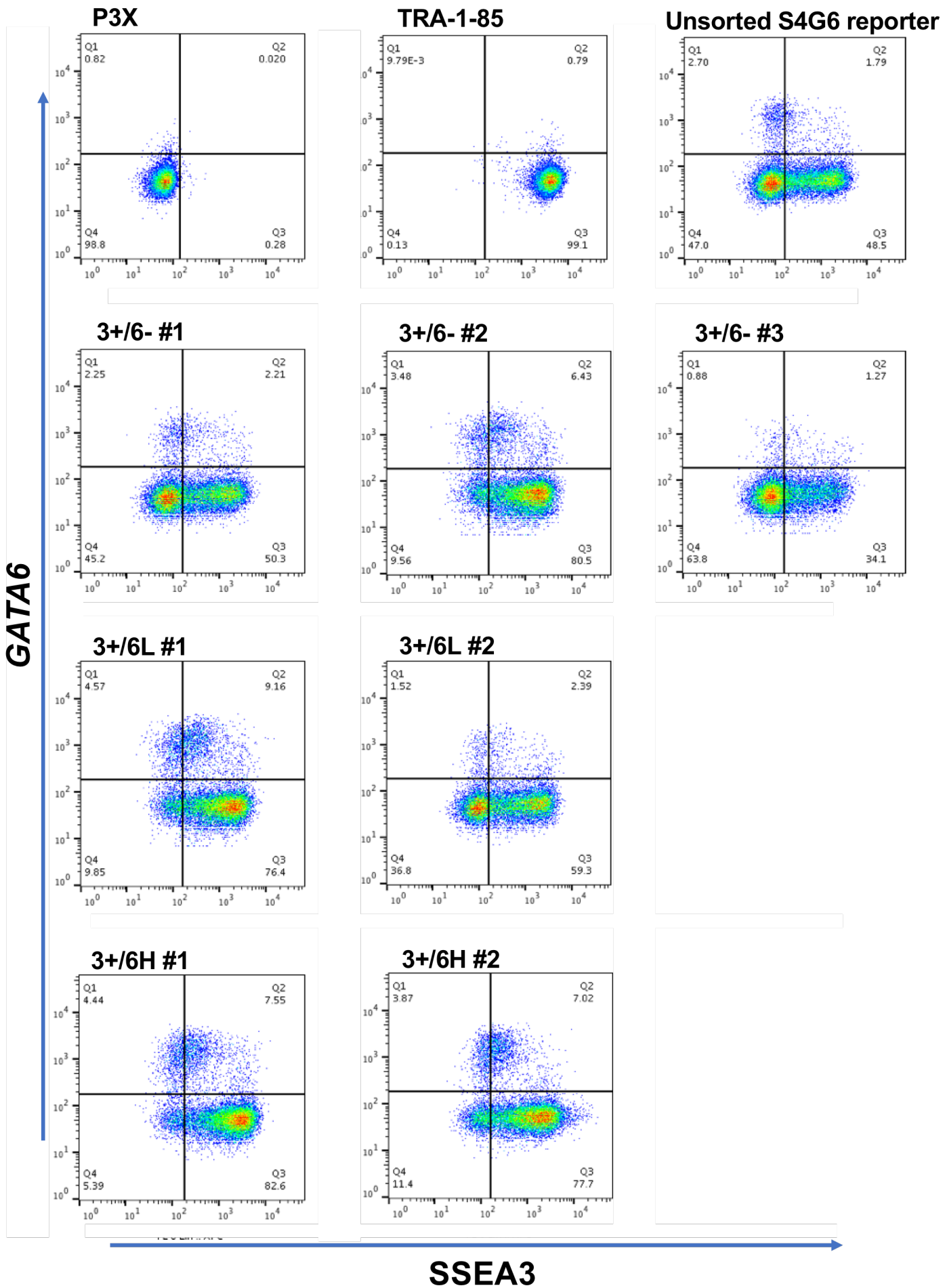
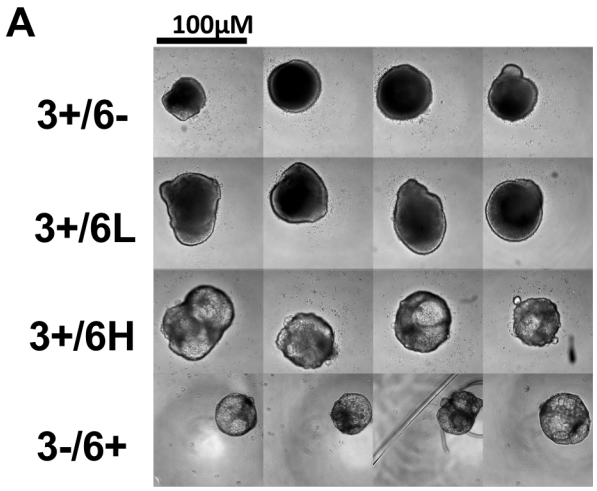
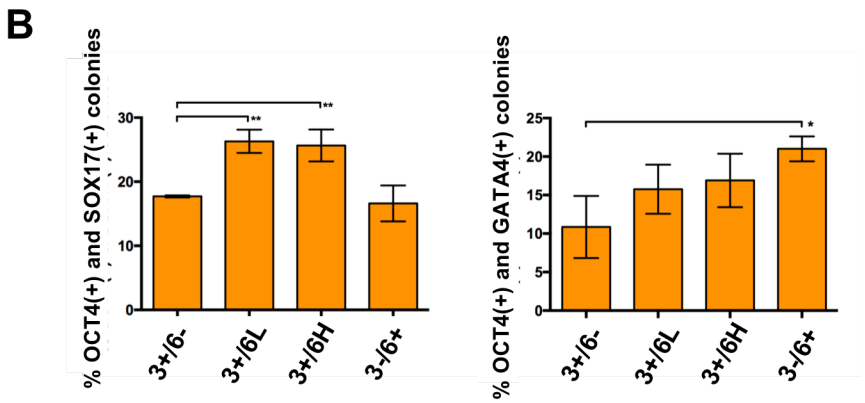

P3X
Q1 0.82
Q2 0.020
Q4 98.8
Q3 0.28
TRA-1-85
Q1 9.79E-3
Q2 0.79
Q4 0.13
Q3 99.1
Unsorted S4G6 reporter
Q1 2.70
Q2 1.79
Q4 47.0
Q3 48.5
3+/6- #1
Q1 2.25
Q2 2.21
Q4 45.2
Q3 50.3
3+/6- #2
Q1 3.48
Q2 6.43
Q4 9.56
Q3 80.5
3+/6- #3
Q1 0.88
Q2 1.27
Q4 63.8
Q3 34.1
3+/6L #1
Q1 4.57
Q2 9.16
Q4 9.85
Q3 76.4
3+/6L #2
Q1 1.52
Q2 2.39
Q4 36.8
Q3 59.3
3+/6H #1
Q1 4.44
Q2 7.55
Q4 5.39
Q3 82.6
3+/6H #2
Q1 3.87
Q2 7.02
Q4 11.4
Q3 77.7
GATA6
SSEA3

Figure S4

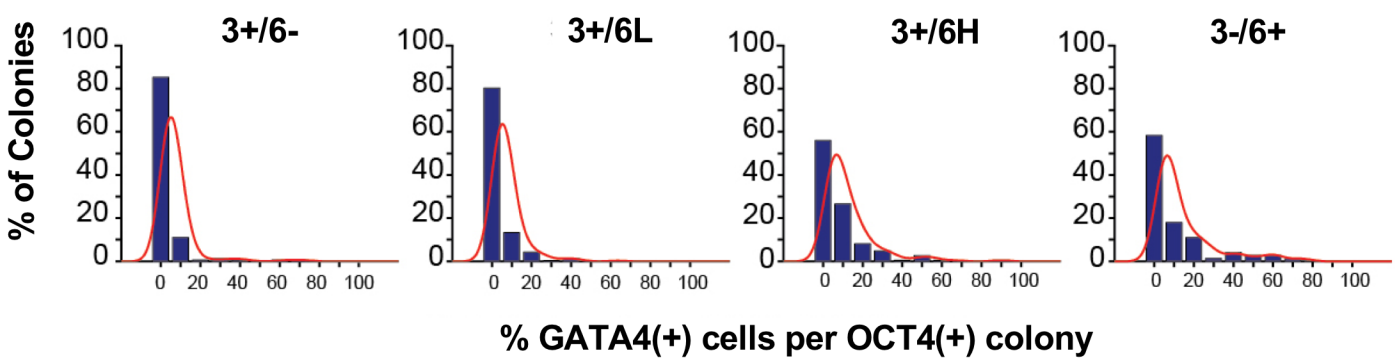
FACS analysis of clones from the 3+/6-, 3+/6L and 3+/6H cell fractions (two clones from each) derived from the single cell deposition experiment, after five passages in self- renewing conditions, showing the expression of SSEA3 (x-axis) vs *GATA6* (y-axis). P3-X serves as the negative, and TRA-1-85 as positive controls for appropriate FACS gating. All long-term clones show the re-distribution of all cell fractions after prolonged passaging.

Figure S5





C



D

	3+/6-	3+/6L	3+/6H	3-/6+
3+/6-	0	0.0251	0.2253	0.2411
3+/6L	0.0251	0	0.1613	0.194
3+/6H	0.0225	0.1613	0	0.043
3-/6+	0.2411	0.194	0.043	0

Figure S5

(A) Representative images to demonstrate the differing morphology of embryoid bodies generated from cells from the 3+/6-, 3+/6L, 3+/6H and 3-/6+ cell fractions after 10 days of differentiation. Images were taken at x4 magnification on the InCell Analyzer 2000. **(B)** Percentage of colonies containing both OCT4(+) and SOX17(+) cells (left graph, 3+/6- to 3+/6L p=0.0012, 3+/6- to 3+/6H p=0.0052) or both OCT4(+) and GATA4(+) cells (right graph, 3+/6- to 3-/6+ p=0.015) from the reporter line S4G6 4/F-9. Significance was calculated using t-test of three biological replicates. **(C)** Histogram showing the distribution of GATA4(+) cells in OCT4-positive colonies resulting from single cells from 3+/6-, 3+/6L, 3+/6H and 3-/6+ fractions. Positive colonies include at least two OCT4(+) cells. Counts are shown as a bar plot (blue) with superimposed estimated nonparametric distribution (red) and are representative of three biological replicates. **(D)** Kullback-Leibler symmetric divergence between GATA4-associated distributions in OCT4-positive colonies. This measure increases with reduced similarity between distributions; zero indicates identical distributions.

Figure S6

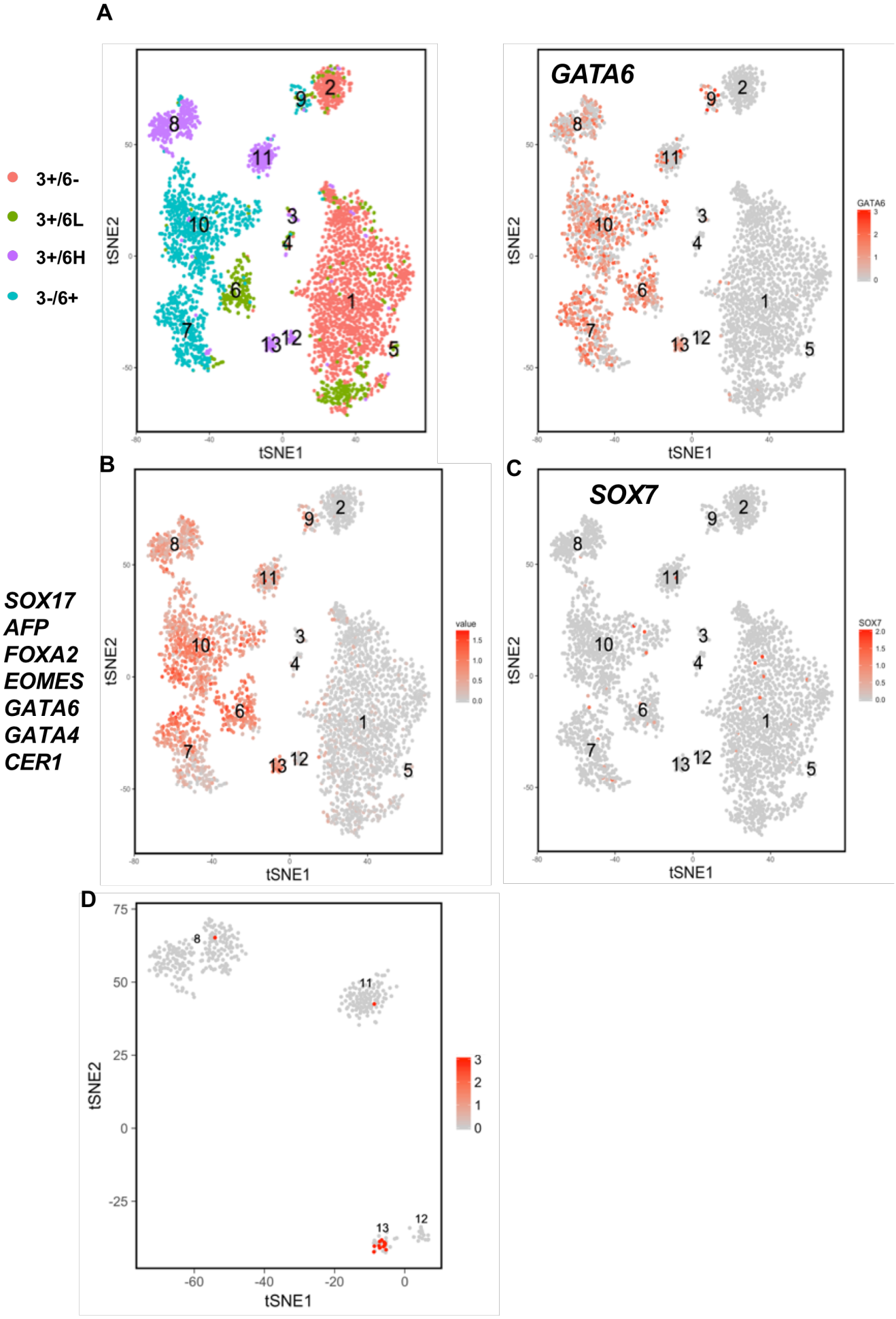

A
3+/6-
3+/6L
3+/6H
3-/6+
tSNE1
tSNE2
GATA6
B
SOX17
AFP
FOXA2
EOMES
GATA6
GATA4
CER1
value
C
SOX7
D
OCT4/SOX2/GATA6 positive cells

Figure S6

(A) The left panel shows a tSNE plot of single cell data from the four cell fractions, colored according to their original fraction. Each dot represents a single cell and putative clusters are numbered at random. The right panel shows the same tSNE plot overlaid with the expression of *GATA6*, showing only cells from the GFP positive sorted fractions had *GATA6* expression. Red represents *GATA6* positive, and grey, *GATA6* negative cells. **(B)** tSNE plot of all single cells to show the average expression of multiple endoderm associated genes indicating that their expression correlates with *GATA6*. Only cells from the GFP positive cell fractions showed expression of these genes. **(C)** tSNE plot of all single cells showing the expression of the primitive endoderm marker *SOX7*. Red represents *SOX7* positive, and grey, *SOX7* negative cells. **(D)** Reduced tSNE plot showing only the clusters composed of cells from the 3+/6H fraction. Red represents cells that are positive for all of the genes *OCT4*, *SOX2*, and *GATA6*, and grey are cells that do not co-express all three of these genes.

Supplemental Experimental Procedures

Cell Culture

Human embryonic stem cells (hESC) were routinely cultured as previously described (Draper et al., 2002) on 0.1% gelatin coated flasks containing mitomycin C inactivated mouse embryonic fibroblasts (MEFs) from e12.5 MF-1 mouse embryos in Knockout DMEM with 20% Knockout serum replacement (KO/SR) (Gibco), 4ng/mL bFGF (R&D Systems), 1% non-essential amino acids (Gibco), 1mM L-Glutamine (Gibco) and 0.1mM β-mercaptoethanol (Gibco). Cells were passaged using 1mg/mL of collagenase IV (ThermoFisher) at a split ratio of 1:3 every 3-4 days. For feeder free culture, cells were grown on rhVitronectin and E8 media (Life Technologies). These cells were passaged using ReLeSR (Stem Cell Technologies) at a ratio of 1:3-1:6 every 3-4 days. For the genetic modification work, the hESC line, Shef4 (Aflatoonian et al., 2010), was routinely grown in mTeSR1 medium (Stem Cell Technologies) on Matrigel (BD Biosciences) and passaged using collagenase IV (Stem Cell Technologies) as above.

Embryoid bodies

To investigate population bias, cells were differentiated using embryoid bodies in the serum free, defined media, APEL (Stem Cell Technologies) (Ng et al., 2008). Single cells were sorted as described below, into the respective cell fractions and single cells were resuspended in APEL media with 10μM Y-27632 (Tocris) (Watanabe et al., 2007) at a concentration of 3000 cells/100μL. 100μL of cell suspension was pipetted into the inner 60 wells of a non-adherent U-shaped 96-well plate (Sigma) to generate EBs of 3000 cells in size. The remaining outer 36 wells were filled with 100μL of sterile PBS to humidify the plate. Plates were centrifuged at 1000rpm for 3 mins to collect cells at the bottom of the wells and EBs were left to develop for 10 days at 37°C under a humidified atmosphere of 5% CO_2 in air.

Generation of a *GATA6*-GFP reporter hESC line

A *GATA6* reporter hESC line was generated using a standard gene targeting replacement vector designed to knock-in an enhanced Green Fluorescent Protein (GFP) reporter cassette by homologous recombination into exon 2 of the human *GATA6* locus at the position of the ATG translational initiation codon. The reporter cassette contained the GFP coding sequence linked via an IRES sequence to a puromycin resistance gene with a polyadenylation signal

sequence; followed by a neomycin resistance gene expressed from a constitutive promoter (PGK*neo*pA) and flanked by *frt* sites. The gene targeting vector was built utilizing recombineering technology (Zhang et al., 1998) according to standard protocols (details available on request to co-author KA) as follows. First, the reporter cassette GFP-IRES-*puro*pA/frt/PGK*neo*pA/frt was inserted directly after the start codon (ATG) of *GATA6* sequence in a BAC vector (RP11-523D21) by recombineering using the oligonucleotides hGata6-GFP-up and hGata6-GFP-dn. For making the targeting vector, a fragment comprised of 2.6 kb 5'- and 4.1 kb 3'- GATA6 homology sequences flanking the GFP cassette was subcloned from the BAC by recombineering (as previously described) into a PCR amplified p15A-HSV*tk*DTA-amp plasmid backbone using the primers hGata6-sub1 and hGata6-sub2.

hGata6-GFP-up: 5'-
ACCCCACCTCAGGAGCTAGACGTCAGCTTGGAGCGGCGCCGGACCGTGGATGGTGAG
CAAGGGCGAGGAGCTGTTCACCGGGGTGGTGC-3'

hGata6-GFP-dn: 5'-
CCCGCGGCCCCGAAGCGCTTCGGCAAGCACCAGCCGCCGTCAGTCAAGGCATCATGA
CCATGATTACGCCAAGCTTGGGCTGCAGGTTCTTTCCGCCT-3'

hGata6-sub1: 5'-
GAATGAGAGAGATTTTATTCAACTAAAAATAAGCAAGCTTCCTAGGTTGTGCGATCGCG
ATATCTTAATTAATAAGATGATCTTCTTGAG-3'

hGata6-sub2: 5'-
GTTCATATACACACCCCCTCTTCGCTCCCTCCAAACAGTTATCACAACTCAGATCTTACG
TATTACCAATGCTTAATCAGTGAGG-3'

The final replacement vector contained homology arms to the *GATA6* locus 5' and 3' of the desired insertion position (6.7 kb total homology), and also a negative selection marker (HSV*tk*DTA) positioned outside the homology arms. A Zinc Finger Nuclease (ZFN) was designed with specificity to the sequence of the *GATA6* gene spanning the ATG translational initiation codon in exon 2; ZFN recognition sequence; 5'AGCTTGGAGCGGCGCCGGACCGTGGATGGCCTTGACTGACGGC (ZFN cleavage site in bold; ATG codon underlined). Shef4 hESC grown in mTeSR1 medium on Matrigel were gently trypsinized and co-electroporated (800 volts, 3 μfarads) in mTeSR1 medium with 50 μg of the replacement vector (linearized at a unique restriction site) and with 25 μg of each of the two plasmids expressing the pair of ZFN polypeptides. Cells were plated onto Matrigel coated dishes and into mTeSR1 medium with ROCK inhibitor. Forty eight hours later selection was

commenced with 50 µg per ml G418. Cells were fed daily for 12 days with mTeSR1 plus G418, and then G418 resistant colonies picked into 96 well plates and into mTeSR1 medium plus G418. Ten days later cells were replica plated into 96 well plates, frozen, and genomic DNA prepared from one of the replicas. Clones with homologous integration of the vector were identified by restriction digestion of genomic DNA from the 96 well plate and Southern blotting/hybridization with appropriate flanking sequence probes. Correct gene targeted integrations were identified at a frequency of ~4% of total G418 resistant clones, and positive clones thawed and expanded. One of these clones (S4G6 4/F-9), which was heterozygous for the reporter knock-in, was further modified to excise the PGK*neo*pA gene by FLP recombination at the flanking *frt* sites in order to avoid any potential unpredicted phenotypic effects arising from the constitutively expressed PGK promoter. Excision was achieved by transient exposure of cells to cell permeable FLP recombinase (TAT-FLP) kindly donated by Frank Edenhofer and applied using an established protocol (Patsch et al., 2011). Cells were plated at clonal density and colonies screened after replica plating for sensitivity to G418. G418 sensitive colonies were further analyzed by Southern blotting/hybridization to confirm correct excision of the PGK*neo*pA selection marker via FLP recombination at the *frt* sites. One of the derivative clones obtained after FLP recombination, called S4G6 A3, and the S4G6 4/F-9 clone, were confirmed to have a normal karyotype (30 metaphases scanned), and furthermore no additional unpredicted vector integration events were detected in either of these clones by Southern blot hybridization analysis. S4G6 A3 and S4G6 4/F-9 gave essentially similar profiles when analyzed by FACS for GFP expression and SSEA3 expression.

Antibodies

The following monoclonal antibodies were derived from the relevant hybridomas grown in-house, pre-titered and used to detect surface antigen expression as previously described (Adewumi et al., 2007; Draper et al., 2002): MC631, anti-Stage Specific Embryonic Antigen-3 (SSEA3) (Shevinsky et al., 1982), MC813-70, anti-Stage Specific Embryonic Antigen-4 (SSEA4) (Kannagi et al., 1983), MC480, anti-Stage Specific Embryonic Antigen-1 (SSEA1) (Solter and Knowles, 1978), TRA-1-60 and TRA-1-81 (Andrews et al., 1984) and TRA-1-85 (Williams et al., 1988). Antibody P3X, from the parental myeloma, P3X63ag8 (Köhler and Milstein, 1975), was used as a negative control (Draper et al., 2002). For intra-cellular staining, commercial antibodies were obtained and used as detailed by the manufacturer as follows. OCT4A, 1:200, SOX2, 1:200 (Cell Signaling Technologies, #C52G3, #D6D9 respectively), SOX17 at 1µg/mL, GATA4 at 1µg/mL and GATA6 at 1µg/mL (R&D Systems, #AF1924, AF2606, #AF1700 respectively). Secondary antibodies were also obtained commercially and

used as per manufacturer instructions. AlexFluor647 conjugated goat-anti-mouse IgG (H+L) (Stratech #209-605-082) 1:100, AlexFluor594 conjugated donkey-anti-goat IgG+IgM (Stratech #708-585149) 1:100, AlexFluor647 conjugated donkey anti-rabbit IgG+IgM (Stratech #609-605-213) 1:100.

Flow Cytofluorimetry and Fluorescence Activated Cell Sorting

hESC were dissociated using trypLE (Gibco), stained for the relevant surface markers and analyzed using a BD FACS Jazz. The same machine was also used to sort cells. After sorting, single cells were seeded at clonogenic densities of 500 cells/cm^2 in standard hESC medium on MEF, in the presence of 10μM Y-27632 (Tocris) (Watanabe et al., 2007). Y-27632 was subsequently removed after 24h. Clonogenic assays had three technical repeats over three biological repeats. For sub-cloning experiments, single cells were sorted directly into single wells of 96-well plates in standard 20% KO/SR, DMEM and MEF conditions with 10μM Y-27632. Resulting colonies were picked initially, and then passaged using 1mg/mL collagenase IV. All sub-clones were passaged at least 5 times before analysis.

Immunocytochemistry

Cells were fixed using 4% paraformaldehyde for 15 minutes at room temperature, and washed twice with Dulbecco's Phosphate Buffered Saline, without calcium and magnesium (PBS). For intracellular staining, cells were permeablized and blocked in PBS with 10% fetal calf serum, 0.3mM glycine, 1% bovine serum albumin (BSA), and 0.1% Tween for 2h at room temperature. Cells were incubated with the primary antibody resuspended in blocking buffer (without 0.3mM Glycine) overnight at 4^{o}C, washed twice in PBS and incubated with the species specific secondary antibody for 2h at 4^{o}C. Finally, cells were washed twice in PBS and imaged on an InCell Analyzer 2000 (GE Healthcare). Images were analyzed using the Developer Toolbox software (GE Healthcare).

Gene expression analysis

Total RNA was extracted using Trizol (Life Technologies) and a centrifugation based column kit (Norgen RNA clean-up and concentration kit) followed by cDNA synthesis using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Quantitative real-time PCR was performed on the QuantStudio 12K Flex Real-Time PCR system (ThermoFisher) using Taqman universal master mix (Invitrogen) in conjunction with the Roche universal probe library system (Roche). Primers and probes were designed by Primer3 software (Roche) and where

possible intron-spanning primers were used. Gene expression was normalized to β-Actin in all experiments.

Clustering, Box Plot and Kullback-Leibler Divergence Analysis

All statistical analyses were implemented using MATLAB or GraphPad. Hierarchical clustering was performed using Spearman's rank correlation. Colormap indicates levels of expression of 1/Delta-CT values standardized by row. Box plots indicate values between the 25th and 75th percentile with level represented by the median and whisker lengths of 2.7SD. Outliers extending beyond whiskers were excluded. The Kruskal-Wallis pairwise test was used to compare distributions at a statistical significance level of <0.05.

Kullback-Leibler analysis was implemented in MATLAB. Kernel density estimation was used to estimate nonparametric probability density functions (PDF) from histograms. Kullback-Leibler divergence was used to quantify shape changes in the histograms. For p and q denoting two PDFs, the Kullback-Leibler symmetric divergence was calculated pairwise.

Processing, Read Alignment and Digital Gene Expression (DGE) Matrix Construction.

Raw sequencing data was quality filtered, adapter- and polyA-trimmed and reads satisfying a length criterion of 30nts were aligned to the human (hg19) genome using Bowtie2 (v2.2.9) with cross-species mapping reads to mouse (mm9) removed. Aligned reads were tagged with gene exons using Bedtools Intersect (v2.26.0). Digital gene expression (DGE) matrices were then generated for each time point from all Drop-seq runs. To digitally count gene transcripts, reads with the same corresponding cell barcode were aggregated together and unique UMIs and cell barcodes were merged within 1 Hamming and 2 Levenshtein distances, respectfully. We filtered all DGEs to exclude cells detecting <500 genes and <1250 transcripts from all downstream analyses.

tSNE analysis of single cell RNA-seq data.

DGE matrices were normalized by the number of transcripts in log space (e.g. ln(transcripts/10,000+1)). Cells were projected onto a 2D embedding using t-Distributed Stochastic Neighbor Embedding (tSNE) with cell loadings associated with 30 principal components utilizing all expressed genes as input. Cluster assignments were computed using a density-based clustering approach (DBSCAN). The degree of similarity between clusters was computed by averaging the gene expression for each cluster, and calculating a Euclidean

distance matrix between all pairs, then using this data as input for hierarchical clustering with optimal leaf ordering. We implement a negative binomial generalized linear model to identify differentially expressed genes (DEGs) enriched in each cluster. Genes satisfying an abs (log(average expression difference))>1 and P-value<0.01 were considered statistically significant. Gene Ontology (GO) enrichments were computed using all DEGs and a subset of the top 50 DEGs by average expression difference per cluster. Significant enrichments satisfying a log10(Benjamini-Hochberg(P-value)) <-3 threshold were graphed. The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (Edgar *et al*., 2002) and are accessible through GEO series accession number GSE113168 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113168).

Supplementary References

Adewumi, O., Aflatoonian, B., Ahrlund-Richter, L., Amit, M., Andrews, P.W., Beighton, G., Bello, P.A., Benvenisty, N., Berry, L.S., and Bevan, S. (2007). Characterization of human embryonic stem cell lines by the International Stem Cell Initiative. Nature biotechnology *25*, 803-816.

Aflatoonian, B., Ruban, L., Shamsuddin, S., Baker, D., Andrews, P., and Moore, H. (2010). Generation of Sheffield (Shef) human embryonic stem cell lines using a microdrop culture system. In Vitro Cellular & Developmental Biology-Animal *46*, 236-241.

Andrews, P.W., Banting, G., Damjanov, I., Arnaud, D., and Avner, P. (1984). Three monoclonal antibodies defining distinct differentiation antigens associated with different high molecular weight polypeptides on the surface of human embryonal carcinoma cells. Hybridoma *3*, 347-361.

Draper, J.S., Pigott, C., Thomson, J.A., and Andrews, P.W. (2002). Surface antigens of human embryonic stem cells: changes upon differentiation in culture. Journal of anatomy *200*, 249-258.

Kannagi, R., Cochran, N.A., Ishigami, F., Hakomori, S.-i., Andrews, P., Knowles, B.B., and Solter, D. (1983). Stage-specific embryonic antigens (SSEA-3 and-4) are epitopes of a unique globo-series ganglioside isolated from human teratocarcinoma cells. The embo journal *2*, 2355.

Köhler, G., and Milstein, C. (1975). Continuous cultures of fused cells secreting antibody of predefined specificity. Nature *256*, 495-497.

Ng, E.S., Davis, R., Stanley, E.G., and Elefanty, A.G. (2008). A protocol describing the use of a recombinant protein-based, animal product-free medium (APEL) for human embryonic stem cell differentiation as spin embryoid bodies. Nature protocols *3*, 768-776.

Patsch, C., Kesseler, D., and Edenhofer, F. (2011). Genetic engineering of mammalian cells by direct delivery of FLP recombinase protein. Methods *53*, 386-393.

Shevinsky, L.H., Knowles, B.B., Damjanov, I., and Solter, D. (1982). Monoclonal antibody to murine embryos defines a stage-specific embryonic antigen expressed on mouse embryos and human teratocarcinoma cells. Cell *30*, 697-705.

Solter, D., and Knowles, B.B. (1978). Monoclonal antibody defining a stage-specific mouse embryonic antigen (SSEA-1). Proceedings of the National Academy of Sciences *75*, 5565-5569.

Watanabe, K., Ueno, M., Kamiya, D., Nishiyama, A., Matsumura, M., Wataya, T., Takahashi, J.B., Nishikawa, S., Nishikawa, S.-i., and Muguruma, K. (2007). A ROCK inhibitor permits survival of dissociated human embryonic stem cells. Nature biotechnology *25*, 681.

Williams, B., Daniels, G., Pym, B., Sheer, D., Povey, S., Okubo, Y., Andrews, P., and Goodfellow, P. (1988). Biochemical and genetic analysis of the Ok a blood group antigen. Immunogenetics *27*, 322-329.

Zhang, Y., Buchholz, F., Muyrers, J.P., and Stewart, A.F. (1998). A new logic for DNA engineering using recombination in Escherichia coli. Nature genetics *20*, 123.