Single-cell analyses of cellular reprogramming and embryonic stem cells

by

Dina Adel Faddah

B.S. Biology
The University of North Carolina at Chapel Hill, 2006

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2014

Signature of author:

Dina Adel Faddah
Department of Biology
May 23, 2014

Certified by:

Rudolf Jaenisch
Professor of Biology
Founding Member, Whitehead Institute
Thesis Supervisor

Accepted by:

Michael T. Hemann
Associate Professor of Biology
Co-Chair, Biology Graduate Committee

Single-cell analyses of cellular reprogramming and embryonic stem cells

by

Dina Adel Faddah

Submitted to the Department of Biology on May 23, 2014
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

Three years before the start of this thesis, Yamanaka and Takahashi published a groundbreaking paper entitled "Induced of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." A mere two scientists reprogrammed somatic cells to an embryonic stem-cell like state (termed induced pluripotent stem cells, iPSCs) by simply overexpressing four transcription factors: Oct4, Sox2, c-Myc, and Klf4. During cellular reprogramming, only a small fraction of cells become iPSCs. Previous analyses of gene expression during reprogramming were based on populations of cells, impeding single-cell level identification of reprogramming events. Using single-cell analysis, we found Esrrb, Utf1, Lin28 and Dppa2 to be predictive markers of reprogramming. We found that single cells exhibit high variation in gene expression early in reprogramming and this heterogeneity decreases are the cell reaches pluripotency. Our results show that a stochastic phase of gene activation is followed by a late hierarchical phase, initiated by activation of the Sox2 locus, leading to the activation of the pluripotency circuitry. Finally, we reprogram cells without Oct4, Klf4, Sox2, c-Myc, and Nanog.

Embryonic stem cells (ESCs) are the gold standard comparison for iPSCs. Our investigation of ESCs must continue in parallel to that of iPSCs since we cannot truly understand iPSCs if we do not understand the molecular mechanisms that regulate ESC pluripotency. The homeodomain transcription factor Nanog is a central part of the core pluripotency transcriptional network and plays a critical role in ESC self-renewal. Several reports have suggested that Nanog expression is allelically regulated and that transient downregulation of Nanog in a subset of pluripotent cells predisposes them toward differentiation. Using single-cell gene expression analyses combined with different reporters for the two alleles of Nanog, we show that Nanog is biallelically expressed in ESCs independently of culture condition. We also show that the overall variation in endogenous Nanog expression in ESCs is very similar to that of several other pluripotency markers. Our analysis suggests that reporter-based studies of gene expression in pluripotent cells can be significantly influenced by the gene-targeting strategy and genetic background employed.

Our results show that single-cell analysis is essential for deciphering the mechanisms of reprogramming and understanding gene regulation of ESCs, exposing important rarities typically masked by population-based assays.

Thesis Supervisor: Rudolf Jaenisch
Title: Professor of Biology and Member of the Whitehead Institute

Dedication

I dedicate this thesis to my beautiful mother Shereen, for her unconditional love and support and my wise father Adel, for buying a one-way ticket from Cairo, Egypt to New York, New York on TWA at age 26 to start a life in the United States.

Acknowledgments

This thesis has come to fruition through the support and mentorship of many people throughout my life.

MIT (2008-2014)

I want to first thank Rudolf for his unwavering support throughout my PhD. I first met Rudolf during my MIT interview in 2008. We had an intense conversation about why the common HGPS mutation, C → T, occurs frequently in the human genome. I left Rudolf's office after this discussion shaking from excitement. Rudolf's intensity was electrifying and I was hooked immediately. I knew I wanted that same intensity and excitement throughout my PhD. Although the Jaenisch lab felt like home from the first day of my rotation, I was a bit hesitant, as I knew Rudolf's lab housed many postdocs and was therefore a challenging environment for a student. However, joining the Jaenisch lab was the best decision I made and I'm so grateful that Rudolf took me as a student. I thank Rudolf for providing me with freedom to pursue my ideas and unconditional support. Rudolf always let me pop into his office to chat, anytime of the day, and I'm so grateful for his patience. I thank Rudolf for supporting me in my Nanog project, especially since my data went against the established dogma. I also am so appreciative of Rudolf's patience and support with my non-traditional graduate student trajectory. I admire Rudolf's intensity, his willingness to learn, and the pace with which he walks through lab. Rudolf has an amazing ability to define the most simple, clean, controlled experiment to answer a biological question. My favorite part of graduate school was writing papers with Rudolf. I thank him for being so fast with writing and submitting papers, which made it easy to share my science with the community. Rudolf, I can't thank you enough for your sharing your wisdom with me and supporting me in every step of my PhD. Importantly, you have ingrained in me the importance of taking risks, as we gain diminishing returns by following the safe path in science.

The good thing about the Jaenisch lab is that it attracts motivated, excited people with loads of personality. I have numerous lab members to thank: Dorothea Maetzel for summer coffee breaks and people watching outside the Whitehead and incredible support, Haoyi Wang for agreeing to help me with my project after a late night talk in the small tissue culture room, John Cassidy for good gossip in the mouse tissue culture room, Meelad Dawlaty for always answering my questions and giving great advice and mentorship, Menno Creyghton for telling me to always take my vacations because nothing in lab will change in the two weeks that you are gone, Mike Lodato for always making me smile in lab and dealing with me in the chemical room, Grant Welstead for his Dr. G jokes, Chris Lengner for answering my millions of questions in lab with a smile, unconditional support and encouragement, providing Room 455 with a great spirit and warmth, and making every experiment look easy, Yossi Buganim for being a truly wonderful collaborator and friend, Sovan Sarkar for giving me good India advice, Albert Wu Cheng for helping on the computational analyses of both of my projects, Kibibi Ganz for always performing blastocyst injections, Bryce Carey for always supporting me in lab (Bryce is one of the reasons I joined the Jaenisch lab, I wanted to be as excited about science as he was), Jacob Hanna for being bold, daring, and never

wavering from those things and people that are important to him, Frank Soldner for being the ultimate "Devil's Advocate"(I only benefitted and grew as a scientist from our long conversations), and Gerry Kemske for her amazing support of the lab.

MIT is such a close community, we are all nestled in Kendall Square and that makes interacting with faculty incredibly easy. I thank Hazel Sive and Nancy Hopkins for long, honest discussion about science, careers, and women in science, Tania Baker for 1st year support, Iain Cheeseman for being a great role model for a young PI on the 4th floor, Bob Weinberg for regular Arabic dialogue in the Whitehead, Peter Reddien, David Page, Dave Bartel, Chris Burge, and David Houseman for fruitful discussions about science and navigating decisions about my career, and Eric Lander and Angelica Amon for supporting me in rotations. Special thanks goes to Alexander van Oudenaarden and his lab, especially Sandy Klemm and Stefan Semrau, for allowing me access and support to microscopes for all FISH experiments.

Thank you to my committee members Phil Sharp, Rick Young, and Konrad Hochedlinger for asking probing questions during my committee meeting, prelim, and thesis defense. Our discussions always made me think twice about what I know and what I don't know.

I loved being a graduate student in the Whitehead, I think it's the best place in the world and I feel so lucky to have spent 5 years of my life in such an amazing and inspirational place. Special thanks goes to David Baltimore and Jack Whitehead for the establishment of the Whitehead. I love thinking about all of the scientists from previous generations who have spent longs nights and weekends in the Whitehead.

Thank you to MIT for providing me with unparalleled opportunities to grow, learn, and develop as a scientist and as a person.

UNC-CH (2003-2006) and NIH (2006-2008)
A short required meeting with my undergraduate research advisor, Gustavo Maroni, is how I "formally" began as a scientist. Dr. Maroni encouraged me to do undergraduate research as a freshman. I assumed I was too young; however, he dismissed this idea and scribbled the name "Jason Lieb" on a yellow post-it note and told me, "This professor is young and new, he must have space, I'm sure he'll take you." These words changed my life forever and I thank Dr. Maroni.

I want to thank my undergraduate research advisors at UNC-CH, Jason Lieb and Todd Vision. I am forever indebted to Todd and Jason for setting my foundation as a scientist. Todd and Jason let me take on my own project as an undergraduate and treated me the same as a graduate student. I didn't realize how special this was at the time; however, looking back, I realize this was an incredible privilege. Todd and Jason had just started their own labs at the time I joined. To this day, I am still in awe with how much I learned from two young PIs. They taught me nearly everything that I have needed to complete this PhD. I thank them for being patient in my training, pushing me

very hard and having high expectations, and spending time with me on all of the details of my experiments, papers, presentations, and proposals. I would not have made it to MIT or NIH without my time in Todd's lab and Jason's lab. I still seek advice from Todd and Jason and will continue to do so as a move forward in my career. Someone once told me that your scientific style is a product of your mentors. This could not hold more truth, as I can see Todd and Jason's style in all that I do, from writing papers, making posters, putting together a talk, or designing an experiment.

I am so lucky (literally) to have worked with my NIH research advisor, Francis Collins. During my junior year at UNC-CH, Francis responded promptly to a short email in which I simply expressed my love for genetics and interest in Hutchinson-Gilford progeria syndrome. I was in a state of transition at NIH, doing research while applying to graduate school and medical school. I thank Francis for *always* making time for regular meetings and doing anything in his power to help me learn and grow, whether it was shadowing in the Clinical Center, discussing ideas for projects, editing my graduate and medical school essays, or providing guidance on a poster. Francis is an incredible leader and I thank him for always helping me move forward in my career, providing his feedback when I am in the midst of making an important decision, cheering on all of my successes, and encouraging me to pursue my dreams. I hope to emulate Francis' care for people when I have my own lab.

I have a number of people from the Lieb, Vision, and Collins lab to thank. From the Lieb lab, I thank Ostranda Williams for teaching me how to pipette, Greg Hogan for setting high goals for himself which in turn taught me to set high goals for myself, Sean Hanlon for support and never making me feel bad for asking questions, and Michael Buck for always forcing me to think about the significance of an experiment. From the Vision lab, I thank Eric Ganko for helping me write my first paper. I would have never made my first Faddah et al. without his guidance and patience. From the Collins lab, I thank Renee Varga for generously taking me on a summer student while she was pregnant and Kan Cao for always reminding me that science should be fun.

My family

Finally, I thank my beautiful family for their support, love, and encouragement. Although no person in my family is a scientist, they "get" being a scientist. They understand when I can't talk on the phone because I'm in tissue culture, they understand the lows of a negative result or failed experiment, they understand the high of having a paper accepted, they understand the anxiety of waiting on a paper or grant review, they understand that going into lab at night or on a weekend isn't a bad thing (as I am excited about an experiment). They understand the life of a scientist, the life that I love.

Dad, thank you for always going out of your way to understand the science that I do, for reading NYT science articles and dropping key words like "pluripotency" and "lincRNAs" in our conversations. I will continue to seek knowledge and question the world around me as you continue to do so. Thank you for being rational when I was irrational and always supporting my decisions with love. I hope I can be as wise and kind as you are in everything that you do.

Mom, thank you for understanding my life as a scientist better than anyone and supporting me to pursue my dreams and a career that makes me happy. Whenever I uttered "I can't" you always told me back "You can." My personality of a scientist is a product of you. Thank you for your unconditional support and the sacrifices you made in your own career to give me the best life possible filled with opportunity. Graduate school wasn't always easy and one phone call with you could make all my problems melt away.

Sister, thank you for being my best friend. Many girls steer away from science at a young age because it's "not cool" or "geeky." We both know that I was definitely a geek as a kid; however, I think having you as a sister and my best friend allowed me to push forward in science and math with confidence. Your presence made me feel secure and therefore it was easy for me to pursue the "uncool" subjects at full force, with no qualms.

Tarik, thank you for always being excited about my work and always asking about my experiments and the backstories of science.

Ameera, thank you for blowing me kisses and learning to say Didi while we Skyped in lab before your 7:30pm bedtime.

Kyle, I cannot wait to start our life together in SF. I would never have imagined how much Kendall Square would truly change my life forever.

I end this journey in science with more questions than answers. I have truly loved MIT, the Whitehead, and the Jaenisch laboratory, and feel fortunate that my enthusiasm for science and research continues to grow and evolve.

Table of Contents

Chapter 1. Introduction

Embryonic stem cells (ESCs), which are derived from the inner cell mass (ICM) of the embryo, are characterized by the ability to self-renew and differentiate into all cell types except those of the extraembryonic lineages (Evans and Kaufman, 1981; Martin, 1981). The developmental potency of ESCs holds great promise for regenerative medicine and the reprogramming of somatic cells to pluripotency could allow for patient-specific stem cells and transplantation therapy (Jaenisch and Young, 2008). Generation of stem cells by somatic cell nuclear transfer and cell fusion has been studied for many years; however, their widespread use has been restricted by limited technical expertise and ethical concerns regarding human oocytes (Cowan et al., 2005; Wakayama et al., 1998; Wilmut et al., 1997; Yamanaka and Blau, 2010).

In 2006, Takahashi and Yamanaka succeeded in reprogramming mouse fibroblasts to induced pluripotent stem cells (iPSCs) by overexpression of four transcription factors: Oct4, Sox2, Klf4 and c-Myc (OSKM) (Takahashi and Yamanaka, 2006). This new method of creating ES-like cells is particularly attractive because it bypasses the use of human oocytes and enables the study of pluripotency and differentiation from readily available somatic cells, like blood (Staerk et al., 2010).

At the start of this thesis, analyses of cellular changes during the reprogramming process have relied on populations of mouse embryonic fibroblasts (MEFs). Microarray data at defined time points during the reprogramming process show that the immediate response to OSKM is characterized by de-differentiation of MEFs and upregulation of proliferative genes, consistent with the expression of c-Myc (Mikkelsen et al., 2008). An upregulation of some lineage specific genes was also observed, probably reflecting responses to Sox2 and Klf4, which function in neural, epidermal, and kidney differentiation (Rowland and Peeper, 2006; Takahashi et al., 2007). Pluripotency markers such as stage-specific embryonic antigen 1 (SSEA1) and alkaline phosphatase are upregulated during reprogramming; however, these markers are not stringent pluripotency markers and only a small fraction of such marker-positive cells will develop later into genuine iPSCs (Brambrink et al., 2008; Stadtfeld et al., 2008). Using knock-in

GFP reporters, reactivation of endogenous Nanog, Oct4, and Sox2 occurs late (day 18-day 25) in the reprogramming process (Brambrink et al., 2008; Stadtfeld et al., 2008).

Transgenic approaches have been developed to circumvent the heterogeneity of virally infected fibroblasts, which were originally used for reprogramming (Wernig et al., 2007). Cells reprogrammed using doxycycline (dox)-inducible lentiviral vectors can be used to make chimeric mice, and cells taken from these chimeras can reprogram upon addition of dox and no further viral transduction (Hanna et al., 2008; Wernig et al., 2008a). The Jaenisch and Hochedlinger labs made transgenic mouse models in which reprogramming factors are expressed from a single genomic locus using a drug-inducible, polycistronic transgene (Carey et al., 2010; Stadtfeld et al., 2010b). Multiple somatic cell types can be directly reprogrammed to generate iPSCs by culture in ESC media and dox.

A report by the Jaenisch lab showed that reprogramming in a monoclonal population of B cells is a stochastic process where almost all mouse donor cells eventually give rise to iPSCs given continued growth and transgene expression (Hanna et al., 2009). B cells, in comparison to MEFs, have a high single-cell cloning efficiency and represent a well-defined lineage-committed cell population. The rearrangement of the immunoglobulin heavy chain locus allows for the unambiguous retrospective identification of the donor cell from which a given iPSC arose.

It is noteworthy that the analyses of cellular changes during reprogramming have relied not on single cells, but rather on populations of cells, only a small and variable fraction of which will eventually become iPSCs with different kinetics. To fully understand the changes that precede iPSC formation, new experimental approaches must be established that allow for molecular analyses on the single cell level. Single-cell analysis can target specific populations and therefore elucidate unknown genes and signaling pathways involved in reprogramming. Many questions still remain unresolved in reprogramming: Does a cell become reprogrammed in a single event or is it a process that evolves over time? What specific steps can be delineated during the process? What factors influence the transitions between these steps?

The primary motivation of this thesis was two-fold: (1) to study gene expression in single cells during the reprogramming in hopes of identifying novel molecular markers that would predict whether a given cell early in the reprograming process would even generate a daughter iPSC and (2) to further understand the nature of the stochastic events that enable reprogramming and determine if any ordered steps occur during the process. Understanding gene expression in single cells is essential to deciphering the molecular events that take place during the reprogramming process.

Part 1. Embryonic stem cells

Pluripotent cells

Embryonic carcinoma (EC) cells, isolated from mouse germ cell tumors, also known as teratocarcinomas, were the first pluripotent cells to be grown *in vitro* (Rossant, 2001). In 1981, two groups welcomed the ESC era. Martin Evans and Matthew Kaufman at the University of Cambridge, published a co-culturing technique in which mouse embryonic fibroblasts and blood serum were used to support the culture of cells derived from the ICM of delayed mouse blastocysts (Evans and Kaufman, 1981). Meanwhile, at the University of California San Francisco, ESCs were isolated from the ICM of blastocysts cultured in medium conditioned by an established teratocarcinoma stem cell line (Martin, 1981). Although this thesis focuses on mouse ESCs, it is important to note that Jamie Thomson at the University of Wisconsin Madison developed the first techniques to isolate and grow human ESCs in culture, making human disease modeling by ESCs a reality (Thomson et al., 1998).

ESCs are pluripotent cells derived from the ICM of the blastocyst, also known as a pre-implantation embryo. ESCs can be cultured *in vitro* for months and years without differentiation. The cells in the ICM (which are explanted *in vitro* to be ESCs) eventually differentiate into the epiblast and the hypoblast. ESCs are of great interest for regenerative medicine because it has been proposed that they can regenerated tissues or cell types ravaged by disease, such as diabetes, blood disorders, and Parkinson's and Alzheimer's disease (Boiani and Scholer, 2005).

Signaling pathways regulating embryonic stem cells

Several extrinsic and intrinsic factors are implicated in the nucleus-directed signaling pathways known to regulate stem cell pluripotency *in vivo* and *in vitro*. Extrinsic factors like leukemia inhibitory factor (LIF), bone morphogenetic protein 4 (BMP4) or basic fibroblast growth factor (bFGF) can be added to ESC cultures to trigger signals that carry through intracellular components and regulate the expression of pluripotency factors. Extracellular signal-regulated kinases (ERK) is an intrinsic signaling factor that mediates mitogen-activated protein kinase (MAPK) pathways, implicated in growth and differentiation of different cell types (Ying et al., 2008). ERK is present within the cell and, in it's active form, will induce the differentiation of mouse ESCs. Therefore, pluripotency could be maintained through inhibition of these signaling pathways (Dutta, 2013; Pera and Tam, 2010).

LIF and JAK/STAT3 signaling

In vitro, LIF is key to maintaining the undifferentiated state of mouse ESCs. It's interesting that LIF is only able to maintain ESCs in the presence of serum, suggesting that additional factors are required. Once supplemented by a feeder layer of MEFs, a human recombinant protein, or a cell line, LIF binds to the LIF receptor (LIFR)-gp130 heterodimer receptor on the cell membrane and activates the signal transducer and activator of transcription-3 (STAT3) (Boiani and Scholer, 2005; Smith et al., 1988). Six STATs have been identified. All but one (STAT4), which is expressed only in the testis and myeloid cells, are expressed ubiquitously (Darnell, 1996). It appears that *in vivo*, the LIF signaling network is not required and mouse embryos without LIF can develop to a stage past that of the ESC derivation. These findings suggest that alternative pathways are potentially involved in maintaining pluripotency *in vivo* and *in vitro* (Nichols et al., 2001).

Moving into the culture milieu, in the presence of LIF, STAT3 binds to phosphotyrosine residues on activated LIFR-gp130 heterodimer receptors and undergoes a phosphorylation and a dimerization. Once phosphorylated, STAT3 dimers translocate to the nucleus and act as transcription factors (Niwa et al., 1998). In addition to

STAT3 nuclear localization, the intracellular domains of LIFR-gp130 heterodimer can recruit the nonreceptor tyrosine kinase Janus (JAK) and the antiphosphotyrosine immunoreactive kinase (TIK) and activate other pathways. ESCs treated with LIF also undergo a phosphorylation of the ERK1 and ERK2, in addition to increasing MAPK activity (Boeuf et al., 1997; Burdon et al., 1999). LIF-STAT3 signaling supports the self-renewal of mouse ESCs, but does not prevent the differentiation of human ESCs, supporting that the LIF-STAT3 signaling pathway is not universal (Humphrey et al., 2004).

BMP4

Knowledge on BMP4 is limited, relative to that of LIF. BMP4 is similar to LIF in that it's a central anti-neurogenesis factor in the embryo. The effect of BMP4 on ESCs is similar to that of LIF—ESCs will differentiate into neurons in the absence of BMP4. Moreover, mouse embryos lacking BMP4 develop past the stage that ESCs can be subsequently derived. In the presence of LIF, BMP4 enhances the pluripotency of ESCs by contributing to the LIF signaling pathway via the activation of SMAD4 (similar to mothers against decapentaplegic homologue-4), which activates members of the *Id* (inhibitor of differentiation) gene family. Serum facilitates this interaction. When LIF is not present, BMP4 resists the LIF cascade, interacting with different SMAD transcription factors (SMAD 1, 5, and 8) that inhibit the *Id* genes. All in all, it appears that a fine balance between LIF and BMP4 is responsible for maintaining the pluripotency and self-renewal of mouse ESCs (Fujiwara et al., 2001; Ying et al., 2003).

It was hypothesized by Austin Smith, professor at the Wellcome Trust Centre for Stem Cell Research at the University of Cambridge, that if a state of pluripotency could be isolated by suppression of general differentiation signals, it would be able to sustain the pluripotent and self-renewal network of ESCs. To identify signaling pathways sufficient for ESC pluripotency and self-renewal, investigation proceeded into previously characterized pathways that were known to induce differentiation (Burdon et al., 1999; Kunath et al., 2007). A key finding was that autoinductive signaling by fibroblast growth factor-4 (FGF4) and the MAPK pathway (via ERK1/2) induces differentiation

of ESCs. Inhibition of ERK1/2 has been previously reported to support the maintenance of pluripotency and neither BMP4 nor LIF inhibits the activation of ERK1/2 (Burdon et al., 1999). Blockade of this pathway by genetic manipulation or small molecules can sustain self-renewal of mouse ESCs in the absence of LIF signaling (Chen et al., 2006). In 2008, Smith and colleagues reported on the "ground state" of pluripotency using defined media with small molecule inhibitors of protein kinases ERK1/2 and (Glycogen Synthase Kinase-3 beta) GSK3β (CHIR99021 and PD0325901), which when combined with LIF, supported the pluripotency of mouse ESCs without feeders or serum ("2i/LIF") (Ying et al., 2008). Inhibition of GSK3β increases the biosynthetic capacity of ESCs and GSK3β has been identified as a key inhibitor of various anabolic processes in the cells. 2i/LIF media conditions inhibit the differentiation towards the neuronal lineage and supports pluripotency and self-renewal of ESCs (Li et al., 2008).

WNT proteins

WNT proteins are secreted glycoproteins that have diverse roles in organogenesis and differentiation (Cadigan and Nusse, 1997). The canonical WNT pathway has been implicated in the pluripotency of mouse ESCs. The WNT pathway is activated when the WNT protein binds to the Frizzled receptor on the cell membrane. Once the pathway is activated, GSK3 is inhibited, nuclear accumulation of β-catenin occurs, and target gens are finally expressed. A study using a specific reverse inhibitor of GSK3, 6-bromoindirubin-3'-oxime (BIO), showed that that activation of the WNT pathway maintains the pluripotent phenotype in both mouse and human ESCs, and sustains expression of pluripotency factors, Oct4, Rex1 (zinc-finger protein-42, Zfp42), and Nanog in the absence of additional LIF (Sato et al., 2004). Modulating WNT signaling in ESCs, either by inactivating the adenomatous polyposis coli (APC) complex (a tumor suppressor that facilitates signals from the cell surface to the nucleus) or by overexpressing β-catenin, results in suppression of neural differentiation *in vitro* (Haegele et al., 2003).

Core transcriptional circuitry of embryonic stem cells

Oct4

Oct4 (also known as Pou5f1), is a POU (Pit-Oct-Unc) domain transcription factor, expressed in oocytes, fertilized embryos, ICM, epiblast, ESCs, ECCs, and germ cells (Boyer et al., 2006a; Okamoto et al., 1990; Rosner et al., 1990; Scholer et al., 1989a; Scholer et al., 1989b). Loss of Oct4 causes aberrant differentiation of the ICM and ESCs into trophectoderm. Overexpression of Oct4 leads to differentiation intro primitive endoderm and mesoderm. These phenotypes suggest that precise Oct4 levels are critical for pluripotency (Nichols et al., 1998; Niwa et al., 2000). Oct4 can regulate gene expression by interacting with other transcription factors within the nucleus, like Sox2 (Boiani and Scholer, 2005).

Sox2

Sox2, SRY-related HMCG box protein, also plays a critical role in the maintenance of pluripotency and lineage specification (Pevny and Lovell-Badge, 1997). Sox2 is expressed in ooyctes, ICM, epiblast, germ cells, ESCs, multipotent cells of the extraembryonic ectoderm, cells of the neural lineage, brachial arches, and gut endoderm (Boyer et al., 2006a). Sox2 is different from Oct4 in that its expression is not restricted to pluripotent cells and is also found in early neural lineages (Avilion et al., 2003). Sox2 has been connected to the regulation of transcription and chromatin (Pevny and Lovell-Badge, 1997). Sox2 deficient embryos die at day E6.5 due to a failure to maintain the epiblast (Avilion et al., 2003). Sox2 null ESCs result in trophectoderm and primitive endoderm-like cells (Yuan et al., 1995). Sox2 binds promoters together with Oct4 and this cooperative event has been shown to be necessary for gene activation at targets, such as Fgf4 (Yuan et al., 1995).

Nanog

Nanog is a homeodomain protein, expressed in the morula, ICM, epiblast, ESCs, ECCs, and germ cells (Boyer et al., 2006a). Nanog was first described by Wang and colleagues as ENK (early embryo-specific NK, NK represents NK-2, a synonym of the fly gene

ventral nervous system defective), due to its homology with members of the *NK* gene family (Wang et al., 2003). Nanog's function as a transcriptional regulator was derived from the presence of a homeodomain.

Two groups cloned Nanog independently. Mitsui and colleagues used digital, *in silico* differential display of expressed sequence tag (EST) libraries in undifferentiated ESCs against somatic tissues, in addition to undifferentiated against differentiated ESCs to identify pluripotency genes, or genes selectively enriched in pluripotent cells. It was then shown that Nanog-null embryos can develop blastocysts but cannot form functional epiblasts (Mitsui et al., 2003). The second group led by Ian Chambers, utilized LIFR null ESCs. They expanded rare ESCs that could survive in culture, and constructed cDNA libraries. Library inserts were subcloned into episomal-plasmid vectors that were based on the replicative function of the polyoma virus. Vectors were used to transfect ESCs that were expressing polyoma large T antigen, and this finally led to the isolation of surviving, LIF-independent cells that were found to express Nanog (Chambers et al., 2003).

The Nanog loss of function phenotype in embryonic development is embryonic lethal at E5.5, lack of the epiblast, and differentiation of ICM into primitive endoderm. The loss of function phenotype in ESCs is loss of pluripotency and differentiation into primitive endoderm. The gain of function phenotype in ESCs is LIF-Stat-3 independent self-renewal and resistance to retinoic acid induced differentiation (Boyer et al., 2006a). Overexpression of Nanog also allows ESCs to be independent of BMP4 (Ying et al., 2003). Ian Chambers, a Scottish professor at the University of Edinburgh, named Nanog after the mythological Celtic "Land of the Ever-Young," since Nanog can maintain ECs in conditions which they would otherwise differentiate (Chambers et al., 2003; Mitsui et al., 2003).

Our understanding of the regulation of Nanog is ever-evolving. In 2007, Chambers and colleagues reported that Nanog appears to fluctuate, meaning it changes back and forth between being monoallelic and biallelic, within ESCs and this downregulation (monoallelism) predisposes cells to differentiation. ESCs in which Nanog is deleted maintain the ability to self renew and contribute to all three germ layers of

chimeras. These data suggest that Nanog is responsible for the establishment of pluripotency but dispensable for maintaining pluripotency (Chambers et al., 2007). Why would a gene, that is dispensable for pluripotency, be regulated in an elegant and sophisticated manner?

Klf4

Klf4, or Kruppel-like factor 4, is a zinc-finger transcription factor expressed in a diverse set of somatic cell types (skin, stomach, small intestine, colon), in addition to ESCs, that shares homology with the *Drosophila* embryonic pattern regulatory gene Kruppel (Garrett-Sinha et al., 1996; Schuh et al., 1986; Shields et al., 1996). Klf2, Klf4 and Klf5 are expressed in ESCs. Klfs harbor redundant functions in ESCs because differentiation occurs upon simultaneous knockdown of all three (Jiang et al., 2008). Klf5 null embryos show early embryonic lethality and ESCs cannot be derived from the ICM (Ema et al., 2008). Niwa and colleagues showed that Klf4, together with Oct4 and Sox2, activates the Lefty1 gene in ESCs, suggesting that Klfs may help enforce gene expression of Oct4 and Sox2 targets (Nakatake et al., 2006).

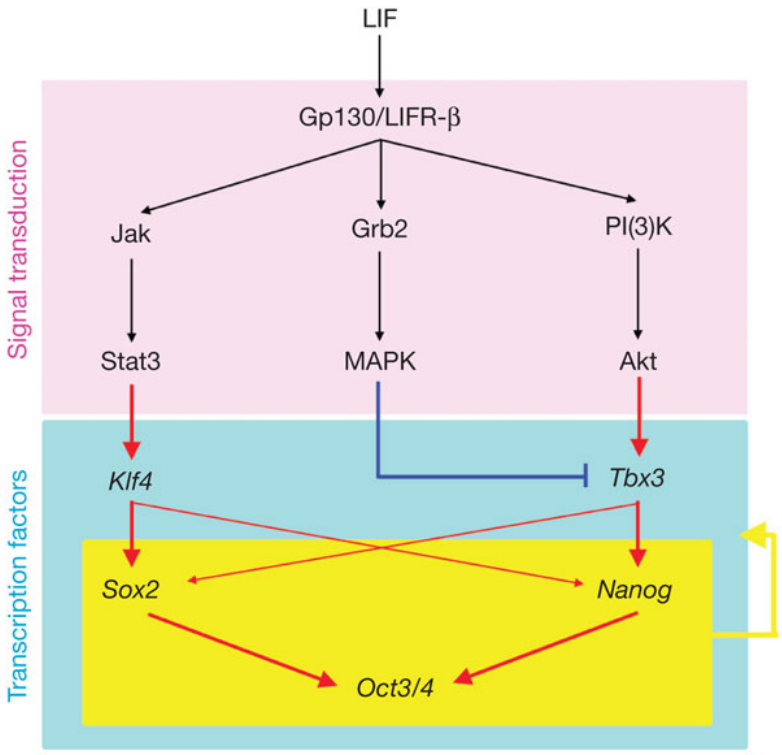


Figure 1. Circuitry of two LIF signaling pathways and pluripotency-associated transcription factors.
Experiments performed by Niwa et al. in 2009 suggested that the Jak-Stat3 pathway activates Klf4 and the PI(3)K-Akt pathway activates Tbx3. The MAPK pathway suppresses nuclear location of Tbx3. Klf4 and Tbx3 primarily activate Sox2 and Nanog, respectively, and maintain expression of Oct3/4. Sox2, Nanog, and Oct3/4 positively regulate transcription of all of these transcription factors. Figure adopted from (Niwa et al., 2009)

Epigenetic and chromatin regulation of pluripotency

Richard Young, Rudolf Jaenisch, Laurie Boyer and colleagues used genome-wide analysis, chromatin immunoprecipitation followed by DNA microarray technology (ChIP-chip), to gain insight onto how these transcription factors contribute to pluripotency in human ESCs (Boyer et al., 2005). These experiments yielded three key findings: (1) Oct4, Sox2, and Nanog bind together at their own promoters, forming an autoregulatory loop (2) Oct4, Sox2, and Nanog co-occupy their target genes (3) Oct4, Sox2, and Nanog target two groups of genes, one that is expressed in ESCs and another that is silent in ESCs, but poised for activation during differentiation. This circuitry suggests that the three genes interact to maintain and enhance their own gene expression (Alon, 2007). Autoregulatory loops are not limited to ESCs, they appear to be a general feature of master regulators of cell states (Odom et al., 2006).

Most of the silent developmental regulators occupied by Oct4, Sox2, and Nanog are also occupied by the Polycomb-group (PcG) proteins (Bernstein et al., 2006; Boyer et al., 2006b; Lee et al., 2006). PcGs are epigenetics regulators that facilitate maintenance of cell state by means of gene silencing. PcGs form multiple Polycomb Repressive Complexes (PRCs), which are conserved from flies to humans. PRC2 catalyzes H3K27 methylation which silences genes (Schuettengruber et al., 2007). Bivalent domains are a feature of silent developmental regulators, occupied by nucleosomes that are marked with both H3K4me3 (activation) and with H3K27me3 (repression) (Bernstein et al., 2006).

Transcription factors can bind enhancers that then coordinate histone and chromatin modifiers to regulate gene expression regarding cell state (Buecker and Wysocka, 2012). “Pioneer” factors are specific types of transcription factors that can reposition nucleosomes (Zaret and Carroll, 2011). H3K4me1 marks all enhancers but active enhancers have H3K27ac, as well. Cells have distinct enhancer patterns and these profiles changes during differentiation. (Creyghton et al., 2010; Rada-Iglesias et al., 2011).

In addition to transcription factors, RNA regulates ESCs (Kanellopoulou et al., 2005; Murchison et al., 2005; Wang et al., 2007). It has been shown by loss of function

experiments in mice that microRNA (miRNAs) play a role in ESC regulation. Dicer-deficient mice fail to develop and ESCs deficient in miRNA processing show defects in differentiation, self-renewal, and viability (Bernstein et al., 2003). Specific miRNAs have been shown to play a role in differentiation, cell patterning, and morphogenesis (Chen et al., 2004; Harfe et al., 2005; Krichevsky et al., 2006; Mansfield et al., 2004). Together, these regulators interact to maintain the pluripotent state.

Part 2. Nuclear reprogramming

During development, the genome undergoes epigenetic alterations to create cell identity and differentiation. For a long time, it was assumed that development was unidirectional and a differentiated cell could not change back to a stem-like cell (Morgan et al., 2005). Due to the pioneering work of John Gurdon, we now know that a differentiated cell nucleus retains the potential to direct the development of an entire organism. The nucleus of a differentiated cell can be erased because epigenetic changes that occur during development are *reversible* (Gurdon, 1962). Three primary strategies have been used to induce the reprogramming of somatic cells to pluripotency: (1) *Reprogramming by nuclear transfer*. This method, also known as "somatic cell nuclear transfer/SCNT", involves the transfer of the nucleus of a somatic cell into an enucleated oocyte, which, when transferred into a pseudopregnant mother, can give rise to a clone (also known as "reproductive cloning") or, once explanted in culture, can produce genetically matched ESCs. (2) *Reprogramming by cell fusion*. This technique involves the fusion of a somatic cell with an ESC that result in a 4n fused cell hybrid that displays all features of a pluripotent ESC. (3) *Reprogramming by defined transcription factors*. Overexpression of transcription factors by infection with viruses can initiate cellular reprograming to a pluripotent state (Jaenisch and Young, 2008).

Reprogramming by nuclear transfer

In 1952 Briggs and King published their article, "Transplantation of Living Nuclei from Blastula Cells into Enucleated Frogs' Eggs," that examined whether nuclei of embryonic cells are differentiated and were the first to conduct a successful nuclear transplantation

experiment with amphibian embryos. They transplanted nuclei from embryonic blastula cells (cells still at an early stage of development) (Briggs and King, 1952). In 1962, Gurdon produced living tadpoles from the adult cells of a frog. His work was a textbook buster, received with skepticism (as any truly groundbreaking science is…) since it contradicted the dogma that adult cells cannot assume new functions. Specifically, Gurdon took the nucleus from a mature intestinal cell and injected the nucleus into a frog's egg whose own nucleus has been removed. Components in the egg were able to reprogram the nucleus; they were able to revert the epigenetic state of the nucleus to a genome that is able to switch from the program of an intestinal cell to that of a developing embryo (Gurdon, 1962). Dolly the sheep (RIP July 5, 1996—February 14, 2003) was the first mammal to be cloned from an adult somatic cell using the process of nuclear transfer (Wilmut et al., 1997). Cloned mice have been generated from mature lymphocytes that carried differentiation-associated immune-receptor rearrangements and from genetically labeled post-mitotic olfactory neurons, demonstrating that the nucleus of a terminally differentiated cell maintains the potential to support development (Eggan et al., 2004; Hochedlinger and Jaenisch, 2002). Nuclear cloning is an incredibly inefficient process in which most clones die soon after implantation or clones are born with serious abnormalities, like obesity and premature death, at all stages of development, (Hochedlinger and Jaenisch, 2003; Ogonuki et al., 2002; Tamashiro et al., 2002; Yang et al., 2007). Although nuclear transfer provides a functional test for reprogramming to totipotency and allows reversibly epigenetic changes to be distinguished from irreversibly epigenetic changes, it is highly controversial due to the ethical concerns of using human oocytes. There is also not an unlimited supply of human oocytes to be used for nuclear transfer experiments.

Reprogramming by cell fusion

Reprogramming of a somatic nucleus to pluripotency has also been shown in hybrids produced by cell fusion of somatic cells and ESCs (Blau and Blakely, 1999). For most hybrids produced by cell fusion, the phenotype of the less-differentiated fusion partner is dominant over the phenotype of the more-differentiated fusion partner (Miller and

Ruddle, 1976). Hybrids between somatic cells and ESCs, embryonic germ cells, or ECCs share features with the parental embryonic cells (Solter, 2006; Tada et al., 2003; Tada et al., 1997; Tada et al., 2001; Zwaka and Thomson, 2005). There is no clear evidence supporting that the somatic nucleus has been fully reprogrammed and has regained the potential to direct development in the absence of the ESC genome (Hochedlinger and Jaenisch, 2006; Jaenisch and Young, 2008). Human ESCs, like mouse, have the potential to reprogram somatic nuclei after cell fusion (Cowan et al., 2005; Yu et al., 2006). Like nuclear transfer, cell fusion is also inefficient, and this has impeded the study of the molecular mechanism. Although fusion bypasses the use of oocytes, tetraploidy of the fused cells is major limitation in using this approach for cell therapy. Generating diploid cells is risky as selective elimination of some ESC-derived chromosomes may trigger genomic instability that can result in cancer (Hochedlinger and Jaenisch, 2006; Jaenisch and Young, 2008; Matsumura et al., 2007).

Reprogramming by defined transcription factors

In 2006 Shinya Yamanaka and Kazutoshi Takahashi stunned the field with a landmark paper in *Cell* entitled, "Induced of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors"(Nakatake et al., 2006). They reprogrammed MEFs and adult fibroblasts to an ESC-like state with viral transduction of four transcription factors: Oct4, Sox2, c-Myc, and Klf4, followed by selection of Fbx15 activation. Cells that had activated Fbx15 were called "induced pluripotent stem cells (iPSCs)" and were demonstrated to be pluripotent by the ability to form teratomas. Importantly, they were unable to generate live chimeras. Now it is generally believed that these Fbx15 iPSCs were incompletely reprogrammed. The pluripotent state was dependent on continual viral expression and endogenous Oct4 and Nanog were expressed at a lower level than in ESCs, with their promoters were mostly methylated. It was well established that Oct4 and Sox2 were vital to pluripotency; however, the use of Klf4 and c-Myc took researchers by surprise (Chambers and Smith, 2004; Ivanova et al., 2006; Masui et al., 2007; Takahashi and Yamanaka, 2006). Was is true? Could the

simple overexpression of four transcription factors in somatic cells really reprogram the cell to pluripotency?

The beauty of Yamanaka and Takahashi's experiments is that they were incredibly easy to reproduce. Global acceptance of novel scientific findings is expedited when other researchers can repeat the experiments with ease. Labs from all over the world raced to complete the most obvious and straightforward experiments. This was an electric time in the stem cell field, similar to the CRISPR/Cas gene-editing craze that is currently occurring at the time of writing this thesis. The Jaenisch, Hochedlinger, and Yamanaka labs used a more stringent selection for pluripotency, cells that had Nanog or Oct4 markers (Maherali et al., 2007; Okita et al., 2007; Wernig et al., 2007). The resulting cells were fully reprogramming by five main criteria. First, Oct4 and Nanog iPSCs gave rise to chimeras, contributed to the germ line, and generated late-stage embryos by tetraploid complementation (Maherali et al., 2007; Okita et al., 2007; Wernig et al., 2007). Second, the inactive X chromosome was reactivated in iPSCs (Maherali et al., 2007). Third, pluripotency marks like alkaline phosphatase (AP), Oct4, Nanog, and stage-specific embryonic antigen 1 (SSEA1) appeared sequentially during the reprogramming process (Brambrink et al., 2008; Stadtfeld et al., 2008; Wernig et al., 2007). Fourth, the pluripotent state of the Nanog and Oct4 iPSCs depended on the activity of the hypomethylated endogenous Oct4 and Nanog promoters and not on the exogenous factors (exogenous factors were Moloney virus vectors that are silenced in ESCs) (Jahner et al., 1982; Okano et al., 1999). Finally, global gene expression of Oct4 and Nanog-selected iPSCs was indistinguishable from ESCs.

Expression of the reprogramming factors in fibroblasts is hypothesized to initiate a series of stochastic events that eventually leads to reprogramming in a small fraction of iPSCs. This is primary supported by two pieces of evidence. First, clonal analyses demonstrated that activation of pluripotency markers can occur at different times after infection in individual daughter cells of the same infected cell (Meissner et al., 2007). Second, clonal analyses of single B cells overtime supported that every cell can give rise to an iPSCs, albeit with different frequencies (Hanna et al., 2009).

Original iPSCs could not be used for therapeutics because they were isolated useing c-Myc, viral induction, and drug-dependent selection for Fbx15, Nanog and Oct4 activation, which led to iPSC-derived mice that developed cancer (Okita et al., 2007). It was shown eventually that c-Myc is dispensable, and genetically unmodified human and mouse fibroblasts could give rise to iPSCs (Meissner et al., 2007; Nakagawa et al., 2008; Park et al., 2008; Takahashi et al., 2007; Wernig et al., 2008b; Yu et al., 2007). Most excitingly, in 2007, a proof of principle experiment was published that demonstrated that iPSCs generated from the skin of a mouse with sickle-cell anemia were able to restore normal blood function when transplanted into diseased mice (Hanna et al., 2007).

Transgenic approaches have been developed to circumvent the heterogeneity of virally infected fibroblasts, which were originally used for reprogramming (Wernig et al., 2007). Cells reprogrammed using dox-inducible lentiviral vectors can be used to make chimeric mice, and cells taken from these chimeras can reprogram upon addition of dox and no further viral transduction (Hanna et al., 2008; Wernig et al., 2008a).. The Jaenisch and the Hochedlinger labs made transgenic mouse models in which reprogramming factors are expressed from a single genomic locus using a drug-inducible, polycistronic transgene (Carey et al., 2010; Stadtfeld et al., 2010b). Multiple somatic cell types can be directly reprogrammed to generate iPSCs by culture in ESC media and dox. Some pieces of the reprogramming puzzle seemed to be coming together; however, due to the inefficiency of reprogramming, the inherent heterogeneity in the process, in addition to the rudimentary single-cell technologies available, the mechanism still largely remains elusive.

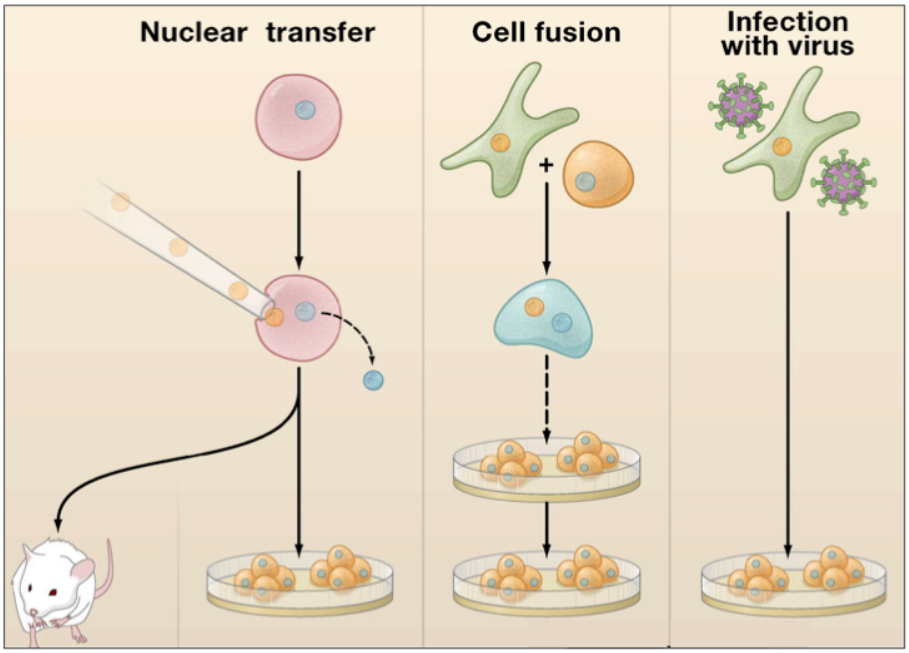


Figure 2. Three strategies to induce reprogramming of somatic cells to pluripotency.
Three primary strategies have been used to induce the reprogramming of somatic cells to pluripotency: (1) *Reprogramming by nuclear transfer*. This method involves the transfer of the nucleus of a somatic cell into an enucleated oocyte, which, when transferred into a pseudopregnant mother, can give rise to a clone (also known as "reproductive cloning") or, once explanted in cultured, can produce genetically matched ESCs (also known as "somatic cell nuclear transfer/SCNT"). (2) *Reprogramming by cell fusion*. This technique involves the fusion of a somatic cell with an ESC that result in a 4n fused cell hybrid that displays all features of a pluripotent ESC. (3) *Reprogramming by defined transcription factors*. Overexpression of transcription factors by infection with viruses can initiate cellular reprograming to a pluripotent state. Figure adopted from (Jaenisch and Young, 2008).

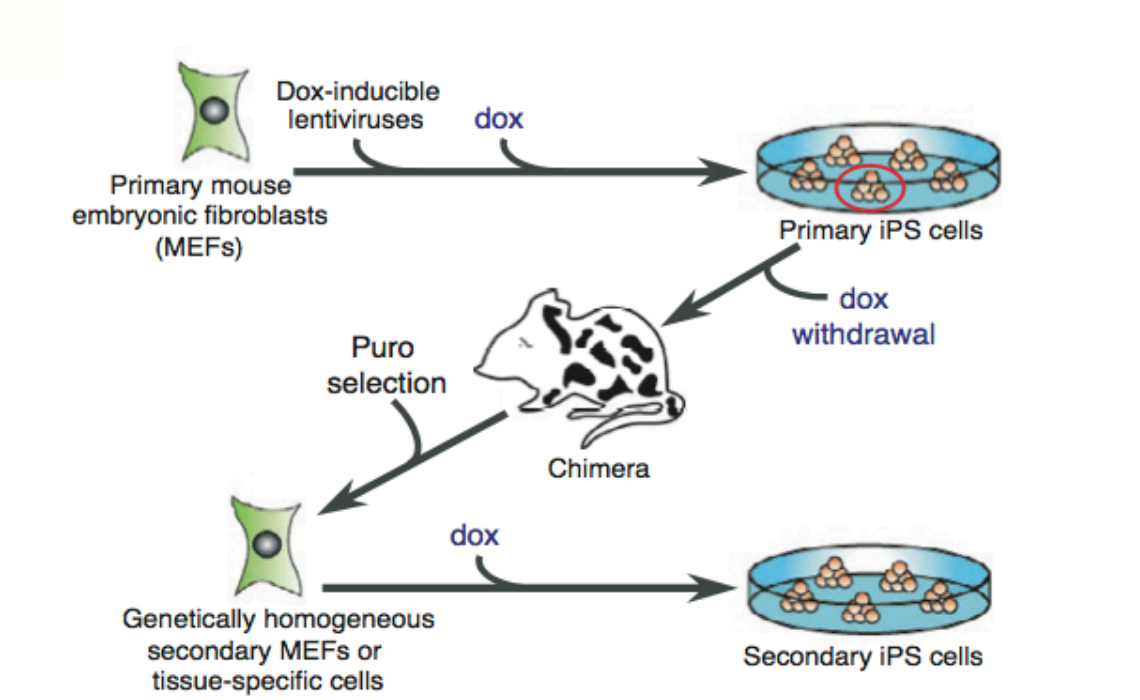


Figure 3. Generation of genetically homogeneous "secondary" cells for reprogramming.
MEFs are infected with dox-inducible lentiviruses encoding the four reprogramming factors followed by induction of reprogramming, "primary" iPSC colony selection, dox withdrawal, chimera formation and selection for iPSC–derived secondary somatic cells. Secondary cells are clonal because they are derived from one iPSC colony. Therefore, heterogeneity observed using single-cell assays is not an artifact of cellular heterogeneity of infected fibroblasts, in terms of transgene copy number and location in a cell. Figure adopted from (Wernig et al., 2008a)

Part 3. Mechanisms of reprogramming by defined factors[1]

Epigenetic changes during reprogramming

The epigenetic signature of a somatic cell must be erased during reprogramming in order to assume a stem cell-like epigenome. These changes include chromatin reorganization, DNA demethylation of promoter regions of pluripotency genes, reactivation of the somatically silenced X chromosome, and genome-wide resetting of histone posttranslational modifications (Fussner et al., 2011; Maherali et al., 2007; Takahashi et al., 2007; Wernig et al., 2007). There are over 100 different histone posttranslational modifications and lysine methylation and acetylation are studied most frequently (Bernstein et al., 2007). The role of histone modifications and chromatin modifiers during reprogramming has been extensively studied (Liang and Zhang, 2013; Schmidt and Plath, 2012; Vierbuchen and Wernig, 2012).

Changes in histone modifications can be seen immediately after factor induction, suggesting that changes in histone marks are an early event that is associated with initiation of the reprogramming process. In contrast, DNA demethylation and X reactivation occur late in the reprogramming process (Koche et al., 2011; Polo et al., 2012). Immediately after OKSM induction, a peak of *de novo* deposition of H3K4me2 is observed at promoter and enhancer regions. H3K4me2 accumulates at the promoters of many pluripotency genes, like *Sall4* and *Fgf4*, which are enriched for Oct4 and Sox2 binding sites and lack H3K4me1 or H3K4me3 marks (Koche et al., 2011). Accumulation of H3K4me2 is also associated with a gradual depletion of H3K27me3 and promoter hypomethylation in regions that are important for reprogramming (Polo et al., 2012). At early time points, however, H3K4me2 does not correlate with the transcription-associated histone mark H3K36me3, occupancy of RNA PolII, or transcriptional activity. These observations suggest that an additional step is required to achieve full activation of these genes and that these loci have not completed chromatin remodeling at early time points (Koche et al., 2011). At the beginning of the reprogramming process, changes in these modifications are almost exclusively restricted

[1] Portions of Part 3 were originally written in a review "Mechanism and models of somatic cell reprogramming" by Yosef Buganim, Dina Faddah, and Rudolf Jaenisch, published in Nature Reviews Genetics PMID: 23681063 and edited for use in this thesis.

to CpG islands, as these regions are more responsive to transcription factor activity and permissive to changes (Ramirez-Carrozzi et al., 2009). At the same time, the promoters of somatic loci begin to lose H3K4me2, consistent with early down-regulation of MEF markers such as *Thy1* and *Postn* (Sridharan et al., 2009; Stadtfeld et al., 2008). A large number of somatic enhancers also lose H3K4me2; hypermethylation and silencing are a result of this change at later stages. Therefore, epigenetic modifications of key MEF-associated genes and early pluripotency genes may represent one of the first steps in the conversion of somatic cells to a pluripotent state.

Role of OSKM factors

Little is known about how ectopic expression of OSKM drives the reprogramming of somatic cells to the pluripotent state. It has been shown that the first transcriptional wave is primarily mediated by c-Myc and occurs in all cells, while the second wave is more restricted to cell amenable to reprogramming and involves a gradual increase of Oct4 and Sox2 targets, leading to the activation of other pluripotency genes that aid in the activation of the pluripotenty circuitry. Klf4 plays a role in both phases. In the first phase it represses somatic genes and in the second phase it facilitates the expression of pluripotency genes (Polo et al., 2012).

Immediately after factor induction, OSKM occupy accessible chromatin, preferentially binding promoters of genes that are active or repressed (Koche et al., 2011; Schmidt and Plath, 2012; Soufi et al., 2012; Sridharan et al., 2009). In addition, OSK become associated with distal elements of numerous genes throughout the genome that display minimal, if any, DNAse hypersensitivity or preexisting histone modifications (Soufi et al., 2012). In turn, the multiple distal genomic sites initially occupied by OSK do not correspond to the distal genomic regions that are bound by these pluripotency factors in ESCs. Based on these observations it has been suggested that OSK may act as "pioneer" factors that open chromatin regions and allow the activation of loci that are essential for establishment and maintenance of the pluripotent state, while c-Myc only facilitates this process (Soufi et al., 2012).

The early promiscuous binding of OSKM, when expressed in fibroblasts, to target sequences present in many genomic regions raises the question of their molecular role in the reprogramming process. Vector transduction-mediated or dox-induced expression of the reprogramming factors in fibroblasts probably does not mimic the expression level or stoichiometry of the endogenous genes in ESCs. It is possible that this flood of OSKM results in widespread and seemingly unrestrained binding of OSKM to multiple regions in the genome, many of which are not occupied by these factors in ESCs. It is possible that OSKM can interact with Mediator/Cohesin complexes, RNA pol II, or elongation factor Ell3 and recruit them to atypical distal enhancers to aid in the opening of these closed regions (Kagey et al., 2010; Lin et al., 2013). Mediator bridges interactions between transcription factors at enhancers and the transcription initiation apparatus at core promoters and, in conjunction with RNA polymerase II and TATA-binding protein (TBP), may gradually initiate transcription from these blocked regions (Kagey et al., 2010). Binding of the "pioneer" factors OSK to super enhancers and the recruitment of the Mediator complex may provide cell type specificity at later stages in the reprogramming process. Transient expression of OKSM is sufficient to open the chromatin and to induce transdifferentiation of fibroblasts to other somatic cells, such as cardiomyocytes and neural progenitor cells, which supports the notion that OSKM are capable of opening chromatin and inducing cell plasticity early in reprogramming (Efe et al., 2011; Kim et al., 2011; Sanyal et al., 2012).

Sometimes OKSM bind jointly to their targets; however, different combinations of the factors regularly occupy non-overlapping genomic regions. For example, Klf4 and c-Myc frequently bind jointly to promoters, while all other OSKM combinations mainly occupy distal elements conserved between human and mouse (Soufi et al., 2012). OSKM bind together at loci that initiate and support the conversion to pluripotency, such as *Glis1*, *mir-302/367 cluster*, *Fbxo15*, *Fgf4*, *Sall4* and *Lin28*, and factors that promote mesenchymal to epithelial transition (MET) (Anokye-Danso et al., 2011; Li et al., 2010; Liao et al., 2011; Maekawa et al., 2011; Soufi et al., 2012; Subramanyam et al., 2011). Half of the enhancers that acquire H3K4me2 in the induced cells are shared enhancers with ESCs and half represent enhancers that are not ESC-specific, supporting the

promiscuous binding of OSKM to various genomic regions that help in the reprogramming process (Koche et al., 2011). Also, in addition to OKSM, activation of other genes early in the reprogramming process may affect the specificity and efficiency of OSKM binding. Binding of the "pioneer" factors OSK in combination with c-Myc to non-ESC specific enhancer regions results in ectopic gene expression. It is possible that this binding may render the initial reprogramming cells susceptible to other gene expression changes, such as activation of genes related to apoptosis, metabolism, MET, and ultimately the silencing of MEF genes and activation of pluripotency genes (Polo et al., 2012).

Factor stoichiometry

The first hint that reprogramming required different expression levels of the individual factors was that the number of proviruses in iPSCs differed widely for individual factors (Wernig et al., 2007). By comparing two genetically highly defined dox-inducible transgenic reprogrammable mouse strains, it has now been shown that factor stoichiometry can influence the epigenetic and biological properties of iPSCs (Carey et al., 2011; Stadtfeld et al., 2010a). Stadtfeld and colleagues showed that ~95% of iPSCs exhibited aberrant methylation of the *Dlk1-Dio3* locus and were unable to generate "all-iPSC" mice by tetraploid complementation, the most stringent test for pluripotency (Stadtfeld et al., 2010a). In contrast, Carey and colleagues used an almost identical reprogrammable transgenic donor mouse strain and found that the majority of iPSCs had retained normal imprinting at the *Dlk1-Dio3* locus and generated "all-iPSC" mice by tetraploid complementation. They showed that the only difference between the two systems was a different stoichiometry of the reprogramming factors, due to a different order within a polycistronic cassette: high quality iPSCs resulted from the donor strain that generated 10 to 20 fold higher levels of Oct4 and Klf4 protein and lower levels of Sox2 and c-Myc than the donor strain that produced only low quality iPSCs (Carey et al., 2011; Stadtfeld et al., 2010a). To further support these data, subsequent studies showed that high levels of Oct4 and low levels of Sox2 are better for iPSC generation (Tiemann et al., 2011; Yamaguchi et al., 2011).

The levels of transgene expression also influence the formation of partially reprogrammed iPSCs. It has been shown that genes expressed in partially reprogrammed colonies are often bound by a higher number of reprogramming factors in the intermediate state than in ESCs (for example, promoter or enhancer regions that are bound by Oct4 and Sox2 solely in ESCs are bound by OSKM in the partially reprogrammed cells) (Sridharan et al., 2009). Alternately, genes that are highly expressed in ESCs are bound by fewer reprogramming factors in the partially reprogrammed cells. Promoter regions bound uniquely by OSKM in partially reprogrammed cells often contain a known DNA binding site for the bound factor. This observation suggests that the excess of factors or factor stoichiometry might influence the targeting of the factors to those regions through direct interactions with their respective DNA binding site. Consistent with this notion is that excess levels of transgenes or factor stoichiometry can cause uncharacteristic binding of OSKM to promoter regions that will result in constant activation of genes that interfere with proper reprogramming. Promiscuous binding of OSKM may be influenced by the stoichiometry of each other and may either facilitate or block reprogramming.

Culture condition and supplements are other parameters known to affect the characteristics of iPSCs (Chen et al., 2011). For example, addition of small molecules and supplements such as valproic acid (VPA), transforming growth factor beta (TGF-β) inhibitors, and vitamin C to the culture medium leads to more efficient derivation of iPSCs (Esteban et al., 2010; Huangfu et al., 2008; Ichida et al., 2009; Maherali and Hochedlinger, 2009). More importantly, iPSCs generated in media without serum and in the presence of vitamin C produced high quality tetraploid complementation-competent iPSCs even when a suboptimal factor stoichiometry was used for inducing pluripotency (Esteban and Pei, 2012; Stadtfeld et al., 2012). In addition, the oxygen level used during isolation of human ESCs was found to affect the state of X chromosome inactivation. While human ESCs isolated under established conditions usually have undergone X inactivation, derivation of the cells under physiological oxygen level led to ESCs with two active X chromosomes, which is similar to mouse ESCs (Lengner et al., 2010). It is

clear that factor stoichiometry and culture conditions affect the efficiency of reprogramming and the quality of iPSCs.

Chromatin modifiers involved in reprogramming

Much information is emerging regarding how chromatin modifiers participate in remodeling the epigenetic program of somatic cells and how they are targeted to genes essential for the reprogramming process. It is reasonable to assume that OSKM binding sites throughout the genome mark regions that eventually undergo epigenetic modification. Consistent with this concept is the finding that Oct4 interacts with the WD-repeat protein-5 (Wdr5), a core member of the mammalian Trithorax (*trxG*) complex, on pluripotency gene promoters and maintains global and localized H3K4me3 distribution (Ang et al., 2011). The H3K27 demethylase enzyme Utx physically interacts with OSK to ablate the repressive mark H3K27me3 from early-activated pluripotency genes such as *Fgf4*, *Sall4*, *Sall1* and *Utf1*. Aberrant H3K27me3 distribution throughout the genome and inhibition of reprogramming is associated with a loss of Utx (Mansour et al., 2012). Tet1 and Tet2, two methylcytosine hydroxylase family members which are important for the early generation of 5-hydroxymethylcytosine (5hmC) during reprogramming, can be recruited by Nanog to enhance the expression of a subset of key reprogramming target genes such as the Nanog locus itself, Esrrb, and Oct4. These data suggest that Tet1 and Tet2 are involved in the demethylation and reactivation of genes and regulatory regions that are important for pluripotency (Costa et al., 2013; Doege et al., 2012; Gao et al., 2013). The poly (ADP-ribose) polymerase-1 (Parp1) has a complementary role in the establishment of early epigenetic marks during reprogramming by regulating 5mC (Doege et al., 2012). Two BAF complex components, Brg1 and Baf155, facilitate reprogramming by establishing a euchromatic chromatin state and facilitating binding of reprogramming factors to important reprogramming gene promoters (Singhal et al., 2010). OSKM-mediated demethylation of pluripotency genes such as *Oct4*, *Nanog* and *Rex1* and enhancement of reprogramming to iPSCs results from overexpression of Brg1 and Baf155.

Like Tet1/2, Utx and the BAF complex, many other chromatin modifiers have been shown to influence the epigenetic remodeling of reprogrammed cells, For example, H3K36me2 demethylases, Kdm2a/2b, act with Oct4 and facilitate the reprogramming process by regulating H3K36me2 levels at the microRNA cluster 302/367, promoters of early-activated genes, and epithelial-associated genes (Liang et al., 2012; Wang et al., 2011). In the conversion of human fibroblasts to human iPSCs, EHMT1 and SETDB1, H3K9 methyltransferases, and polycomb repressive complexes PRC, PRC1, PRC2, are required to reset the epigenome of the somatic cells; depletion of these genes significantly reduces iPSC formation (Onder et al., 2012). SUV39H, another H3K9 methyltransferase, contributes to heterochromatin formation and hinders the reprogramming process (Schotta et al., 2003). This suggests that (a) loss of SUV39H may have a global effect on chromatin organization that leads to aberrant transcriptional regulation or that (b) H3K9 methyltransferases have various specificities, with some targeting somatic-associated loci and others targeting pluripotency-associated loci. Similarly, DOT1L, a H3K79me2 methyltransferase, inhibits the reprogramming process in the early to middle phase. Loss of DOT1L increases reprogramming efficiency by enabling the loss of H3K79me2 from fibroblast-associated genes like the mesenchymal master regulators, *SNAI1*, *SNAI2*, *ZEB1*, and *TGFB2*. Silencing of these genes indirectly increases the expression of the pluripotency genes *NANOG* and *LIN28* and is essential for proper reprogramming (Onder et al., 2012).

Markers of reprogramming

Ectopic expression of OSKM induces a heterogeneous population of cells with each cell embarking on different fates such as apoptosis, senescence, uncontrolled proliferation, and partial or full reprogramming. It is relatively easy to differentiate between non-reprogrammed and reprogrammed cells; however, it is more challenging to distinguish between partially and fully reprogrammed cells. Unfortunately, partially reprogrammed cells can be morphologically identical to ESCs with many pluripotency genes being expressed. Also, no molecular markers have been identified that would predict whether a given cell early in the process will ever generate a daughter iPSC because

reprogramming is a stochastic process (Hanna et al., 2009). Changes such as loss of MEF markers, activation of the MET program or appearance of markers such as SSEA1 or AP are nonspecific, more global, and not restricted to cells destined to become iPSCs (Hansson et al., 2012; Subramanyam et al., 2011).

Global gene expression analyses and proteomic patterns of clonal cell populations or enriched populations at different stages after factor induction have been performed to molecularly characterize the various phases of the reprogramming process (Golipour et al., 2012; Hansson et al., 2012; Mikkelsen et al., 2008; Polo et al., 2012; Samavarchi-Tehrani et al., 2010). These analyses suggested that genes such as *Fbxo15* mark the initiation phase and genes including *Nanog*, *Oct4*, and *Sox2* are activated during the late phase. Importantly, gene expression and proteomic analyses of heterogeneous populations provide limited insight because the rare cells destined to become iPSCs are masked.

Models of reprogramming

Somatic stem cells versus differentiated donor cells

The generation of cloned animals by nuclear transfer was so inefficient, therefore it was hypothesized that clones may not have been derived from differentiated cells as assumed but rather from rare somatic stem cells present in the heterogeneous donor cell population (Pennisi and Williams, 1997). As mentioned in Part 2, this issue was resolved when mature B and T cells were used as donors to create monoclonal mice that carried the Ig and TCR rearrangements of the B and T cell donors, respectively, in all tissues, unambiguously proving the origin from a terminally differentiated donor cell (Hochedlinger and Jaenisch, 2002). Similarly, because reprogramming by defined factors is also inefficient, it was hypothesized that only a fraction of cells are able to generate iPSCs, consistent with an "elite model" where only rare somatic stem cells present in the donor population could generate iPSCs (Wakao et al., 2013; Yamanaka, 2009). Several studies rule out the elite model and support that all cells, including terminally differentiated cells, have the potential to generate iPSC daughter cells. First, iPSCs have been derived from terminally differentiated cells such as liver, spleen, T, and B

cells (Aoi et al., 2008; Hanna et al., 2008; Seki et al., 2010; Stadtfeld et al., 2012). As was performed with nuclear transfer, specific genomic rearrangement of the Ig locus or the T cell receptor in iPSCs proved unambiguously that the cells were indeed derived from mature B or T cells and excluded the possibility of mesenchymal stem cell contamination (Hanna et al., 2008; Hochedlinger and Jaenisch, 2002). Second, clonal analysis of single B cells overtime showed that almost all somatic cells have the potential to generate a daughter iPSC (Hanna et al., 2009).

Stochastic and deterministic models of reprogramming

Reprogramming of somatic cells to pluripotency could occur by two mechanisms: (1) a *stochastic* model in which iPSCs appear with variable latencies, or (2) a *deterministic* model in which reprogrammed cells would be generated with a fixed latency. The stochastic model postulates that it cannot be predicted when or if a given cell would generate an iPSC daughter. Single-cell cloning experiments supported the stochastic model by demonstrating that some sister cells from an early colony generated iPSCs with variable latency and other sister cells never gave rise to iPSCs (Meissner et al., 2007).

It has been suggested that the initial response to ectopic expression of OSKM in somatic cells may be a deterministic response involving epigenetically events that activate loci critical for pluripotency (Polo et al., 2012; Smith et al., 2010). It is possible that the initial promiscuous interaction of OSKM with the genome is initiated by any factor that destabilizes the compacted chromatin typical of somatic cells. It is this destabilization that may render the somatic chromatin susceptible to become hyperdynamic chromatin, which has been shown to be the hallmark of the ESC epigenetic state (Meshorer et al., 2006; Zhu et al., 2013). Consistent with this notion are the findings that general chromatin remodeling complexes (BAF), global basal transcription machinery components, transcription factor IID (TFIID) complex, and exposure of cells to broad DNA methyltransferase and histone deacetylase inhibitors like 5-azacytidine and VPA can substantially enhance reprogramming in combination with OSKM. Also, reprogramming efficiency has reported to increase by means of down-

regulation of Lamin A in fibroblasts, a global chromatin organization modulator, which is not expressed in ESCs (Singhal et al., 2010; Takeuchi and Bruneau, 2009) (Huangfu et al., 2008; Mikkelsen et al., 2008; Pijnappel et al., 2013) (Mattout et al., 2011; Zuo et al., 2012). Thus, although OSKM are highly efficient in inducing pluripotency, any chromatin remodeler or transcription factor, even those that do not normally function in ESCs, might be able to initiate the process leading to pluripotency, albeit with an efficiency too low to be detected by standard reprogramming techniques.

It has been suggested that reprogramming by nuclear transfer or by cell fusion is deterministic because it leads to activation of the somatic *Oct4* within two cell divisions (nuclear transfer) or in the absence of DNA replication (fusion) (Jaenisch and Young, 2008; Yamanaka and Blau, 2010). As mentioned in Part 2, mechanistic insight into the cloned embryo or in the ESC/somatic cell hybrid has been difficult. Therefore, it is still unknown whether nuclear transfer or cell fusion activates the pluripotency circuitry by a deterministic rather than stochastic mechanism. It may be that both, deterministic and stochastic mechanisms, drive the reprogramming of somatic cells by transcription factors as well as by nuclear transfer and cell fusion.

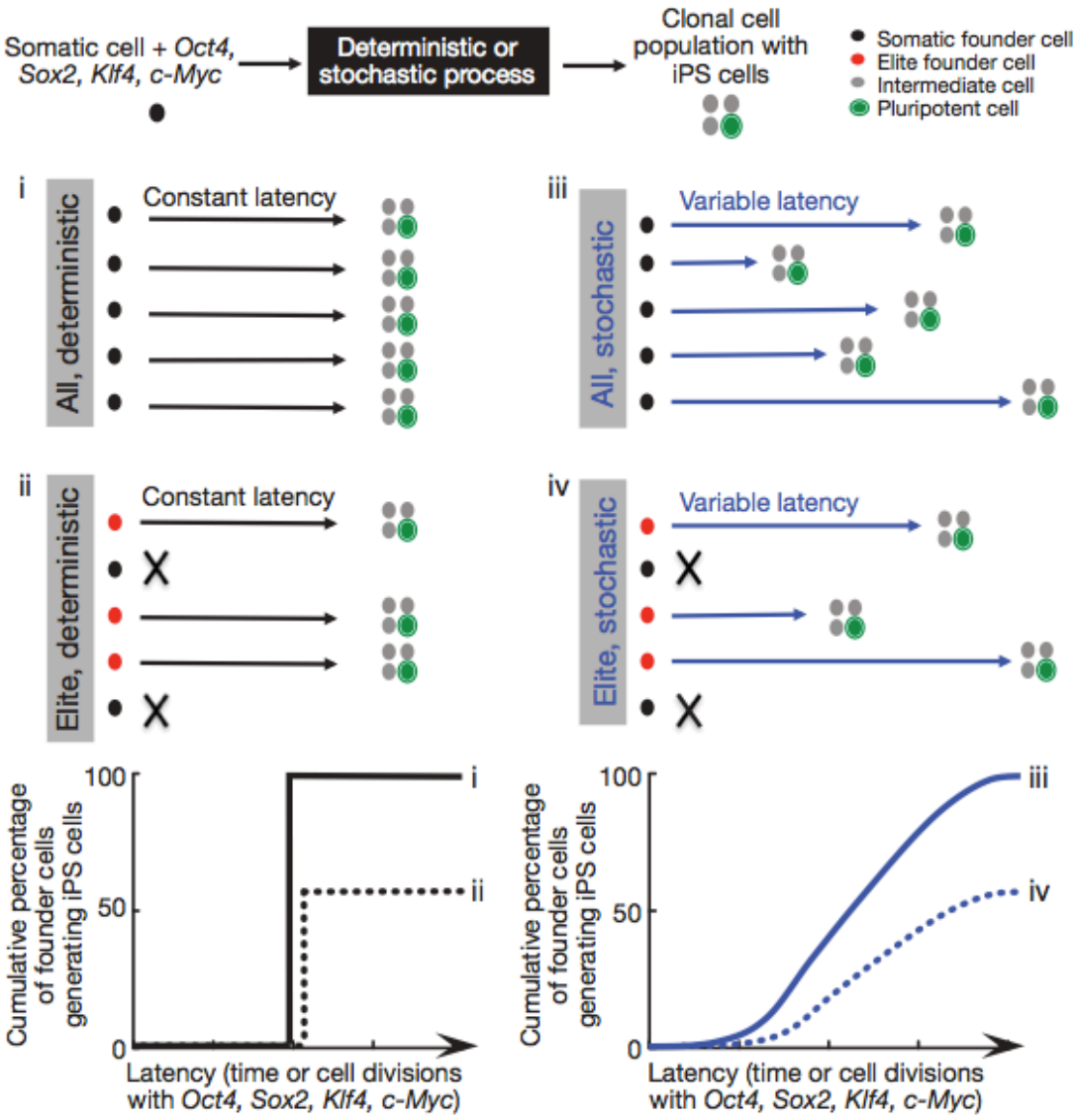


Figure 4. Models of cellular reprogramming to a pluripotent state.
Four diffferent models have been proposed to account for the latency of somatic cells in generating iPSCs following overexpression of OSKM. *Deterministic* models suggest that either all (model i) or a subset of elite, stem-like cells (model ii) within a donor population have the potential to generate iPSCs with a fixed latency. *Stochastic* models sugest that all cells (model iii) or only a subset of elite, stem-like cells (model iv) within a donor population have the potential to generated iPSCs, with variable latencies. Latency is defined as the time or number of cell divisions a donor cell undergoes until it generates a daughter iPSC. Figure adapted from (Hanna et al., 2009).

Mechanisms from population-based studies of reprogramming

After Yamanaka's landmark paper, groups worked at unprecedented speed to study the reprogramming process by analyzing transcriptional and epigenetic changes in cell populations at different time points after factor induction (Takahashi et al., 2007; Takahashi and Yamanaka, 2006). These are the most straightforward experiments to perform in hopes of understanding this complicated process. Populations of MEFs were primarily used to analyze cellular changes during the reprogramming process.

Microarray data at defined time points during the reprogramming process showed that the immediate response to OSKM is characterized by de-differentiation of MEFs and upregulation of proliferation genes, consistent with the expression of c-Myc (Mikkelsen et al., 2008). 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 MET transition. Also, BMP signaling has been shown to act with OSKM to stimulate a miRNA expression signature associated with MET through the initiation phase (Li et al., 2010; Samavarchi-Tehrani et al., 2010).

In an attempt to overcome the problem of cell heterogeneity, reprogramming has been traced at single-cell resolution using time-lapse microscopy (Araki et al., 2010; Smith et al., 2010). Single-cell tracking by real time microscopy has given insights into morphological changes during reprogramming but the approach has not provided information on molecular events driving the process at the single-cell level. These studies showed that the cells underwent a shift in their proliferation rate and reduction in cell size soon after factor induction. These events occurred within the first cell division and with the same kinetics in all cells that give rise to iPSCs.

Single cells and cellular reprogramming?

Prior to the start of this thesis, all molecular analyses of cellular changes during reprogramming had relied on populations of cells. Population-based studies were initially useful for understanding the global changes that occur in cells during the reprogramming process; however, overtime they provided less and less relevant insight. Since only a small fraction of the induced cells become reprogrammed, gene expression profiles of cell populations at different time points after factor induction were not detecting changes in rare cells destined to become iPSCs. When I joined the Janiesch lab in September 2009, it was clear that in order to fully understand the changes that precede iPSC formation, we must study single cells. It is with this general idea that I embarked upon my PhD.

References

Alon, U. (2007). Network motifs: theory and experimental approaches. Nat Rev Genet *8*, 450-461.

Ang, Y.S., Tsai, S.Y., Lee, D.F., Monk, J., Su, J., Ratnakumar, K., Ding, J., Ge, Y., Darr, H., Chang, B.*, et al.* (2011). Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell *145*, 183-197.

Anokye-Danso, F., Trivedi, C.M., Juhr, D., Gupta, M., Cui, Z., Tian, Y., Zhang, Y., Yang, W., Gruber, P.J., Epstein, J.A.*, et al.* (2011). Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. Cell Stem Cell *8*, 376-388.

Aoi, T., Yae, K., Nakagawa, M., Ichisaka, T., Okita, K., Takahashi, K., Chiba, T., and Yamanaka, S. (2008). Generation of pluripotent stem cells from adult mouse liver and stomach cells. Science *321*, 699-702.

Araki, R., Jincho, Y., Hoki, Y., Nakamura, M., Tamura, C., Ando, S., Kasama, Y., and Abe, M. (2010). Conversion of ancestral fibroblasts to induced pluripotent stem cells. Stem Cells *28*, 213-220.

Avilion, A.A., Nicolis, S.K., Pevny, L.H., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. Genes Dev *17*, 126-140.

Bernstein, B.E., Meissner, A., and Lander, E.S. (2007). The mammalian epigenome. Cell *128*, 669-681.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K.*, et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bernstein, E., Kim, S.Y., Carmell, M.A., Murchison, E.P., Alcorn, H., Li, M.Z., Mills, A.A., Elledge, S.J., Anderson, K.V., and Hannon, G.J. (2003). Dicer is essential for mouse development. Nat Genet *35*, 215-217.

Blau, H.M., and Blakely, B.T. (1999). Plasticity of cell fate: insights from heterokaryons. Seminars in cell & developmental biology *10*, 267-272.

Boeuf, H., Hauss, C., Graeve, F.D., Baran, N., and Kedinger, C. (1997). Leukemia inhibitory factor-dependent transcriptional activation in embryonic stem cells. J Cell Biol *138*, 1207-1217.

Boiani, M., and Scholer, H.R. (2005). Regulatory networks in embryo-derived pluripotent stem cells. Nat Rev Mol Cell Biol *6*, 872-884.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., *et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947-956.

Boyer, L.A., Mathur, D., and Jaenisch, R. (2006a). Molecular control of pluripotency. Curr Opin Genet Dev *16*, 455-462.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., *et al.* (2006b). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Brambrink, T., Foreman, R., Welstead, G.G., Lengner, C.J., Wernig, M., Suh, H., and Jaenisch, R. (2008). Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. Cell Stem Cell *2*, 151-159.

Briggs, R., and King, T.J. (1952). Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. Proc Natl Acad Sci U S A *38*, 455-463.

Buecker, C., and Wysocka, J. (2012). Enhancers as information integration hubs in development: lessons from genomics. Trends Genet *28*, 276-284.

Burdon, T., Stracey, C., Chambers, I., Nichols, J., and Smith, A. (1999). Suppression of SHP-2 and ERK signalling promotes self-renewal of mouse embryonic stem cells. Dev Biol *210*, 30-43.

Cadigan, K.M., and Nusse, R. (1997). Wnt signaling: a common theme in animal development. Genes Dev *11*, 3286-3305.

Carey, B.W., Markoulaki, S., Beard, C., Hanna, J., and Jaenisch, R. (2010). Single-gene transgenic mouse strains for reprogramming adult somatic cells. Nat Methods *7*, 56-59.

Carey, B.W., Markoulaki, S., Hanna, J.H., Faddah, D.A., Buganim, Y., Kim, J., Ganz, K., Steine, E.J., Cassady, J.P., Creyghton, M.P., *et al.* (2011). Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. Cell Stem Cell *9*, 588-598.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643-655.

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

Chambers, I., and Smith, A. (2004). Self-renewal of teratocarcinoma and embryonic stem cells. Oncogene *23*, 7150-7160.

Chen, C.Z., Li, L., Lodish, H.F., and Bartel, D.P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. Science *303*, 83-86.

Chen, J., Liu, J., Chen, Y., Yang, J., Chen, J., Liu, H., Zhao, X., Mo, K., Song, H., Guo, L.*, et al.* (2011). Rational optimization of reprogramming culture conditions for the generation of induced pluripotent stem cells with ultra-high efficiency and fast kinetics. Cell Res *21*, 884-894.

Chen, S., Do, J.T., Zhang, Q., Yao, S., Yan, F., Peters, E.C., Scholer, H.R., Schultz, P.G., and Ding, S. (2006). Self-renewal of embryonic stem cells by a small molecule. Proc Natl Acad Sci U S A *103*, 17266-17271.

Costa, Y., Ding, J., Theunissen, T.W., Faiola, F., Hore, T.A., Shliaha, P.V., Fidalgo, M., Saunders, A., Lawrence, M., Dietmann, S.*, et al.* (2013). NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. Nature *495*, 370-374.

Cowan, C.A., Atienza, J., Melton, D.A., and Eggan, K. (2005). Nuclear reprogramming of somatic cells after fusion with human embryonic stem cells. Science *309*, 1369-1373.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A.*, et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931-21936.

Darnell, J.E., Jr. (1996). Reflections on STAT3, STAT5, and STAT6 as fat STATs. Proc Natl Acad Sci U S A *93*, 6221-6224.

Doege, C.A., Inoue, K., Yamashita, T., Rhee, D.B., Travis, S., Fujita, R., Guarnieri, P., Bhagat, G., Vanti, W.B., Shih, A.*, et al.* (2012). Early-stage epigenetic modification during somatic cell reprogramming by Parp1 and Tet2. Nature *488*, 652-655.

Dutta, D. (2013). Signaling pathways dictating pluripotency in embryonic stem cells. Int J Dev Biol *57*, 667-675.

Efe, J.A., Hilcove, S., Kim, J., Zhou, H., Ouyang, K., Wang, G., Chen, J., and Ding, S. (2011). Conversion of mouse fibroblasts into cardiomyocytes using a direct reprogramming strategy. Nat Cell Biol *13*, 215-222.

Eggan, K., Baldwin, K., Tackett, M., Osborne, J., Gogos, J., Chess, A., Axel, R., and Jaenisch, R. (2004). Mice cloned from olfactory sensory neurons. Nature *428*, 44-49.

Ema, M., Mori, D., Niwa, H., Hasegawa, Y., Yamanaka, Y., Hitoshi, S., Mimura, J., Kawabe, Y., Hosoya, T., Morita, M.*, et al.* (2008). Kruppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs. Cell Stem Cell *3*, 555-567.

Esteban, M.A., and Pei, D. (2012). Vitamin C improves the quality of somatic cell reprogramming. Nat Genet *44*, 366-367.

Esteban, M.A., Wang, T., Qin, B., Yang, J., Qin, D., Cai, J., Li, W., Weng, Z., Chen, J., Ni, S.*, et al.* (2010). Vitamin C enhances the generation of mouse and human induced pluripotent stem cells. Cell Stem Cell *6*, 71-79.

Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature *292*, 154-156.

Fujiwara, T., Dunn, N.R., and Hogan, B.L. (2001). Bone morphogenetic protein 4 in the extraembryonic mesoderm is required for allantois development and the localization and survival of primordial germ cells in the mouse. Proc Natl Acad Sci U S A *98*, 13739-13744.

Fussner, E., Djuric, U., Strauss, M., Hotta, A., Perez-Iratxeta, C., Lanner, F., Dilworth, F.J., Ellis, J., and Bazett-Jones, D.P. (2011). Constitutive heterochromatin reorganization during somatic cell reprogramming. EMBO J *30*, 1778-1789.

Gao, Y., Chen, J., Li, K., Wu, T., Huang, B., Liu, W., Kou, X., Zhang, Y., Huang, H., Jiang, Y.*, et al.* (2013). Replacement of Oct4 by Tet1 during iPSC induction reveals an important role of DNA methylation and hydroxymethylation in reprogramming. Cell Stem Cell *12*, 453-469.

Garrett-Sinha, L.A., Eberspaecher, H., Seldin, M.F., and de Crombrugghe, B. (1996). A gene for a novel zinc-finger protein expressed in differentiated epithelial cells and transiently in certain mesenchymal cells. The Journal of biological chemistry *271*, 31384-31390.

Golipour, A., David, L., Liu, Y., Jayakumaran, G., Hirsch, C.L., Trcka, D., and Wrana, J.L. (2012). A late transition in somatic cell reprogramming requires regulators distinct from the pluripotency network. Cell Stem Cell *11*, 769-782.

Gurdon, J.B. (1962). The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. Journal of embryology and experimental morphology *10*, 622-640.

Haegele, L., Ingold, B., Naumann, H., Tabatabai, G., Ledermann, B., and Brandner, S. (2003). Wnt signalling inhibits neural differentiation of embryonic stem cells by controlling bone morphogenetic protein expression. Mol Cell Neurosci *24*, 696-708.

Hanna, J., Markoulaki, S., Schorderet, P., Carey, B.W., Beard, C., Wernig, M., Creyghton, M.P., Steine, E.J., Cassady, J.P., Foreman, R.*, et al.* (2008). Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. Cell *133*, 250-264.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature *462*, 595-601.

Hanna, J., Wernig, M., Markoulaki, S., Sun, C.W., Meissner, A., Cassady, J.P., Beard, C., Brambrink, T., Wu, L.C., Townes, T.M., *et al.* (2007). Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. Science *318*, 1920-1923.

Hansson, J., Rafiee, M.R., Reiland, S., Polo, J.M., Gehring, J., Okawa, S., Huber, W., Hochedlinger, K., and Krijgsveld, J. (2012). Highly coordinated proteome dynamics during reprogramming of somatic cells to pluripotency. Cell reports *2*, 1579-1592.

Harfe, B.D., McManus, M.T., Mansfield, J.H., Hornstein, E., and Tabin, C.J. (2005). The RNaseIII enzyme Dicer is required for morphogenesis but not patterning of the vertebrate limb. Proc Natl Acad Sci U S A *102*, 10898-10903.

Hochedlinger, K., and Jaenisch, R. (2002). Monoclonal mice generated by nuclear transfer from mature B and T donor cells. Nature *415*, 1035-1038.

Hochedlinger, K., and Jaenisch, R. (2003). Nuclear transplantation, embryonic stem cells, and the potential for cell therapy. The New England journal of medicine *349*, 275-286.

Hochedlinger, K., and Jaenisch, R. (2006). Nuclear reprogramming and pluripotency. Nature *441*, 1061-1067.

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A.E., and Melton, D.A. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. Nat Biotechnol *26*, 795-797.

Humphrey, R.K., Beattie, G.M., Lopez, A.D., Bucay, N., King, C.C., Firpo, M.T., Rose-John, S., and Hayek, A. (2004). Maintenance of pluripotency in human embryonic stem cells is STAT3 independent. Stem Cells *22*, 522-530.

Ichida, J.K., Blanchard, J., Lam, K., Son, E.Y., Chung, J.E., Egli, D., Loh, K.M., Carter, A.C., Di Giorgio, F.P., Koszka, K., *et al.* (2009). A small-molecule inhibitor of tgf-Beta signaling replaces sox2 in reprogramming by inducing nanog. Cell Stem Cell *5*, 491-503.

Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I.R. (2006). Dissecting self-renewal in stem cells with RNA interference. Nature *442*, 533-538.

Jaenisch, R., and Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell *132*, 567-582.

Jahner, D., Stuhlmann, H., Stewart, C.L., Harbers, K., Lohler, J., Simon, I., and Jaenisch, R. (1982). De novo methylation and expression of retroviral genomes during mouse embryogenesis. Nature *298*, 623-628.

Jiang, J., Chan, Y.S., Loh, Y.H., Cai, J., Tong, G.Q., Lim, C.A., Robson, P., Zhong, S., and Ng, H.H. (2008). A core Klf circuitry regulates self-renewal of embryonic stem cells. Nat Cell Biol *10*, 353-360.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., *et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Kanellopoulou, C., Muljo, S.A., Kung, A.L., Ganesan, S., Drapkin, R., Jenuwein, T., Livingston, D.M., and Rajewsky, K. (2005). Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes Dev *19*, 489-501.

Kim, J., Efe, J.A., Zhu, S., Talantova, M., Yuan, X., Wang, S., Lipton, S.A., Zhang, K., and Ding, S. (2011). Direct reprogramming of mouse fibroblasts to neural progenitors. Proc Natl Acad Sci U S A *108*, 7838-7843.

Koche, R.P., Smith, Z.D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B.E., and Meissner, A. (2011). Reprogramming factor expression initiates widespread targeted chromatin remodeling. Cell Stem Cell *8*, 96-105.

Krichevsky, A.M., Sonntag, K.C., Isacson, O., and Kosik, K.S. (2006). Specific microRNAs modulate embryonic stem cell-derived neurogenesis. Stem Cells *24*, 857-864.

Kunath, T., Saba-El-Leil, M.K., Almousailleakh, M., Wray, J., Meloche, S., and Smith, A. (2007). FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. Development *134*, 2895-2902.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K., *et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Lengner, C.J., Gimelbrant, A.A., Erwin, J.A., Cheng, A.W., Guenther, M.G., Welstead, G.G., Alagappan, R., Frampton, G.M., Xu, P., Muffat, J., *et al.* (2010). Derivation of pre-X inactivation human embryonic stem cells under physiological oxygen concentrations. Cell *141*, 872-883.

Li, P., Tong, C., Mehrian-Shai, R., Jia, L., Wu, N., Yan, Y., Maxson, R.E., Schulze, E.N., Song, H., Hsieh, C.L., *et al.* (2008). Germline competent embryonic stem cells derived from rat blastocysts. Cell *135*, 1299-1310.

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

Liang, G., He, J., and Zhang, Y. (2012). Kdm2b promotes induced pluripotent stem cell generation by facilitating gene activation early in reprogramming. Nat Cell Biol *14*, 457-466.

Liang, G., and Zhang, Y. (2013). Embryonic stem cell and induced pluripotent stem cell: an epigenetic perspective. Cell Res *23*, 49-69.

Liao, B., Bao, X., Liu, L., Feng, S., Zovoilis, A., Liu, W., Xue, Y., Cai, J., Guo, X., Qin, B., *et al.* (2011). MicroRNA cluster 302-367 enhances somatic cell reprogramming by accelerating a mesenchymal-to-epithelial transition. The Journal of biological chemistry *286*, 17359-17364.

Lin, C., Garruss, A.S., Luo, Z., Guo, F., and Shilatifard, A. (2013). The RNA Pol II elongation factor Ell3 marks enhancers in ES cells and primes future gene activation. Cell *152*, 144-156.

Maekawa, M., Yamaguchi, K., Nakamura, T., Shibukawa, R., Kodanaka, I., Ichisaka, T., Kawamura, Y., Mochizuki, H., Goshima, N., and Yamanaka, S. (2011). Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1. Nature *474*, 225-229.

Maherali, N., and Hochedlinger, K. (2009). Tgfbeta signal inhibition cooperates in the induction of iPSCs and replaces Sox2 and cMyc. Current biology : CB *19*, 1718-1723.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., *et al.* (2007). Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell Stem Cell *1*, 55-70.

Mansfield, J.H., Harfe, B.D., Nissen, R., Obenauer, J., Srineel, J., Chaudhuri, A., Farzan-Kashani, R., Zuker, M., Pasquinelli, A.E., Ruvkun, G., *et al.* (2004). MicroRNA-responsive 'sensor' transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. Nat Genet *36*, 1079-1083.

Mansour, A.A., Gafni, O., Weinberger, L., Zviran, A., Ayyash, M., Rais, Y., Krupalnik, V., Zerbib, M., Amann-Zalcenstein, D., Maza, I., *et al.* (2012). The H3K27 demethylase Utx regulates somatic and germ cell epigenetic reprogramming. Nature *488*, 409-413.

Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc Natl Acad Sci U S A *78*, 7634-7638.

Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., Takahashi, K., Okochi, H., Okuda, A., Matoba, R., Sharov, A.A., *et al.* (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol *9*, 625-635.

Matsumura, H., Tada, M., Otsuji, T., Yasuchika, K., Nakatsuji, N., Surani, A., and Tada, T. (2007). Targeted chromosome elimination from ES-somatic hybrid cells. Nat Methods *4*, 23-25.

Mattout, A., Biran, A., and Meshorer, E. (2011). Global epigenetic changes during somatic cell reprogramming to iPS cells. Journal of molecular cell biology *3*, 341-350.

Meissner, A., Wernig, M., and Jaenisch, R. (2007). Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. Nat Biotechnol *25*, 1177-1181.

Meshorer, E., Yellajoshula, D., George, E., Scambler, P.J., Brown, D.T., and Misteli, T. (2006). Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. Dev Cell *10*, 105-116.

Mikkelsen, T.S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B.E., Jaenisch, R., Lander, E.S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. Nature *454*, 49-55.

Miller, R.A., and Ruddle, F.H. (1976). Pluripotent teratocarcinoma-thymus somatic cell hybrids. Cell *9*, 45-55.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell *113*, 631-642.

Morgan, H.D., Santos, F., Green, K., Dean, W., and Reik, W. (2005). Epigenetic reprogramming in mammals. Hum Mol Genet *14 Spec No 1*, R47-58.

Murchison, E.P., Partridge, J.F., Tam, O.H., Cheloufi, S., and Hannon, G.J. (2005). Characterization of Dicer-deficient murine embryonic stem cells. Proc Natl Acad Sci U S A *102*, 12135-12140.

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat Biotechnol *26*, 101-106.

Nakatake, Y., Fukui, N., Iwamatsu, Y., Masui, S., Takahashi, K., Yagi, R., Yagi, K., Miyazaki, J., Matoba, R., Ko, M.S., *et al.* (2006). Klf4 cooperates with Oct3/4 and Sox2 to activate the Lefty1 core promoter in embryonic stem cells. Molecular and cellular biology *26*, 7772-7782.

Nichols, J., Chambers, I., Taga, T., and Smith, A. (2001). Physiological rationale for responsiveness of mouse embryonic stem cells to gp130 cytokines. Development *128*, 2333-2339.

Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Scholer, H., and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. Cell *95*, 379-391.

Niwa, H., Burdon, T., Chambers, I., and Smith, A. (1998). Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. Genes Dev *12*, 2048-2060.

Niwa, H., Miyazaki, J., and Smith, A.G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat Genet *24*, 372-376.

Niwa, H., Ogawa, K., Shimosato, D., and Adachi, K. (2009). A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. Nature *460*, 118-122.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Nekludova, L., Rolfe, P.A., Danford, T.W., Gifford, D.K., Fraenkel, E., Bell, G.I., and Young, R.A. (2006). Core transcriptional regulatory circuitry in human hepatocytes. Molecular systems biology *2*, 2006 0017.

Ogonuki, N., Inoue, K., Yamamoto, Y., Noguchi, Y., Tanemura, K., Suzuki, O., Nakayama, H., Doi, K., Ohtomo, Y., Satoh, M.*, et al.* (2002). Early death of mice cloned from somatic cells. Nat Genet *30*, 253-254.

Okamoto, K., Okazawa, H., Okuda, A., Sakai, M., Muramatsu, M., and Hamada, H. (1990). A novel octamer binding transcription factor is differentially expressed in mouse embryonic cells. Cell *60*, 461-472.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.

Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. Nature *448*, 313-317.

Onder, T.T., Kara, N., Cherry, A., Sinha, A.U., Zhu, N., Bernt, K.M., Cahan, P., Mancarci, O.B., Unternaehrer, J., Gupta, P.B.*, et al.* (2012). Chromatin-modifying enzymes as modulators of reprogramming. Nature.

Park, I.H., Zhao, R., West, J.A., Yabuuchi, A., Huo, H., Ince, T.A., Lerou, P.H., Lensch, M.W., and Daley, G.Q. (2008). Reprogramming of human somatic cells to pluripotency with defined factors. Nature *451*, 141-146.

Pennisi, E., and Williams, N. (1997). Will Dolly send in the clones? Science *275*, 1415-1416.

Pera, M.F., and Tam, P.P. (2010). Extrinsic regulation of pluripotent stem cells. Nature *465*, 713-720.

Pevny, L.H., and Lovell-Badge, R. (1997). Sox genes find their feet. Curr Opin Genet Dev *7*, 338-344.

Pijnappel, W.W., Esch, D., Baltissen, M.P., Wu, G., Mischerikow, N., Bergsma, A.J., van der Wal, E., Han, D.W., Bruch, H., Moritz, S.*, et al.* (2013). A central role for TFIID in the pluripotent transcription circuitry. Nature *495*, 516-519.

Polo, J.M., Anderssen, E., Walsh, R.M., Schwarz, B.A., Nefzger, C.M., Lim, S.M., Borkent, M., Apostolou, E., Alaei, S., Cloutier, J.*, et al.* (2012). A molecular roadmap of reprogramming somatic cells into iPS cells. Cell *151*, 1617-1632.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279-283.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. Cell *138*, 114-128.

Rosner, M.H., Vigano, M.A., Ozato, K., Timmons, P.M., Poirier, F., Rigby, P.W., and Staudt, L.M. (1990). A POU-domain transcription factor in early stem cells and germ cells of the mammalian embryo. Nature *345*, 686-692.

Rossant, J. (2001). Stem cells from the Mammalian blastocyst. Stem Cells *19*, 477-482.

Rowland, B.D., and Peeper, D.S. (2006). KLF4, p21 and context-dependent opposing forces in cancer. Nat Rev Cancer *6*, 11-23.

Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H.K., Beyer, T.A., Datti, A., Woltjen, K., Nagy, A., and Wrana, J.L. (2010). Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell *7*, 64-77.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109-113.

Sato, N., Meijer, L., Skaltsounis, L., Greengard, P., and Brivanlou, A.H. (2004). Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. Nat Med *10*, 55-63.

Schmidt, R., and Plath, K. (2012). The roles of the reprogramming factors Oct4, Sox2 and Klf4 in resetting the somatic cell epigenome during induced pluripotent stem cell generation. Genome biology *13*, 251.

Scholer, H.R., Balling, R., Hatzopoulos, A.K., Suzuki, N., and Gruss, P. (1989a). Octamer binding proteins confer transcriptional activity in early mouse embryogenesis. EMBO J *8*, 2551-2557.

Scholer, H.R., Hatzopoulos, A.K., Balling, R., Suzuki, N., and Gruss, P. (1989b). A family of octamer-specific proteins present during mouse embryogenesis: evidence for germline-specific expression of an Oct factor. EMBO J *8*, 2543-2550.

Schotta, G., Ebert, A., and Reuter, G. (2003). SU(VAR)3-9 is a conserved key function in heterochromatic gene silencing. Genetica *117*, 149-158.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., and Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. Cell *128*, 735-745.

Schuh, R., Aicher, W., Gaul, U., Cote, S., Preiss, A., Maier, D., Seifert, E., Nauber, U., Schroder, C., Kemler, R., *et al.* (1986). A conserved family of nuclear proteins containing structural elements of the finger protein encoded by Kruppel, a Drosophila segmentation gene. Cell *47*, 1025-1032.

Seki, T., Yuasa, S., Oda, M., Egashira, T., Yae, K., Kusumoto, D., Nakata, H., Tohyama, S., Hashimoto, H., Kodaira, M., *et al.* (2010). Generation of induced pluripotent stem cells from human terminally differentiated circulating T cells. Cell Stem Cell *7*, 11-14.

Shields, J.M., Christy, R.J., and Yang, V.W. (1996). Identification and characterization of a gene encoding a gut-enriched Kruppel-like factor expressed during growth arrest. The Journal of biological chemistry *271*, 20009-20017.

Singhal, N., Graumann, J., Wu, G., Arauzo-Bravo, M.J., Han, D.W., Greber, B., Gentile, L., Mann, M., and Scholer, H.R. (2010). Chromatin-Remodeling Components of the BAF Complex Facilitate Reprogramming. Cell *141*, 943-955.

Smith, A.G., Heath, J.K., Donaldson, D.D., Wong, G.G., Moreau, J., Stahl, M., and Rogers, D. (1988). Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature *336*, 688-690.

Smith, Z.D., Nachman, I., Regev, A., and Meissner, A. (2010). Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. Nat Biotechnol *28*, 521-526.

Solter, D. (2006). From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research. Nat Rev Genet *7*, 319-327.

Soufi, A., Donahue, G., and Zaret, K.S. (2012). Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell *151*, 994-1004.

Sridharan, R., Tchieu, J., Mason, M.J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q., and Plath, K. (2009). Role of the murine reprogramming factors in the induction of pluripotency. Cell *136*, 364-377.

Stadtfeld, M., Apostolou, E., Akutsu, H., Fukuda, A., Follett, P., Natesan, S., Kono, T., Shioda, T., and Hochedlinger, K. (2010a). Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature *465*, 175-181.

Stadtfeld, M., Apostolou, E., Ferrari, F., Choi, J., Walsh, R.M., Chen, T., Ooi, S.S., Kim, S.Y., Bestor, T.H., Shioda, T.*, et al.* (2012). Ascorbic acid prevents loss of Dlk1-Dio3 imprinting and facilitates generation of all-iPS cell mice from terminally differentiated B cells. Nat Genet *44*, 398-405, S391-392.

Stadtfeld, M., Maherali, N., Borkent, M., and Hochedlinger, K. (2010b). A reprogrammable mouse strain from gene-targeted embryonic stem cells. Nat Methods *7*, 53-55.

Stadtfeld, M., Maherali, N., Breault, D.T., and Hochedlinger, K. (2008). Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. Cell Stem Cell *2*, 230-240.

Staerk, J., Dawlaty, M.M., Gao, Q., Maetzel, D., Hanna, J., Sommer, C.A., Mostoslavsky, G., and Jaenisch, R. (2010). Reprogramming of human peripheral blood cells to induced pluripotent stem cells. Cell Stem Cell *7*, 20-24.

Subramanyam, D., Lamouille, S., Judson, R.L., Liu, J.Y., Bucay, N., Derynck, R., and Blelloch, R. (2011). Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells. Nat Biotechnol *29*, 443-448.

Tada, M., Morizane, A., Kimura, H., Kawasaki, H., Ainscough, J.F., Sasai, Y., Nakatsuji, N., and Tada, T. (2003). Pluripotency of reprogrammed somatic genomes in embryonic stem hybrid cells. Developmental dynamics : an official publication of the American Association of Anatomists *227*, 504-510.

Tada, M., Tada, T., Lefebvre, L., Barton, S.C., and Surani, M.A. (1997). Embryonic germ cells induce epigenetic reprogramming of somatic nucleus in hybrid cells. EMBO J *16*, 6510-6520.

Tada, M., Takahama, Y., Abe, K., Nakatsuji, N., and Tada, T. (2001). Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. Current biology : CB *11*, 1553-1558.

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

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Takeuchi, J.K., and Bruneau, B.G. (2009). Directed transdifferentiation of mouse mesoderm to heart tissue by defined factors. Nature *459*, 708-711.

Tamashiro, K.L., Wakayama, T., Akutsu, H., Yamazaki, Y., Lachey, J.L., Wortman, M.D., Seeley, R.J., D'Alessio, D.A., Woods, S.C., Yanagimachi, R.*, et al.* (2002). Cloned mice have an obese phenotype not transmitted to their offspring. Nat Med *8*, 262-267.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145-1147.

Tiemann, U., Sgodda, M., Warlich, E., Ballmaier, M., Scholer, H.R., Schambach, A., and Cantz, T. (2011). Optimal reprogramming factor stoichiometry increases colony numbers and affects molecular characteristics of murine induced pluripotent stem cells. Cytometry Part A : the journal of the International Society for Analytical Cytology *79*, 426-435.

Vierbuchen, T., and Wernig, M. (2012). Molecular roadblocks for cellular reprogramming. Molecular cell *47*, 827-838.

Wakao, S., Kitada, M., and Dezawa, M. (2013). The elite and stochastic model for iPS cell generation: multilineage-differentiating stress enduring (Muse) cells are readily reprogrammable into iPS cells. Cytometry Part A : the journal of the International Society for Analytical Cytology *83*, 18-26.

Wakayama, T., Perry, A.C., Zuccotti, M., Johnson, K.R., and Yanagimachi, R. (1998). Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. Nature *394*, 369-374.

Wang, S.H., Tsai, M.S., Chiang, M.F., and Li, H. (2003). A novel NK-type homeobox gene, ENK (early embryo specific NK), preferentially expressed in embryonic stem cells. Gene expression patterns : GEP *3*, 99-103.

Wang, T., Chen, K., Zeng, X., Yang, J., Wu, Y., Shi, X., Qin, B., Zeng, L., Esteban, M.A., Pan, G.*, et al.* (2011). The histone demethylases Jhdm1a/1b enhance somatic cell reprogramming in a vitamin-C-dependent manner. Cell Stem Cell *9*, 575-587.

Wang, Y., Medvid, R., Melton, C., Jaenisch, R., and Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet *39*, 380-385.

Wernig, M., Lengner, C.J., Hanna, J., Lodato, M.A., Steine, E., Foreman, R., Staerk, J., Markoulaki, S., and Jaenisch, R. (2008a). A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. Nat Biotechnol *26*, 916-924.

Wernig, M., Meissner, A., Cassady, J.P., and Jaenisch, R. (2008b). c-Myc is dispensable for direct reprogramming of mouse fibroblasts. Cell Stem Cell *2*, 10-12.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature *448*, 318-324.

Wilmut, I., Schnieke, A.E., McWhir, J., Kind, A.J., and Campbell, K.H. (1997). Viable offspring derived from fetal and adult mammalian cells. Nature *385*, 810-813.

Yamaguchi, S., Hirano, K., Nagata, S., and Tada, T. (2011). Sox2 expression effects on direct reprogramming efficiency as determined by alternative somatic cell fate. Stem cell research *6*, 177-186.

Yamanaka, S. (2009). Elite and stochastic models for induced pluripotent stem cell generation. Nature *460*, 49-52.

Yamanaka, S., and Blau, H.M. (2010). Nuclear reprogramming to a pluripotent state by three approaches. Nature *465*, 704-712.

Yang, X., Smith, S.L., Tian, X.C., Lewin, H.A., Renard, J.P., and Wakayama, T. (2007). Nuclear reprogramming of cloned embryos and its implications for therapeutic cloning. Nat Genet *39*, 295-302.

Ying, Q.L., Nichols, J., Chambers, I., and Smith, A. (2003). BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. Cell *115*, 281-292.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519-523.

Yu, J., Vodyanik, M.A., He, P., Slukvin, II, and Thomson, J.A. (2006). Human embryonic stem cells reprogram myeloid precursors following cell-cell fusion. Stem Cells *24*, 168-176.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R.*, et al.* (2007). Induced pluripotent stem cell lines derived from human somatic cells. Science *318*, 1917-1920.

Yuan, H., Corbi, N., Basilico, C., and Dailey, L. (1995). Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. Genes Dev *9*, 2635-2645.

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes Dev *25*, 2227-2241.

Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L.*, et al.* (2013). Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell *152*, 642-654.

Zuo, B., Yang, J., Wang, F., Wang, L., Yin, Y., Dan, J., Liu, N., and Liu, L. (2012). Influences of lamin A levels on induction of pluripotent stem cells. Biology open *1*, 1118-1127.

Zwaka, T.P., and Thomson, J.A. (2005). A germ cell origin of embryonic stem cells? Development *132*, 227-233.

Chapter 2. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase

Yosef Buganim[1*], Dina A. Faddah [1,2*], Albert W. Cheng[1,3], Elena Itskovich[1], Styliani Markoulaki[1], Kibibi Ganz[1], Sandy L. Klemm[5], Alexander van Oudenaarden[2,4,6], 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

* These authors contributed equally to this work.

Published as:
Buganim Y*, Faddah DA*, Cheng AW, Itskovich E, Markoulaki S, Ganz K, Klemm SL, van Oudenaarden A, Jaenisch R. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell.* 150(6):1209-22. 2012.
*indicates equal contributions

DAF and YB shared all experiments and analyses equally, except DAF performed all FISH experiments and analyses, YB performed all lentiviral infections, AWC performed computational analyses of Fluidigm data, and SM and KG performed blastocyst injections. EI assisted YB. SLK helped optimized FISH protocol. AvO provided support of FISH experiments. DAF, YB, and RJ wrote the paper and designed and conceived of experiments.

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.

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.

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.

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 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: (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 dox-withdrawal. 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 qRT-PCR (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 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 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.

Acknowledgements

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.

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 25mg 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 C_t values from the BioMark System into log-based expression values are described in Supplemental Methods.

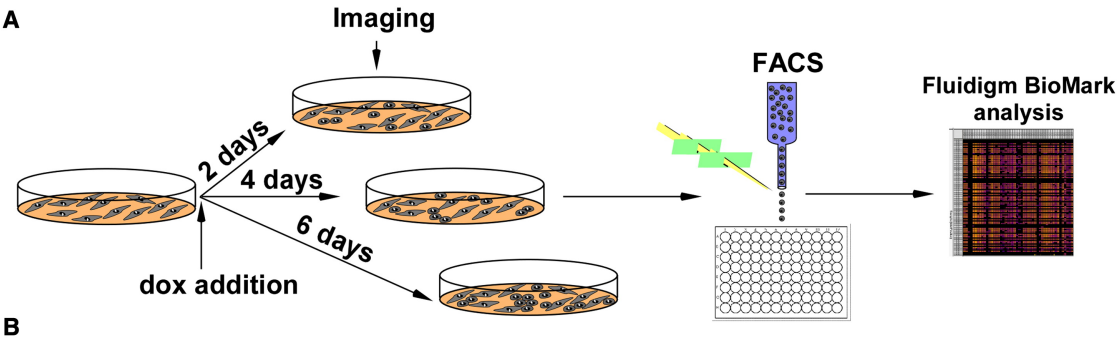
Single-cell gene expression qPCR

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.

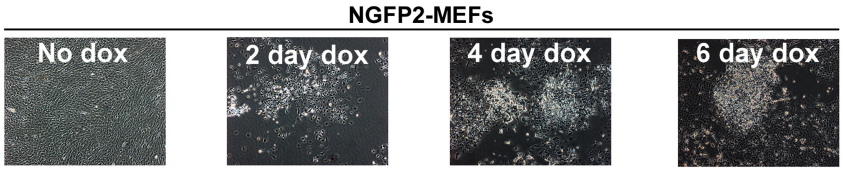
Jensen-Shannon Divergence

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

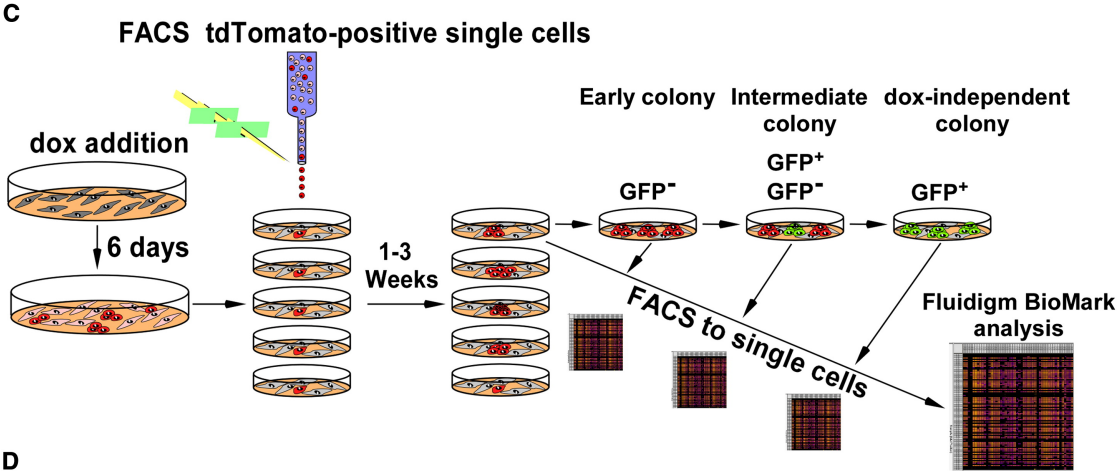
A



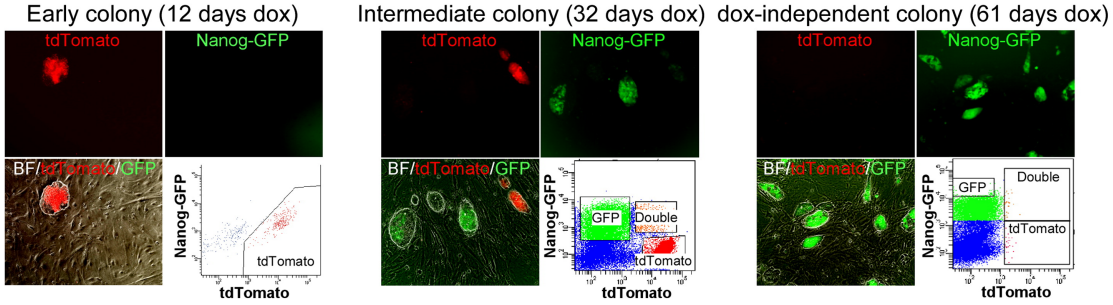
B



C



D



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

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.

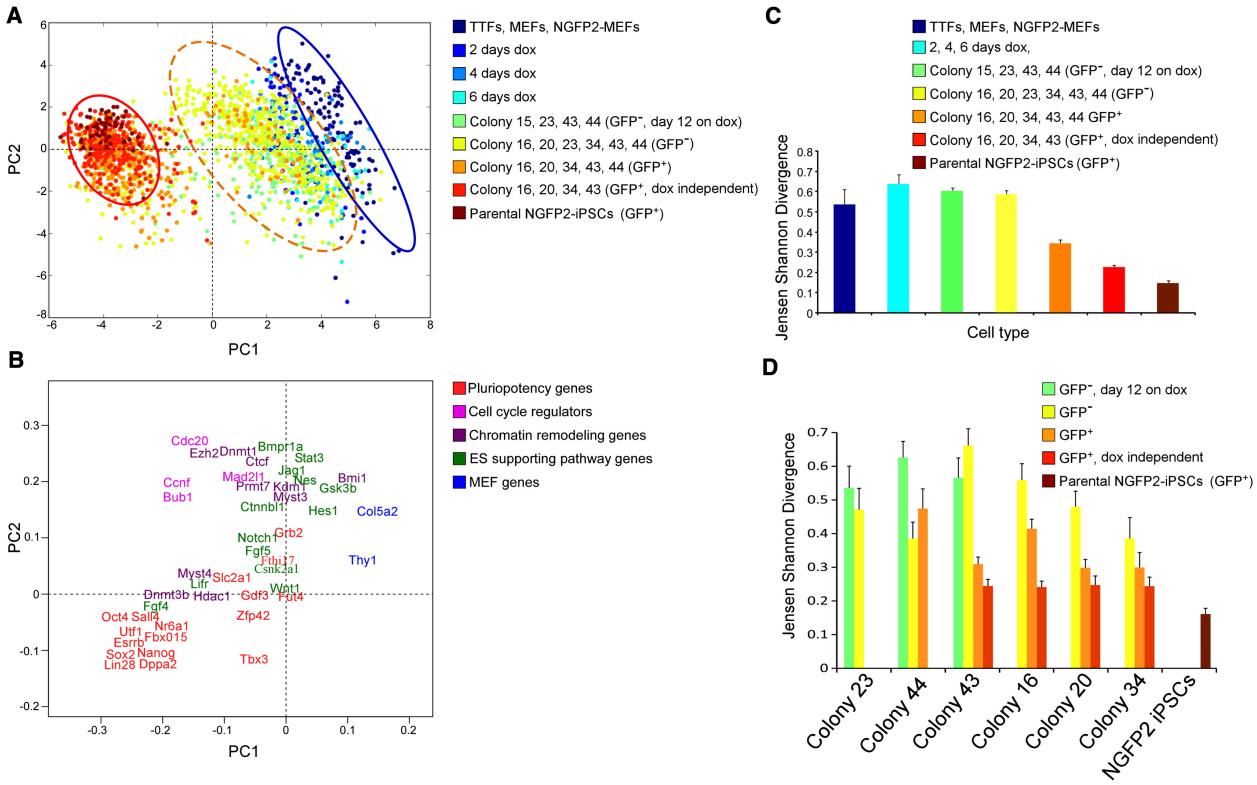


Figure 2. Three reprogramming states

(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.

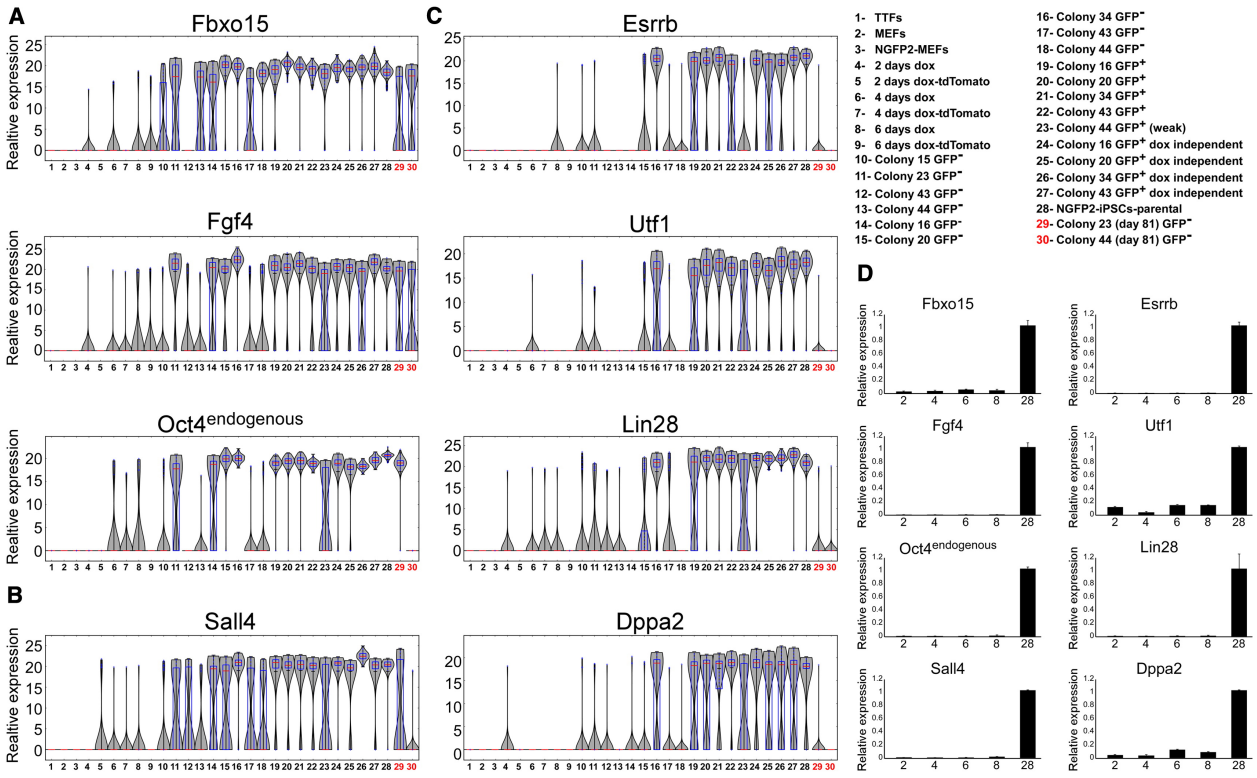


Figure 3. Established early markers are not sufficient to mark cells that will become iPSCs

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 RT-PCR of Fbxo15, Fgf4, Oct4, Sall4, Esrrb, Utf1, Lin28, and Dppa2 expression in non-clonal 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.

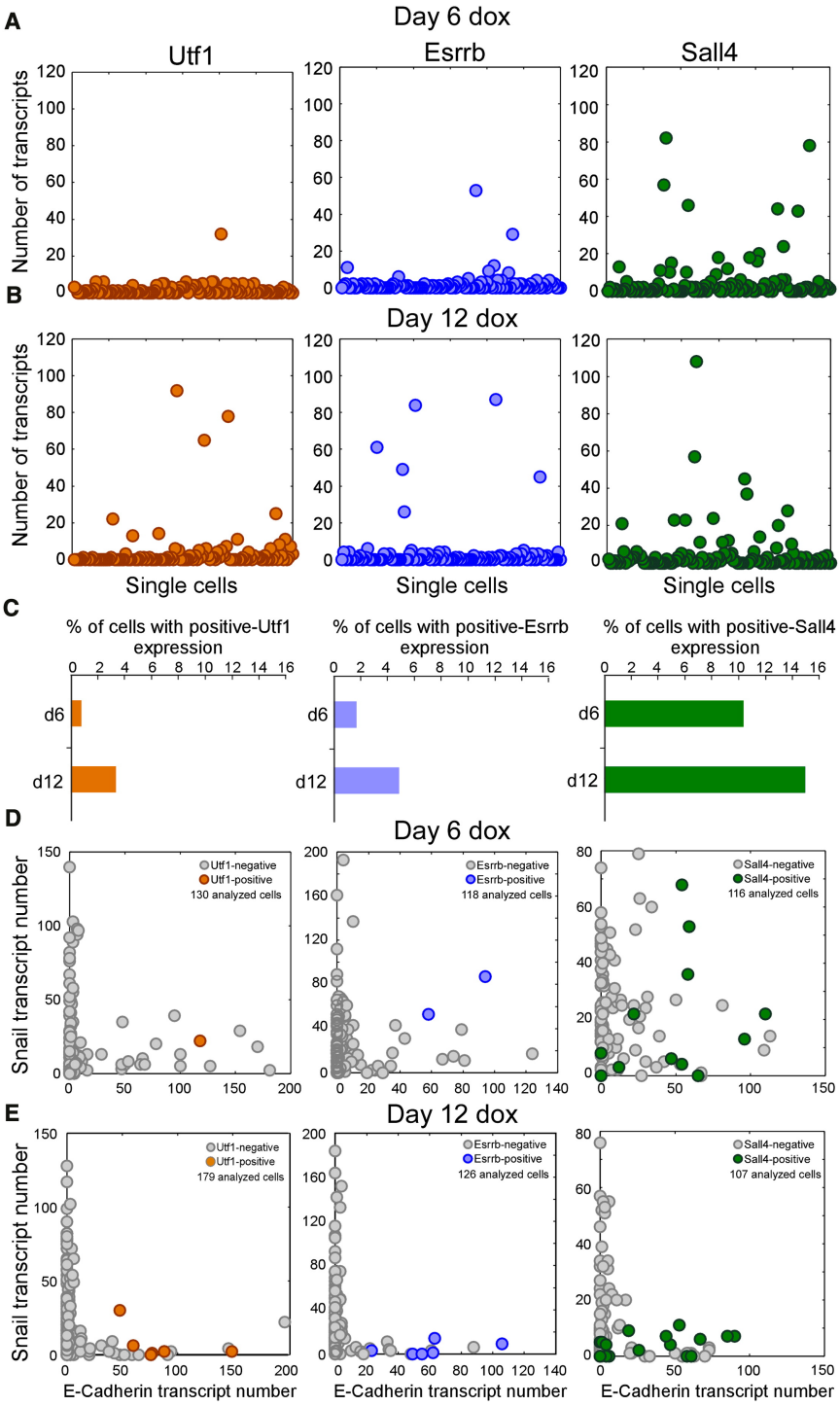


Figure 4. Early markers for reprogramming

(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.

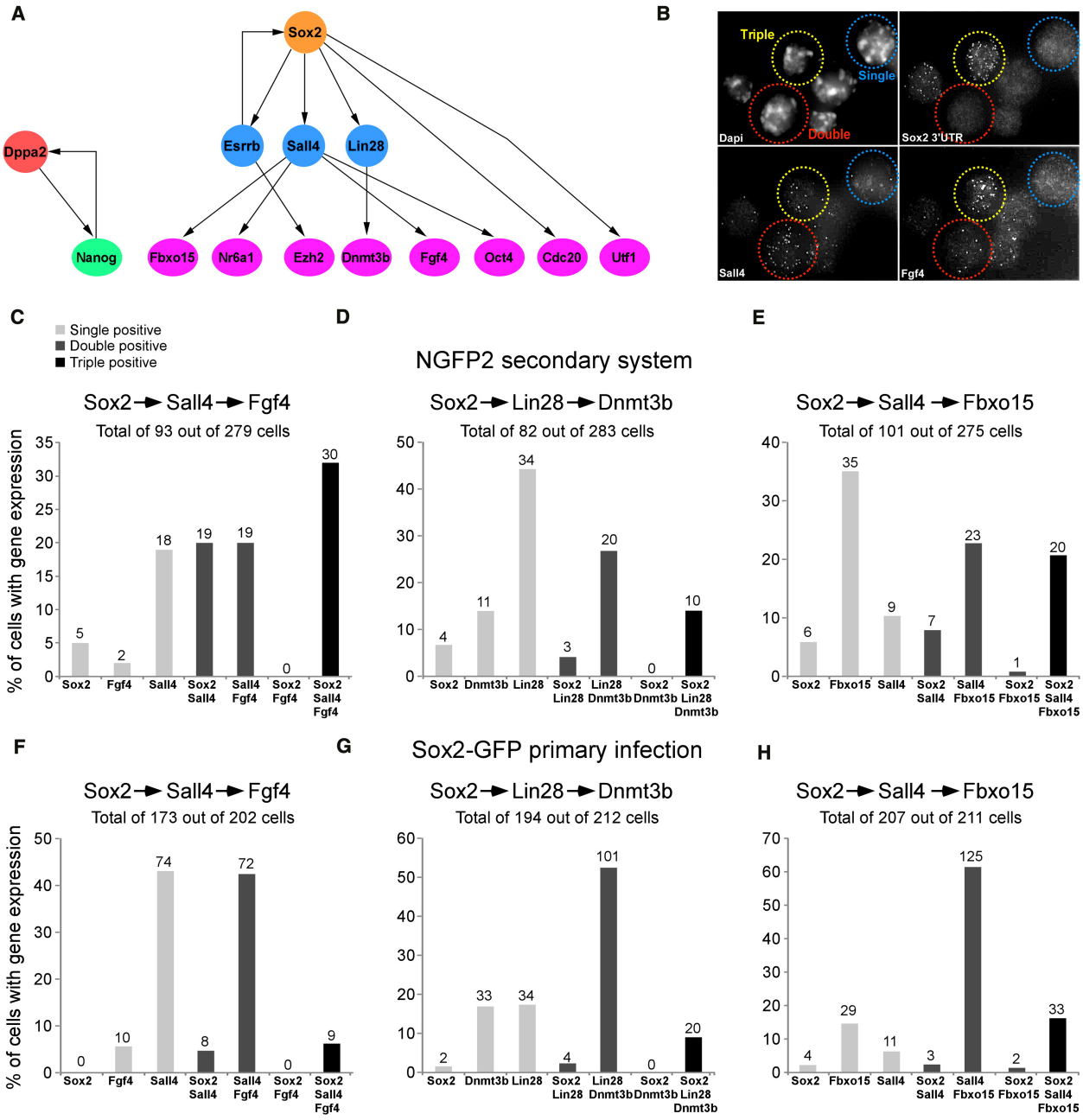


Figure 5. Model to predict the order of transcriptional events in single cells

(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.

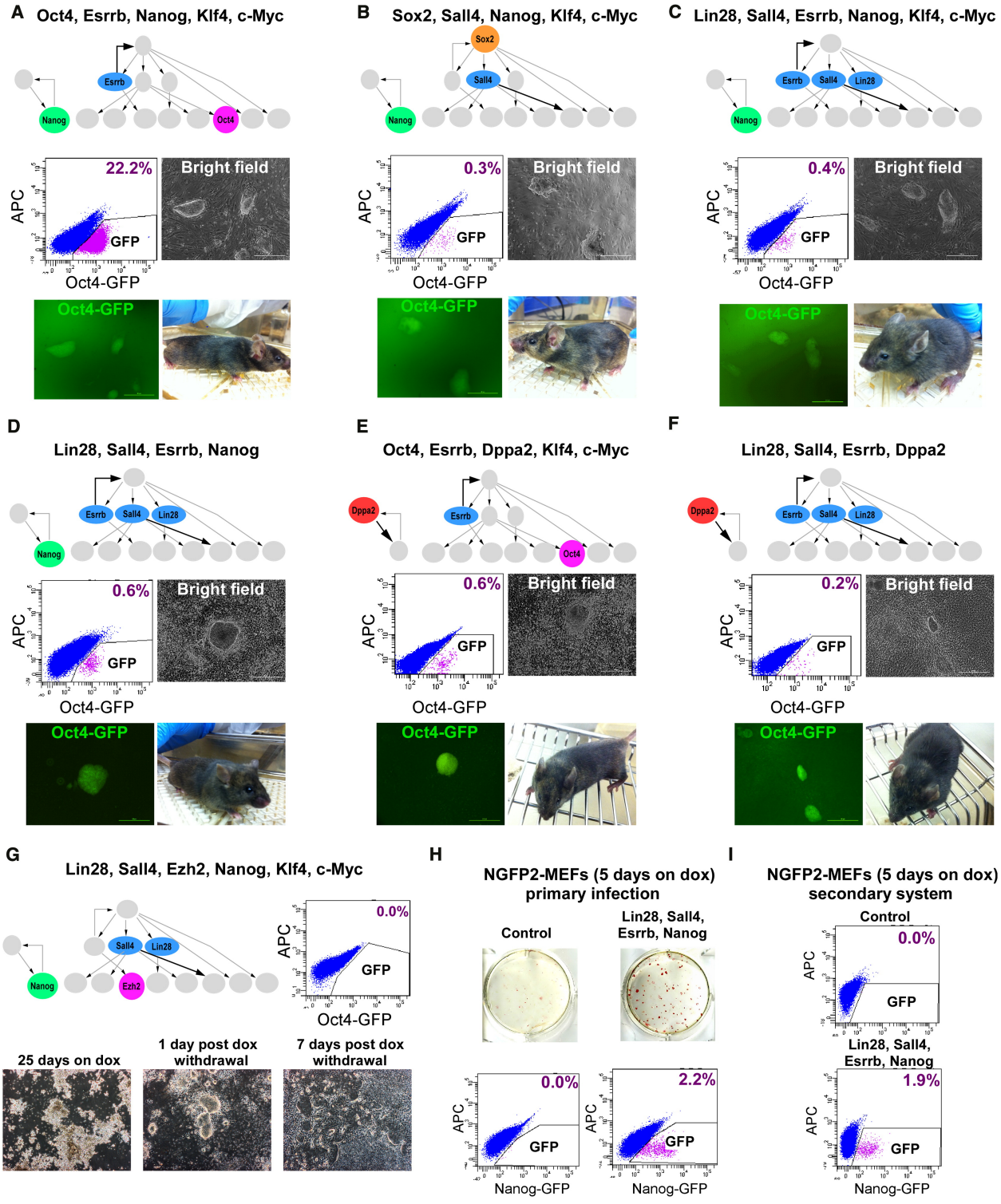


Figure 6. Cellular reprogramming with factors derived from Bayesian network

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.

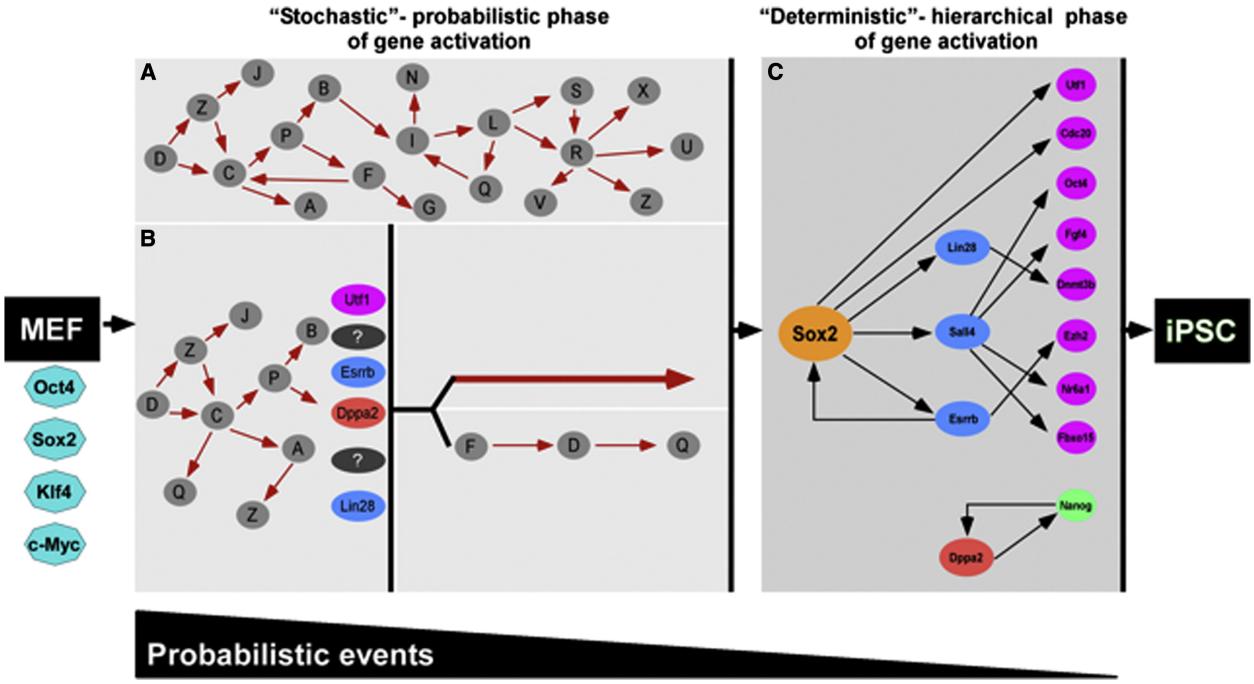


Figure 7. Two phases in reprogramming

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.

References

Bhutani, N., Brady, J.J., Damian, M., Sacco, A., Corbel, S.Y., and Blau, H.M. (2010). Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature 463, 1042-1047.

Boiani, M., Eckardt, S., Scholer, H.R., and McLaughlin, K.J. (2002). Oct4 distribution and level in mouse clones: consequences for pluripotency. Genes Dev 16, 1209-1219.

Boiani, M., and Scholer, H.R. (2005). Regulatory networks in embryo-derived pluripotent stem cells. Nat Rev Mol Cell Biol 6, 872-884.

Brambrink, T., Foreman, R., Welstead, G.G., Lengner, C.J., Wernig, M., Suh, H., and Jaenisch, R. (2008). Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. Cell Stem Cell 2, 151-159.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P.S., Rothenberg, M.E., Leyrat, A.A., Sim, S., Okamoto, J., Johnston, D.M., Qian, D., et al. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat Biotechnol 29, 1120-1127.

Diehn, M., Cho, R.W., Lobo, N.A., Kalisky, T., Dorie, M.J., Kulp, A.N., Qian, D., Lam, J.S., Ailles, L.E., Wong, M., et al. (2009). Association of reactive oxygen species levels and radioresistance in cancer stem cells. Nature 458, 780-783.

Egli, D., Birkhoff, G., and Eggan, K. (2008). Mediators of reprogramming: transcription factors and transitions through mitosis. Nat Rev Mol Cell Biol 9, 505-516.

Feng, B., Jiang, J., Kraus, P., Ng, J.H., Heng, J.C., Chan, Y.S., Yaw, L.P., Zhang, W., Loh, Y.H., Han, J., et al. (2009). Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. Nat Cell Biol 11, 197-203.

Guo, G., Huss, M., Tong, G.Q., Wang, C., Li Sun, L., Clarke, N.D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev Cell 18, 675-685.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature 462, 595-601.

Hanna, J.H., Saha, K., and Jaenisch, R. (2010). Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. Cell 143, 508-525.

Hong, H., Takahashi, K., Ichisaka, T., Aoi, T., Kanagawa, O., Nakagawa, M., Okita, K., and Yamanaka, S. (2009). Suppression of induced pluripotent stem cell generation by the p53-p21 pathway. Nature 460, 1132-1135.

Itzkovitz, S., Lyubimova, A., Blat, I.C., Maynard, M., van Es, J., Lees, J., Jacks, T., Clevers, H., and van Oudenaarden, A. (2011). Single-molecule transcript counting of stem-cell markers in the mouse intestine. Nat Cell Biol 14, 106-114.

Jaenisch, R., and Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell 132, 567-582.

Koche, R.P., Smith, Z.D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B.E., and Meissner, A. (2011). Reprogramming factor expression initiates widespread targeted chromatin remodeling. Cell Stem Cell 8, 96-105.

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

Lin, J.H. (1991). Divergence Measures Based on the Shannon Entropy. Ieee T Inform Theory 37, 145-151.

Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B., et al. (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev 25, 594-607.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., et al. (2007). Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. Cell Stem Cell 1, 55-70.

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

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

Meissner, A., Wernig, M., and Jaenisch, R. (2007). Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. Nat Biotechnol 25, 1177-1181.

Mikkelsen, T.S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B.E., Jaenisch, R., Lander, E.S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. Nature 454, 49-55.

Narsinh, K.H., Sun, N., Sanchez-Freire, V., Lee, A.S., Almeida, P., Hu, S., Jan, T., Wilson, K.D., Leong, D., Rosenberg, J., et al. (2011). Single cell transcriptional profiling

reveals heterogeneity of human induced pluripotent stem cells. J Clin Invest 121, 1217-1221.

Ng, H.H., and Surani, M.A. (2011). The transcriptional and signalling networks of pluripotency. Nat Cell Biol 13, 490-496.

Okita, K., Ichisaka, T., and Yamanaka, S. (2007). Generation of germline-competent induced pluripotent stem cells. Nature 448, 313-317.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell 137, 1194-1211.

Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature 463, 913-918.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods 5, 877-879.

Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R.C., and Melton, D.A. (2002). "Stemness": transcriptional profiling of embryonic and adult stem cells. Science 298, 597-600.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447, 425-432.

Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H.K., Beyer, T.A., Datti, A., Woltjen, K., Nagy, A., and Wrana, J.L. (2010). Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. Cell Stem Cell 7, 64-77.

Silva, J., Nichols, J., Theunissen, T.W., Guo, G., van Oosten, A.L., Barrandon, O., Wray, J., Yamanaka, S., Chambers, I., and Smith, A. (2009). Nanog is the gateway to the pluripotent ground state. Cell 138, 722-737.

Smith, Z.D., Nachman, I., Regev, A., and Meissner, A. (2010). Dynamic single-cell imaging of direct reprogramming reveals an early specifying event. Nat Biotechnol 28, 521-526.

Stadtfeld, M., Maherali, N., Breault, D.T., and Hochedlinger, K. (2008). Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. Cell Stem Cell 2, 230-240.

Surani, M.A., Hayashi, K., and Hajkova, P. (2007). Genetic and epigenetic regulators of pluripotency. Cell 128, 747-762.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663-676.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M.A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6, 468-478.

Wernig, M., Lengner, C.J., Hanna, J., Lodato, M.A., Steine, E., Foreman, R., Staerk, J., Markoulaki, S., and Jaenisch, R. (2008). A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. Nat Biotechnol 26, 916-924.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature 448, 318-324.

Yamanaka, S. (2009). Elite and stochastic models for induced pluripotent stem cell generation. Nature 460, 49-52.

Zhang, J., Tam, W.L., Tong, G.Q., Wu, Q., Chan, H.Y., Soh, B.S., Lou, Y., Yang, J., Ma, Y., Chai, L., et al. (2006). Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. Nat Cell Biol 8, 1114-1123.

Zhao, X.Y., Li, W., Lv, Z., Liu, L., Tong, M., Hai, T., Hao, J., Guo, C.L., Ma, Q.W., Wang, L., et al. (2009). iPS cells produce viable mice through tetraploid complementation. Nature 461, 86-90.

Extended Methods

Quantitative real-time PCR

Total RNA was isolated using Rneasy Kit (QIAGEN). One microgram RNA was reversed transcribed using a First Strand Synthesis kit (Invitrogen). Quantitative PCR analysis was performed in duplicate using 1/100 of the reverse transcription reaction in an ABI Prism 7300 (Applied Biosystems) with Platinum SYBR green qPCR SuperMix-UDG with ROX (Invitrogen). Specific primers flanking an intron were designed to the different genes. Error bars represent s.d. of the mean of duplicate reactions.

Viral preparation and infection

Construction of lentiviral vectors containing Klf4, Sox2, Oct4 and Myc under control of the tetracycline operator and a minimal CMV promoter has been described previously (Brambrink et al., 2008). Construction of lentiviral vectors containing the following factors (Lin28, Sall4, Ezh2, Esrrb, Nanog, and Utf1) under control of the tetracycline operator and a minimal CMV promoter were generated by cloning the open reading frame of the factors, obtained by reverse transcription with specific primers, into the TOPO-TA vector (Invitrogen), and then restricted with EcoRI or MfeI and inserted into the FUW-teto expressing vector. Replication-incompetent lentiviral particles were packaged in 293T cells with a VSV-G coat and used to infect MEFs containing M2rtTA and Oct4-GFP or NGFP2-MEFs. Viral supernatants from cultures were filtered through a 0.45 mM filter and added to the cells. To initiate reprogramming the cells were grown in ES cell medium + 2mg/ml Doxycycline (DMEM supplemented with 15% FBS (Hyclone), leukemia inhibitory factor, beta-mercaptoethanol (Sigma-Aldrich), penicillin/streptomycin, L-glutamine and nonessential amino acid.

Chimera Formation

All animal procedures were performed according to NIH guidelines and were approved by the Committee on Animal Care at MIT. All 2n injections were performed using B6D2F2 embryos. Oct4-GFP or NGFP-2 iPSCs were derived from an agouti mouse and

could be identified by coat color as adults. Diploid blastocysts (94–98 hr after hCG injection) were placed in a drop of HEPES-CZB medium under mineral oil. A flat tip microinjection pipette with an internal diameter of 16 μm was used for iPS cell injections. Each blastocyst received 8–10 iPS cells. After injection, blastocysts were cultured in potassium simplex optimization medium (KSOM) and placed at 37°C until transferred to recipient females. About 15-20 injected blastocysts were transferred to each uterine horn of 2.5-day-postcoitum pseudopregnant B6D2F1 female.

Flow cytometry

Cells were trypsinized, washed once in PBS and resuspended in fluorescence-activated cell sorting (FACS) buffer (PBS + 5% FBS). The percentage of GFP-positive cells (Nanog-GFP or Oct4-GFP) was analyzed using FACS- calibur.

Secondary somatic cell isolation and culture

Primary NGFP2 iPSCs were electroporated with 25mg of linearized FUW-TetO-tdTomato construct. The transduced cells were selected using the Zeocin (400ug/ml) antibiotic. For MEF isolation, chimeric embryos were isolated at E13.5, and the head and internal organs were removed. The remaining tissue was physically dissociated and incubated in trypsin at 37 °C for 20 min, after which cells were resuspended in MEF media containing puromycin (2mg/ml, selection against the M2rTtA) and expanded for two passages before freezing. Secondary MEFs used for the described experiments were thawed and experiments plated 2 days before dox addition. Cells were plated at optimal density of 50,000 cells per 6-well plate and reprogrammed with mouse ES medium supplemented with 2mg/ml doxycycline (Sigma).

FISH and imaging

We performed FISH as outlined in (Raj et al., 2010; Raj et al., 2008). All hybridizations were performed in solution using probes coupled to either tetramethylrhodamine (TMR) (Invitrogen), Alexa 594 (Invitrogen) or Cy5 (GE Amersham). We used TMR for the probes against Esrrb, Utf1, Sox2 3'UTR, and Dnmt3b mRNA, Alexa 594 for Sall4, E-

cadherin, and Lin28 mRNA and Cy5 for Fgf4, Fbxo15, Snail, and Sox2 3'UTR. Optimal probe concentrations during hybridization were determined empirically. Imaging involved taking stacks of images spaced 0.3 μm apart using filters appropriate for DAPI, TMR, Alexa 594 and Cy5. All images were taken with a Nikon Ti-E inverted fluorescence microscope (Donatello) equipped with a 100X oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software (Molecular Devices, Downington, PA). During imaging, we minimized photobleaching through the use of an oxygen-scavenging solution using glucose oxidase.

Image analysis

We segmented the cells manually and counted the number of fluorescent spots, each of which corresponds to an individual mRNA, using a combination of a semi-automated method described in (Itzkovitz et al., 2011; Raj et al., 2008) and custom software written in MATLAB (Mathworks). We estimate our mRNA counts to be accurate to within 10–20%.

Single-cell Data Visualization

Principal component analysis (PCA) was performed in R using Bayesian Principal Component Analysis (bpca) function with missing value estimation (MVE) provided in the pcaMethods module. The PCA scores of the principle component 1 (PC1) and PC2 are color coded according to the cell types. And the loadings of each variable (genes) are represented in scatter plots.

Single-cell gene expression qPCR

Inventoried TaqMan assays (Applied Biosystem) were pooled to a final concentration of 0.2 for each of the 48 assays. Individual cells were sorted directly into 5ml RT-PreAmp Master Mix (2.5ml CellsDirect Reaction Mix (Invitrogen); 1.25 ml 0.2 pooled assays; 0.1 ml RT/Taq enzyme [CellsDirect qRT-PCR kit, Invitrogen]; 1.15 ml water). Cell lysis and sequence- specific reverse transcription were performed at 50°C for 15 min. The reverse transcriptase was inactivated by heating to 95°C for 2 min. Subsequently, in the

same tube, cDNA went through sequence-specific amplification by denaturing at 95C for 15s, and annealing and amplification at 60°C for 4 min for 18 cycles. These preamplified products were diluted 5-fold prior to analysis with Universal PCR Master Mix and inventoried TaqMan gene expression assays (ABI) in 96.96 Dynamic Arrays on a BioMark System (Fluidigm). Ct values were calculated from the system's software (BioMark Real-time PCR Analysis; Fluidigm). Each assay was performed in replicate.

Jensen-Shannon Divergence

Jensen-Shannon Divergence (JSD) was calculated to assess within-group similarity of gene expression within each cell line according to [Lin, J. (1991). "Divergence measures based on the shannon entropy". IEEE Transactions on Information Theory 37 (1): 145–151. doi:10.1109/18.61115]. Expression values of genes were transformed so that they sum up to 1 in each cell. Each cell is thus represented as a vector of probabilities P_i. Cells from the same line were grouped together and for each group, the Jensen-Shannon Divergence (JSD) was calculated from the probability vectors (P_1,P_2,...P_n) of cells in each group.

$$\mathrm{JSD}(P_1,P_2,...,P_n) = \mathrm{H}\left(\frac{1}{n}\sum_{i=1}^{n} P_i\right) - \frac{1}{n}\sum_{i=1}^{n}\mathrm{H}(P_i)$$

where H(P) is the Shannon entropy given by:

$$\mathrm{H}(P) = -\sum_{i=1}^{k} P(x_i)\log_2 P(x_i)$$

Confidence intervals (CIs) were estimated by bootstrapping (sampling with replacement). The 95% CIs were shown as error bars.

Single-cell Data Processing

C_t values obtained from the BioMark System were converted into log-based expression values according to a set of rules provided below. Briefly, for each gene, inconsistent readings or "Failed" quality control readings were filtered out. Cells with failed or inconsistent detection of control genes (Hprt, Gapdh) were removed from the analysis.

Expression values were calculated by subtracting the average gene C_t values from the average control C_t values in the corresponding cell. An arbitrary value of 20 was added to make all values non-negative. These values are called AC_{20} (Average Control at 20) to reflect the property that this quantity is a log-based representation of gene expression values such that the average control gene values are rescaled to 20. Expression values of pluripotency-associated genes (Oct4, Sox2, Nanog, Lin28, Fbxo15, Zfp42, Fut4, Tbx3, Esrrb, Dppa2, Utf1, Sall4, Gdf3 and Fgf4) which were lower than the maximum values observed in MEF samples are potential false positives and are thus set to zeros.

Ct value processing filters (in order of execution)

Primary filter:

1) For each gene,
    a. For each gene, including controls, remove data with CtCall = FAILED and CtQuality < threshold (--Ct-quality-threshold , Default No Threshold)
    b. For each gene, including controls, remove CtValues >= CtValueThreshold (--Ct-value-threshold, Default: 30.0) to filter out low expression genes (they will be not expressed)
    c. Here: No more values that are FAILs.
    d. For each gene, including controls, set all the CtCall to "INC" (inconsistent) if the difference between the maximum CtValue and the minimum CtValue > MaxCtRepDev (--max-Ct-deviation-between-replicates, Default: 2.0)

Sample filter:

2) For each control gene (in control gene list):
    a. if it is not found, remove the whole sample row.
    b. if that gene is marked as "INC" (in 2 of primary filter), remove the whole sample row

c. if no more CtValues are retained after primary filter or the number of CtValues < minValidReplicatesControl (--min-number-of-valid-data-point-per-control, Default: 1), remove the whole sample row.
d. If the mean of the CtValues > CtValueThresholdPerControl (--Ct-value-threshold-for-per-control-average, default: 25.0), remove the whole sample row.

Gene filter:

3) For each non-control gene:
    a. if that gene is marked as "INC" (in 2 of primary filter) or has all Ctvalues removed, don't do anything here. Do not continue to next step.
    b. If number of CtValues retained after primary filter is not zero but is < minValidReplicates (--min-number-of-valid-data-point-per-gene, default: 2), mark gene as "INC"
    c. If the mean of the CtValues > CtValueThresholdPerData (--Ct-value-threshold-for-data-average, default: 30.0), remove gene (remove all CtValues)

ACx Output:

4) For each sample:
    a. If sample is invalidated by sample filter, don't continue to next step.
    b. For each non-control gene:
        i. If gene not found for this sample, output NA
        ii. If "INC", output NA
        iii. If No CtValues (i.e., removed by primary filter or gene filter), output 0.0 (for genes that don't express, CtCall will be highly likely to FAIL in most/all replicates).
        iv. Else output ACx(g,s) x (--offset-output, default: 20)

Average Control at x (ACx) values

$$\mathrm{AC}_x(\text{gene } g, \text{controls } c \in C, \text{sample } s) = x + \overline{\mathrm{Ct}(c,s)} - \overline{\mathrm{Ct}(g,s)}$$

Property:

$$\text{When } \overline{\text{Ct}(c,s)} = \overline{\text{Ct}(g,s)},\ \text{AC}_x(g,C,s) = x$$

Larger value => higher expression

To find fold change of the gene from sample1 (s1) to sample 2 (s2):

$$\frac{\text{expression of g in sample 2}}{\text{expression of g in sample 1}} = R = 2^{\text{AC}_x(g,C,s2)-\text{AC}_x(g,C,s1)}$$

Group Distance Method

Distance between samples X, Y were defined as the average linkage Pearson distance of the single cell expression profiles of X and that of Y, as given by:

$$\text{AvgLinkage}(X,Y) = \frac{1}{|X||Y|}\sum_{x\in X}\sum_{y\in Y}\text{d}(x,y)$$

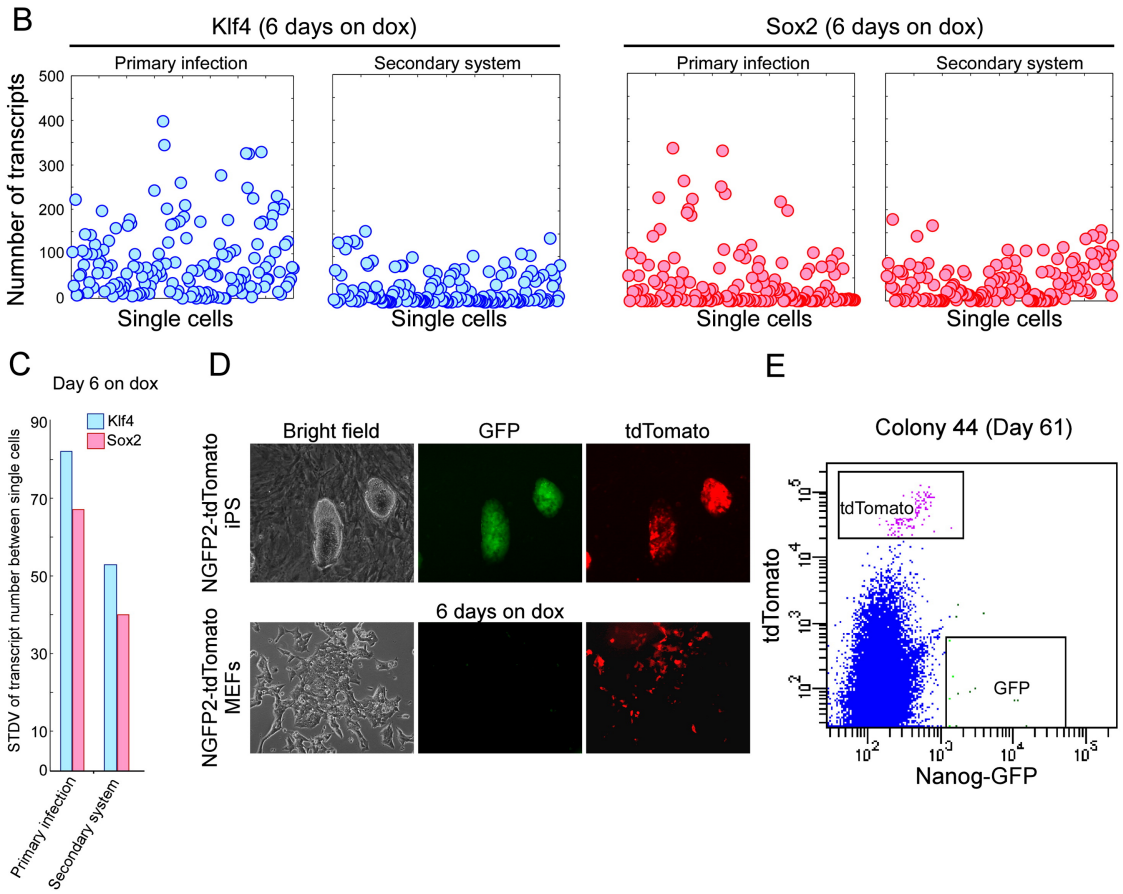
where,

$$\text{d}(x,y) = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right)$$

These were visualized as a distance matrix heatmap with d

A

Class	Genes
House keeping/control	Gapdh, Hprt
Mouse embryonic fibroblast	Thy1, Col5a2
Pluripotency	Pou5f1, Sox2, Nanog, Lin28, Fbxo15, Zfp42, Fut4, Tbx3, Esrrb, Dppa2, Utf1, Sall4, Gdf3, Grb2, Slc2a1, Fthi17, Nr6a1
Mouse embryonic stem cell signaling pathway	Bmpr1a, Stat3, Ctnnbl1, Nes, Wnt1, Gsk3b, Csnk2a1, Lifr ,Hes1, Jag1, Notch1, Fgf5, Fgf4
Chromatin modulators	Myst3, Kdm1, Hdac1, Dnmt1, Prmt7, Ctcf, Myst4, Dnmt3b, Ezh2, Bmi1
Cell cycle regulators	Bub1, Cdc20, Mad2l1, Ccnf



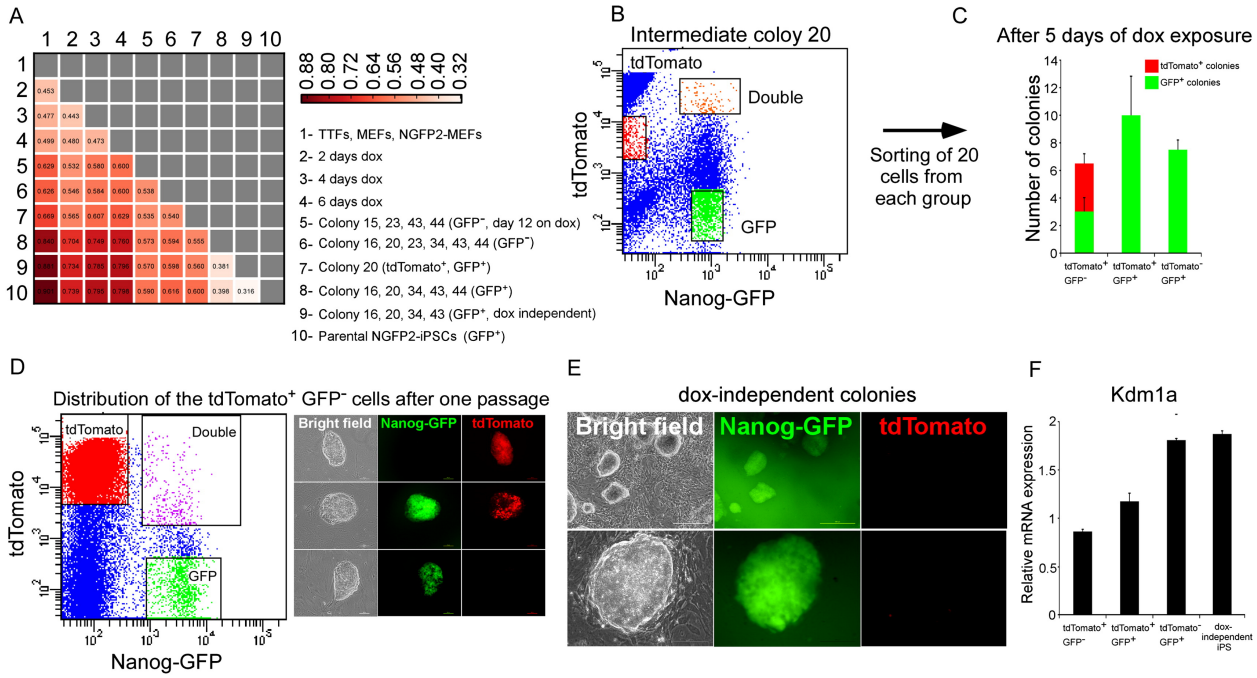
Supplemental Figure 1. Selection of 48 candidate genes and NGFP2-tdTomato system

(A) Six classes of genes are represented: House keeping control, mouse embryonic fibroblast, pluripotency, mouse embryonic stem cells supporting pathway, chromatin modulators, and cell cycle regulators.

(B) single molecule mRNA FISH of Klf4 (blue) and Sox2 (red) expression in single primary-infected NGFP MEFs and secondary NGFP2 MEFs on dox for six days. Each cell is represented as a single dot. In total, 160 single cells are displayed for each plot.

(C) Bar plot of the standard deviation of Klf4 (blue) and Sox2 (red) and transcripts between single cells of primary-infected and secondary MEFs.

(D) Representative bright field, GFP, and tdTomato images of NGFP2-tdTomato-iPSCs and NGFP2-tdTomato-MEFs after six days of dox exposure.
(E) Flow cytometric analysis of GFP and tdTomato in NGFP2 cells of colony 44 on dox for 61 days. See also Figure 1.



Supplemental Figure 2. Analysis of intermediate cells

(A) Heatmap of the Pearson distance and average linkage between populations listed in Figure 2. See Supplemental Methods (Group Distance Method) for detailed explanation of normalization and data analysis.

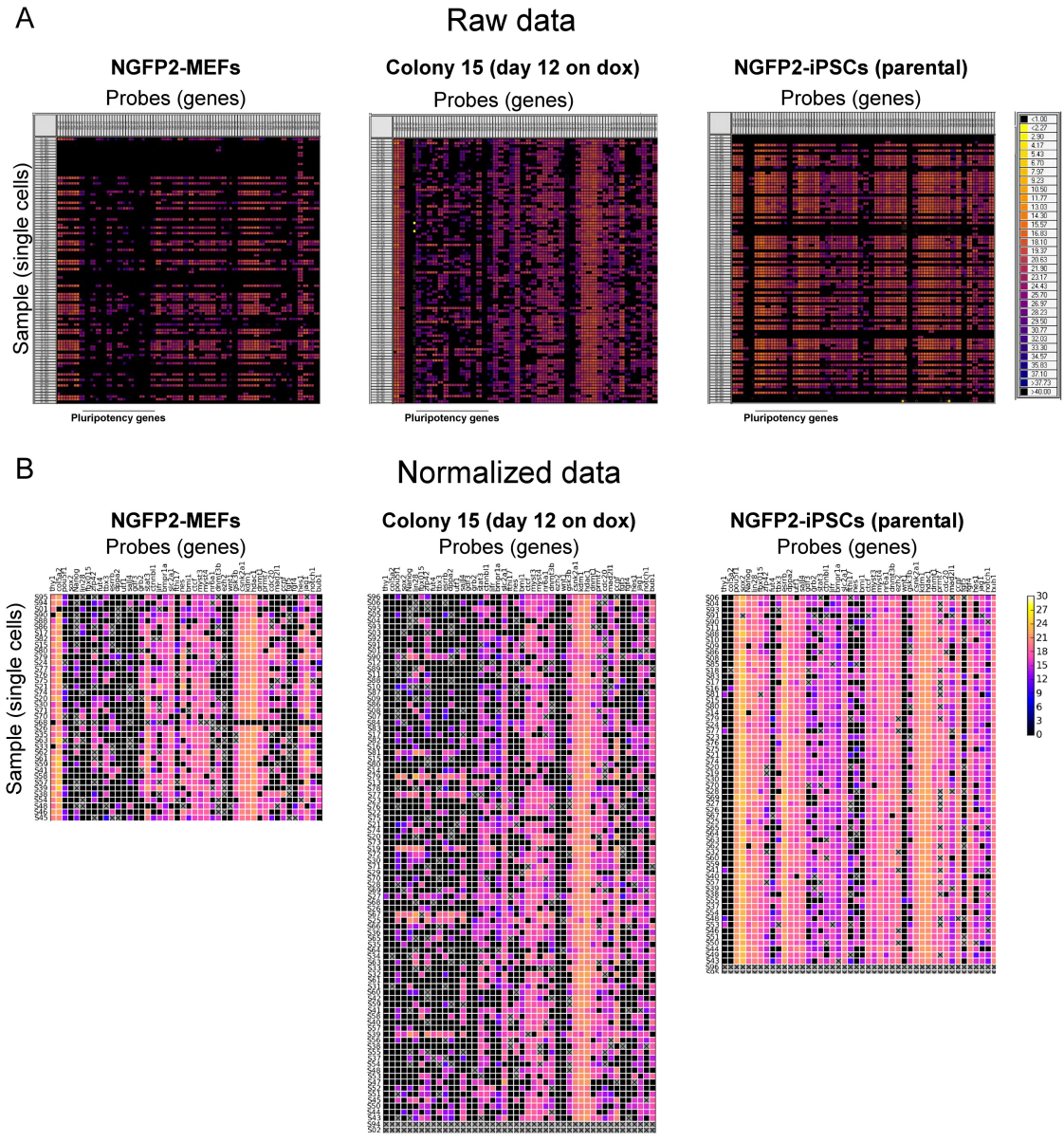
(B) Flow cytometric analysis of tdTomato and GFP in colony 20 at day 32 on dox.

(C) Bar plot of the number of tdTomato+ and GFP+ colonies derived from twenty cells of each fraction boxed in (A) after 5 days of dox exposure.

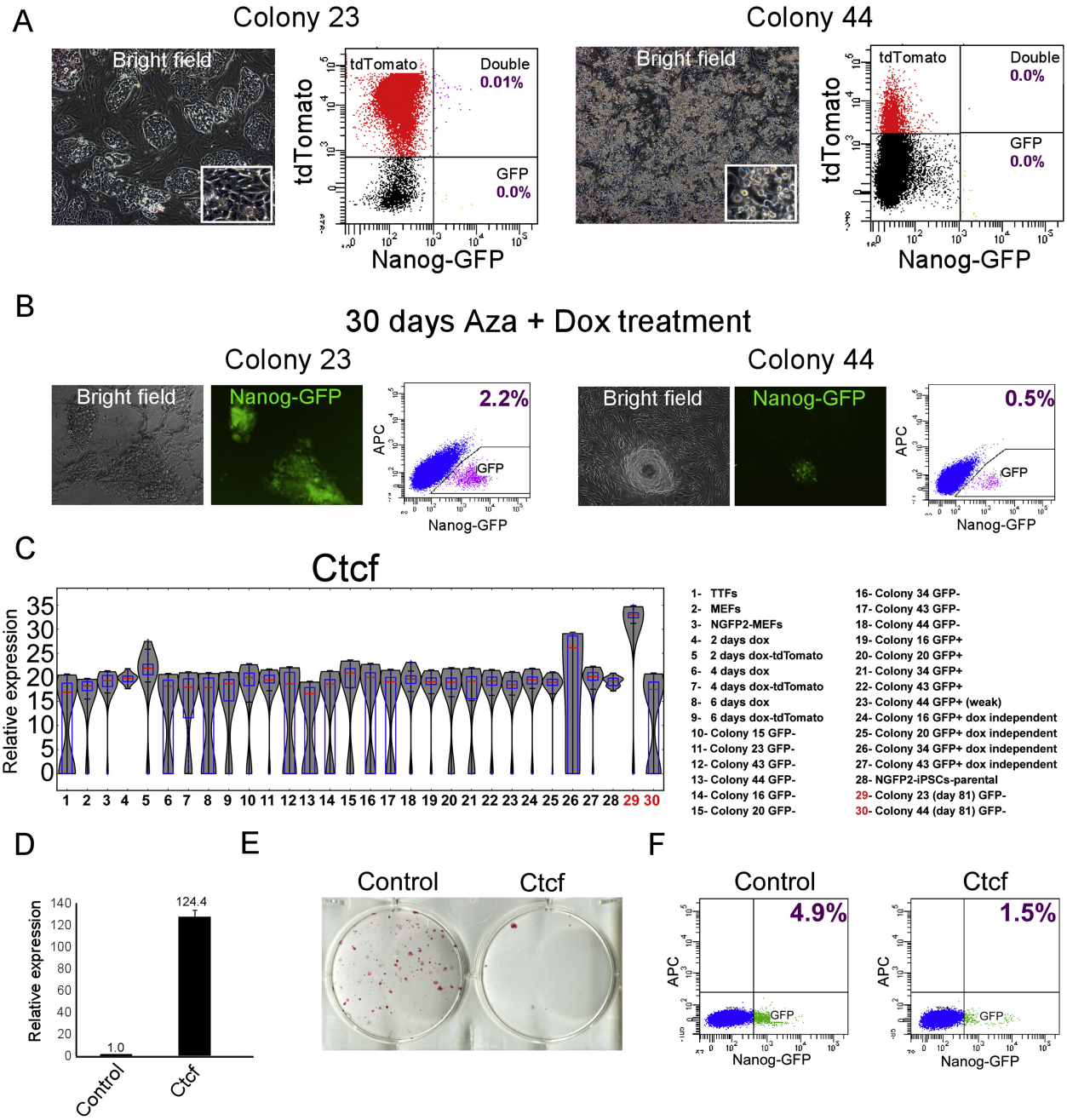
(D) Flow cytometric analysis and representative images of bright field, GFP, and tdTomato cells derived from tdTomato+/GFP- cells after one passage.

(E) Representative bright field, GFP, and tdTomato images of dox-independent colonies derived from intermediate tdTomato+/GFP- cells.

(F) Quantitative RT-PCR of Kdm1a in tdTomato+/GFP-, tdTomato+/GFP+, tdTomato-/GFP+, and dox-independent iPSCs 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 1.



Supplemental Figure 3. Raw and normalized Fluidigm data
Representative (A) raw and (B) normalized Fluidigm data for NGFP2-MEFs, colony 15-day 12 on dox, NGFP2-iPSCs. See also Figure 2.

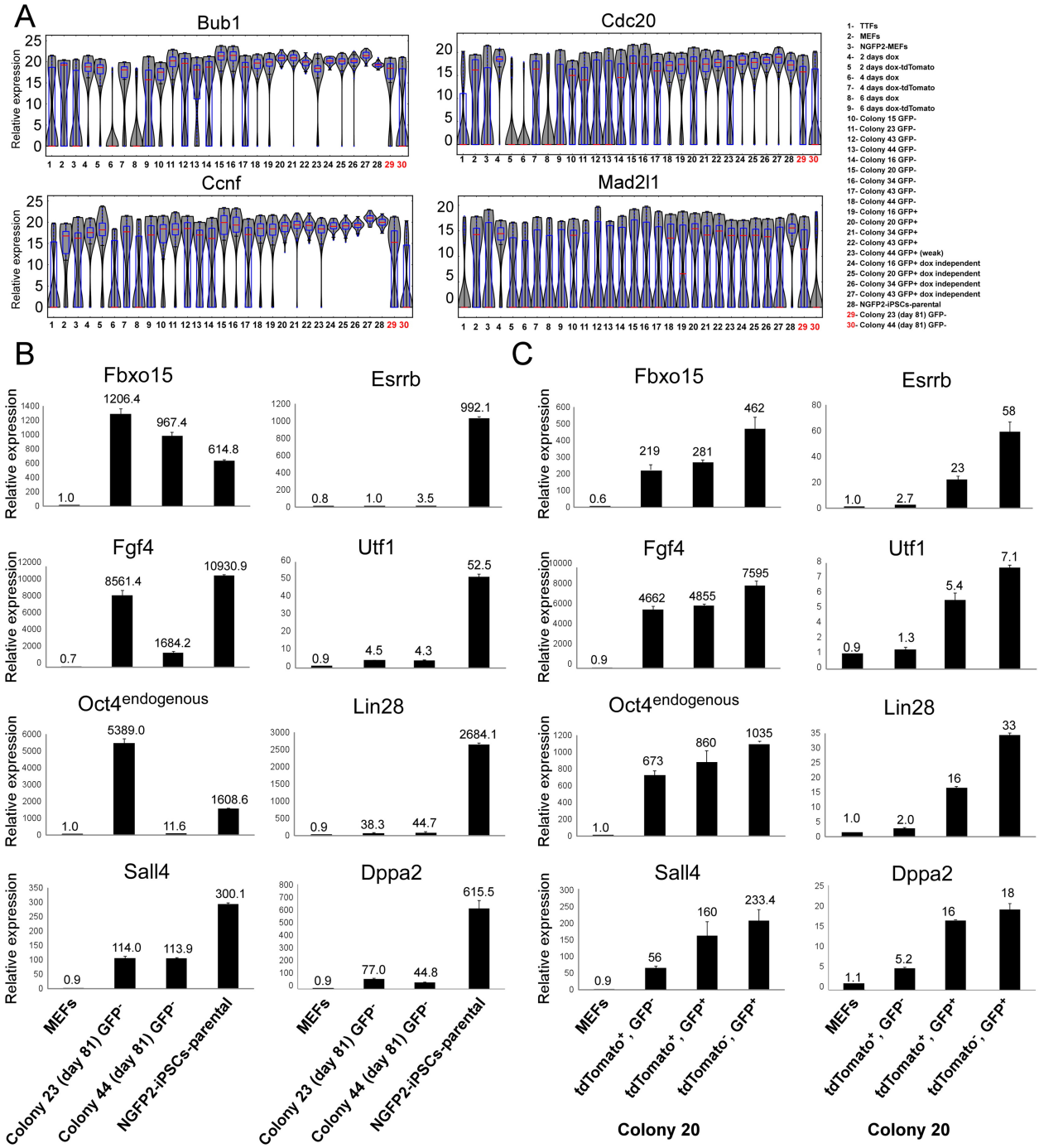


Supplemental Figure 4. Analysis of partially reprogrammed populations

(A) Representative bright field images of Colonies 23 and 44 and flow cytometric analysis of tdTomato and GFP at day 81. Colony 23 failed to activate GFP in the majority of cells upon continual passaging to day 81 (0.01% tdTomato+/GFP+). Colony 44 contained a few cells with a low level of GFP that disappeared upon continual passaging and dox-withdrawal.

(B) Representative bright field (left) and GFP (middle) images of stable dox-independent GFP+ colonies after 30 days of treatment with AZA. Flow cytometric analysis of GFP in colony 23 (2.2% GFP+) and colony 44 (0.5% GFP+) after 30 days of treatment with AZA (right).

(C) mRNA expression levels of Ctcf in populations noted in legend (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.
(D) Quantitative RT-PCR of Ctcf overexpression in NGFP2 MEFs at day 13 of dox exposure followed by 3 days of dox withdrawal. Error bars are presented as a mean $\pm$ standard deviation of two duplicate runs from a typical experiment.
(E) Alkaline phosphatase immunostaining of NGFP2 cells upon overexpression of Ctcf.
(F) Flow cytometric analysis of GFP in NGFP2 cells upon overexpression of Ctcf. See also Figure 3.

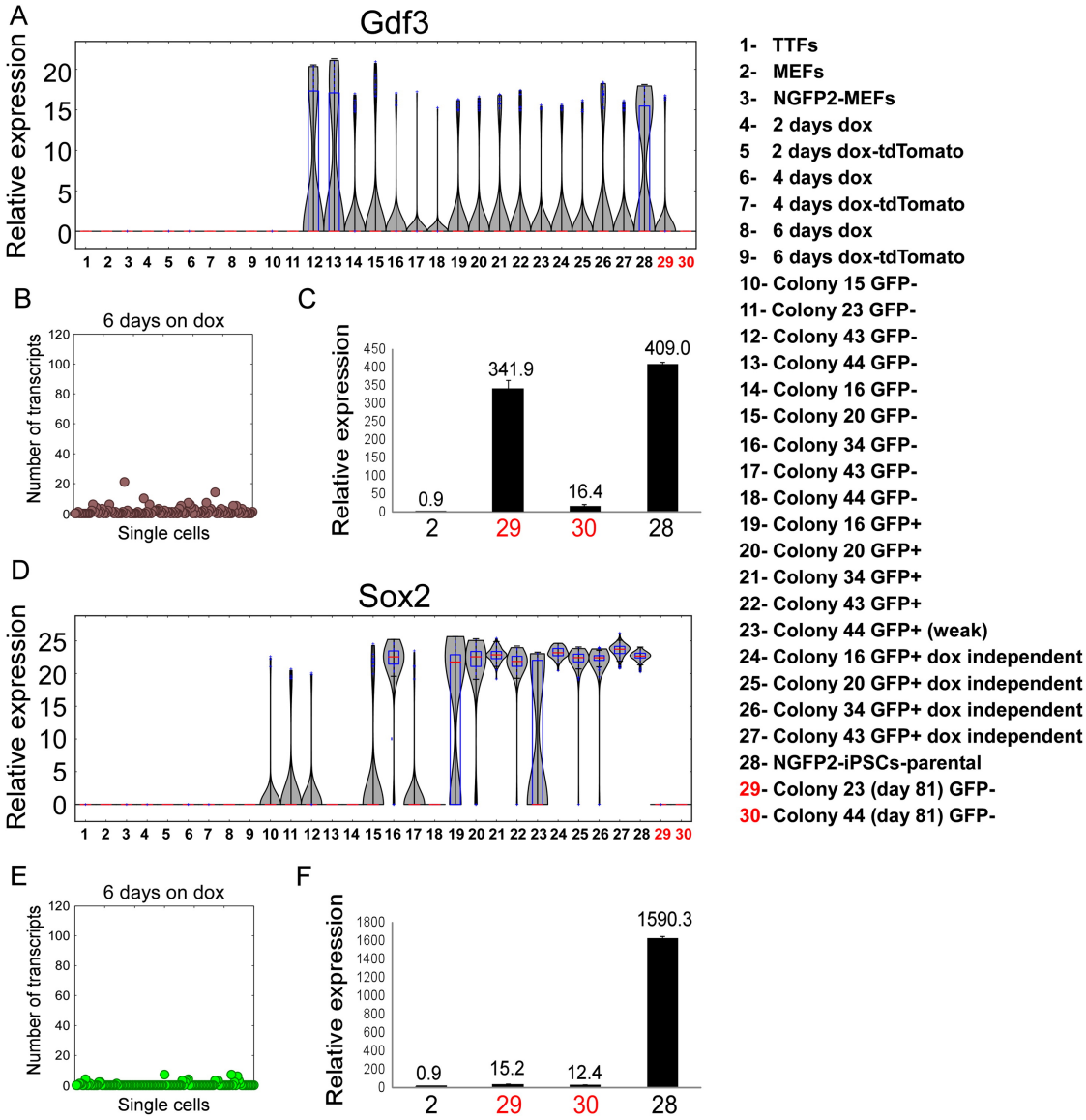


Supplemental Figure 5. Cell cycle regulators and gene expression of early candidate markers in partially reprogrammed and intermediate cell populations

(A) mRNA expression levels of Bub1, Ccnf, Cdc20, and Mad2l1 in populations noted in legend (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.

(B) Quantitative RT-PCR of Fbxo15, Fgf4, Oct4 endogenous, Sall4, Esrrb, Utf1, Lin28, and Dppa2 expression in MEFs, NGFP2 iPSCs, colony 23, and colony 44, normalized to the Hprt house keeping control gene. Error bars are presented as a mean ± standard

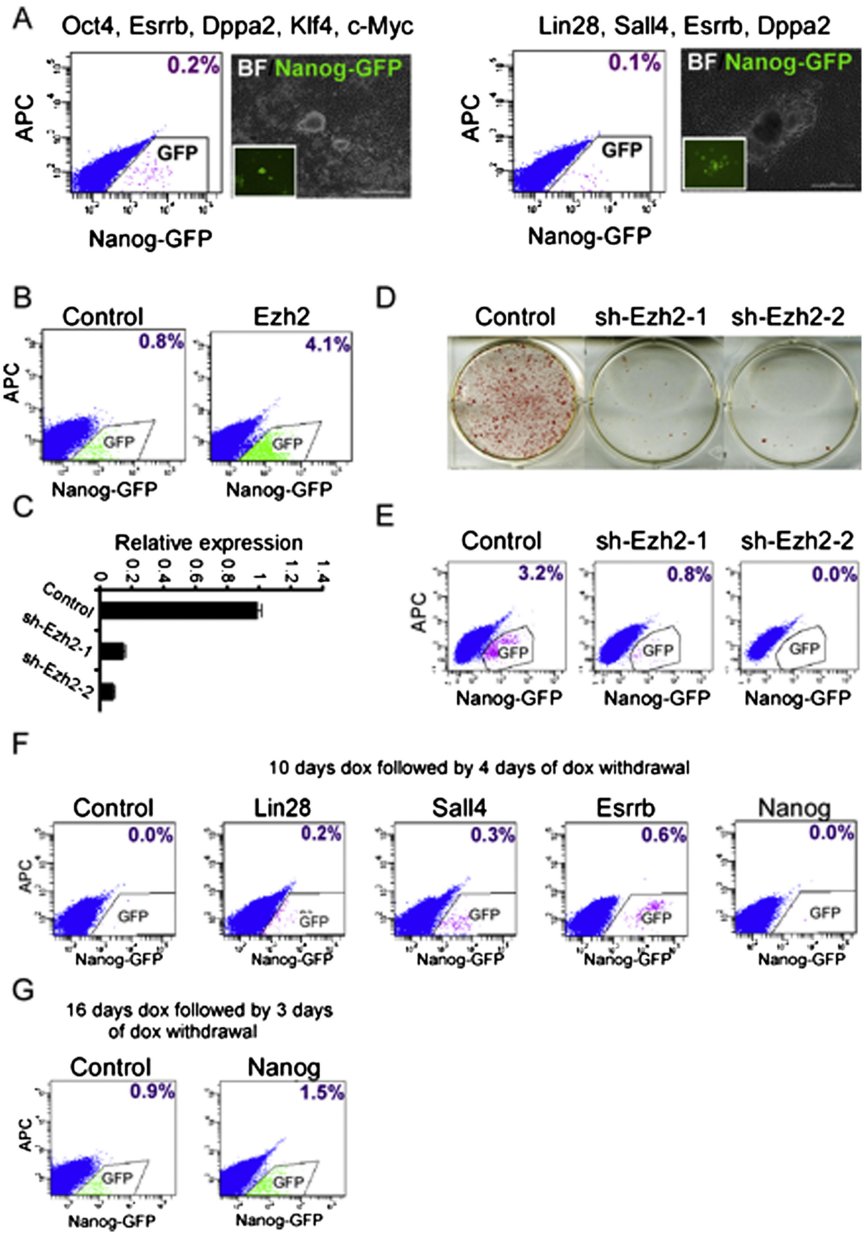
deviation of two duplicate runs from a typical experiment. Samples are numbers according to legend in Figure 3. (C) Quantitative RT-PCR of Fbxo15, Fgf4, Oct4 endogenous, Sall4, Esrrb, Utf1, Lin28, and Dppa2 expression in the fractions boxed in Supplemental Figure 1F (tdTomato-positive/GFP-negative, tdTomato-positive/GFP-positive, tdTomato-negative/GFP-positive), 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. Samples are numbers according to legend in Figure 3. See also Figure 3.



Supplemental Figure 6. Late candidate markers

(A and D) mRNA expression levels of Gdf3 and Sox2 in populations noted in legend (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. (C and F) Quantitative RT-PCR of Gdf3 and Sox2 expression in MEFs, NGFP2 iPSCs, colony 23, and colony 44, 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. Samples are numbers according to legend in Figure 3 and to the (right).

(B and E) single molecule mRNA FISH of Gdf3 (brown) and Sox2 (green) expression in NGFP2 cells at day 6 on dox. Each cell is represented as a single dot. In total, 120 cells are shown for each plot. See also Figure 5.



Supplemental Figure 7. Analysis of Ezh2 and individual factor contributions

Flow cytometric analysis of GFP in Nanog-GFP cells reprogrammed with (A) Oct4, Esrrb, Dppa2, Klf4, and c-Myc (left) and Lin28, Sall4, Esrrb, Dppa2 (right), 16 days on dox, 5 days without dox. Representative images of stable dox-independent GFP+ colonies and bright-field pictures derived from these iPSCs are shown.

(B) Flow cytometric analysis of GFP upon overexpression of Ezh2 and dox exposure for 7 days followed by 3 days of dox withdrawal.

(C) Quantitative RT-PCR of Ezh2 expression in NGFP2 cells, three days post shRNA knockdown. Two hairpins were used and expression levels were normalized for Hprt. Error bars are presented as a mean ± standard deviation of two duplicate runs from a typical experiment. (D) Alkaline phosphatase immunostaining of NGFP2 cells after 16 days of shRNA knockdown and dox addition.

(E) Flow cytometric analysis of GFP in NGFP2 cells at day 16 upon shRNA knockdown and dox addition. GFP+ cells are gated.
(F) Flow cytometric analysis of GFP upon overexpression of Lin28, Sall4, Esrrb, and Nanog individually in NGFP2 MEFs on dox for 10 days followed by 4 days dox wilthdrawl.
(G) Flow cytometric analysis of GFP upon overexpression of Nanog individually in NGFP2 MEFs on dox for 16 days followed by 3 days dox withdrawal. See also Figure 6.

Table S1. Quantitative real-time PCR primers

	Forward	Reverse
Fbxo15	CGAGAATGGTGGACTAGCTTTTG	GGCCATGGGAATGAATATTTG
Fgf4	GCAGACACGAGGGACAGTCT	ACTCCGAAGATGCTCACCAC
Oct4^{endogenous}	TCAGTGATGCTGTTGATCAGG	GCTATCTACTGTGTGTCCCAGTC
Sall4	GCAAGTCACCAGGGCTCTT	CCTCCTTAGCTGACAGCAATC
Gdf3	GGACCTGGGTTGGCACAAG	TTTGCCATGAACCCCTTAGG
Esrrb	CACCTGCTAAAAAGCCATTGACT	CAACCCCTAGTAGATTCGAGACGAT
Utf1	GTCCCTCTCCGCGTTAGC	GGCAGGTTCGTCATTTTCC
Sox2^{endogenous}	CCGTTTTCGTGGTCTTGTTT	TCAACCTGCATGGACATTTT
Lin28	GAAGAACATGCAGAAGCGAAGA	CCGCAGTTGTAGCACCTGTCT
Nanog	AAACCAGTGGTTGAAGACTAGCAA	GGTGCTGAGCCCTTCTGAATC
Ezh2	GAGGGCTATCCAGACTGGTG	TTCGATGCCCACATACTTCA
Hprt	GCAGTACAGCCCCAAAATGG	GGTCCTTTTCACCAGCAAGCT

Table S2. Primers using for cloning of cDNA for lentiviral vectors

	Forward	Reverse
Sall4-cDNA	GCAAGTCACCAGGGCTCTT	CCTCCTTAGCTGACAGCAAT
Esrrb-cDNA	GCTGGAACACCTGAGGGTAA	GGTCTCCACTTGGATCGTGT
Utf1-cDNA	CTACCTGGCTCAGGGATGCT	GACTGGGAGTCGTTTCTGGA
Lin28-cDNA	HANNA ET AL. 2009 NATURE	HANNA ET AL. 2009 NATURE
Nanog-cDNA	CGCCATCACACTGACATGA	TGGAAGAAGGAAGGAACCTG
Dppa2-cDNA	AAAGAAGTCGGCATTCATTCA	ATTCTTCCATTCCCTTTAGATCA
Ezh2-cDNA	GAAGAATAATCATGGGCCAGAC	TGCCCACAGTACTCAAGGTTC

Table S3. 48 Inventoried TaqMan assays (obtained from Applied Biosystems)

1	Mm00650983_g1	Inventoried	Cdc20
2	Mm00660135_m1	Inventoried	Bub1
3	Mm00432385_m1	Inventoried	Ccnf
4	Mm00786984_s1	Inventoried	Mad2l1
5	Mm02391771_g1	Inventoried	Hdac1
6	Mm00484020_m1	Inventoried	Ctcf
7	Mm01211941_m1	Inventoried	Myst3
8	Mm03053249_g1	Inventoried	Myst4
9	Mm03053308_g1	Inventoried	Bmi1
10	Mm01181033_m1	Inventoried	Kdm1
11	Mm00599763_m1	Inventoried	Dnnmt1
12	Mm03053759_s1	Inventoried	Nr6a1
13	Mm03053810_s1	Inventoried	Sox2
14	Mm03053707_s1	Inventoried	Bmpr1a
15	Mm03053495_s1	Inventoried	Dnmt3b
16	Mm00487448_s1	Inventoried	Fut4 (SSEA-1)
17	Mm00442942_m1	Inventoried	Lifr
18	Mm00473214_s1	Inventoried	Lin28
19	Mm02384862_g1	Inventoried	Nanog
20	Mm03053917_g1	Inventoried	Pou5f1 (OCT4)
21	Mm01265526_m1	Inventoried	Fbxo15
22	Mm01192270_m1	Inventoried	Slc2a1
23	Mm00447703_g1	Inventoried	Utf1
24	Mm03053975_g1	Inventoried	Zfp42 (REX-1)
25	Mm03053853_s1	Inventoried	Esrrb
26	Mm03053490_s1	Inventoried	Stat3
27	Mm03023989_g1	Inventoried	Grb2
28	Mm00809779_s1	Inventoried	Tbx3
29	Mm01343391_gH	Inventoried	Dppa2
30	Mm01615680_sH	Inventoried	Fthl17
31	Mm00453037_s1	Inventoried	Sall4
32	Mm03023988_m1	Inventoried	Gdf3
33	Mm00499427_m1	Inventoried	Ctnnbl1(β-catenin)
34	Mm01243796_g1	Inventoried	Csnk2a1
35	Mm03053261_s1	Inventoried	Gsk3b
36	Mm00810320_s1	Inventoried	Wnt1
37	Mm01342805_m1	Inventoried	Hes1
38	Mm03053874_s1	Inventoried	Jag1
39	Mm03053614_s1	Inventoried	Notch1
40	Mm01159248_m1	Inventoried	Ezh2

41	Mm03053244_s1	Inventoried	Nes
42	Mm03053741_s1	Inventoried	Fgf4
43	Mm03053745_s1	Inventoried	Fgf5
44	Mm00493681_m1	Inventoried	Thy1
45	Mm00483675_m1	Inventoried	Col5a2
46	Mm00446968_m1	Inventoried	Hprt
47	Mm03302249_g1	Inventoried	Gapdh
48	Mm01250624_m1	Inventoried	Prmt7

Table S4. shRNA lentiviral vector set (obtained from open biosystems):

Ezh2	RMM4534-NM_007971

Chapter 3. Single-cell analysis reveals that expression of Nanog is biallelic and equally variable as that of other pluripotency factors in mouse embryonic stem cells

Dina A. Faddah[1,2], Haoyi Wang[1], Albert Wu Cheng[1,3], Yarden Katz[1,4], Yosef Buganim[1], Rudolf Jaenisch[1,2]

1. Whitehead Institute for Biomedical Research, Cambridge, MA 02142
2. Department of Biology
3. Department of Computational and Systems Biology
4. Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology, Cambridge, MA 02139

Published as:
Faddah DA, Wang H, Cheng AW, Katz Y, Buganim Y, Jaenisch R. Single-Cell Analysis Reveals that Expression of Nanog is Biallelic and Equally Variable as that of Other Pluripotency Factors in Mouse ESCs. *Cell Stem Cell*. 13 (1): 23-9. 2013.

DAF performed all experiments and analyses, except HW designed and cloned the targeting vectors, AWC analyzed data in Figure 2F, 2G, 2H, and YB performed experiment in Supplemental Figure 1K and 1L. YK contributed to experimental design for Supplemental Figure 2H. DAF conceived and designed all experiments. DAF and RJ wrote the paper.

The homeodomain transcription factor Nanog is a central part of the core pluripotency transcriptional network and plays a critical role in embryonic stem (ES) cell self-renewal. Several reports have suggested that Nanog expression is allelically regulated and that transient down-regulation of Nanog in a subset of pluripotent cells predisposes them towards differentiation. Using single-cell gene expression analyses combined with different reporters for the two alleles of Nanog, we show that Nanog is biallelelically expressed in ES cells independently of culture condition. We also show that the overall variation in endogenous Nanog expression in ES cells is very similar to that of several other pluripotency markers. Our analysis suggests that reporter-based studies of gene expression in pluripotent cells can be significantly influenced by the gene targeting strategy and genetic background employed.

Embryonic stem cells, derived from the inner cell mass of the embryo, have the ability to divide indefinitely while maintaining the capacity to differentiate into different cell types with core transcription factors being known to regulate the pluripotent state (Jaenisch and Young, 2008; Orkin et al., 2008). Nanog is important for this network but the mechanisms governing Nanog regulation are unclear (Chambers et al., 2003; Mitsui et al., 2003).

Several studies have proposed that Nanog protein expression fluctuates in ES cells suggesting that allelic regulation of the gene itself contributes to this heterogeneity (Chambers et al., 2007; Kalmar et al., 2009; Macarthur et al., 2012; Miyanari and Torres-Padilla, 2012; Singh et al., 2007; Wray et al., 2010). These allelic fluctuations were seen in medium containing serum/leukemia inhibitory factor (LIF) and to a lesser extent, if at all, in 2i/LIF (inhibition of MAPK and GSK-3) (Silva et al., 2008; Silva et al., 2009; Wray et al., 2010; Ying et al., 2008). It has been suggested that fluctuating levels of Nanog mediate ES cell self-renewal vs. differentiation with low or no Nanog expression thought to render cells susceptible to intrinsic or extrinsic signals inducing differentiation and generating functional heterogeneity within pluripotent cell populations. Recently, it has been shown that Nanog activity is autorepressive and may regulate allelelic switching (Fidalgo et al., 2012; Navarro et al., 2012). Surprisingly, Nanog can be deleted in ES cells without affecting their potential to generate chimeras (Chambers et al., 2007).

In this study, we investigated variation in Nanog expression using single-cell analysis in mouse ES cells. To monitor the two alleles of Nanog in single cells using

single-molecule-mRNA-FISH (sm-mRNA-FISH) (Buganim et al., 2012; Raj et al., 2008), we generated a V6.5 ES cell line where GFP was inserted immediately downstream of the Nanog coding region with the selectable marker being deleted. Sequences encoding mCherry were inserted by a similar targeting strategy into the second Nanog allele (Figure 1A, S1A). In this construct GFP and mCherry dissociate from Nanog by self-cleavage of a 2A peptide and do not alter Nanog function. We quantified transcripts of Nanog, mCherry, and GFP in single Nanog-2A-GFP/Nanog-2A-mCherry ES cells (cells termed NGNC here) by sm-mRNA-FISH and found that all cells expressed mCherry and GFP transcripts (Figure 1B) with the total level of Nanog transcripts in a given cell being approximately equal to the sum of the GFP and mCherry transcripts (Figure 1C). Boxplot analysis revealed GFP expression and mCherry expression to be equal and approximately half that of Nanog expression (Figure 1D). We quantified mCherry+/GFP+, GFP+, and mCherry+ cells grown in serum/LIF by flow cytometric analysis and found 96% mCherry+/GFP+, 0.6% GFP+, and 0.1% mCherry+ (Figure 1E). Finally, all NGNC cells grown in serum/LIF or 2i/LIF were GFP+ and mCherry+ by immunostaining (Figure S1B). In summary, our results indicate that both Nanog alleles are expressed in the great majority of cells regardless of culture condition.

To compare the variability of Nanog expression to that of other pluripotency factors, we used sm-mRNA-FISH to quantify transcripts of 9 pluripotency genes (Nanog, Dnmt3b, Utf1, Sox2, Lin28, Sall4, Tet1, Klf2, Fbx15), 1 housekeeping gene (Gapdh), and a known heterogeneously expressed gene (Stella) each in combination with Oct4 in single cells (Figure 1F-1O, S1C-D). Out of 899 cells analyzed, we only identified 1% that were Nanog-/Oct4+ (Figure S1C). Klf2 and Fbx15 were not always co-expressed with Oct4 with 10% of Klf2-/Oct4+ cells and 14% Fbx15-/Oct4+ cells (Figure 1N-S1D). Figure 1O shows 40% Stella-/Oct4+ negative cells, a number slightly lower than the 70-80% Stella negative cells identified by immunofluorescence in a previous report (Hayashi et al., 2008). All genes examined had different levels of expression and ranges of expression levels in single cells (Figure 1P). Importantly, Stella had the highest coefficient of variation, while all other genes, including Nanog and Gapdh, had similar coefficients of variation. These data suggest that Nanog is just as

variable in gene expression as any other pluripotency factor and even a housekeeping gene, like Gapdh (Figure 1Q). Thus, our data, based upon single cell expression studies, do not support the concept that Nanog is more heterogeneously expressed than most other pluripotency genes.

Our conclusions about Nanog expression differ from those seen in prior studies, so we investigated potential explanations. The majority of studies characterizing heterogeneity in Nanog expression have used heterozygous loss-of-function knock-in GFP reporters. Specifically, in the Nanog GFP +/- allele generated by (Hatano et al., 2005), the coding sequences were replaced with a GFP-IRES-puro-pA reporter and a selection cassette in the targeted allele (we designate these cells as "NHET" ES cells), whereas the TNGA allele was generated by inserting the eGFP marker at the Nanog AUG codon (Chambers et al., 2007). In a third study a triplicate GFP sequence had been inserted into one and a corresponding mCherry construct into the other Nanog allele resulting in "NGR" ES cells. The GFP and mCherry allele also contained an IRES-Neo or IRES-Hygro selection cassette, respectively (Miyanari and Torres-Padilla, 2012). Both fluorescent proteins dissociate from Nanog by self-cleavage of a 2A peptide and thus were not expected to interfere with Nanog function. Using time-lapse analysis, dynamic fluctuations of Nanog expression were observed in agreement with previous reports (Chambers et al., 2007; Kalmar et al., 2009). In addition, RNA-FISH and allele-specific single cell-RT-PCR found that about 80% of the cells expressed Nanog monoallelically, a fraction that decreased to about 30% when the cells were cultured in 2i/LIF condition.

In an effort to reconcile our data with the published Nanog expression patterns we used sm-mRNA-FISH to measure Nanog, Oct4, and GFP expression in V6.5 ES cells targeted with an identical vector as previously described (NHET ES cells (Hatano et al. 2005)) or by using the published targeted E14Tg2a ES cells (TNGA: (Chambers et al., 2007)) (Figure S1E).

To assess the influence of culture condition we compared gene expression in ES cells that were grown under three different conditions: (i) on feeders in serum and LIF; (ii) on feeders in 2i and LIF but no serum; (iii) on gelatin (no feeders) in 2i and LIF and

no serum. Using sm-mRNA-FISH we found that the culture conditions (ii) and (iii) did not significantly affect the level of Nanog (between 140 and 145 transcripts), of Oct4 (between 190 and 205 transcripts) and of GFP (between 175 and 180 transcripts). In the following experiments we used only cells grown on feeders that were either cultured in serum and LIF or in 2i and LIF and no serum.

Confirming the published data, this analysis revealed that the majority of NHET ES cells cultured in serum/LIF or 2I/LIF were GFP- (79% and 69%, respectively) (Figure 2A). However, the great majority of the GFP- cells grown in 2i/LIF (98%) and 100% of GFP- cells in serum/LIF expressed Nanog RNA. Similarly, most TNGA GFP- ES cells cultured in serum/LIF condition were Nanog+ (Figure 2B). These data, summarized in Figure 2C, indicate that GFP+ and GFP- NHET and TNGA ES cells expressed Nanog and Oct4 mRNA at comparable levels. Cultivation of NHET cells in 2i/LIF substantially increased the number of Nanog transcripts in NHET but not in TNGA cells. Quantification of GFP+ and GFP- fractions in both cells lines cultured in serum/LIF by flow cytometry was consistent with the sm-mRNA-FISH analysis (Figure 2D). Immunostaining of each cell line revealed that both the GFP+ and GFP- cells expressed Nanog and Oct4 protein (Figure 2E and S1F). In both TNGA and NHET cell lines we found GFP-, GFP+, and 'speckled' colonies containing both GFP+ and GFP- cells (Figure 2E and S1F). We also found that GFP- cells can give rise to GFP+ cells and GFP+ can generate GFP- cells within 1 or 2 passages (Figure S1G), consistent with previous reports (Chambers et al., 2007).

To monitor the non-targeted allele of NHET ES cells, mCherry was inserted immediately downstream of the Nanog coding region (using Nanog-2A-mCherry construct). We found the NHET GFP- cells to be mCherry+, further supporting that the other allele of Nanog is active in the GFP- cells (Figure S1H). Importantly, western blotting was performed on protein derived from the GFP+ and GFP- fractions of NHET and TNGA and confirmed that GFP expression did not reflect Nanog protein expression (Figure S1I-J), a result different from published data (Chambers et al., 2007). In summary, these observations demonstrate (a) that only a fraction of NHET and TNGA cells express GFP in agreement with previous reports (Chambers et al.,

2007), (b) that the NHET and TNGA GFP- cells also express Nanog, (c) that 2i/LIF affects Nanog, Oct4, and GFP expression differently in TNGA and NHET ES cells and (d) that the GFP reporter targeting strategies that disrupt one allele may not be a faithful indicator of endogenous Nanog expression.

To compare the GFP+ and GFP- cells in terms of their pluripotent state, we analyzed the transcriptional profiles of NHET GFP+ and GFP- cells by single-cell gene expression quantitative RT-PCR using Fluidigm Biomark (Buganim et al., 2012) (Figure 2F-H). The genes tested in this analysis included ES cell-associated chromatin remodeling genes and modification enzymes, ES cell cell-cycle regulator genes, pluripotency markers, MEF markers, and genes active in signal transduction pathways important for ES cell maintenance and differentiation (see list of genes in legend). Expression of all of the genes analyzed showed similar distributions of expression levels in single GFP+ and GFP- cells, supporting the notion that GFP+ and GFP- ES cells have a very similar expression profile (Figure 2F). In agreement with this conclusion, hierarchical clustering (Figure 2G) and principal component analysis (Figure 2H) did not separate the GFP positive and negative cells. Only 3% of GFP- cells were separated from the majority of cells, and these likely represent differentiating cells as they differed in cell cycle regulators and some pluripotency markers. We conclude that the GFP+ and GFP- cells have very similar gene expression profiles, suggesting that they are equivalent in terms of their pluripotency status.

To test whether haploinsufficiency of Nanog was responsible for the large proportion of GFP-/Nanog+ cells in NHET and TNGA ES cells we overexpressed Nanog (Figure S1K-L). NHET and TNGA ES cells were infected with M2rtTA and tetO-Nanog-2A-Blue Fluorescent Protein (BFP). Dox was added to the cells and high BFP+/GFP+ cells were sorted onto feeder MEFs. Equal numbers of cells from single BFP+/GFP+ colonies were plated in the presence and absence of dox and analyzed for GFP and BFP. In three lines from both TNGA and NHET backgrounds none exhibited an increase in GFP+ cells upon Nanog overexpression (Figure S1K-L). The presence of GFP-/BFP+ cells and the observation that over-expression of Nanog did not increase

the fraction of GFP+ cells (Figure S1K-L) is consistent with previous reports (Fidalgo et al., 2012; Navarro et al., 2012).

It seemed possible that the different Nanog expression patterns in NGNC cells vs. TNGA and NHET cells were a result of the gene targeting strategy used, which in the latter two cell lines resulted in a Nanog null allele and may have disturbed normal Nanog regulation. To directly test if gene targeting of Nanog was responsible for "GFP fluctuations" of Nanog expression, we targeted V6.5 (C57Bl/6 x 129) cells, the background of NHET, and E14Tg2a (129/Ola) cells, the background of TNGA, with our Nanog-2A-GFP vector (Figure S2A). We found that all V6.5 and E14Tg2a Nanog-2A-GFP cells expressed GFP and Nanog by sm-mRNA FISH, immunostaining and flow cytometry and that GFP expression faithfully reflected Nanog expression with GFP expression (48 transcripts/cell) approximately half of Nanog expression (112 transcripts/cell) in single cells (Figures 2I-L, S2B, S2C). To assay for pluripotency of TNGA and NHET GFP+ and GFP- cells and our V6.5 + Nanog-2A-GFP cells, we sorted 150 of the lowest GFP- cells and 150 of the highest GFP+ cells from TNGA and NHET and counted the number of undifferentiated colonies at one week after plating. We also sorted 150 of the lowest GFP+ cells and 150 of the highest GFP+ cells from our V6.5 + Nanog-2A-GFP line. The low GFP+ cells are prone to differentiation generating only 16 undifferentiated colonies as compared to 44 from the high GFP+ cells. TNGA and NHET GFP+ and GFP- cells gave rise to approximately the same number of undifferentiated colonies, further supporting that the cells are in equivalent states of pluripotency (Figure S2D). V6.5 + Nanog-2A-GFP ES cells were induced to differentiate by treatment with retinoic acid for 48 hours and, as expected, all GFP was lost (Figure S2E). Similarly to NHET and TNGA, a Nanog-GFP human ES cell reporter line generated by inserting GFP into the 5' untranslated region of the Nanog gene upstream of the Nanog start codon (ATG) yielded many GFP-, ES cell-like cells, suggesting similar regulation of Nanog expression in humans (Fischer et al., 2010) (Figure S2F).

The targeting strategy for NGR cells (Miyanari and Torres-Padilla, 2012) did not disrupt the coding sequences of the Nanog alleles but nevertheless showed monoallelic

expression in a significant fraction of the cells. We considered two possibilities to explain the difference between these results and ours. First, the targeting of the Nanog alleles in NGR cells involved the insertion of a ~4kb transgene containing a selectable marker in addition to three repeats of the GFP or mCherry coding sequences into the 3' UTR, resulting in a ~4kb insert compared to our construct that comprised only ~700bp with the selection cassette removed. It is possible that the larger insert disrupted Nanog regulation of the NGR alleles. We tested whether deletion of the selectable marker affected expression of the inserted transgene and, using sm-mRNA-FISH to measure Nanog expression, found that deletion of the selectable marker reduced the proportion of GFP negative cells from ~20% to 0, suggesting that the size of the genetic construct used may influence the results for this type of reporter (compare Figure 2J with S2G). We also noticed that Miyanari et al. used C57BL/6 x cas (BC1) ES cells and C57BL/6 (BD10) ES cells, while we used C57BL/6 x 129 (V6.5) ES cells. To examine if genetic background could affect Nanog and Oct4 expression heterogeneity, we measured Nanog and Oct4 expression in single ES cells from different genetic backgrounds cultured in serum/LIF and 2i/LIF using sm-mRNA-FISH (Figure S2H and 2C (contains V6.5 and E14Tg2a data)). Out of 1113 single cells analyzed from the 6 ES cell lines, we only found 3 cells with no Nanog transcripts, consistent with our previous data in Figure S1C. However, we also found that V6.5 had fewer low Nanog-expressing cells (0%) as compared to V26.2 (C57BL/6) (9%) and ESC1 (C57BL/6 x cas) (13%) in serum/LIF condition (Figure S2I). Importantly, these low expressing Nanog cells were not differentiated and had high expression of Oct4 (~150 transcripts). Thus, genetic background does appear to influence the pattern of Nanog expression.

Filipczyk et al., in this issue of Cell Stem Cell, generated ES cells that carried different fluorescent reporters in both alleles of Nanog, similar to the construct described in Figure 1A (Filipczyk et al., 2013). In agreement with our results (Figure 1B-E) they observed that most cells expressed both reporters, although with greater variability in expression level that may in part be a result of their use of a larger size insert.

In summary, we have found using single-cell analysis that Nanog is biallelically expressed in mouse ES cells and that the degree of variation in expression level is very

similar to that of many other pluripotency factors. We do not see evidence of a distinct subpopulation of cells with low Nanog expression, although it is possible that such a population exists in some circumstances. Our analysis of a range of Nanog-GFP reporters suggests that disruption of one of the two alleles or insertion of a large downstream cassette may disturb normal transcriptional control and thus not give a faithful reflection of endogenous Nanog expression. More broadly, our findings also suggest that these issues are important to take into account when designing reporter constructs to monitor other factors, in the pluripotency network and beyond.

Acknowledgements

We thank Austin Smith for sharing TNGA cells, Bryce Carey, Meelad Dawlaty, Frank Soldner, and Ameera Salama for helpful discussions, Sandy Klemm for discussions and Stella FISH probe, Thor Theunissen for help with sorting, and Alexander van Oudenaarden for facilities. 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. Y.B. was supported by a NIH Kirschstein NRSA (1 F32 GM099153-01A1). R.J. is an adviser to Stemgent and a cofounder of Fate Therapeutics. This work was supported by NIH grants HD 045022 and R37CA084198 to RJ.

Methods

Generation and culture of Nanog-2A-GFP/2A-mCherry (NGNC) and E14Tg2a + Nanog-2A-GFP ES cells

To generate Nanog-2A-GFP and Nanog-2A-mCherry alleles, 2kb region upstream of Nanog stop codon was amplified from V6.5 ESC DNA and cloned into multiple cloning site 1 (SbfI and NheI) of OCT4-2A-eGFP-PGK-Puro vector (Hockemeyer et al., 2011). This was followed by cloning 3 kb region downstream of Nanog stop codon into multiple cloning site 2 (AscI and FseI) of this vector. Vector was linearized with NcoI and electroporated into V6.5 ES cells [mix background: 129/sv(M) x C57/BL6(F)] or E14Tg2a ES cells (129/Ola) following standard procedures. Puro resistant clones were

picked and screened with 5′ and 3′ external probes and GFP sequence internal probe. pPAC-Cre plasmid was electroporated into properly targeted clones to excise PGK-Puro cassette and generate Nanog-2A-GFP allele. These cells were retargeted using Nanog-2A-mCherry-PGK-Neo vector to generate NGNC ES cells. ES cells for serum/LIF experiments were cultured on gamma-irradiated DR4 feeders using standard ES cell media containing LIF. ES cells for 2i/LIF experiments were cultured on gamma-irradiated DR4 feeders and also on gelatin without feeders using standard 2i media containing LIF. Both conditions (2i/LIF on and off feeders) yielded the same results.

Southern blots

10-15μg of genomic DNA was digested with restriction enzymes overnight and southern blot was performed as previously explained (Carey et al., 2011).

Immunostaining

ES cells cultured on feeders were washed with PBS and fixed with 4% paraformaldehyde for 15mins at room temperature. Cells were then permeabilized and blocked in 0.2% TritonX 5% BSA in PBS for 30 min at room temperature. Primary and secondary incubations were performed as described previously (Hanna et al., 2009). Following antibodies were used: Affinity purified Nanog Rabbit polyclonal antibody (Bethyl Laboratories, 1:250) and mouse monoclonal Oct4 clone C10 antibody (Santa Cruz, 1:100), mouse monoclonal GFP antibody (Abcam, ab1218, 1:2000), rabbit polyclonal GFP antibody (Abcam, ab6556, 1:2000), and mouse monoclonal [1C51] mCherry antibody (Abcam, ab125096, 1:2000).

Flow cytometry

Cells were trypsinized, washed once in PBS and resuspended in fluorescence-activated cell sorting (FACS) buffer (PBS + 5% FBS). The percentage of GFP-positive cells (Nanog-GFP), mCherry-positive (Nanog-mCherry) was analyzed using FACSCalibur.

Retinoic acid-induced differentiation

ES cells were trypsinized and plated on gelatin in ES cell medium (+LIF). The next day, the media was replaced with ES cell media minus LIF plus 1 × 10-7 M retinoic acid. Cells were cultured for an additional 48 hours with one media exchange.

FISH and imaging

We performed FISH as outlined in ((Raj et al., 2010; Raj et al., 2008)). All hybridizations were performed in solution using probes coupled to either tetramethylrhodamine (TMR) (Invitrogen), Alexa 594 (Invitrogen) or Cy5 (GE Amersham). We used TMR for the probes against Nanog and Dnmt3b mRNA, Alexa 594 for the probes against GFP, Rex1, Stella, Lin28, Sox2, Sall4, Tet1, Utf1, Klf2, Fbx15, and Gapdh mRNA and Cy5 for the probes against Oct4 and mCherry mRNA. Optimal probe concentrations during hybridization were determined empirically. Imaging involved taking stacks of images spaced 0.4 μm apart using filters appropriate for DAPI, TMR, Alexa 594 and Cy5. All images were taken with a Nikon Ti-E inverted fluorescence microscope equipped with a 100X oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software (Molecular Devices, Downington, PA). During imaging, we minimized photobleaching through the use of an oxygen-scavenging solution using glucose oxidase.

Image analysis

We segmented the cells manually and counted the number of fluorescent spots, each of which corresponds to an individual mRNA, using a combination of a semi-automated method described in ((Itzkovitz et al., 2011; Raj et al., 2008)) and custom software written in MATLAB (Mathworks). We estimate our mRNA counts to be accurate to within 10–20%.

Single-cell Data Processing and Visualization

Processing of Fluidigm data was as described previously (Buganim et al., 2012). Briefly, cells with non-consistent duplicates in control genes were discarded. Ct values of genes

were normalized to control of the same cell into AC20 value which is in the log2 space and increases with expression level. Hierarchical clustering (center gene by median, gene-normalized, Pearson correlation average-linkage) was done with BioPython and visualized in Java TreeView. PCA was done with bpca missing value estimation in R. For heatmap, normalized expression values of single cells from each sample is binned and the density/fraction in each bin is represented by color intensity in the heatmap.

Single-cell gene expression qPCR

Single-cell gene expression qPCR was performed as previously described (Buganim et al., 2012). Briefly, 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 were performed. Products were analyzed, and Ct values were calculated from the system's software.

Viral preparation and infection

Construction of the Nanog-2A-EBFP lentiviral vector under the control of the tetracycline operator and a minimal CMV promoter was generated by cloning the open reading frame of the mouse Nanog gene (without STOP codon), obtained by reverse transcription with specific primers (see Buganim et al, 2012),into the TOPO-TA vector (Invitrogen), and then restricted with MfeI and inserted into the FUW-TetO-2A-EBFP expressing vector. Replication-incompetent lentiviral particles were packaged in 293T cells with a VSV-G coat and used to infect the Nanog GFP+/- and TNGA ES cells. Viral supernatants from cultures were filtered through a 0.45 mM filter and added to the cells. To initiate Nanog and EBFP expression cells were grown in ES cell medium + 2 mg/ml doxycycline (DMEM supplemented with 15% FBS (Hyclone), leukemia inhibitory factor, beta-mercaptoethanol (Sigma-Aldrich), penicillin/streptomycin, L-glutamine, and nonessential amino acids.

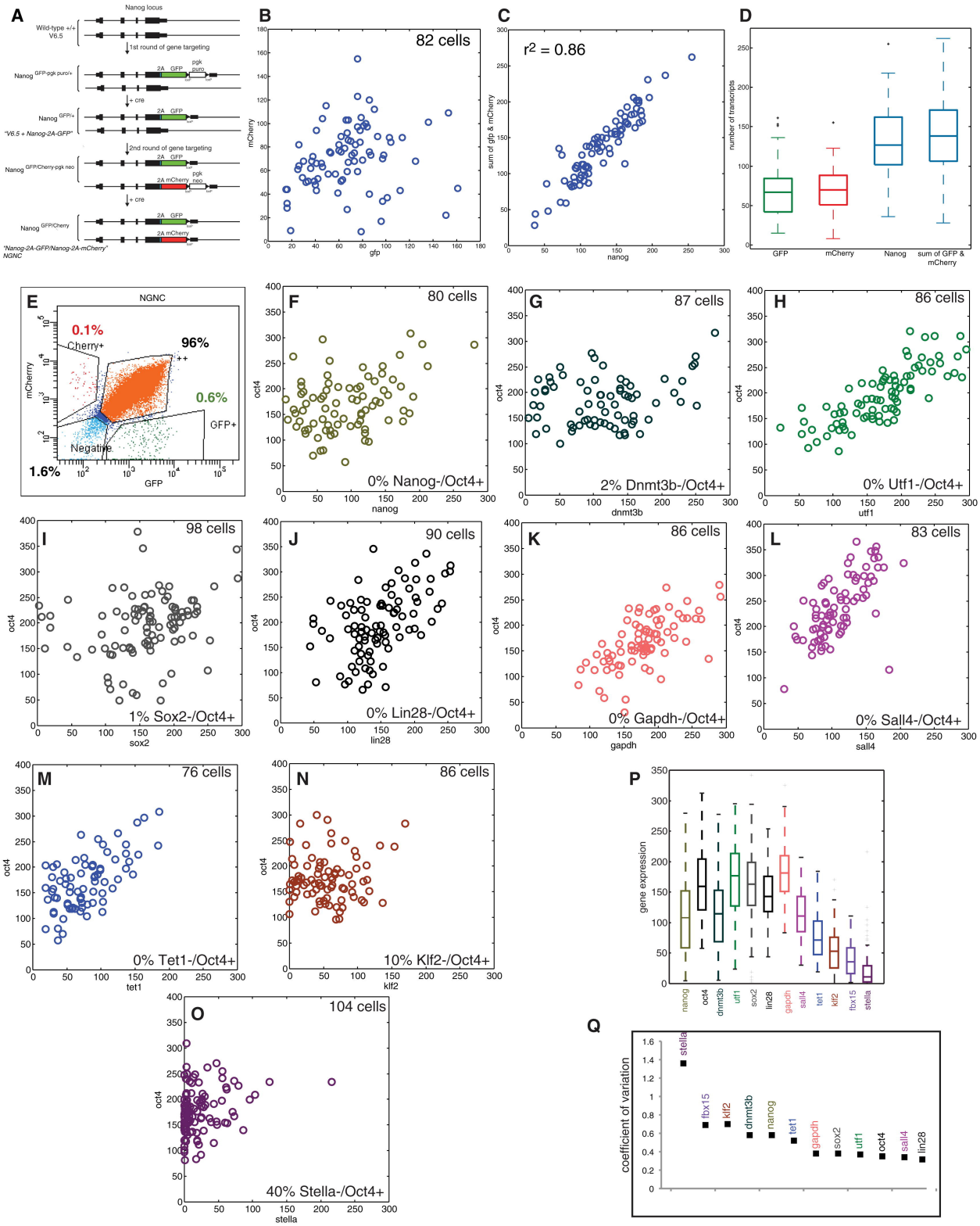


Figure 1. Nanog is biallelically expressed in ES cells and equally variable as that of other pluripotency factors

(A) Schematic of the NGNC reporter targeting. Four rounds of gene targeting were performed: (1) V6.5 ES cells targeted with Nanog-2A-GFP floxed pgk puro (2) Cre excision of the floxed pgk puro (3) Nanog-2A-GFP ES cells targeted with Nanog-2A-mCherry pgk neo (4) Cre excision of the floxed pgk neo.
(B) sm-mRNA-FISH of mCherry vs. GFP expression in single NGNC ES cells cultured with serum/LIF, 82 cells analyzed.
(C) sm-mRNA-FISH of sum of mCherry and GFP vs. Nanog expression in single NGNC ES cells cultured with serum/LIF.
(D) Box plot of GFP (green), mCherry (red), Nanog (Blue), and sum of GFP and mCherry (blue) transcripts in single cells, quantified by sm-mRNA-FISH. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme datapoints not considered to be not outliers, and the outliers (+) are plotted individually. Points are drawn as outliers if they are larger than Q3+W*(Q3-Q1) or smaller than Q1-W*(Q3-Q1).
(E) Flow cytometric analysis of NGNC ES cells in serum/LIF.
(F-O) sm-mRNA-FISH of Oct4 vs. Nanog (F), Dnmt3b (G), Utf1 (H), Sox2 (I), Lin28 (J), Gapdh (K), Sall4 (L), Tet1 (M), Klf2 (N), and Stella (O) expression in single V6.5 ES cells cultured with serum/LIF.
(P) Box plot of transcripts in single cells, quantified by sm-mRNA-FISH, of the genes in (F-O). On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme datapoints not considered to be not outliers, and the outliers (+) are plotted individually. Points are drawn as outliers if they are larger than Q3+W*(Q3-Q1) or smaller than Q1-W*(Q3-Q1).
(Q) Coefficient of variation of the genes in (F-O). See also Figure S1A and S1B.

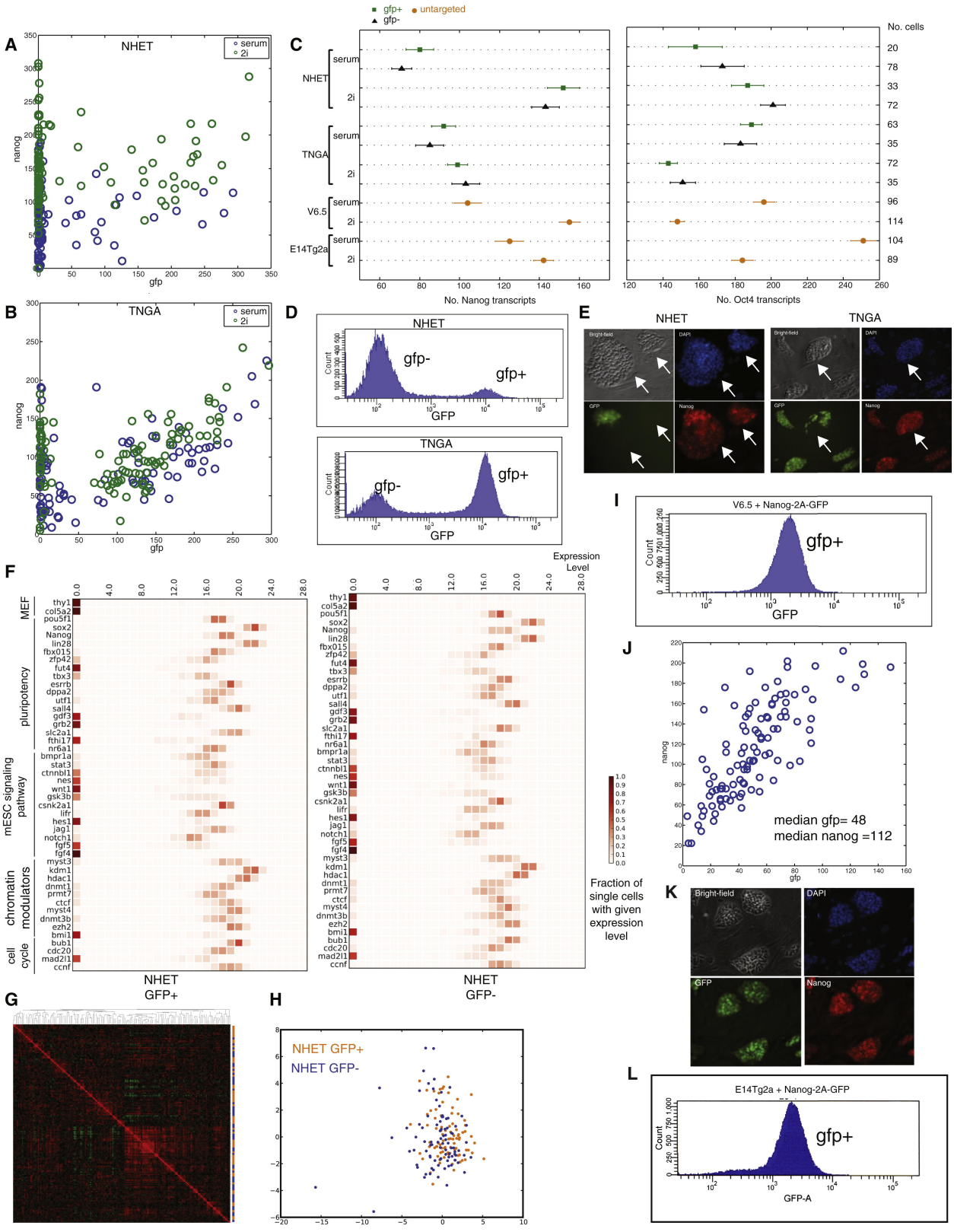


Figure 2. Nanog heterozygous loss of function knock-in reporters do not reflect Nanog expression

sm-mRNA-FISH of Nanog vs. GFP expression in single (A) NHET ES cells and (B) TNGA ES cells cultured in serum/LIF (blue) and 2i/LIF (green) condition. (NHET serum-102, NHET 2i -105, TNGA serum- 98, TNGA 2i-107 cells analyzed).
(C) Plot of the median number of Nanog (left) and Oct4 (right) transcripts, quantified by sm-mRNA-FISH, in GFP+ (square, green) and GFP- (triangle, black) fractions of NHET and TNGA ES cells and V6.5 and E14Tg2a (untargeted ES cells) cultured in serum/LIF (serum) and 2i/LIF (2i) condition. Error bars represented standard error of the mean.
(D) Flow cytometric analysis of GFP in NHET ES cells (top) and TNGA ES cells (bottom).
(E) Representative bright-field image (upper left), DAPI (upper right), and immunostaining of GFP protein (bottom left) and Nanog protein (bottom right) of NHET (left) and TNGA (right) ES cells cultured in serum/LIF. White arrows indicate GFP-/Nanog+ cells.
(F) Heatmap of gene expression values of single NHET GFP+ (left) and GFP- (right) ES cells. Fraction of single-cells with an expression level (top number) is indicated by color of the box (see key on right). The genes tested in this analysis included ES cell-associated chromatin remodeling genes and modification enzymes (Myst3, Kdm1, Hdac1, Dnmt1, Prmt7, Ctcf, Myst4, Dnmt3b, Ezh2, Bmi1), ES cell cell-cycle regulator genes (Bub1, Cdc20, Mad2l1, Ccnf), pluripotency markers (Oct4, Sox2, Nanog, Lin28, Fbxo15, Zfp42, Fut4, Tbx3, Esrrb, Dppa2, Utf1, Sall4, Gdf3, Grb2, Slc2a1, Fthi17, Nr6a1), MEF markers (Thy1 and Col5a2), and genes active in signal transduction pathways important for ES cell maintenance and differentiation (Bmpr1a, Stat3, Ctnnbl1, Nes, Wnt1, Gsk3b, Csnk2a1, Lifr, Hes1, Jag1, Notch1, Fgf5, Fgf4). (G) Hierarchical clustering of single NHET GFP+ and GFP- ES cells. Bar on right displays GFP+ (orange dot) and GFP- cells (blue dot).
(H) Principal component (PC) projections of single NHET GFP+ (orange) and GFP- (blue) ES cells, colored by their sample identification.
(I) Flow cytometric analysis of GFP in V6.5 + Nanog-2A-GFP ES cells cultured with serum/LIF.
(J) sm-mRNA-FISH of Nanog vs. GFP expression in single V6.5 + Nanog-2A-GFP ES cells (pgk puro looped out) cultured with serum/LIF, 107 cells analyzed.
(K) Representative bright-field image (upper left), DAPI (upper right), and immunostaining of GFP protein (bottom left) and Nanog protein (bottom right) of V6.5 + Nanog-2A-GFP ES cells cultured with serum/LIF.
(L) Flow cytometric analysis of GFP in E14Tg2a + Nanog-2A-GFP (pgk puro looped out) ES cells cultured with serum/LIF. See also Figures S1 and S2.

References

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. Cell 150, 1209-1222.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 113, 643-655.

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

Fidalgo, M., Faiola, F., Pereira, C.F., Ding, J., Saunders, A., Gingold, J., Schaniel, C., Lemischka, I.R., Silva, J.C., and Wang, J. (2012). Zfp281 mediates Nanog autorepression through recruitment of the NuRD complex and inhibits somatic cell reprogramming. Proc Natl Acad Sci U S A 109, 16202-16207.

Filipczyk, A., Gkatzis, K., Fu, J., Hoppe, P., Lickert, H., Anastassiadis, K., and Schroeder, T. (2013). Biallelic expression of Nanog protein in mouse embryonic stem cells. Cell Stem Cell.

Fischer, Y., Ganic, E., Ameri, J., Xian, X., Johannesson, M., and Semb, H. (2010). NANOG reporter cell lines generated by gene targeting in human embryonic stem cells. PLoS One 5, e12533.

Hatano, S.Y., Tada, M., Kimura, H., Yamaguchi, S., Kono, T., Nakano, T., Suemori, H., Nakatsuji, N., and Tada, T. (2005). Pluripotential competence of cells associated with Nanog activity. Mech Dev 122, 67-79.

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

Jaenisch, R., and Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell 132, 567-582.

Kalmar, T., Lim, C., Hayward, P., Munoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biol 7, e1000149.

Macarthur, B.D., Sevilla, A., Lenz, M., Muller, F.J., Schuldt, B.M., Schuppert, A.A., Ridden, S.J., Stumpf, P.S., Fidalgo, M., Ma'ayan, A., et al. (2012). Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. Nat Cell Biol.

Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell 113, 631-642.

Miyanari, Y., and Torres-Padilla, M.E. (2012). Control of ground-state pluripotency by allelic regulation of Nanog. Nature 483, 470-473.

Navarro, P., Festuccia, N., Colby, D., Gagliardi, A., Mullin, N.P., Zhang, W., Karwacki-Neisius, V., Osorno, R., Kelly, D., Robertson, M., et al. (2012). OCT4/SOX2-independent Nanog autorepression modulates heterogeneous Nanog gene expression in mouse ES cells. EMBO J 31, 4547–4562.

Orkin, S.H., Wang, J., Kim, J., Chu, J., Rao, S., Theunissen, T.W., Shen, X., and Levasseur, D.N. (2008). The transcriptional network controlling pluripotency in ES cells. Cold Spring Harb Symp Quant Biol 73, 195-202.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods 5, 877-879.

Silva, J., Barrandon, O., Nichols, J., Kawaguchi, J., Theunissen, T.W., and Smith, A. (2008). Promotion of reprogramming to ground state pluripotency by signal inhibition. PLoS Biol 6, e253.

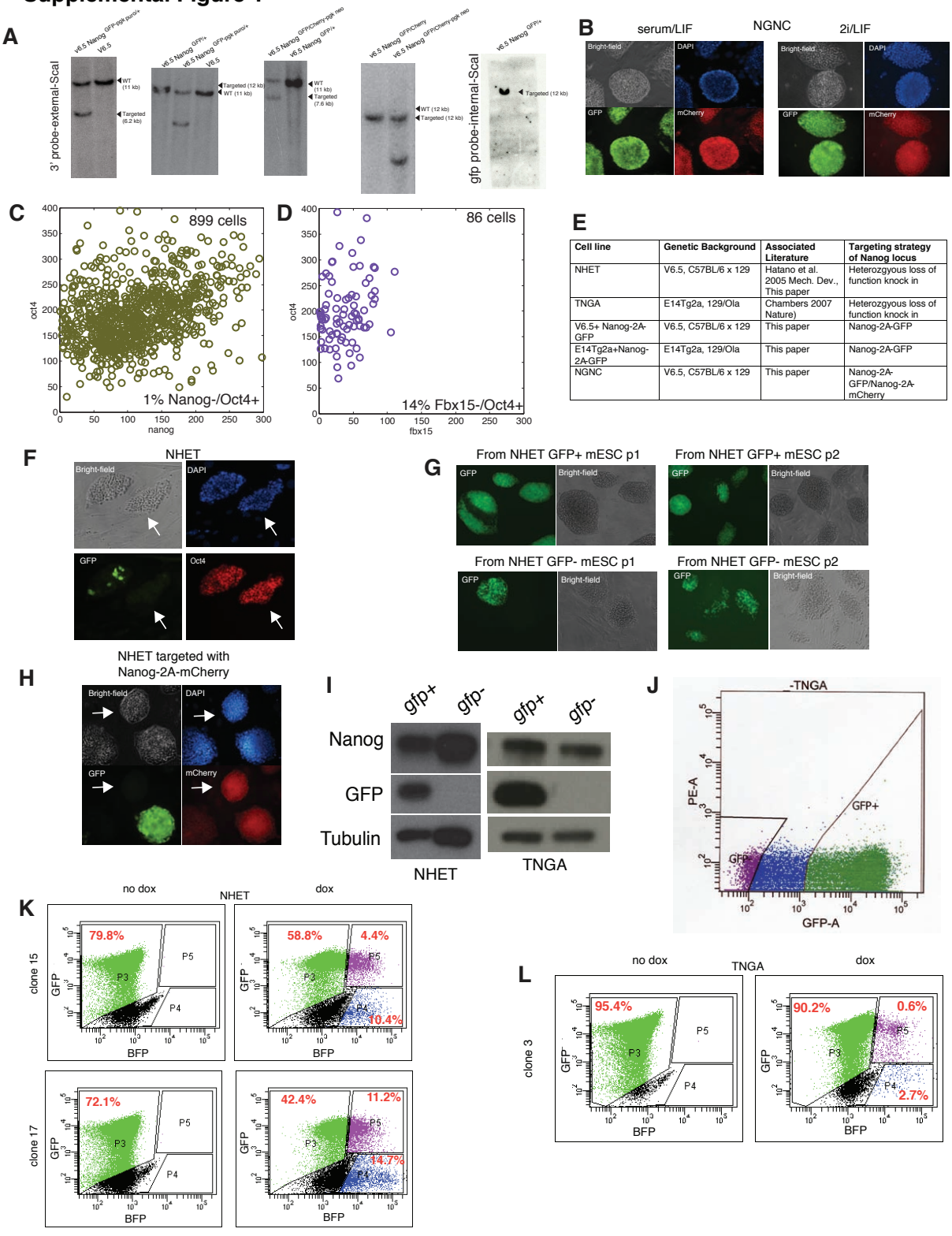
Silva, J., Nichols, J., Theunissen, T.W., Guo, G., van Oosten, A.L., Barrandon, O., Wray, J., Yamanaka, S., Chambers, I., and Smith, A. (2009). Nanog is the gateway to the pluripotent ground state. Cell 138, 722-737.

Singh, A.M., Hamazaki, T., Hankowski, K.E., and Terada, N. (2007). A heterogeneous expression pattern for Nanog in embryonic stem cells. Stem Cells 25, 2534-2542.

Wray, J., Kalkan, T., and Smith, A.G. (2010). The ground state of pluripotency. Biochem Soc Trans 38, 1027-1032.

Ying, Q.L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature 453, 519-523.

Supplemental Figure 1

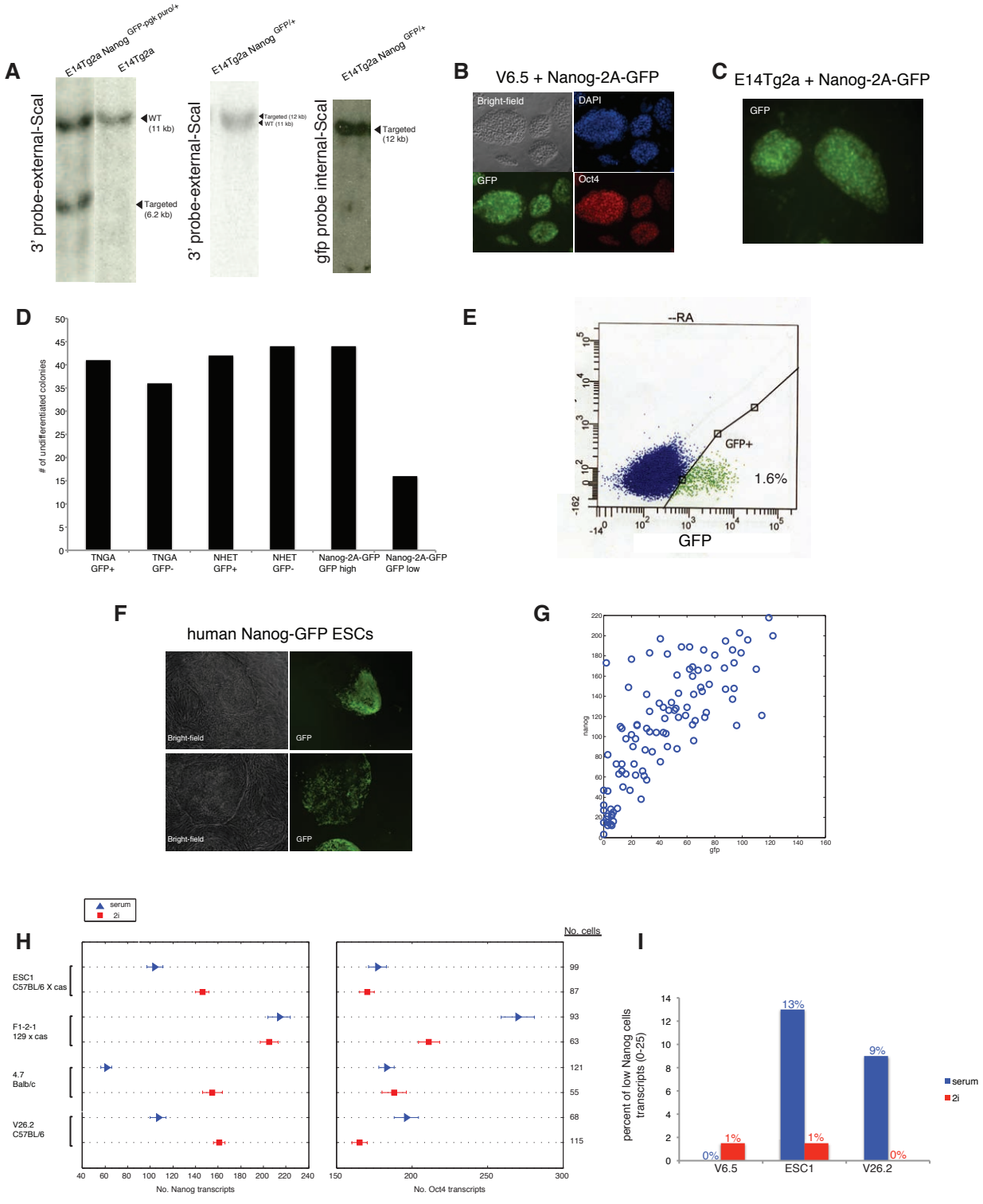


Cell line	Genetic Background	Associated Literature	Targeting strategy of Nanog locus
NHET	V6.5, C57BL/6 x 129	Hatano et al. 2005 Mech. Dev., This paper	Heterozgyous loss of function knock in
TNGA	E14Tg2a, 129/Ola	Chambers 2007 Nature)	Heterozgyous loss of function knock in
V6.5+ Nanog-2A-GFP	V6.5, C57BL/6 x 129	This paper	Nanog-2A-GFP
E14Tg2a+Nanog-2A-GFP	E14Tg2a, 129/Ola	This paper	Nanog-2A-GFP
NGNC	V6.5, C57BL/6 x 129	This paper	Nanog-2A-GFP/Nanog-2A-mCherry

Supplemental Figure 1-Related to Figure 1 and Figure 2

(A) Southern blot analysis of correctly targeted NGNC clones.
(B) Representative bright-field images (upper left), DAPI (upper right), and immunostaining of GFP protein (bottom left) and mCherry protein (bottom right) of NGNC ES cells cultured in serum/LIF (left) and 2i/LIF (right).
(C-D) sm-mRNA-FISH of Oct4 vs. Nanog (C, 899 cells), Fbx15 (D) expression in single V6.5 ES cells cultured with serum/LIF.
(E) Table of the cell lines, genetic background, associated literature, and targeting strategy of Nanog locus used in Figure 2.
(F) Representative bright-field image (upper left), DAPI (upper right), and immunostaining of GFP protein (bottom left) and Oct4 protein (bottom right) of NHET ES cells cultured in serum/LIF. White arrows indicate GFP-/Oct4+ cells. Note speckled colony containing GFP+ and GFP- cells in the upper left corner.
(G) Representative GFP and bright-field images of cells arising from GFP+ (top) and GFP- (bottom) NHET ES cells after one (left) and two (right) passages. Note the appearance of GFP- cells from GFP+ cells (top panels) and GFP+ cells from GFP- cells (bottom panels).
(H) Representative bright-field image (upper left), DAPI (upper right), and immunostaining of GFP protein (bottom left) and mCherry protein (bottom right) of NHET ES cells targeted with Nanog-2A-mCherry and cultured with serum/LIF. White arrows indicate GFP-/mCherry+ cells.
(I) Western blot of Nanog, GFP, and Tubulin protein from NHET ES cells (left) and TNGA ES cells (right).
(J) Gates for sorting GFP+ and GFP- fractions from TNGA cells. GFP+ and GFP- cells from TNGA cells were sorted, fractions were lysed and analyzed by immunoblotting.
(K) Flow cytometry of BFP+/GFP+ NHET ESC clones 15 and clones 17 on dox and without dox. (L) Flow cytometry of GFP of TNGA clone 3 on dox and without dox.

Supplemental Figure 2



Supplemental Figure 2-Related to Figure 2

(A) Southern blot analysis of correctly targeted E14Tg2a + Nanog-2A-GFP clones.
(B) Representative bright-field image (upper left), DAPI (upper right), and immunostaining of GFP protein (bottom left) and Oct4 protein (bottom right) of V6.5 + Nanog-2A-GFP ES cells cultured in serum/LIF.
(C) Representative GFP image of E14Tg2a + Nanog-2A-GFP ES cells.
(D) Barplot of the number of undifferentiated colonies derived from TNGA GFP+, TNGA GFP-, NHET GFP+, NHET GFP-, V6.5 + Nanog-2A-GFP high, and V6.5 + Nanog-2A-GFP low ES cells.
(E) Flow cytometric analysis of GFP in V6.5 + Nanog-2A-GFP (pgk puro looped out) ES cells treated with retinoic acid for 48 hours.
(F) Bright-field and GFP images of Nanog-GFP human ESC reporter line from Fischer et al. 2010.
(G) sm-mRNA-FISH of Nanog vs. GFP expression in single V6.5+ Nanog-2A-GFP-pgk puro (pgk puro not looped out) ES cells cultured in serum/LIF, 107 cells analyzed.
(H) Plot of the median number of Nanog (left) and Oct4 (right) transcripts, quantified by sm-mRNA-FISH, in ESC1 (C57BL/6 x cas), F1-2-1 (129 x cas), 4.7 (Balb/c), and V26.2 (C57BL/6) ES cells cultured in serum/LIF (triangle, blue) and 2i/LIF (square, red). Error bars represented standard error of the mean.
(I) Histogram of percent of low Nanog-expressing cells (those cells with Nanog transcripts between 0 and 25) for V6.5, ESC1, V26.2 cultured in serum/LIF (blue) and 2i/LIF (red).

Chapter 4. Future Directions
Single cells and cellular reprogramming

Deciphering the mechanisms of cellular reprogramming has been hindered by the inability to track the few cells that will become iPSCs. Using single-cell analysis, we found Esrrb, Utf1, Lin28 and Dppa2 to be predictive markers of reprogramming. We found that single cells exhibit high variation in gene expression early in reprogramming and this heterogeneity decreases are the cell reaches pluripotency. Our results show that a stochastic phase of gene activation is followed by a late hierarchical phase, initiated by activation of the Sox2 locus, leading to the activation of the pluripotency circuitry. Finally, cells were reprogrammed without the generic "Yamanaka" factors Oct4, Sox2, Klf4, c-Myc and Nanog (Buganim et al., 2012).

Recently, Jacob Hanna and colleagues performed a small interfering RNA (siRNA) screen for epigenetic regulators whose knockdown would turn epiblast stem cells (EpiSCs) into ESCs (Rais et al., 2013). EpiSCs are a developmentally more advanced cell type, relative to ESCs (Bao et al., 2009). They found that depletion of methyl-binding protein 3 (Mbd3), a core unit of the Mbd3/nucleosome remodeling and deacetylation (NuRD) complex enhanced the reversion of EpiSCs to ESCs, with almost every cell reverting, in addition to facilitating the conversion of primordial germ cells to ESCs. Ultimately, depletion of Mbd3 in mouse and human somatic cells, in addition to OSKM overexpression and reprogramming in naïve pluripotency conditions, enables "all iPSC reprograming." Clonal analyses supported that the dynamics of reprogramming was consistent with a deterministic model and live-cell imaging documented synchronous activation of an Oct4-GFP reporter in all cells. Total time for the reprogramming was a mere seven days (Rais et al., 2013). This system provides a powerful tool to study the mechanism of reprogramming in a homogenous population, which would most likely eliminate the need for single-cell analyses. Since 100% of the cells are reprogramming in a synchronous fashion, it should be possible to define the sequence of transcriptional and epigenetic changes with standard population-based assays like RNA-sequencing and quantitative PCR (qPCR). It would be most ideal to study reprogramming in a synchronized population without genetic modification.

Therefore, future studies should be aimed at reproducing the same results with chemical inhibitors of Mbd3, like histone deacetylase inhibitors (Huangfu et al., 2008).

It will be interesting to explore whether specific combinations of chromatin modifiers are able to reset the epigenome of a somatic cell and reprogram it to pluripotency in the absence of pluripotency factors. Our analysis of reprogramming is lacking at the level of translation. Therefore, future studies should be aimed at understanding cells beyond the level of gene expression.

Future applications of iPSC technology

The potential of iPSC technology for the study of complex diseases and eventually for stem cell therapy may fundamentally change our approach to medicine. Alzheimer's and Parkinson's disease are the most common neurodegenerative disorders that have puzzled scientists for centuries. They are not inherited in a clear Mendelian fashion and appear to result from complex interactions between genetic and environmental factors ("sporadic disease") (Soldner et al., 2011). The combination of poorly understood genetics and the inability to manipulate and study primary human neuronal tissue has hindered the development of a reliable *in vitro* cell culture model critical to study the genetic components of sporadic disease (Nussbaum and Ellis, 2003). Recently, genome-wide association studies (GWAS) have identified common genetic variants in less than 100 genes for Alzheimer's and Parkinson's disease (Satake et al., 2009). Furthermore, an *in vitro* cell culture model is now feasible with iPSC technology and genome-editing tools that enable the transformation of skin from a patient into an embryonic stem-like cell, whose genes can be manipulated, and then into a diseased neuron (Soldner and Jaenisch, 2012). Due to these technological advancements it is finally conceivable to model sporadic disease in a dish.

Prior to utilizing human iPSCs in a clinical setting, investigation must be focused into controlling for the genetic background in which the disease occurs and establishing robust differentiation protocols (Saha and Jaenisch, 2009; Soldner and Jaenisch, 2012). Recent advances in gene-editing technologies, such as CRISPR/Cas, enable the targeted modification of human cells for gene disruptions, genetic repair, or insertion of reporters

that will allow the generation of reporter lines and the perturbation of genetic elements harboring genetic variants (Wang et al., 2013). The trifecta of iPSC, gene-editing, and genome-wide technologies provides the opportunity to systematically examine how genetic variants contribute to sporadic disease with an *in vitro* cell culture model.

Nanog and heterogeneity of pluripotency factors in ESCs

Prior to the publication of the data in Chapter 3, it was widely believed that Nanog expression is heterogeneous in mouse ESCs. It was suggested that this heterogeneity results from allelic regulation and that downregulation of the gene predisposed cells towards differentiation. Four conclusions can be drawn from the data in Chapter 3: (1) Single-cell analysis reveals Nanog is biallelically expressed in mouse ESCs (2) Variation in Nanog expression is similar to that of other pluripotency markers (3) Heterozygous loss of function knock-in reporters do not reflect Nanog expression, and (4) Genetic background can influence the range of Nanog expression (Faddah et al., 2013).

Interestingly, Rex1 is also suggested to fluctuate in mouse ESCs but it was also created using a loss of function knock-in reporter (Toyooka et al., 2008). It appears that Stella, Esrrb, and Fbx15 are clearly heterogeneously expressed and have expression profiles in ESC cultures very distinct from Nanog (Faddah et al., 2013; Payer et al., 2006). This heterogeneity has been termed "metastability" and has been suggested to be an essential component of ESC cultures. It remains unknown what the biological relevance is of this heterogeneity. It is important to remember that ESCs are a culture artifact and do not occur *in vivo* and therefore the biological relevance and function of heterogeneity is unknown (Smith, 2013).

Austin Smith, Ian Chambers and colleagues published the paper creating the idea that Nanog fluctuates in mouse ESCs (Chambers et al., 2007). Smith published a response to the data in Chapter 3 in the same issue of *Cell Stem Cell*. His explanation of the data was the following: "Remarkably, however, Faddah et al. did not examine ESCs without feeders in serum and LIF, and therefore cannot draw conclusions pertinent to the circumstance in which heterogeneity has been documented. It would be intriguing if their reporter remained homogeneously expressed in these conditions." (Smith, 2013).

Unfortunately, Smith and colleagues neglected to publish their culture conditions in the Methods of their original paper (Chambers et al., 2007). Finally, it is incredibly rare to hear of an ESC lab that culture ESCs without feeders (personal communication).

To support the data in Chapter 3, and address Smith's point, Filipczyk et al. generated dual reporters using fusion proteins, used feeder-free conditions, and also concluded that Nanog is biallelically expressed in mouse ESCs (Filipczyk et al., 2013). The findings in Chapter 3, in addition to Filipczyk et al. directly challenge the idea that Nanog fluctuates in mouse ESCs, arguing against monoallelic expression of Nanog. It is possible that transcriptional bursting, in which a gene stochastically transitions between states of transcriptional inactivity and activity may explain the previously reported monoallelic expression of Nanog (Chubb et al., 2006; Dar et al., 2012; Golding et al., 2005; Kaern et al., 2005; Raj et al., 2006; Raj and van Oudenaarden, 2008). This concept suggests that transcription is discontinuous and mRNA molecules are produced in transcriptional bursts, possibly by randomly switching back and forth between transcriptionally active and inactive states. Little is known about how these bursts originate, but it has been suggested that the basis for bursting may be stochastic events of chromatin remodeling (Raj and van Oudenaarden, 2008). A recent report suggested that a fraction of genes in human cells are monoallelically expressed due to bursting. The phenomenon is particularly interesting because these findings force us to rethink established dogma of how alleles are expressed in cells and the consequences in regard to cellular decision making and how this may contribute to disease (Deng et al., 2014).

There is currently no evidence that fluctuating expression of pluripotency factors occurs during epiblast development and differentiation *in vivo*. (Smith, 2013). Nanog has been termed the "gateway to pluripotency," meaning it's essential for the establishment of pluripotency during the derivation of ESCs and iPSCs. Recently, two groups found that Nanog is dispensable for the generation of iPSCs by producing iPSCs from Nanog (-/-) fibroblasts that contributed to the germline of chimeric mice (Carter et al., 2014; Schwarz et al., 2014). Thus, Nanog may be important during the reprogramming process; however, it is not required for establishing pluripotency in the mouse.

Concluding remarks

The stem cell field was truly electrified by Yamanaka's discovery. At the time, there were many obvious unanswered questions; however, the field has slowed down a bit as we sit and reflect on the important questions regarding the molecular mechanism that remain unsolved. It is imperative that we address the challenges associated with human disease modeling in order to move forward with patient-specific cell therapies. It is equally important that we fully understand the genetic landscape of human and mouse ESCs because they are the gold standard comparison to iPSCs. We cannot understand iPSCs without understanding ESCs. Since the ESC field is incredibly exciting and promising, it's critically important that we ensure high-quality research and reproducible results in all scientific journals. Finally, we must continue to nurture innovative science in the face of increasingly limited funding.

References

Bao, S., Tang, F., Li, X., Hayashi, K., Gillich, A., Lao, K., and Surani, M.A. (2009). Epigenetic reversion of post-implantation epiblast to pluripotent embryonic stem cells. Nature *461*, 1292-1295.

Buganim, Y., Faddah, D.A., Cheng, A.W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S.L., van Oudenaarden, A., and Jaenisch, R. (2012). Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell *150*, 1209-1222.

Carter, A.C., Davis-Dusenbery, B.N., Koszka, K., Ichida, J.K., and Eggan, K. (2014). Nanog-Independent Reprogramming to iPSCs with Canonical Factors. Stem cell reports *2*, 119-126.

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

Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193-196.

Faddah, D.A., Wang, H., Cheng, A.W., Katz, Y., Buganim, Y., and Jaenisch, R. (2013). Single-cell analysis reveals that expression of nanog is biallelic and equally variable as that of other pluripotency factors in mouse ESCs. Cell Stem Cell *13*, 23-29.

Filipczyk, A., Gkatzis, K., Fu, J., Hoppe, P.S., Lickert, H., Anastassiadis, K., and Schroeder, T. (2013). Biallelic expression of nanog protein in mouse embryonic stem cells. Cell Stem Cell *13*, 12-13.

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A.E., and Melton, D.A. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. Nat Biotechnol *26*, 795-797.

Jaenisch, R. (2012). Nuclear cloning and direct reprogramming: the long and the short path to Stockholm. Cell Stem Cell *11*, 744-747.

Nussbaum, R.L., and Ellis, C.E. (2003). Alzheimer's disease and Parkinson's disease. The New England journal of medicine *348*, 1356-1364.

Payer, B., Chuva de Sousa Lopes, S.M., Barton, S.C., Lee, C., Saitou, M., and Surani, M.A. (2006). Generation of stella-GFP transgenic mice: a novel tool to study germ cell development. Genesis *44*, 75-83.

Rais, Y., Zviran, A., Geula, S., Gafni, O., Chomsky, E., Viukov, S., Mansour, A.A., Caspi, I., Krupalnik, V., Zerbib, M., *et al.* (2013). Deterministic direct reprogramming of somatic cells to pluripotency. Nature *502*, 65-70.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216-226.

Saha, K., and Jaenisch, R. (2009). Technical challenges in using human induced pluripotent stem cells to model disease. Cell Stem Cell *5*, 584-595.

Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Kawaguchi, T., Tsunoda, T., Watanabe, M., Takeda, A.*, et al.* (2009). Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nat Genet *41*, 1303-1307.

Schwarz, B.A., Bar-Nur, O., Silva, J.C., and Hochedlinger, K. (2014). Nanog is dispensable for the generation of induced pluripotent stem cells. Current biology : CB *24*, 347-350.

Smith, A. (2013). Nanog heterogeneity: tilting at windmills? Cell Stem Cell *13*, 6-7.

Soldner, F., and Jaenisch, R. (2012). Medicine. iPSC disease modeling. Science *338*, 1155-1156.

Soldner, F., Laganiere, J., Cheng, A.W., Hockemeyer, D., Gao, Q., Alagappan, R., Khurana, V., Golbe, L.I., Myers, R.H., Lindquist, S.*, et al.* (2011). Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. Cell *146*, 318-331.

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

Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F., and Jaenisch, R. (2013). One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell *153*, 910-918.

Curriculum vitae

Dina Adel Faddah

Education

2008-2014: PhD in Biology, Massachusetts Institute of Technology, Cambridge, MA
2002-2006: BS in Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC

Research Experience

2009-2014: Graduate Studies
Laboratory of Prof. Rudolf Jaenisch
Massachusetts Institute of Technology, Whitehead Institute
Single-cell analyses of cellular reprogramming and embryonic stem cells

2006-2008: Postbaccalaureate Studies
Laboratory of Dr. Francis Collins
National Institutes of Health, National Human Genome Research Institute
Development and establishment of therapies for Hutchinson-Gilford progeria syndrome in mice and humans

Summer 2005: Summer Studies
Laboratory of Dr. Francis Collins
National Institutes of Health, National Human Genome Research Institute
Characterization of a mouse model of Hutchinson-Gilford progeria syndrome

2003-2006: Undergraduate Studies
Laboratory of Prof. Jason Lieb and Prof. Todd Vision
University of North Carolina at Chapel Hill
Systematic identification of balanced transposition polymorphisms in Saccharomyces cerevisiae

Honors and Awards

2010	Jerome and Florence Brill Graduate Student Fellowship, Whitehead Institute for Biomedical Research
2012	Quantitative Developmental Biology Fellowship, RIKEN Center for Developmental Biology
2011	US delegation, 61st Lindau Meeting of Nobel Laureates in Physiology and Medicine
2010	Vertex Pharmaceuticals Scholar
2010	Graduate Woman of Excellence, MIT
2010-2013	Graduate Research Fellowship, National Science Foundation
2008	Postbaccalaureate Intramural Research Training Award, NIH
2004	Excellent Poster Award, Sigma Xi
2004	Smallwood Fellowship for Undergraduate Research, UNC-CH
2002-2006	Dean's List, UNC-CH

Publications

Theunissen TW*, Powell BE*, Wang H*, Mitalipova M, **Faddah DA**, Maetzel D, Ganz K, Shi L, Stelzer Y, Zhang J, Lungjangwa T, Imsoonthornruska S, Rangarajan S, Fan ZP, Young RA, Gray N, Jaenisch R. Systematic Identification of Kinase Inhibitors that Induce and Maintain Naive Human Pluripotency. *To be submitted to Cell Stem Cell.* (*co-first authors)

Buganim Y*, Markoulak S*, van Wietmarschen N, Hoke H, Wu T, Ganz K, Akhtar-Zaidi B, He Y, Abraha BJ, Porubsky D, Kulenkampff E, **Faddah DA**, Shi L, Gao Q, Sarkar S, Cohen M, Goldman J, Nery JR, Schultz MD, Ecker JR, Xiao A, Young RA, Lansdorp PM. Jaenisch R. The developmental potential of iPSCs is greatly influenced by the selection of the reprogramming factors. *Cell Stem Cell*. In revision. (*co-first authors)

Klemm SL*, Semrau S*, Wiebrands K*, Mooijman D, **Faddah DA**, Jaenisch R, van Oudenaarden A. Transcriptional profiling of cells sorted by RNA abundance. *Nature Methods*. 2014. (*co-first authors)

Faddah DA, Wang H, Cheng AW, Katz Y, Buganim Y, Jaenisch R. Single-Cell Analysis Reveals that Expression of Nanog is Biallelic and Equally Variable as that of Other Pluripotency Factors in Mouse ESCs. *Cell Stem Cell*. 13 (1): 23-9. 2013.

- "Nanog Heterogeneity: Tilting at Windmills?" By Austin Smith. Cell Stem Cell preview. July 3, 2013.
- "Study challenges long-held assumption of gene expression in embryonic stem cells." By Nicole Giese. Whitehead Institute News. July 3, 2013.

Buganim Y, **Faddah DA**, Jaenisch R. Mechanisms and models of somatic cell reprogramming. *Nature Reviews Genetics*. 14(6): 427-39. 2013.

Buganim Y*, **Faddah DA***, Cheng AW, Itskovich E, Markoulaki S, Ganz K, Klemm SL, van Oudenaarden A, Jaenisch R. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*. 150(6):1209-22. 2012. (*co-first authors)

- *Science* cover. December 6, 2013.
- "Single-cell analysis of reprogramming." *Nature Methods*. Research Highlights. November 6, 2012.
- "Stem Cells: Revealing the reprogramming program." By Mary Muers. *Nature Reviews Genetics*. November 2012.
- "Order from chaos: single cell reprogramming in two phases." By Guangjin Pan and Duanquing Pei. *Cell Stem Cell* preview. October 5, 2012.
- "Turning back time for cells." By Anne Trafton. MIT homepage cover. September 17, 2012.
- "Tracking stem cell reprogramming." By Anne Trafton. MIT News Office. September 13, 2012.
- "Whitehead scientists bring new efficiency to stem cell reprogramming." By Nicole Giese and Anne Trafton. MIT News Office and Whitehead Institute News. September 13, 2012.
- Selected for Faculty of 1000 Biology.
- Best of *Cell* 2012. 1 of 12 papers chosen, out of ~400 papers, for the year.

Carey BW, Markoulaki S, Hanna J, **Faddah DA**, Buganim Y, Kim J, Ganz K, Steine EJ, Cassady JP. Creyghton MP, Welstead GG, Gao Q, Jaenisch R. Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. *Cell Stem Cell*. 9(6): 588-598. 2011.

Cao K, Blair CD, **Faddah DA**, Olive M, Erdos MR, Nabel EG, Collins FS. Lamin A and telomerase collaborate to trigger cellular senescence. *Journal of Clinical Investigation*. 121(7): 2833-44. 2011.

Kim JP, Lengner CJ, Kirak O, Hanna J, Cassady JP, Lodato MA, Wu S, **Faddah DA**, Steine E, Gao Qing, Fu D, Dawlaty MM, Jaenisch R. Reprogramming of postnatal neurons into induced pluripotent stem cells by defined factors. *Stem Cells*. 29(6): 992-1000. 2011.

Faddah DA, Ganko EW, McCoach C, Pickrell JK, Hanlon SE, Mann FG, Mieczkowska A, Jones CD, Lieb JD, Vision TJ. Systematic identification of balanced transposition polymorphisms in Saccharomyces cerevisiae. *PLoS Genetics*. 5(6) e1000502. 2009.

Capell BC, Olive M, Erdos MR, Cao K, **Faddah DA**, Whipperman M, Conneely KN, Song H, Qu X, Ganesh S, Avallone H, Kolodgie F, Virmani R, Nabel EG, Collins FS. A farnesyltransferase inhibitor prevents both the onset and late progression of cardiovascular disease in a progeria mouse model. *PNAS*. 105(41): 15902-7. 2008.

Varga R, Eriksson M, Erdos MR, Olive M, Harten I, Kolodgie F, Capell BC, Cheng J, **Faddah D**, Perkins S, Avallone H, San H, Qu X, Ganesh S, Gordon LB, Virmani R, Wight TN, Nabel EG, Collins FS. Progressive vascular smooth muscle cell defects in a mouse model of Hutchinson-Gilford progeria syndrome. *PNAS*.103 (9) 3250-3255. 2006.

Patents

"Programming and Reprogramming of Cells." PCT/US2013/037623, filed in 2013.

Oral Presentations

Invited speaker, Fluidigm single-cell symposium | The Paradigm of the Single Cell, September 2013, Boston, MA, USA. "A Single-Cell View Of Pluripotency In Mouse Embryonic Stem Cells."

Invited speaker, MITRA Biotech, May 2013, Bangalore, India. "Single cell analysis of cellular reprogramming and mouse ESCs."

Invited speaker, RIKEN CDB Quantitative Developmental Biology Symposium, March 2012, Kobe, Japan. "Single-cell analysis of cellular reprogramming."

Teaching Experience

2012	Teaching Assistant, Introductory Biology Lab (7.02), MIT
2012	Nobelprize.org student interviewer. Dr. Bruce Beutler, June 12, 2012.
2009-2012	Whitehead Labs for High School Students.
2009	Teaching Assistant, Introductory Biology (7.01) MIT
2004-2008	Trained one undergraduate at UNC-CH (Biff Mann, now in graduate school at Stanford).