Published in final edited form as:
Cell. 2012 September 14; 150(6): 1209–1222. doi:10.1016/j.cell.2012.08.023.

Single-cell gene expression analyses of cellular reprogramming reveal a stochastic early and hierarchic late 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, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [3]Department of Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [4]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [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

Abstract

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 Fbxo15, Fgf4, and Oct4 previously suggested to be reprogramming markers. 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 the pluripotency circuitry.

Introduction

Differentiated cells can be reprogrammed to a pluripotent state by overexpression of Oct4, Sox2, Klf4, and c-Myc (OSKM) (Takahashi and Yamanaka, 2006). Fully reprogrammed induced pluripotent stem cells (iPSCs) can contribute to the three germ layers and give rise to fertile mice by tetraploid complementation (Okita et al., 2007; Zhao et al., 2009). The reprogramming process is characterized by widespread epigenetic changes (Koche et al., 2011; Maherali et al., 2007; Mikkelsen et al., 2008) that generate iPSCs that functionally and molecularly resemble embryonic stem (ES) cells.

*Correspondence to Rudolf Jaenisch (jaenisch@wi.mit.edu).
[7]These authors contributed equally to this work.

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 de-differentiation 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 were 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 reprogramming factors initiate a sequence of probabilistic events that eventually lead to the small and unpredictable fraction of iPSCs. Clonal analyses support the stochastic model, demonstrating that activation of pluripotency markers occurs at different times after infection in individual daughters of the same fibroblast (Meissner et al., 2007). However, since the molecular changes occurring at the different stages during the reprogramming process were based upon the analysis of heterogeneous cell populations, it has not been possible to clarify the events that occur in the rare single cells that eventually form an iPSC. Moreover, there has been little insight into the sequence of events that drive the process.

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* do 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 “non-stochastic” or more ”hierarchical” phase of gene expression responsible for the activation of the endogenous pluripotent 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: (i) 96.96 Dynamic Array chips (Fluidigm), which allows quantitative analysis of 48 genes in duplicate in 96 single cells (Guo et al., 2010), and (ii) 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). Because reprogramming requires a vast number of epigenetic changes, we chose a group of ES-associated chromatin remodeling genes and modification enzymes [*Myst3, Kdm1, Hdac1, Dnmt1, Prmt7, Ctcf, Myst4, Dnmt3b, Ezh2, Bmi1*] (Reik, 2007;

Surani et al., 2007). Since 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 ES cells maintenance and differentiation [*Bmpr1a, Stat3, Ctnnbl1, Nes, Wnt1, Gsk3b, Csnk2a1, Lifr, Hes1, Jag1, Notch1, Fgf5, Fgf4*] (Boiani and Scholer, 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, Fthi17, 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 six days. Because transgene expression between single cells was more variable in the virus-infected MEFs we used the secondary system for all analyses (Figure 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 drug addition with the first colonies appearing around seven days after dox addition. Thus, to detect early transcriptional changes in the reprogramming process, non-clonal populations of NGFP2 MEFs were exposed to dox for two, four and six days. At each time point, the cells were imaged, 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 six days, we utilized a modified experimental setup. Because most cells senesced, became contact inhibited or transformed after exposure to dox for six 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 both for cell-cell interactions enabling proliferation of single cells and calibration of the FACS machine (i.e tdTomato+ cells vs tdTomato− cells). This system allowed tracing 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 six days, sorted for tdTomato and seeded each as a single cell in one well 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 the Fluidigm. Each parental cell was passaged to test its capacity to generate doxindependent, 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: (i) early dox-dependent GFP− cells (ii) intermediate dox-dependent GFP− and GFP+ cells and (iii) 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 passage without dox. A few cells (0.01%) from colony 23 activated GFP only at day 81.

To gain insight into intermediate clonal cell populations, we analyzed single tdTomato+/GFP+ double-positive cells from colony 20 at day 32 in dox by Fluidigm. 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 towards iPSCs or are ‘stuck’, we sorted twenty 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 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 9216 qRT-PCR reactions. The output of one run on the Biomark is a 96x96 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 principle components. 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 1864 cells from different stages during reprogramming we found that the first principal component (PC1) explains 22.5% of the observed variance while 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 2 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), mouse embryonic fibroblasts (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 ES-like morphology may not be dox-independent. Cells on dox for four 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 cluster 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 ES cell 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* 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 further examined the variation within and between colonies using JSD (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 iPS 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 de-differentiating 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 doxwithdrawal. Because colonies 23 and 44 did not generate iPSCs, they were designated as ‘partially reprogrammed colonies’. We tested whether methylation of pluripotency genes contributed to the partially reprogrammed state by treating colonies 23 and 44 with the DNA methyltransferase inhibitor 5-aza-cytidine (azaC) (Mikkelsen et al., 2008). After thirty days of azaC and dox treatment followed by

eight days of azaC and dox withdrawal, GFP+ cells appeared at a frequency of 2.2% in colony 23 and 0.5% in colony 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 if 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 day 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 analyzed the expression of four well-known mESC cell cycle regulators, *Bub1, Ccnf, Cdc20* and *Mad2l1* using violin plots. 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 (day 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 and 23 at levels similar to those seen in iPS cells (Figure 3A and Figure 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 predictive markers 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: (i) non-predictive, like *Sall4* that was activated very early but was also activated robustly in partially reprogrammed cells (Figure 3B and Figure S5B) and (ii) 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 non-clonal cell populations as detected by qRTPCR (Figure 3D).

To validate the Fluidigm results, we utilized the sm-mRNA-FISH technique and quantified transcripts of the non-predictive marker, *Sall4*, and two potential predictive markers, *Esrrb* and *Utf1*, in single NGFP2 MEFs on dox for six and twelve 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 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 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 NGFP 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 co-expressed with both E-cadherin and Snail, while at day 12 Utf1 and Esrrb were only co-expressed with E-cadherin. Sall4 was co-expressed with Snail and E-cadherin at day 6 similarly to Utf1 and Esrrb but also in many cells at day 12. These data support the notion that MET and Sall4 represent non-predictive markers, while 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 and S6D–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 using a subset of cells that expressed all 48 genes taken at different times in the reprogramming process. 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 on 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: (i) Sox2 activates Sall4 and then activates the downstream gene *Fgf4*. (ii) Sox2 first activates Lin28 and then induces the downstream gene *Dnmt3b*. (iii) 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 1 transcript of a given gene. *Combination 1:* While 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 co-expressed *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 one 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 co-expressed *Sox2* and *Fgf4* (Combination 1) and *Sox2* and *Dnmt3b* (Combination 2) (Figures 5F and 5G). We identified two cells co-expressing *Sox2* and *Fbxo15*; however, similar to the one *Sox2/Fbxo15* co-expressing 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 doublepositive 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 genes 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 three combinations omitted either Sox2 or Oct4 or both. Combination (1) replaced Sox2 with Esrrb because the network predicted that Esrrb could activate Sox2 (Figure 6A). Combination (2) replaced Oct4 with Sall4 because Sall4 was predicted to be upstream of Oct4 (Figure 6B). Combination (3) omitted both Sox2 and Oct4 because the model predicted that Lin28, Sall4, Esrrb, and Nanog could drive the cells to pluripotency independently of the two master regulators Sox2 and Oct4 (Figure 6C). Nanog was co-transduced in all combinations because the model predicted that it functioned independently of Sox2 and Oct4 (Figure 5A). MEFs were transduced with the three different combinations as well as with Klf4 and c-Myc to induce proliferation. After 25 days on dox, GFP was detected by flow cytometry at a frequency of 22.2%, 0.3%, and 0.4%, respectively in the three combinations (Figures 6A–6C). These data are consistent with exogenous *Oct4* facilitating the activation of the endogenous circuitry but not being essential. Finally, we transduced the cells with combination (3) but without Klf4 and c-Myc. GFP was detected by flow cytometry after 25 days on dox at a frequency of 0.6%, indicating that Klf4 and c-Myc were not required to drive the cells toward pluripotency (Figure 6D).

To test whether Dppa2 has a role in the activation of the core pluripotency as predicted by the model, we infected both Oct4-GFP and Nanog-GFP MEFs with modified combination (1) and (4), whereby Nanog was replaced by Dppa2 (Figures 6E, 6F, and S7A). For modified combination 1 (Oct4, Esrrb, Dppa2, Klf4, c-Myc), GFP was detected by flow

cytometry after 16 days on dox followed by five days of dox withdrawal at a frequency of 0.6% and 0.2% in the Oct4-GFP MEFs and Nanog-GFP MEFs, respectively. For modified combination 4 (Lin28, Sall4, Esrrb, Dppa2), GFP was detected by flow cytometry after 16 days on dox followed by five days of dox withdrawal at a frequency of 0.2% and 0.1% in the Oct4-GFP MEFs and Nanog-GFP MEFs, respectively. Dox-independent iPSCs from all combinations were GFP+ as detected by microscopy and generated chimeras (Figures 6A–6F).

To determine the importance of a particular functional link in the network, we transduced the Oct4-GFP MEFs with Lin28, Sall4, Ezh2, Nanog, Klf4 and c-Myc (a modified combination 3), replacing Esrrb with its downstream target Ezh2 as predicted from the model. After 25 days on dox, abundant amounts of transformed cells were found on the plate, and 1-day post dox withdrawal there appeared to be some cells that morphologically resembled iPSCs. However, 7 days after dox withdrawal, no stable iPS colonies were found, suggesting incomplete reactivation of the core circuitry required for fully reprogrammed iPSCs consistent with failure to detect GFP+ cells (Figure 6G). It is tempting to speculate that the absence of Esrrb from the combination prevented the activation of endogenous Sox2 and the pluripotency circuitry. To test whether Ezh2 has a negative effect on the reprogramming process that might be responsible for the observed incomplete reprogramming process, we transduced NGFP2 MEFs with a viral construct expressing Ezh2 and monitored its effect on the reprogramming process. In parallel, we transduced the cells with shRNA for Ezh2 and monitored its effect on the reprogramming process. Overexpressing Ezh2 enhanced reprogramming and knocking down inhibited reprogramming, consistent with a positive effect of Ezh2 (Figures S7B–S7E).

To test the synergistic effects of our and the Yamanaka factors, we transduced NGFP2 MEFs that harbor OSKM with *Lin28, Sall4, Esrrb*, and *Nanog* and found stable dox-independent iPS colonies with GFP+ cells with a frequency of 2.2% after only five days of dox exposure (Figure 6H). Flow cytometric analysis of secondary cells carrying these factors generated 1.9% GFP+ cells after 5 days of growth in dox followed by 3 days without dox but none in the controls (Figure 6I). To examine the effect of each of the four transcription factors in facilitating the reprogramming process, we transduced NGFP2 MEFs with Lin28, Sall4, Esrrb or Nanog individually. The factors had different effects with Lin28, Sall4, and Esrrb facilitating the reprogramming after 10 days of dox exposure followed by 4 days of dox withdrawal and Nanog enhancing the process after 13 days of dox followed by 3 days of dox withdrawal (Figures S7F and S7G). Our results show that various factor combinations can activate the pluripotency circuitry even in the absence of exogenous Oct4, Sox2, and Nanog, and support our model of activation that drives the cell toward transgene independency.

Discussion

While single-cell gene expression analysis has been applied previously to studies in the mouse intestine (Itzkovitz et al., 2011), human colon tumors (Dalerba et al., 2011), the mouse zygote and blastocyst (Guo et al., 2010; Tang et al., 2010), and human iPSCs (Narsinh et al., 2011), such an approach has not been used to define the cell states and molecular transitions during the conversion of somatic cells to iPSCs.

Two models, designated as a ‘stochastic’ or a ‘deterministic’ process, have been proposed to explain the mechanism of reprogramming (Hanna et al., 2009; Yamanaka, 2009). A number of studies are most consistent with the stochastic model (Hanna et al., 2009) posing that the reprogramming factors in fibroblasts initiate a sequence of stochastic events that eventually leads to the small and unpredictable fraction of iPS cells (Jaenisch and Young, 2008). In

contrast, nuclear transfer (Boiani et al., 2002) or cell fusion (Bhutani et al., 2010) induce reprogramming rapidly and possibly as a single event with little heterogeneity observed in somatic cells, possibly consistent with a deterministic process (Hanna et al., 2010). So far the molecular analyses of reprogramming were 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 represent 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 ES-cell pluripotency because Sox2-null ES cells differentiated 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 cells to enter the pluripotent state.

We show that the activation of the pluripotent 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 pluripotent 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, Sall4, Esrrb, and Dppa2 were sufficient to generate fully reprogrammed iPSCs, albeit with lower efficiency than OSKM. It has been shown that Sall4 can activate the distal enhancer of Oct4 and together with Sall1, Utf1, Nanog, and c-Myc, can generate iPSCs in 2i condition, and that Esrrb can upregulate Sox2 and other pluripotency genes (Feng et al., 2009; Mansour et al., 2012; Zhang et al., 2006). Our Bayes model is consistent with these data.

Single-cell technology is in its infancy and our conclusions were based on the expression of 48 genes in approximately 7000 single cells. Clearly, genome-wide expression analyses in single cells would be highly informative. We chose MEFs as donor cell type as has been used in most previous studies and it is possible that other donor cell types may reveal different expression profiles.

In summary, single-cell gene expression analysis revealed an unanticipated heterogeneity in gene expression between sister cells, consistent with stochastic epigenetic alterations during

the early phase of the reprogramming process. This was followed by a more hierarchal mechanism late in the process where activation of some key genes predicts the expression of downstream genes and the establishment of the pluripotency circuitry. It will be of great interest to define the molecular determinants that drive the epigenetic changes during the early stochastic phase and the later more consecutive stage of reprogramming.

Materials and Methods

Quantitative real-time PCR

Total RNA was isolated using Rneasy Kit (QIAGEN) and reversed transcribed using a First Strand Synthesis kit (Invitrogen). Analysis was performed in an ABI Prism 7300 (Applied Biosystems) with SYBR green and ROX (Invitrogen). Details in Supplemental Methods.

Viral preparation and infection

Construction of lentiviral vectors containing OSKM under control of the tetracycline operator and a minimal CMV promoter has been described previously (Brambrink et al., 2008). Production of Lin28, Sall4, Ezh2, Esrrb, Nanog, and Utf1 in Supplemental Methods.

Chimera formation

All animal procedures were performed according to NIH guidelines and approved by the Committee on Animal Care at MIT. Blastocyst injections were performed as described previously (Wernig et al., 2007) and in Supplemental Methods.

Flow cytometry

Cells were trypsinized, washed once in PBS and resuspended in PBS + 5% FBS. The percentage of GFP+ cells was analyzed using FACS-LSR.

Secondary somatic cell isolation and culture

Primary NGFP2 iPSCs were electroporated with 25μg of linearized FUW-TetO-tdTomato construct. The transduced cells were selected using Zeocin (400ug/ml). MEF isolation and culturing was performed as described previously (Wernig 2008) and in Supplemental Methods.

FISH imaging and analysis

We performed FISH imaging and analysis as described previously (Raj et al., 2010; Raj et al., 2008) and in Supplemental Methods. Hybridizations were performed in solution using probes coupled to TMR, Alexa 594 (Invitrogen) or Cy5 (GE Amersham). Stacks of images spaced 0.3 um apart were taken with Nikon Ti-E inverted fluorescence microscope (Donatello) equipped with 100x oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software (Molecular Devices).

Single-cell data processing and visualization

PCA analysis and conversion of Ct values from the BioMark System into log-based expression values are described in Supplemental Methods.

Single-cell gene expression qPCR

Single-cell qPCR was performed as described previously (Diehn et al., 2009) and in Supplemental Methods. Single cells were sorted directly into RT-PreAmp Master Mix (CellsDirect) and pooled assays. Cell lysis, sequence-specific RT, and then sequence-

specific amplification of cDNA was performed. Products were analyzed and Ct values were calculated from the system's software.

Jensen-Shannon Divergence

Analysis was calculated to assess within-group similarity of gene expression within each cell line according to (Lin, 1991) and in Supplemental Methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Yarden Katz, Sovan Sarkar, and Jonathan Friedman for fruitful discussions, Patti Wisniewski and Chad Araneo for their help with cell sorting, and Stuart Levine for help with pilot Fluidigm experiments. Y.B. was supported by a NIH Kirschstein NRSA (1 F32 GM099153-01A1). D.A.F. is a Vertex Scholar and was supported by a NSF Graduate Research Fellowship and Jerome and Florence Brill Graduate Student Fellowship. A.W.C was supported by a Croucher and Ludwig Research Fellowship. R.J. is an adviser to Stemgent and cofounder of Fate Therapeutics. This work was supported by NIH grants HD 045022 and R37CA084198 to RJ and the NIH/NCI Physical Sciences Oncology Center at MIT (U54CA143874) and a NIH Pioneer award (1DP1OD003936) to A.v.O.

References

Bhutani N, Brady JJ, Damian M, Sacco A, Corbel SY, Blau HM. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature. 2010; 463:1042–1047. [PubMed: 20027182]

Boiani M, Eckardt S, Scholer HR, McLaughlin KJ. Oct4 distribution and level in mouse clones: consequences for pluripotency. Genes Dev. 2002; 16:1209–1219. [PubMed: 12023300]

Boiani M, Scholer HR. Regulatory networks in embryo-derived pluripotent stem cells. Nat Rev Mol Cell Biol. 2005; 6:872–884. [PubMed: 16227977]

Brambrink T, Foreman R, Welstead GG, Lengner CJ, Wernig M, Suh H, Jaenisch R. Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. Cell Stem Cell. 2008; 2:151–159. [PubMed: 18371436]

Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, Sim S, Okamoto J, Johnston DM, Qian D, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat Biotechnol. 2011; 29:1120–1127. [PubMed: 22081019]

Diehn M, Cho RW, Lobo NA, Kalisky T, Dorie MJ, Kulp AN, Qian D, Lam JS, Ailles LE, Wong M, et al. Association of reactive oxygen species levels and radioresistance in cancer stem cells. Nature. 2009; 458:780–783. [PubMed: 19194462]

Egli D, Birkhoff G, Eggan K. Mediators of reprogramming: transcription factors and transitions through mitosis. Nat Rev Mol Cell Biol. 2008; 9:505–516. [PubMed: 18568039]

Feng B, Jiang J, Kraus P, Ng JH, Heng JC, Chan YS, Yaw LP, Zhang W, Loh YH, Han J, et al. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. Nat Cell Biol. 2009; 11:197–203. [PubMed: 19136965]

Guo G, Huss M, Tong GQ, Wang C, Li Sun L, Clarke ND, Robson P. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell. 2010; 18:675–685. [PubMed: 20412781]

Hanna J, Saha K, Pando B, van Zon J, Lengner CJ, Creyghton MP, van Oudenaarden A, Jaenisch R. Direct cell reprogramming is a stochastic process amenable to acceleration. Nature. 2009; 462:595–601. [PubMed: 19898493]

Hanna JH, Saha K, Jaenisch R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. Cell. 2010; 143:508–525. [PubMed: 21074044]

Hong H, Takahashi K, Ichisaka T, Aoi T, Kanagawa O, Nakagawa M, Okita K, Yamanaka S. Suppression of induced pluripotent stem cell generation by the p53-p21 pathway. Nature. 2009; 460:1132–1135. [PubMed: 19668191]

Itzkovitz S, Lyubimova A, Blat IC, Maynard M, van Es J, Lees J, Jacks T, Clevers H, van Oudenaarden A. Single-molecule transcript counting of stem-cell markers in the mouse intestine. Nat Cell Biol. 2011; 14:106–114. [PubMed: 22119784]

Jaenisch R, Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell. 2008; 132:567–582. [PubMed: 18295576]

Koche RP, Smith ZD, Adli M, Gu H, Ku M, Gnirke A, Bernstein BE, Meissner A. Reprogramming factor expression initiates widespread targeted chromatin remodeling. Cell Stem Cell. 2011; 8:96–105. [PubMed: 21211784]

Li R, Liang J, Ni S, Zhou T, Qing X, Li H, He W, Chen J, Li F, Zhuang Q, et al. A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. Cell Stem Cell. 2010; 7:51–63. [PubMed: 20621050]

Lin JH. Divergence Measures Based on the Shannon Entropy. Ieee T Inform Theory. 1991; 37:145–151.

Macfarlan TS, Gifford WD, Agarwal S, Driscoll S, Lettieri K, Wang J, Andrews SE, Franco L, Rosenfeld MG, Ren B, et al. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev. 2011; 25:594–607. [PubMed: 21357675]

Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, Arnold K, Stadtfeld M, Yachechko R, Tchieu J, Jaenisch R, et al. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell Stem Cell. 2007; 1:55–70. [PubMed: 18371336]

Mansour AA, Gafni O, Weinberger L, Zviran A, Ayyash M, Rais Y, Krupalnik V, Zerbib M, Amann-Zalcenstein D, Maza I, et al. The H3K27 demethylase Utx regulates somatic and germ cell epigenetic reprogramming. Nature. 2012

Masui S, Nakatake Y, Toyooka Y, Shimosato D, Yagi R, Takahashi K, Okochi H, Okuda A, Matoba R, Sharov AA, et al. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol. 2007; 9:625–635. [PubMed: 17515932]

Meissner A, Wernig M, Jaenisch R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. Nat Biotechnol. 2007; 25:1177–1181. [PubMed: 17724450]

Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, Bernstein BE, Jaenisch R, Lander ES, Meissner A. Dissecting direct reprogramming through integrative genomic analysis. Nature. 2008; 454:49–55. [PubMed: 18509334]

Narsinh KH, Sun N, Sanchez-Freire V, Lee AS, Almeida P, Hu S, Jan T, Wilson KD, Leong D, Rosenberg J, et al. Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. J Clin Invest. 2011; 121:1217–1221. [PubMed: 21317531]

Ng HH, Surani MA. The transcriptional and signalling networks of pluripotency. Nat Cell Biol. 2011; 13:490–496. [PubMed: 21540844]

Okita K, Ichisaka T, Yamanaka S. Generation of germline-competent induced pluripotent stem cells. Nature. 2007; 448:313–317. [PubMed: 17554338]

Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009; 137:1194–1211. [PubMed: 19563753]

Raj A, Rifkin SA, Andersen E, van Oudenaarden A. Variability in gene expression underlies incomplete penetrance. Nature. 2010; 463:913–918. [PubMed: 20164922]

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods. 2008; 5:877–879. [PubMed: 18806792]

Ramalho-Santos M, Yoon S, Matsuzaki Y, Mulligan RC, Melton DA. "Stemness": transcriptional profiling of embryonic and adult stem cells. Science. 2002; 298:597–600. [PubMed: 12228720]

Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature. 2007; 447:425–432. [PubMed: 17522676]

Samavarchi-Tehrani P, Golipour A, David L, Sung HK, Beyer TA, Datti A, Woltjen K, Nagy A, Wrana JL. Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell. 2010; 7:64–77. [PubMed: 20621051]

Silva J, Nichols J, Theunissen TW, Guo G, van Oosten AL, Barrandon O, Wray J, Yamanaka S, Chambers I, Smith A. Nanog is the gateway to the pluripotent ground state. Cell. 2009; 138:722–737. [PubMed: 19703398]

Smith ZD, Nachman I, Regev A, Meissner A. Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. Nat Biotechnol. 2010; 28:521–526. [PubMed: 20436460]

Stadtfeld M, Maherali N, Breault DT, Hochedlinger K. Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. Cell Stem Cell. 2008; 2:230–240. [PubMed: 18371448]

Surani MA, Hayashi K, Hajkova P. Genetic and epigenetic regulators of pluripotency. Cell. 2007; 128:747–762. [PubMed: 17320511]

Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006; 126:663–676. [PubMed: 16904174]

Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell. 2010; 6:468–478. [PubMed: 20452321]

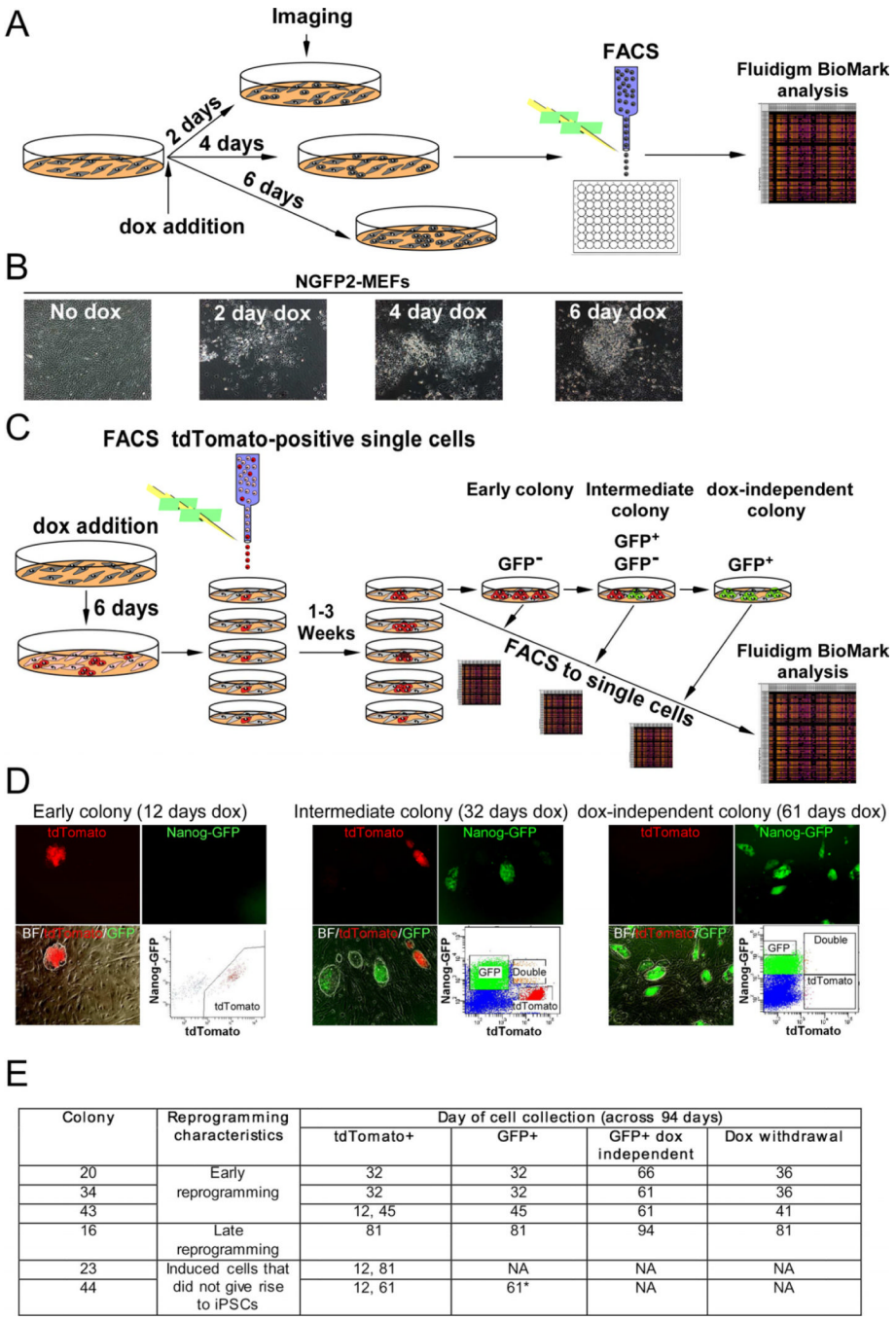
Wernig M, Lengner CJ, Hanna J, Lodato MA, Steine E, Foreman R, Staerk J, Markoulaki S, Jaenisch R. A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. Nat Biotechnol. 2008; 26:916–924. [PubMed: 18594521]

Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature. 2007; 448:318–324. [PubMed: 17554336]

Yamanaka S. Elite and stochastic models for induced pluripotent stem cell generation. Nature. 2009; 460:49–52. [PubMed: 19571877]

Zhang J, Tam WL, Tong GQ, Wu Q, Chan HY, Soh BS, Lou Y, Yang J, Ma Y, Chai L, et al. Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. Nat Cell Biol. 2006; 8:1114–1123. [PubMed: 16980957]

Zhao XY, Li W, Lv Z, Liu L, Tong M, Hai T, Hao J, Guo CL, Ma QW, Wang L, et al. iPS cells produce viable mice through tetraploid complementation. Nature. 2009; 461:86–90. [PubMed: 19672241]



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

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 Figure S1 and S2.

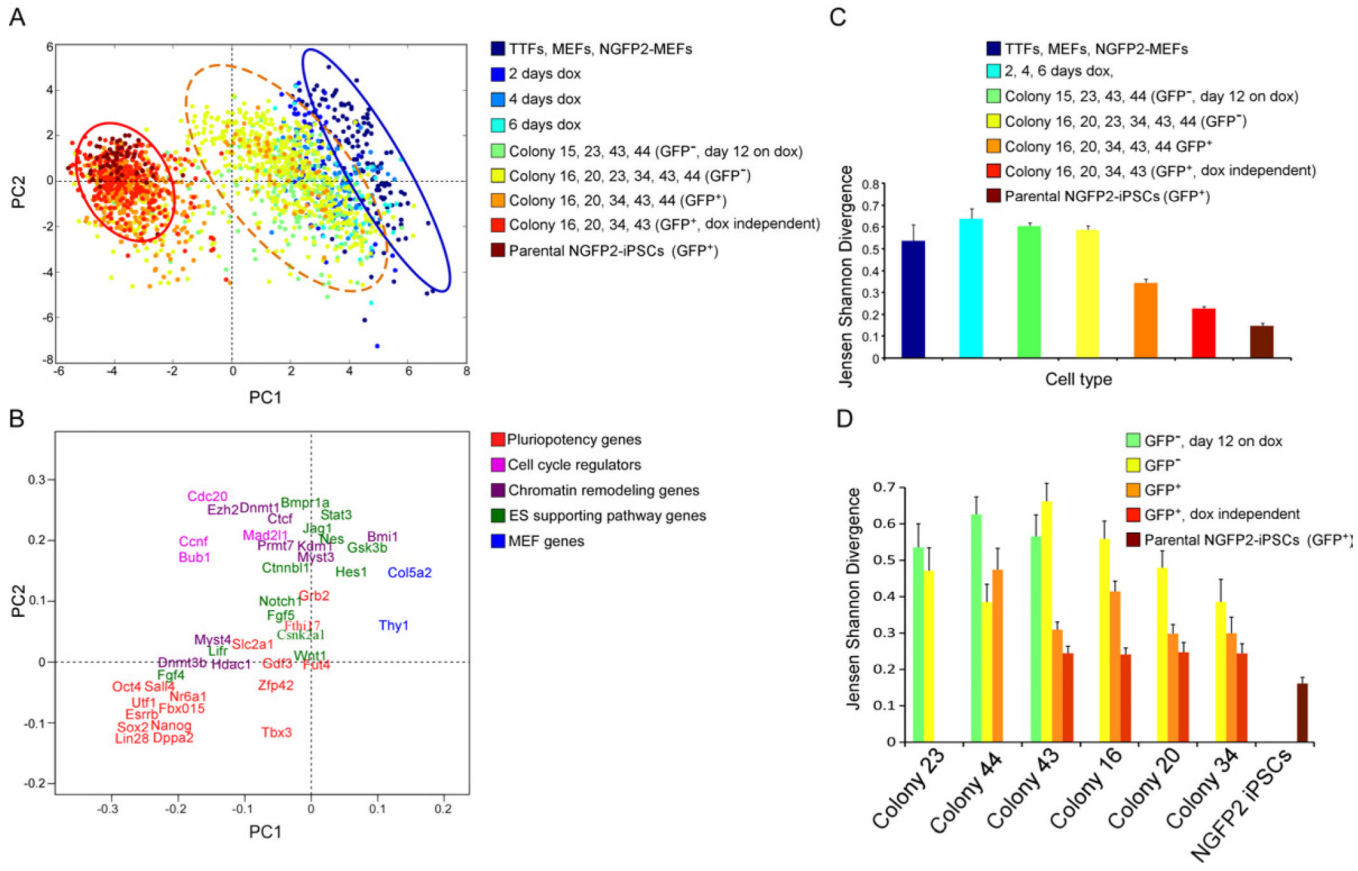


Figure 2. Three reprogramming states

(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-D) Jensen Shannon Divergence 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.

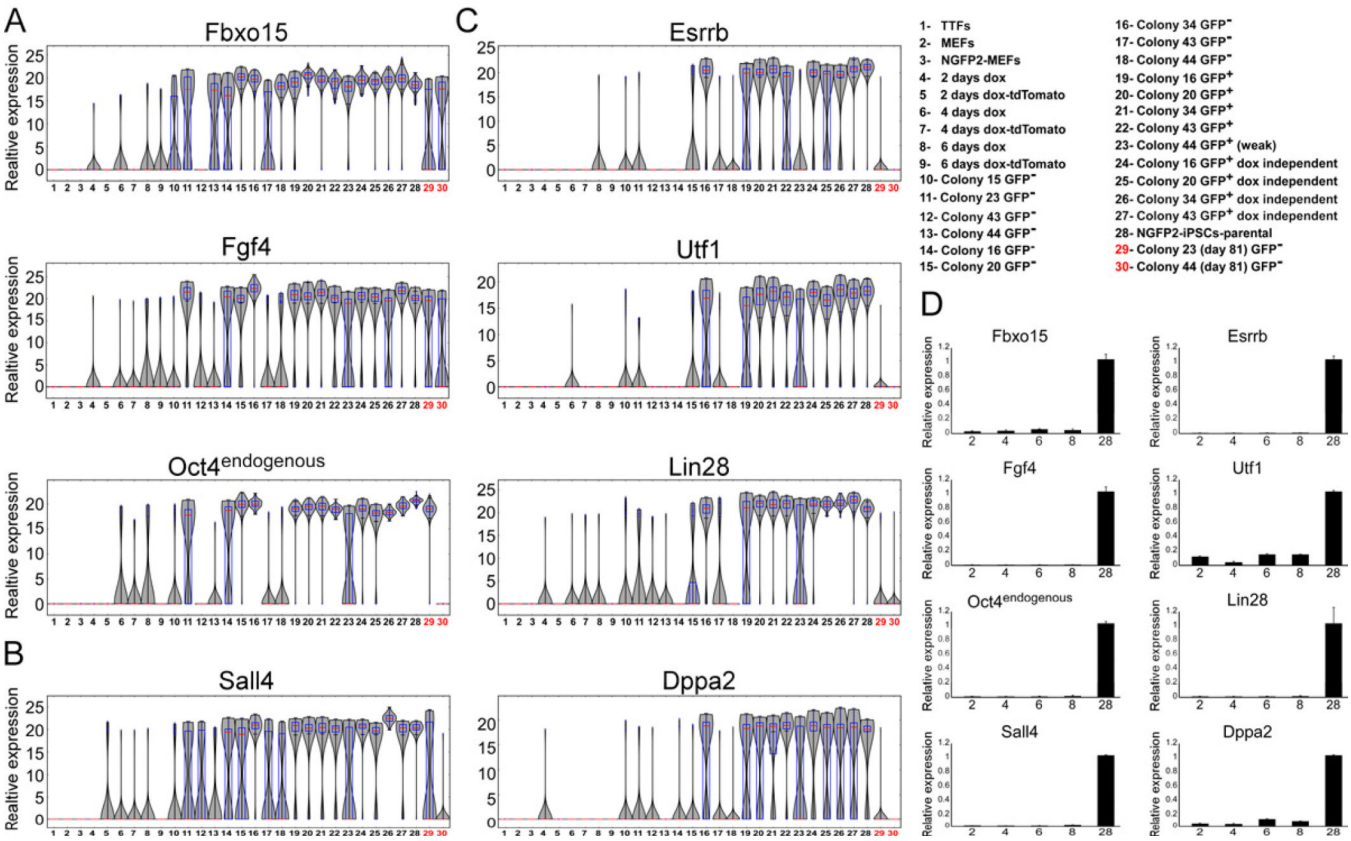


Figure 3. Established early markers are not sufficient to mark cells that will become iPSCs

mRNA expression levels of (A) Fbxo15, Fgf4 and Oct4 (B) Sall4 and (C) Esrrb, Utf1, Lin28, Dppa2 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 RTPCR 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 of two duplicate runs from a typical experiment. See also Figure S4 and S5.

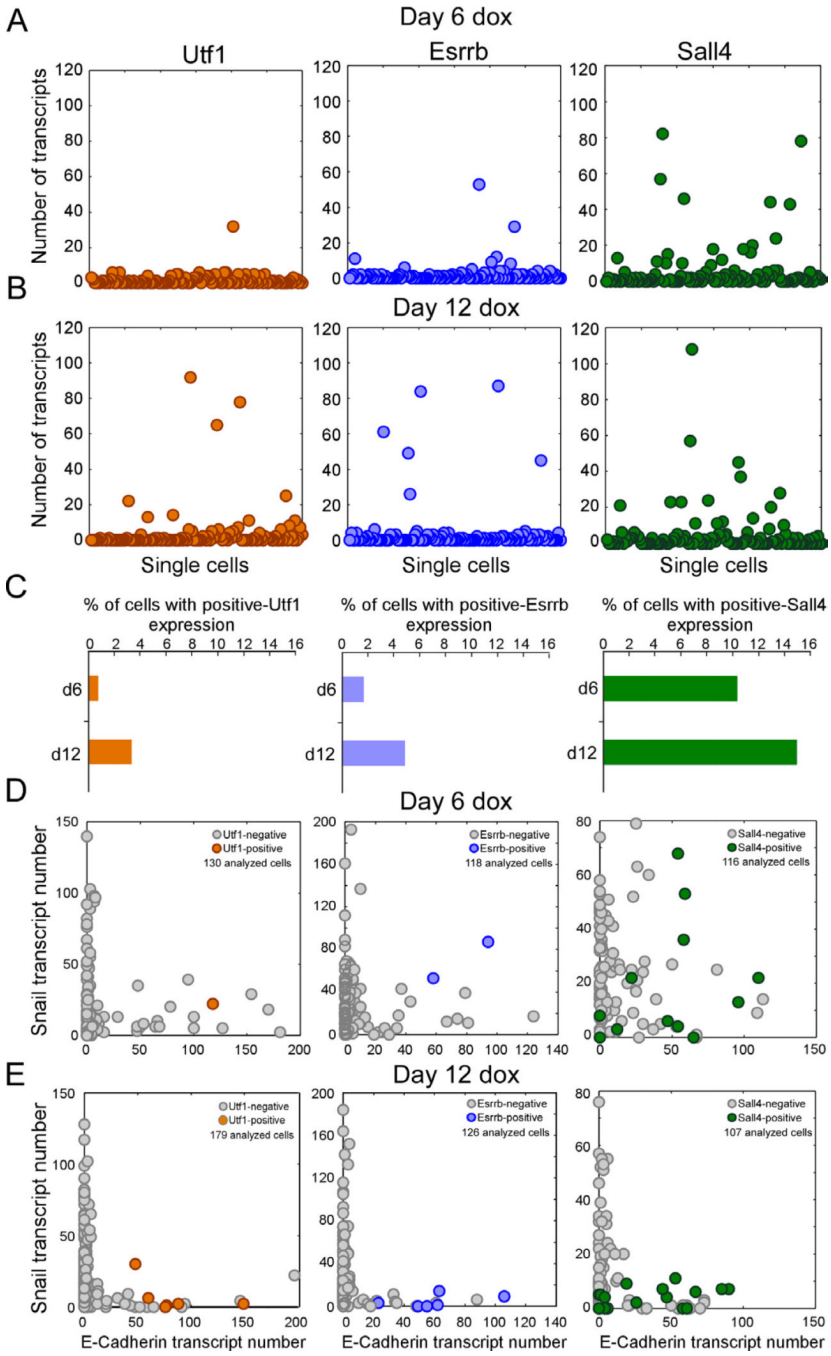


Figure 4. Early markers for reprogramming

(A and B) sm-mRNA-FISH of Utf1 (orange), Esrrb (blue), Sall4 (green) expression in NGFP2 cells at day (A) 6 and (B) 12 on dox. Each cell is represented as a single dot. 120 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 vs. E-cadherin expression in single NGFP2 cells at day (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.

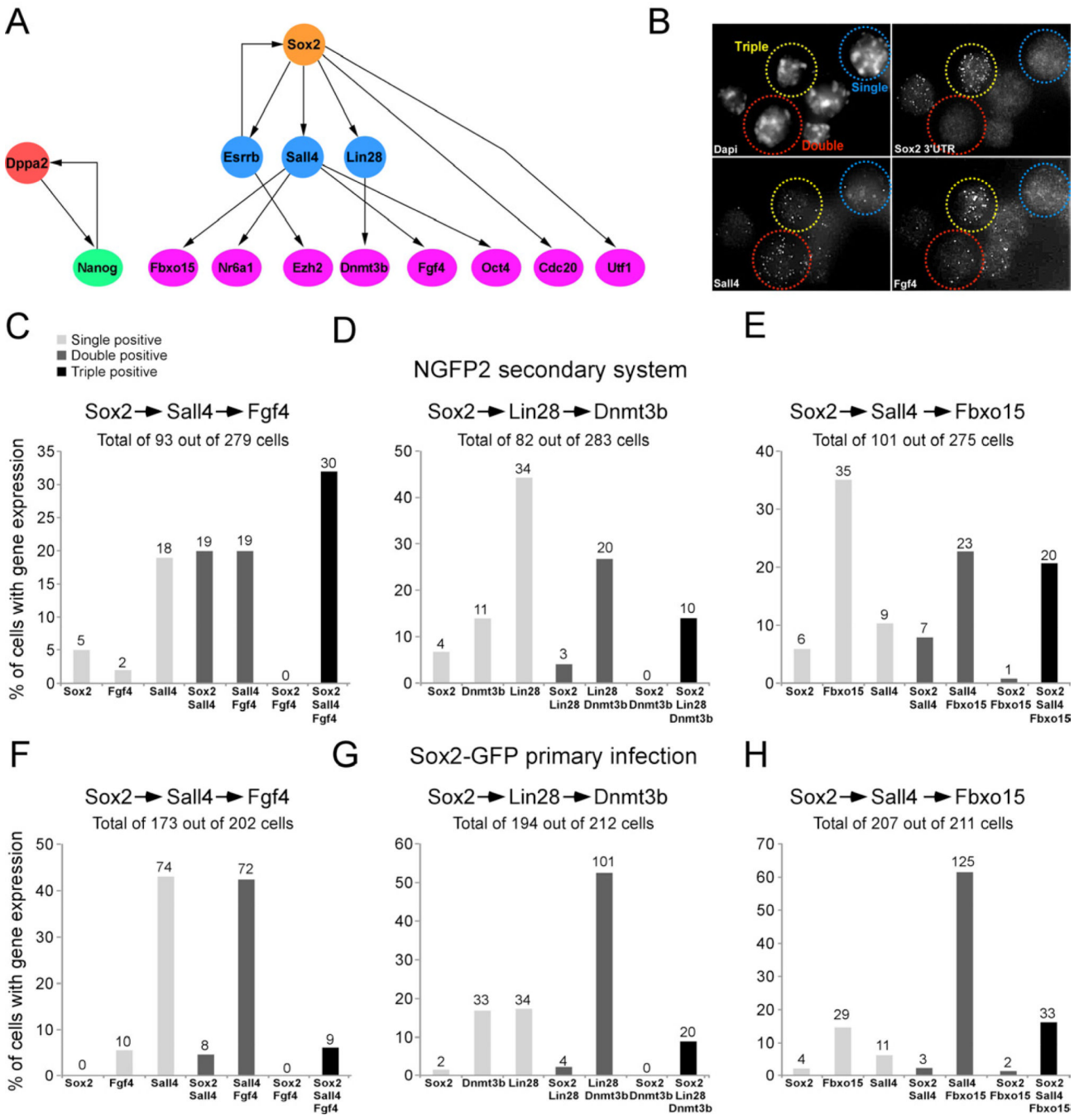


Figure 5. Model to predict the order of transcriptional events in single cells

(A) Bayesian network to describe the hierarchy of transcriptional events among a subset of pluripotent genes. (B) sm-mRNA-FISH representative image of combination in Figure 5C showing a single positive cell (blue, Sall4), double positive cell (red, Sall4/Fgf4), and triple positive cell (yellow, Sox2/Sall4/Fgf4). (C–E) Bar plot of the percent of cells with transcripts, quantified by single molecule mRNA FISH, of single positive (light grey), double positive (dark grey), and triple positive (black) expression in single NGFP2 cells at day 12 on dox and in (F-H) single primary infected Sox2-GFP cells at day 12 on dox. The numbers of cells in each category is indicated on top of each bar. See also Figure S6 and Table S5.

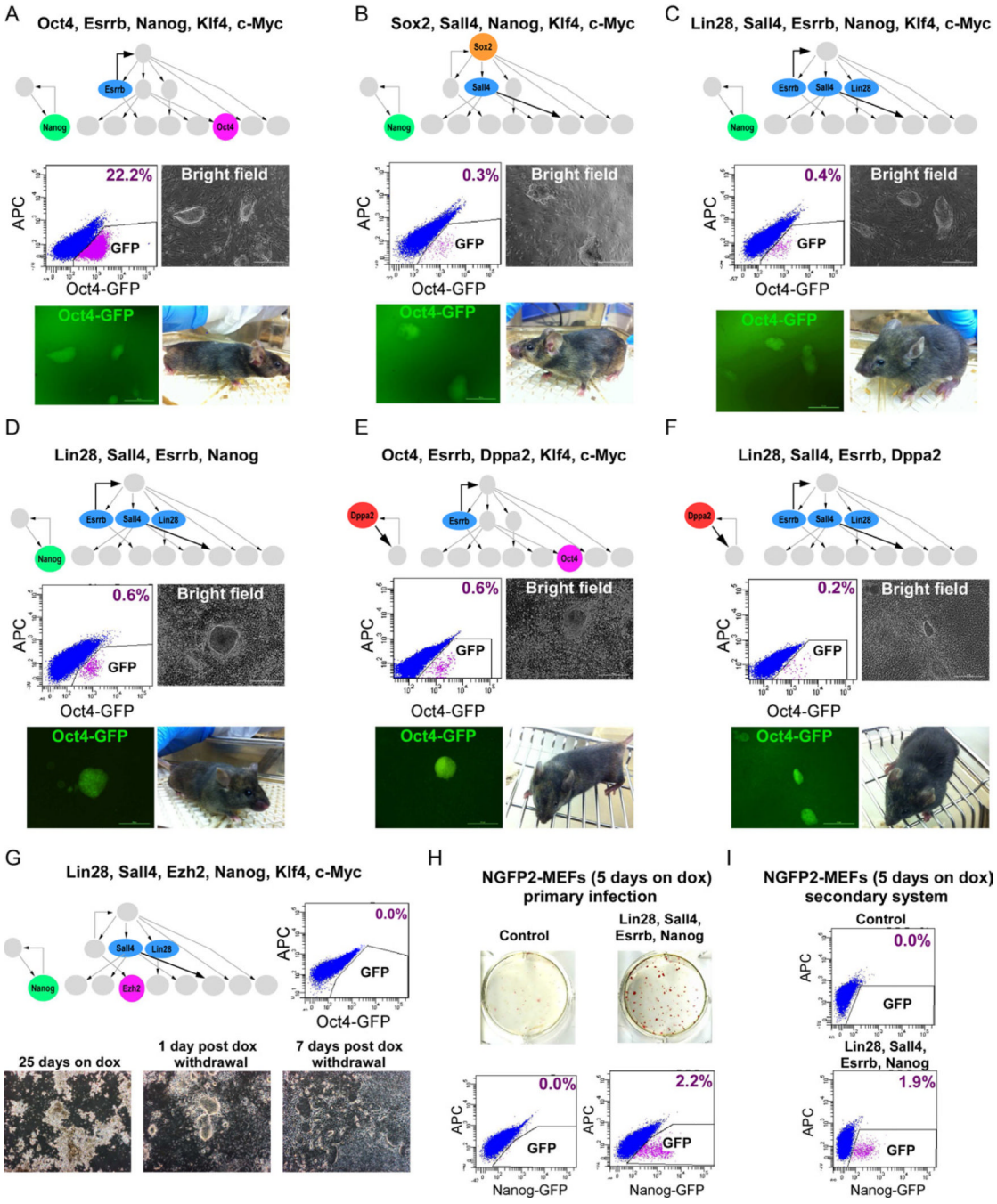


Figure 6. Cellular reprogramming with factors derived from Bayesian network

Flow cytometric analysis of GFP in Oct4-GFP cells reprogrammed with (A) Oct4, Esrrb, Nanog, Klf4, and c-Myc, (B) Sox2, Sall4, Nanog, Klf4, and c-Myc (C) Lin28, Sall4, Esrrb, Nanog, Klf4, and c-Myc (D) Lin28, Sall4, Esrrb, and Nanog, 25 days on dox, 5 days without dox. (E) Oct4, Esrrb, Dppa2, Klf4, and c-Myc (F) Lin28, Sall4, Esrrb, Dppa2, 16 days on dox, 5 days without dox. Representative images of stable dox-independent GFP+ colonies and bright-field pictures of chimeras derived from the iPSCs are shown. (G) Flow cytometric analysis of GFP in Oct4-GFP cells reprogrammed with Lin28, Sall4, Ezh2, Nanog, Klf4 and c-Myc, 7 days post dox withdrawl (upper right). Representative bright-field pictures of the cells 25 days on dox, 1 day post dox withdrawal, and 7 days post dox

withdrawal are shown (bottom). (H) AP immunostaining and flow cytometric analysis of GFP in control NGFP2 MEFs (upper left) and NGFP2 MEFs reprogrammed with Lin28, Sall4, Esrrb, and Nanog by primary infection (upper right), 5 days on dox, 3 days without dox. Flow cytometric analysis of GFP is shown (bottom). (I) Flow cytometric analysis of GFP in control NGFP2 MEFs (upper) and secondary NGFP2- Lin28, Sall4, Esrrb, and Nanog MEFs (bottom), 5 days on dox, 3 days without dox. See also Figure S7.

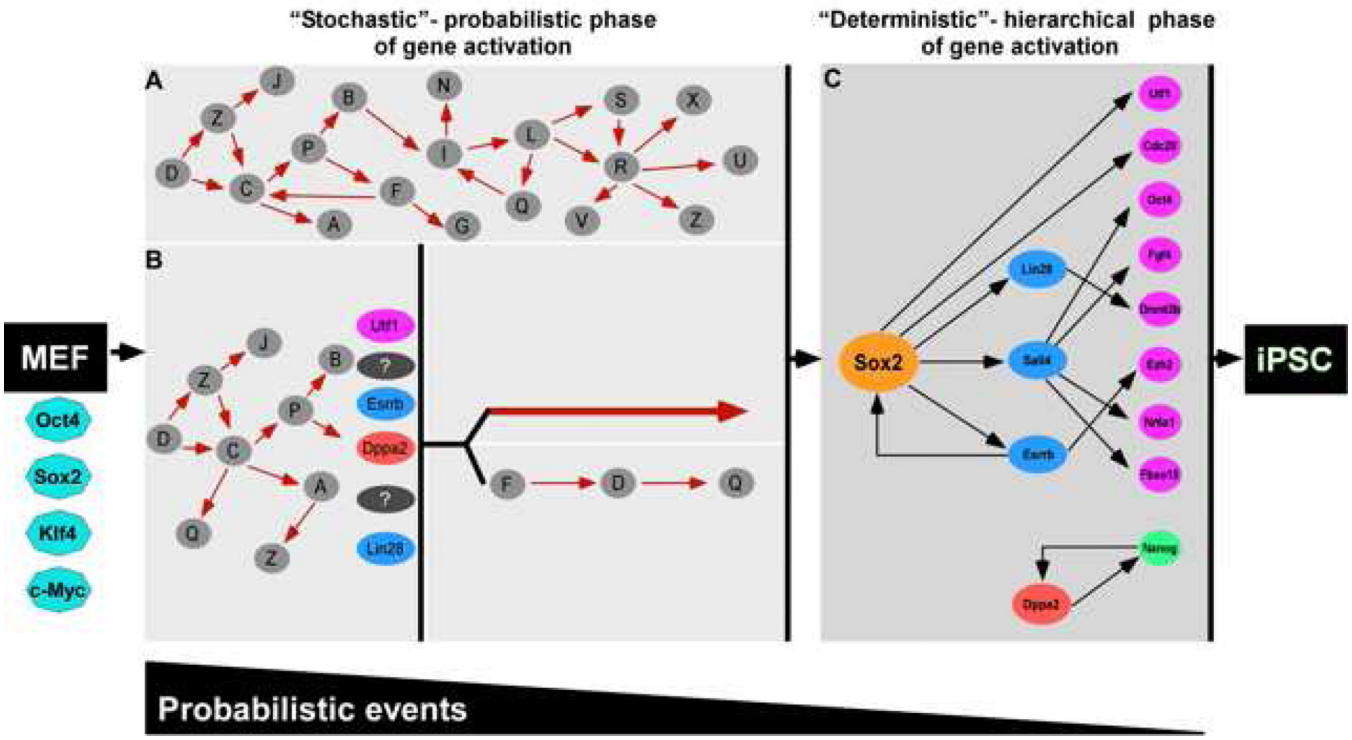


Figure 7. Two phases in reprogramming

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, 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.