Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase

Cell

Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase

Yosef Buganim,[1,7] Dina A. Faddah,[1,2,7] Albert W. Cheng,[1,3] Elena Itskovich,[1] Styliani Markoulaki,[1] Kibibi Ganz,[1] Sandy L. Klemm,[5] Alexander van Oudenaarden,[2,4,6] and Rudolf Jaenisch[1,2,*]
[1]The Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[2]Department of Biology
[3]Department of Computational and Systems Biology
[4]Department of Physics
[5]Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[6]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT, Utrecht, The Netherlands
[7]These authors contributed equally to this work
*Correspondence: jaenisch@wi.mit.edu
http://dx.doi.org/10.1016/j.cell.2012.08.023

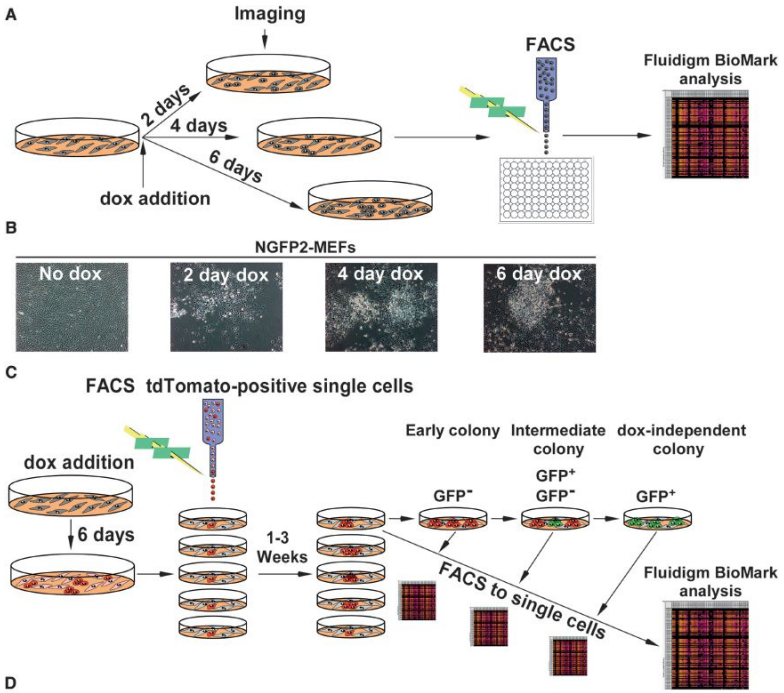
SUMMARY

During cellular reprogramming, only a small fraction of cells become induced pluripotent stem cells (iPSCs). Previous analyses of gene expression during reprogramming were based on populations of cells, impeding single-cell level identification of reprogramming events. We utilized two gene expression technologies to profile 48 genes in single cells at various stages during the reprogramming process. Analysis of early stages revealed considerable variation in gene expression between cells in contrast to late stages. Expression of Esrrb, Utf1, Lin28, and Dppa2 is a better predictor for cells to progress into iPSCs than expression of the previously suggested reprogramming markers Fbxo15, Fgf4, and Oct4. Stochastic gene expression early in reprogramming is followed by a late hierarchical phase with Sox2 being the upstream factor in a gene expression hierarchy. Finally, downstream factors derived from the late phase, which do not include Oct4, Sox2, Klf4, c-Myc, and Nanog, can activate

generate iPSCs that functionally and molecularly resemble embryonic stem cells (ESCs).

To further understand the reprogramming process, transcriptional and epigenetic changes in cell populations were analyzed at different time points after factor induction. For example, microarray data showed that the immediate response to the reprogramming factors was characterized by dedifferentiation of mouse embryonic fibroblasts (MEFs) and upregulation of proliferative genes, consistent with c-Myc expression (Mikkelsen et al., 2008). It has been shown that the endogenous pluripotency markers Sox2 and Nanog are activated after early markers such as alkaline phosphatase (AP) and SSEA1 (Stadtfeld et al., 2008). Recently, gene expression profiling and RNAi screening in fibroblasts revealed three phases of reprogramming termed initiation, maturation, and stabilization, with the initiation phase marked by a mesenchymal-to-epithelial transition (MET) (Li et al., 2010; Samavarchi-Tehrani et al., 2010).

Given these data, a stochastic model has emerged to explain how forced expression of the transcription factors initiates the process that eventually leads to the pluripotent state in only a small fraction of the transduced cells (Hanna et al., 2009; Yamanaka, 2009). Most data have been interpreted to support a stochastic model (Hanna et al., 2009) posing that the reprog-



E

Colony	Reprogramming characteristics	Day of cell collection (across 94 days)			
		tdTomato+	GFP+	GFP+ dox independent	Dox withdrawal
20	Early reprogramming	32	32	66	36
34		32	32	61	36
43		12, 45	45	61	41
16	Late reprogramming	81	81	94	81
23	Induced cells that did not give rise to iPSCs	12, 81	NA	NA	NA
44		12, 61	61*	NA	NA

To understand the changes that precede iPSC formation, we used gene expression analysis to profile 48 genes in single cells derived from early time points, intermediate cells, and fully reprogrammed iPSCs, demonstrating that cells at different stages of the reprogramming process can be separated into two defined populations with high variation in gene expression at early time points. We also demonstrate that activation of genes such as *Fbxo15, Fgf4*, and *Oct4* does not stringently predict successful reprogramming, in contrast to Esrrb, Utf1, Lin28, and Dppa2, which more rigorously mark the rare cells that are destined to become iPSCs. Moreover, our results suggest that stochastic gene expression changes early in the reprogramming process are followed by a ''nonstochastic'' or more ''hierarchical'' phase of gene expression responsible for the activation of the endogenous pluripotency circuitry. Finally, based on the events that occur in this late consecutive phase, we show that the activation of the pluripotency core circuitry is possible by various combinations of factors and even in the absence of the generic ''Yamanaka'' factors.

RESULTS

Single-Cell Expression Profiling at Defined Time Points

To measure gene expression in single cells at defined time points during the reprogramming process, we combined two complimentary tools: (1) 96.96 Dynamic Array chips (Fluidigm), which allows quantitative analysis of 48 genes in duplicate in 96 single cells (Guo et al., 2010), and (2) single-molecule-mRNA fluorescent in situ hybridization (sm-mRNA-FISH), which allows the quantification of mRNA transcripts of up to three genes in hundreds to thousands of cells (Raj et al., 2008).

We selected gene candidates based on the major events that occur during reprogramming (Figure S1A available online). Because reprogramming requires a vast number of epigenetic changes, we chose a group of ESC-associated chromatin-remodeling genes and modification enzymes (*Myst3*, *Kdm1*, *Hdac1*, *Dnmt1*, *Prmt7*, *Ctcf*, *Myst4*, *Dnmt3b*, *Ezh2*, *Bmi1*) (Reik, 2007; Surani et al., 2007). Because high proliferative capacity is essential to facilitate the reprogramming process, we selected ESC cell-cycle regulator genes (*Bub1*, *Cdc20*, *Mad2l1*, *Ccnf*) (Hong et al., 2009). We also included key genes that are active in signal transduction pathways important for ESC maintenance and differentiation (*Bmpr1a*, *Stat3*, *Ctnnbl1*, *Nes*, *Wnt1*, *Gsk3b*, *Csnk2a1*, *Lifr*, *Hes1*, *Jag1*, *Notch1*, *Fgf5*, *Fgf4*) (Boiani and Schöler, 2005; Samavarchi-Tehrani et al., 2010). Finally, we chose a large number of pluripotency marker genes in an attempt to detect early and late markers in reprogramming (*Oct4*, *Sox2*, *Nanog*, *Lin28*, *Fbxo15*, *Zfp42*, *Fut4*, *Tbx3*, *Esrrb*, *Dppa2*, *Utf1*, *Sall4*, *Gdf3*, *Grb2, Slc2a1*, *Fthl17*, *Nr6a1*) (Ng and Surani, 2011; Ramalho-Santos et al., 2002). We used *Gapdh* and *Hprt* as control genes and *Thy1* and *Col5a2* as markers for MEFs.

To circumvent the genetic heterogeneity of ''primary'' virus-transduced fibroblasts, we utilized previously characterized clonal doxycycline (dox)-inducible secondary NGFP2 MEFs (Wernig et al., 2008). Briefly, these cells are derived from a homogenous donor cell population containing preselected proviral integrations of OSKM, each under the TetO promoter, reverse tetracycline transactivator (rtTA) in the Rosa26 locus, and a GFP reporter knocked into the *Nanog* locus. To compare variability between systems, we quantified *Sox2* and *Klf4* transcripts by sm-mRNA-FISH in single virus-infected MEFs and single secondary MEFs on dox for 6 days. Because transgene expression between single cells was more variable in the virus-infected MEFs, we used the secondary system for all analyses (Figures S1B and S1C).

We analyzed clonal populations (cells derived from a single cell) throughout the process of dox-independent iPSC formation, beginning at day 2 of dox addition with the first colonies appearing around 7 days after dox addition. Thus, to detect early transcriptional changes in the reprogramming process, nonclonal populations of NGFP2 MEFs were exposed to dox for 2, 4, and 6 days. At each time point, the cells were imaged and sorted to single cells, and gene expression was profiled using the Fluidigm system (Figures 1A and 1B). To profile clonal populations of cells on dox for more than 6 days, we utilized a modified experimental setup. Because most cells senesced, became contact inhibited, or transformed after exposure to dox for 6 days, which interfered with single-cell sorting to identify those rare cells that were destined to become iPSCs, we generated secondary cells that, in addition to the *Nanog-GFP* gene, carried a tdTomato reporter. tdTomato was electroporated into NGFP2 iPSCs, and a single colony was picked and expanded. Cells derived from this colony were injected into blastocysts, and secondary MEFs were derived (Figure S1D). The presence of the tdTomato reporter enabled us to sort single secondary cells in the presence of unmarked feeder cells, which were important for both cell-cell interactions enabling proliferation of single cells and calibration of the FACS machine (i.e., tdTomato$^+$ cells versus tdTomato$^-$ cells). This system allowed tracing of the tdTomato$^+$ rare cells that bypassed senescence and contact inhibition and continued to proliferate forming colonies on the feeder layer.

Initially, labeled NGFP2 MEFs were exposed to dox for 6 days, sorted for tdTomato, and seeded each as a single cell in one well

Figure 1. Experimental Scheme Used to Monitor Transcriptional Profiles of Single Cells at Defined Time Points during the Reprogramming Process

(A) Scheme used for single-cell gene expression analysis with Fluidigm.

(B) Representative images of NGFP2 MEFs without dox and at days 2, 4, and 6 on dox.

(C) Scheme of NGFP2/tdTomato secondary system used to measure single-cell gene expression of clonal dox-dependent (GFP$^-$, GFP$^+$) and -independent (GFP$^+$) cells.

(D) Representative images and FACS analysis of dox-dependent and -independent cells at days 12, 32, and 61 on dox.

(E) Six colonies were profiled over the course of 94 days. Colony 44 (starred) contained a few cells with a low level of GFP that were sorted at day 61 and disappeared upon continual passaging and dox withdrawal.

See also Figures S1 and S2.

of four 24-well plates containing unmarked feeders. At different times between 1 and 3 weeks during the reprogramming process, tdTomato$^+$ colonies derived from single cells were imaged, split to another plate, sorted to single cells, and analyzed for their transcriptional profile using Fluidigm. Each parental cell was passaged to test its capacity to generate dox-independent, fully reprogrammed iPSCs. This system allowed tracing gene expression changes in multiple clonally related single sister cells over different times during reprogramming. Clonal populations were passaged, and gene expression was profiled as a function of time in three subpopulations: (1) early dox-dependent GFP$^-$ cells, (2) intermediate dox-dependent GFP$^-$ and GFP$^+$ cells, and (3) dox-independent GFP$^+$ cells (Figures 1C and 1D).

Out of 96 tdTomato$^+$ single cells, only seven cells generated a colony reflecting the low efficiency of the process. Single cells in these seven clonal populations (colonies 15, 16, 20, 23, 34, 43, and 44) were profiled over the course of 94 days (Figure 1E). Cells were sorted for GFP after detection on the inverted fluorescence microscope. Colonies 34, 20, and 43 gave rise to dox-independent cells relatively early in the process, whereas colony 16 gave rise to dox-independent cells very late in the process. Colonies 23 and 44 did not generate stable GFP colonies for 81 days of continuous culture in dox. Colony 44 contained a few cells with a very low level of GFP (Figure S1E) that disappeared upon further passaging. A few cells (0.01%) from colony 23 activated GFP only at day 81.

To gain insight into intermediate clonal cell populations, we analyzed by Fluidigm single tdTomato$^+$/GFP$^+$ double-positive cells from colony 20 at day 32 in dox. Using Pearson distance and average linkage of the gene expression data, we found that these double-positive cells represented an intermediate state between tdTomato$^+$/GFP$^-$ and tdTomato$^-$/GFP$^+$ cells (Figure S2A). To test whether tdTomato$^+$/GFP$^-$ cells present at day 32 are on the path toward iPSCs or are "stuck," we sorted 20 cells from colony 20 tdTomato$^+$/GFP$^-$, tdTomato$^+$/GFP$^+$, and tdTomato$^-$/GFP$^+$ cells onto three different feeder plates in dox (Figure S2B). After 5 days, the tdTomato$^+$/GFP$^-$ cells gave rise to tdTomato$^-$/GFP$^+$ colonies (Figures S2C and S2D). All groups generated stable, dox-independent tdTomato$^-$/GFP$^+$ iPSCs, albeit with different latencies (Figure S2E). Of the genes examined, Kdm1, a lysine-specific demethylase involved in silencing of viral sequences in mouse ESCs (mESCs) (Macfarlan et al., 2011), was found differentially expressed between tdTomato$^+$/GFP$^-$, tdTomato$^+$/GFP$^+$, and tdTomato$^-$/GFP$^+$ cells (Figure S2F). These data support the notion that silencing of viral sequences is a common late step in reprogramming.

Behavior of Single Cells during Reprogramming

For each profiled subpopulation, we obtained replicate gene expression data for 48 genes in 96 single cells. The Fluidigm microfluidics system combines samples and primer-probe sets for 9,216 quantitative RT-PCR (qRT-PCR) reactions. The output of one run on the Biomark is a 96 × 96 matrix of cycle threshold (Ct) values (Figure S3).

To globally visualize the data, we used principal component analysis (PCA). PCA is a technique used to reduce dimensionality of the data by finding linear combinations (dimensions; in this case, the number of genes) of the original data ranked by their importance. The data are projected to PC1 and PC2, the two most important PCs. In Figure 2A, the gene expression space is 48 dimensional because of the 48 genes, and each of the data points is a cell. The coordinate in each dimension is the normalized gene expression value for a given gene in that cell. Each component has contributions from all of the 48 genes since the components cut across this 48D space. Applied to the expression data derived from 1,864 cells from different stages during reprogramming, we found that the first principal component (PC1) explains 22.5% of the observed variance, whereas the second principal component (PC2) explains 5.8%. These values are lower than in a recent single-cell study of 64-celled embryos (Guo et al., 2010) and may reflect the substantially higher number of cells analyzed and the high degree of cell heterogeneity during reprogramming. A projection of the expression patterns onto PC1 and PC2 separates individual cells into two distinct clusters (blue and red circles) as well as a third cluster (orange dotted circle) representing the early transition from fibroblasts to iPSC precursors (Figure 2A). The first cluster (dark blue, enclosed in the blue circle) contains the three control groups, tail tip fibroblasts (TTF), MEFs, and NGFP2 MEFs. The second cluster (orange, red, brown, enclosed in the red circle) contains dox-dependent and -independent GFP$^+$ cells and the parental NGFP2 iPSCs. The third rather heterogeneous cluster (lighter blue(s), turquoise, green, and yellow, enclosed in the orange dotted circle) contains the GFP$^-$ cells exposed to dox for 2, 4, and 6 days and dox-dependent later GFP$^-$ cells. This cluster contains induced cells prior to the activation of the *Nanog-GFP* locus, possibly representing an early intermediate state. Importantly, a few cells from earlier time points (green and yellow dots) showed a similar pattern of expression as in the second cluster. This agrees with the observation that iPS colonies appear with different latencies and that early colonies with ESC-like morphology may not be dox independent. Cells on dox for 4 days cluster very closely to the MEFs, suggesting that the epigenetic changes that characterize a fully reprogrammed iPSC do not occur early in reprogramming (Guo et al., 2010). The gap between the orange dotted and red clusters reflects the transition from induced fibroblast to iPSC (Figure 2A).

Because PCA components consist of contributions from all 48 genes, it is possible to identify the most information-rich genes in classifying the two clusters (Figure 2B). Of the genes examined, *Thy1*, *Col5a2*, *Bmi1*, *Gsk3b*, and *Hes1* were the most specific markers of the first cluster. For the second cluster, it was *Dppa2*, *Sox2*, *Nanog*, *Esrrb*, *Oct4*, *Sall4*, *Utf1*, *Lin28*, and *Nr6a1*, whereas several other pluripotency genes were not strictly associated. For example, *Fut4* and *Grb2* do not significantly differentiate between the two clusters. Similarly, genes such as *Stat3*, *Hes1*, *Jag1*, *Gsk3b*, *Bmpr1a*, *Nes*, and *Wnt1*, which are known to be important for the ESC state, are less indicative of the second cluster (Figure 2B).

To examine within-group variability combining all genes, we used Jensen-Shannon Divergence (JSD) (Figures 2C and 2D). The parental NGFP2 iPSCs were the least variable group. An increase in variation was seen in MEFs when dox was added followed by a steep decrease after the activation of the *Nanog*

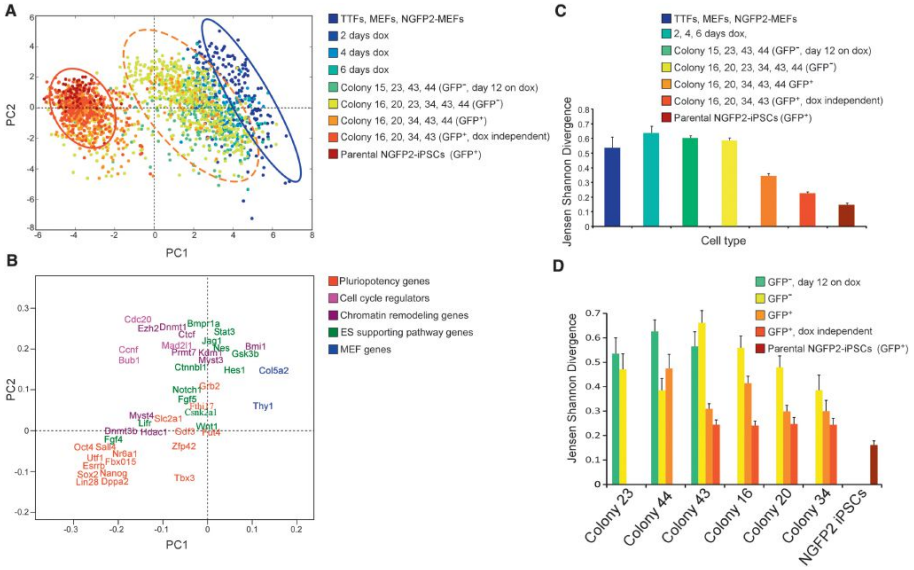


Figure 2. Three Reprogramming States

(A) Principal component (PC) projections of individual cells, colored by their sample identification. The blue circle surrounds one population, and the red circle surrounds another population. The orange dotted circle surrounds a third intermediate population.

(B) PC projections of the 48 genes, showing the contribution of each gene to the first two PCs. The first PC can be interpreted as discriminating between cluster 1 and cluster 2; the second between pluripotency genes and cell-cycle regulators.

(C and D) JSD analysis of within-group (C) and within-colony (D) variability, colored by the same sample identification as in (A). Error bars represent the 95% confidence interval.

See also Figure S3.

locus (GFP$^+$ cells), suggesting that the activation of the endogenous *Nanog* locus marks events that drive the cells to pluripotency (Silva et al., 2009). Notably, although the dox-independent cells were derived from the same parental cells, they exhibited a higher variation (red) than their parental cells (brown), indicating that each reprogramming event (colony) results in a slightly different epigenetic state (Figure 2C).

We used JSD to further examine the variation within and between colonies (Figure 2D) and found that the variation between GFP$^-$ and GFP$^+$ cells within a colony was similar to that among all colonies (Figure 2C). Colony 44, which contained only a few cells with low GFP (Figure S1E), exhibited high variation between the GFP$^+$ cells. Colonies 20 and 34, which gave rise to early stable dox-independent iPSC colonies, showed low variation between late GFP$^-$ cells (Figure 2D) even early in the process. Notably, all of the colonies that gave rise to fully reprogrammed iPSCs (colonies 43, 16, 20, 34) exhibited a similarly low variation between GFP$^+$ dox-independent cells, indicating significantly reduced variation between single cells after core circuitry activation.

Analysis of Induced Cells that Do Not Give Rise to iPSCs

Upon retrospective tracing, we found two colonies, 23 and 44, that failed to give rise to stable iPSCs (Figure S4A). Both exhibited early dedifferentiating morphological changes associated with reprogramming (Smith et al., 2010), with colony 23 producing homogenous cultures of cells with epiblast stem cell-like morphology (flat colonies), and colony 44 producing transformed-like cells. Colony 23 failed to activate GFP in most cells with only a small fraction activating the endogenous *Nanog* locus (0.01% GFP$^+$) even after 81 days of culture. Colony 44 contained a few cells with a low level of GFP that appeared at day 61 and disappeared upon continued passaging and dox withdrawal. Because colonies 23 and 44 did not generate iPSCs, they were designated as "partially reprogrammed colonies." We treated colonies 23 and 44 with the DNA methyltransferase inhibitor 5-aza-cytidine (azaC) to test whether methylation of pluripotency genes contributed to the partially reprogrammed state (Mikkelsen et al., 2008). After 30 days of azaC and dox treatment followed by 8 days of azaC and dox withdrawal, GFP$^+$ cells appeared at a frequency of 2.2% in colony 23 and 0.5% in colony

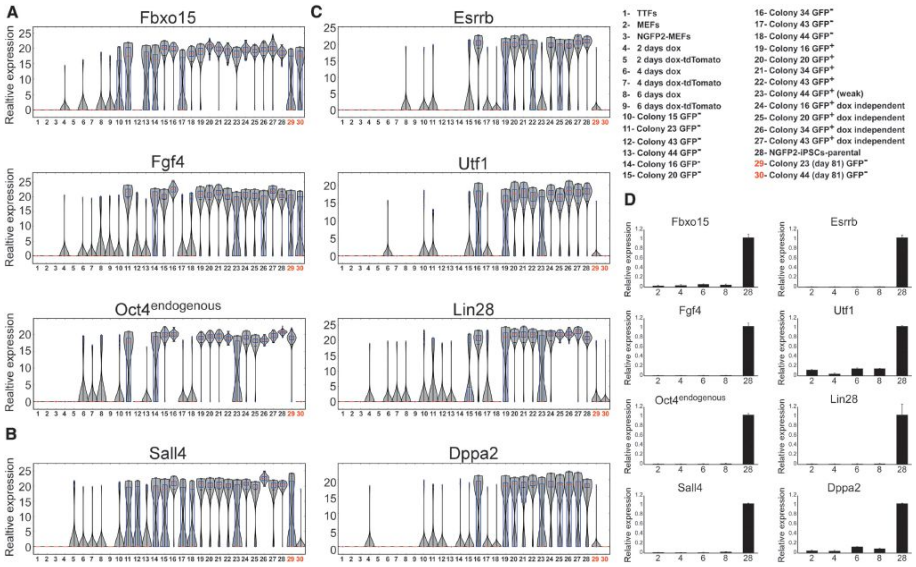


Figure 3. Established Early Markers Are Not Sufficient to Mark Cells that Will Become iPSCs

(A–C) mRNA expression levels of Fbxo15, Fgf4, and Oct4 (A); Sall4 (B); and Esrrb, Utf1, Lin28, and Dppa2 (C) in populations noted in Figure 1 and legend (upper right) are shown in violin plots. Median values are indicated by red line, lower and upper quartiles by blue rectangle, and sample minima/maxima by black line. The two partially reprogrammed colonies (colonies 23 and 44) are marked in red.

(D) Quantitative RT-PCR of Fbxo15, Fgf4, Oct4, Sall4, Esrrb, Utf1, Lin28, and Dppa2 expression in nonclonal cell populations noted in legend (upper right numbers correspond to x axis), normalized to the Hprt house-keeping control gene. Error bars are presented as a mean ± standard deviation (SD) of two duplicate runs from a typical experiment.

See also Figures S4 and S5.

44, compared to none in untreated cells (Figure S4B). These partially reprogrammed colonies were used as a control for fully reprogrammed colonies.

To determine whether the variability in single-cell gene expression was a result of differences between distinct cell populations or just stochastic noise, we analyzed our data with violin plots. Population noise and gene expression noise should exhibit unimodal distribution around a reference level in these density plots, whereas a multimodal distribution is indicative of distinct gene expression differences between cell populations. Of the genes examined, we identified a highly conserved zinc finger protein, Ctcf (Phillips and Corces, 2009), exhibiting unimodal distributions of extremely high expression only in the partially reprogrammed colony 23 tdTomato$^+$/GFP$^-$ cells (Figure S4C). To determine whether Ctcf interfered with reprogramming, we overexpressed Ctcf in NGFP2 MEFs (Figure S4D). This resulted in reduced AP staining and fewer GFP$^+$ cells (seen by FACS) after 13 days of dox exposure followed by 3 days of dox withdrawal, suggesting that controlled levels of Ctcf may be important for the reprogramming process (Figures S4E and S4F).

Early Markers of Reprogramming

High proliferation is one of the hallmarks of mESCs. As an initial control, we used violin plots to analyze the expression of four well-known mESC cell-cycle regulators, *Bub1*, *Ccnf*, *Cdc20*, and *Mad2l1*. As expected, the expression levels of these genes in single cells were upregulated and were most uniformly expressed in later stage cells and in dox-independent iPSCs (Figure S5A). To examine the expression of established early markers in reprogramming, we analyzed the expression profiles of three well-known markers, Fbxo15, Fgf4, and endogenous Oct4 (Brambrink et al., 2008; Takahashi and Yamanaka, 2006) (Figure 3A). Of the genes examined, all three genes exhibited high expression levels very early in the process (days 2, 4, 6) in a few cells (1 to 8 cells) and were highly expressed in the GFP$^+$ cells as expected for potential early markers. Very early and late in the process, the expression levels of *Fbxo15*, *Fgf4*, and endogenous *Oct4* were unimodal, with a very narrow peak indicating low variation between individual cells.

We noted that *Fbxo15*, *Fgf4*, and endogenous *Oct4* were expressed in some of the partially reprogrammed colonies, 44

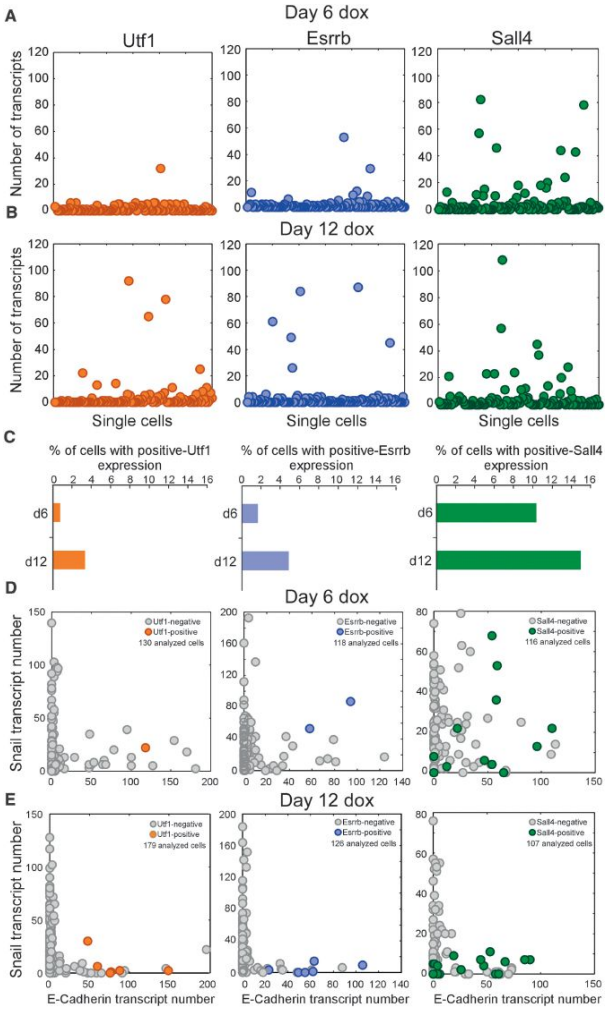


Figure 4. Early Markers for Reprogramming

(A and B) sm-mRNA-FISH of Utf1 (orange), Esrrb (blue), and Sall4 (green) expression in NGFP2 cells at days (A) 6 and (B) 12 on dox. Each cell is represented as a single dot. One hundred and twenty cells were analyzed for each one of the six plots.

(C) Percent of total cell population with high Utf1, Esrrb, and Sall4 at day 6 and day 12.

(D and E) sm-mRNA-FISH of Snail versus E-cadherin expression in single NGFP2 cells at days (D) 6 and (E) 12 on dox. High Utf1 (orange), Esrrb (blue), and Sall4 (green) cells are highlighted. The number of cells analyzed is noted on each plot.

and 23, at levels similar to those seen in iPSCs (Figures 3A and S5B) with *Fbxo15* and *Fgf4* showing a bimodal distribution. Of particular interest is the observation that endogenous *Oct4* was highly expressed in the partially reprogrammed colony 23, suggesting that activation of *Oct4* can occur in partially reprogrammed cells with incomplete reactivation of the core regulatory circuitry. Although exogenous Oct4 is one of the key factors in the reprogramming process, its endogenous activation was insufficient to identify cells as fully reprogrammed and thus cannot be used as a predictive marker for reprogramming.

Also, five additional genes, *Sall4, Esrrb, Utf1, Lin28,* and *Dppa2*, were activated early in a few cells and were highly expressed in GFP$^+$ cells (Figures 3B and 3C). We separated these genes into two classes: (1) nonpredictive, like *Sall4* that was activated very early but was also activated robustly in partially reprogrammed cells (Figures 3B and S5B), and (2) more predictive, like *Esrrb*, *Utf1*, *Lin28*, and *Dppa2* that were activated early in a small fraction of cells but exhibited only low if any expression in partially reprogrammed cells. The distribution of *Esrrb*, *Utf1*, *Lin28*, and *Dppa2* expression was unimodal early and late in the reprogramming process, with a narrow peak indicative of low variation between individual cells (Figure 3C). The expression of the predictive markers also distinguished between tdTomato$^+$/GFP$^-$, tdTomato$^+$/GFP$^+$, and tdTomato$^-$/GFP$^+$ cells (Figure S5C). Of note is that the variability between single cells in early time points was masked in nonclonal cell populations, as detected by qRT-PCR (Figure 3D).

To validate the Fluidigm results, we utilized the sm-mRNA-FISH technique and quantified transcripts of the nonpredictive marker, *Sall4*, and two potential predictive markers, *Esrrb* and *Utf1*, in single NGFP2 MEFs on dox for 6 and 12 days. At day 6, only 1 to 2 cells out of 125 examined cells showed relatively high levels of *Utf1* and *Esrrb*, reflecting the low efficiency of the reprogramming process (Figure 4A), consistent with the Fluidigm analysis. In contrast, *Sall4* exhibited the highest number of cells with high expression levels, which is in agreement with the violin plots

(Figures 3B and 3C). Our analysis found that only 1%–2% of the cells sampled at day 6 and 2%–5% of the cells sampled at day 12 had high expression of *Utf1* and *Esrrb*, whereas 10%–14% of the cells sampled at day 6 and day 12 had high expression of *Sall4* (Figures 4A and 4B). As expected, the number of high Utf1, Esrrb, and Sall4 cells increased by day 12 (Figure 4C). These data suggest that *Esrrb* and *Utf1* are expressed in a few cells very early in the process and thus may represent early markers that predict an eventual reprogramming event of a given cell.

To gain insight into the early markers and MET at the single-cell level, we quantified transcripts of (1) *Snail*, *E-cadherin*, and *Esrrb*; (2) *Snail*, *E-cadherin*, and *Utf1*; and (3) *Snail*, *E-cadherin*, and *Sall4* in single NGFP2 MEFs on dox for 6 and 12 days. Figures 4D and 4E show that the number of E-cadherin$^+$/Snail$^+$ cells decreased whereas the number of E-cadherin$^+$/Snail$^-$ cells increased between day 6 and day 12. At day 6, *Utf1* and *Esrrb* were coexpressed with both *E-cadherin* and *Snail*, whereas at day 12, *Utf1* and *Esrrb* were only primarily coexpressed with *E-cadherin*. *Sall4* was coexpressed with *Snail* and *E-cadherin* at day 6 similarly to *Utf1* and *Esrrb* but also in many cells at day 12. Whereas all high Utf1 and Esrrb cells at day 6 and day 12 on dox expressed *E-cadherin*, only 70% of high Sall4 cells expressed *E-cadherin*. Further, the fraction of Sall4$^+$/E-cadherin$^-$ cells increased from 15% to 37% from day 6 to day 12 on dox, suggesting an accumulation of Sall4$^+$ cells that have not acquired epithelial properties. These data support the notion that MET and Sall4 represent nonpredictive markers, whereas Utf1 and Esrrb represent early and predictive markers.

Activation of Endogenous Sox2 Is a Late Phase in Reprogramming that Initiates a Series of Consecutive Steps toward Pluripotency

To investigate the later phases of reprogramming, we searched for potential late markers. Late markers would be expected to express no or very low transcript levels at early time points and high levels as the cells mature and become iPSCs. We identified *Gdf3* and *Sox2* as genes that appeared late in the process with very low early expression levels as measured by Fluidigm and sm-mRNA-FISH (Figures S6A, S6B, S6D, and S6E). However, *Gdf3*, but not *Sox2*, was activated also in partially reprogrammed cells, identifying only Sox2 as a discriminating late marker for iPSCs (Figures S6C and S6F).

To examine whether reprogramming involves random or sequential activation of marker genes, we derived a Bayes network with a subset of cells that expressed all 48 genes taken at different times in the reprogramming process (Table S5). A Bayes network is a probabilistic model that represents a set of variables and their conditional dependencies. The Bayes network predicted that the activation of the endogenous *Sox2* locus initiates a series of consecutive steps leading to the activation of many pluripotency genes (Figure 5A). For example, given that *Sall4* is expressed, the expression of *Oct4*, *Fgf4*, *Nr6a1*, and *Fbxo15* is conditionally independent of whether *Sox2* is expressed or not. In contrast, if Sox2 initiates a sequence of gene activation and first turns on *Sall4*, which then activates the four downstream targets, one should not find cells that express *Sox2* and one of the four downstream genes (*Oct4*, *Fgf4*, *Nr6a1*, and *Fbxo15*) without *Sall4*. To examine whether the Bayes network predicted true consecutive steps in reprogramming, we investigated three scenarios: (1) Sox2 activates Sall4 and then activates the downstream gene *Fgf4*; (2) Sox2 first activates Lin28 and then induces the downstream gene *Dnmt3b;* (3) Sox2 activates Sall4 and then activates the downstream gene *Fbxo15*. To test these possibilities, we quantified transcripts by sm-mRNA-FISH (Figure 5B) of the three combinations of genes simultaneously in single secondary NGFP2 MEFs (Figures 5C–5E) and single primary-infected Sox2-GFP MEFs (Figures 5F–5H) kept on dox for 12 days, a time point when both fully reprogrammed cells and intermediate colonies have appeared. We designated a cell as "positive" if it expressed at least one transcript of a given gene. *Combination 1:* Although 186 cells out of a total of 279 cells examined were negative, 25 cells expressed one gene, 38 cells expressed two genes, and 30 cells expressed all three genes. Notably, no double-positive cells were seen that coexpressed *Sox2* and *Fgf4* (Figure 5C). *Combination 2:* Out of a total of 283 cells examined, 82 cells were positive for any of the genes with 49 cells expressing one, 23 cells expressing two, and 10 cells expressing all three genes, but no cells expressed just *Sox2* and *Dnmt3b* (Figure 5D). *Combination 3:* Of 275 cells examined, 101 cells were positive for either of the three genes with 50 cells expressing one, 30 cells expressing two, and 20 cells expressing all three genes, but only 1 cell expressed just *Sox2* and *Fbxo15* at a very low level (Figure 5E). The combinations examined in primary-infected cells were similar to the secondary cells in that no cells were seen that coexpressed *Sox2* and *Fgf4* (combination 1) and *Sox2* and *Dnmt3b* (combination 2) (Figures 5F and 5G). We identified two cells coexpressing *Sox2* and *Fbxo15*; however, similar to the one *Sox2*/*Fbxo15* coexpressing cell in the secondary system, these two cells each expressed only one *Sox2* transcript (Figure 5H). The primary infected cells had a significantly lower number of negative cells compared to the secondary system, probably due to high transgene levels in the primary infected cells. Generally, the largest fraction of cells with gene expression in each combination was that of the double-positive cells, *Sall4*/*Fgf4, Lin28*/*Dnmt3b*, and *Sall4*/*Fbx015*, indicating that the activation of *Sall4* and *Lin28* is more promiscuous than the activation of the *Sox2* locus (Figures 5F–5H). These data support the sequential activation of *Sall4* and *Lin28* by Sox2 followed by the activation of *Fgf4*, *Fbxo15*, and *Dnmt3b,* respectively, consistent with a model of a hierarchical activation of key pluripotency genes.

The Hierarchical Model of Gene Activation Predicts Downstream Transcription Factor Combinations Capable of Inducing Reprogramming

To assess whether sequential activation of key pluripotency genes can predict their role in inducing reprogramming, we infected *Oct4-GFP* MEFs with transcription factor combinations derived from the top node of the network (Sox2), the middle nodes (Esrrb, Sall4, Lin28), and the bottom nodes (Oct4 and Nanog). We chose three combinations of factors that were predicted to induce activation of the pluripotency circuitry and generate fully reprogrammed iPSCs: (1) Oct4, Esrrb, Nanog; (2) Sox2, Sall4, Nanog; and (3) Lin28, Sall4, Esrrb, Nanog. These

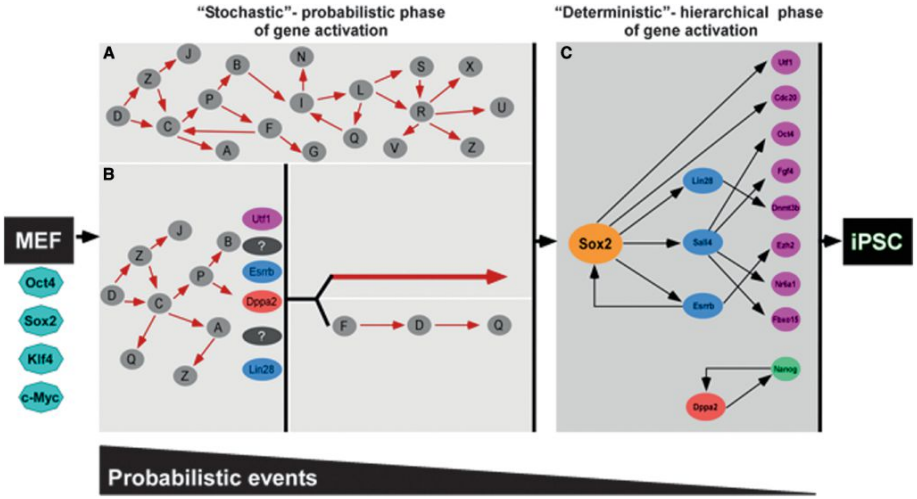


Figure 7. Two Phases in Reprogramming

The reprogramming process can be split into two phases: an early stochastic phase (A and B) of gene activation followed by a later more deterministic phase (C) of gene activation that begins with the activation of the Sox2 locus. After a fibroblast is induced with OSKM, the cell can proceed into either one of two stochastic phases. In (A), stochastic gene activation can lead to the activation of the Sox2 locus. In (B), stochastic gene activation can lead to the activation of "predictive markers" like Utf1, Esrrb, Dppa2, and Lin28, which then mark cells that have a higher probability of activating the Sox2 locus. Activation of the Sox2 locus can be via two potential paths: (1) direct activation of the Sox2 locus or (2) sequential gene activation that leads to the activation of the Sox2 locus. In this model, probabilistic events decrease and hierarchal events increase as the cell progresses from fibroblast to iPSC. Solid red arrows and black arrows denote hypothetical interactions and interactions supported by our data, respectively. The white gap shown between the stochastic (A and B) and deterministic (C) panels represents the transition from induced fibroblast to iPSC illustrated between the orange dotted cluster and red cluster in Figure 2A.

reprogramming have been based on gene expression measurements over heterogeneous populations of cells, precluding insight into events that occur in the rare single cells that ultimately become iPSCs.

Our data are in agreement with the stochastic model but also suggest a sequence of gene activation at later stages (Figure 7). The significant variation between sister cells of initial colonies that does not reveal a specific sequential order of gene expression supports a stochastic mechanism of gene activation early in the process (Figure 7A). Based on the Bayes network model derived from single-cell data, a second later phase of reprogramming seems to be governed by a more sequential or hierarchical mechanism of gene activation with activation of Sox2 initiating consecutive steps that lead to the pluripotent state (Figure 7C). However, our data are also consistent with the possibility that the activation of ''predictive'' markers such as Esrrb or Utf1 represents a key event that either directly activates the *Sox2* locus or initiates a sequence of gene activations eventually resulting in *Sox2* activation (Figure 7B).

Sox2 is indispensable for maintaining ESC pluripotency because Sox2 null ESCs differentiate primarily into trophoectoderm-like cells, and it was suggested, consistent with our hypothesis, that Sox2 contributes to the activation of *Oct4* by maintaining high levels of orphan nuclear receptors like *Nr5a2 (Lrh1)* (Masui et al., 2007). In agreement with this observation, removing Esrrb from a cocktail of transcription factors (Lin28, Sall4, Nanog, Ezh2, Klf4, and c-Myc) yielded iPS-like colonies that were unstable due to their failure to activate the core pluripotency circuitry. Thus, early in the reprogramming process, the four factors induce the somatic cells to acquire epigenetic changes by a stochastic mechanism, leading to an intermediate or partially reprogrammed state (Egli et al., 2008). Activation of endogenous *Sox2* represents a late cell state and can be considered as a first step that drives a consecutive chain of events that allow the cell to enter the pluripotent state.

We show that the activation of the pluripotency circuitry is possible by various subsets of transcription factors even without Oct4, Sox2, Nanog, c-Myc, and Klf4. It is important to note the difference between timing or promiscuity of promoter reactivation during reprogramming and reprogramming potency of the transcription factors. Not all genes that facilitate reprogramming will be predictors of iPSCs. Although Oct4 is very efficient in the reactivation of the core pluripotency circuitry, its own activation does not necessarily predict which cells will become iPSCs (Figure 3). Similarly, Sall4 is a strong inducer of reprogramming but is not predictive of future iPSCs. Lin28,