Roadblocks and Facilitators of Reprogramming to Pluripotency

Herprogrammeren tot pluripotentie: stimuleren of blokkeren

Marti Anne Borkent

The work presented in this thesis was performed at the Cancer Center / Center for Regenerative Medicine at Massachusetts General Hospital and in affiliation with the Harvard Stem Cell Institute and Harvard Medical School in Boston, USA.

The studies described in this thesis were supported by a PhD fellowship from the Boehringer Ingelheim Fonds and from the Gerald and Darlene Jordan Endowment, the NIH (HD058013) and an Early Career Award from the Howard Hughes Medical Institute.

Financial support by the Erasmus MC and the Jurriaanse Stichting for the publication of this thesis is gratefully acknowledged.

This thesis was designed by Marti and printed by Ridderprint.

Cover pictures: 'Myrinx', (front) and 'Loom' (rear) by Rick Berry. Myrinx shows a strong female figure, as often depicted by Rick. She inspires, as she is jumping into the unknown with grace and confidence. Loom depicts a weave of hairpins, being unraveled to reveal their nature. Loom was inspired by the short hairpin RNA screen described in chapters 4 and 5, and appeared on the cover of Stem Cell Reports when chapter 4 was published.
Chapter dividers are filled with details of Myrinx and Loom.

Roadblocks and Facilitators of Reprogramming to Pluripotency

Herprogrammeren tot pluripotentie: stimuleren of blokkeren

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus
Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 22 november 2017 om 15.30 uur

door

Marti Anne Borkent
geboren te Oegstgeest

Erasmus University Rotterdam

Promotiecommissie:

Promotoren:	Prof.dr. J. Gribnau
	Prof.dr. K. Hochedlinger
Overige leden:	Prof.dr. J.N.J. Philipsen
	Dr. R. A. Poot
	Prof.dr. N. Geijsen

Table of contents

Chapter 1	Introduction	7
Chapter 2	The reprogrammable mouse	47
Chapter 3	A roadmap of reprogramming to induced pluripotent stem cells	67
Addendum to chapter 3:	*Gain-of-function screen for enhancers identifies ERas*	109
Chapter 4	A screen for inhibitors of reprogramming identifies SUMO2	127
Chapter 5	Alternative polyadenylation factor Nudt21 inhibits cellular reprogramming	159
Addendum to chapter 4&5:	*Considerations when performing a serial, genome-wide RNA inhibition screen*	177
Chapter 6	Reprogramming to pluripotency as a model for oncogenic IDH mutations	195
Chapter 7	Failure to replicate stress based reprogramming of somatic cells	217
Chapter 8	Discussion and future implications	245
Chapter 9	Summary / Samenvatting	257
	Appendices	265
	Acknowledgements	275

Chapter 1

Introduction

The work described in this thesis revolves around reprogramming cells into stem cells, to understand what happens during this transition and use this to further understand cell fate transformation and its pathological counterpart, cancer. In this introduction, we will discuss the background and impact of stem cells by highlighting the following topics:

- Definitions of stem cells
- Factors required to maintain embryonic stem cells
- History and recent discoveries of cellular reprogramming
- Impact of research into stem cells and reprogramming
- Events during reprogramming of somatic cells to iPSCs
- A view of oncogenic transformation from a reprogramming perspective

Definitions of stem cells

A stem cell can self-renew and give rise to one or multiple differentiated lineages. It can thus divide and give rise to a stem cell that is the same as its mother cell, but also give rise to a daughter cell that differentiates into one or more types of specialized cells[1]. A stem cell can be immortal; it normally retains its ability to proliferate. It has the privilege of expressing factors such as telomerase, which reduces loss of genetic integrity over time, increasing longevity[2].

There is an important distinction between adult- and embryonic stem cells. Adult stem cells are probably best known; these are the stem cells that persist in the mature body, and continuously replenish the cells of the gut, the skin and blood. For example, hematopoietic stem cells reside in bone marrow throughout adult life to replenish all lineages of blood cells through increasingly more differentiated progenitor cells. Hematopoietic stem cells, however, will never give rise to skin or gut epithelium; they themselves are committed such that they are limited to producing blood. They are called multipotent; they can differentiate into 'several' but not many or all lineages. Pluripotent stem cells (PSCs) can differentiate into 'many' lineages. Pluripotency means that all cells of an embryo can be derived from this stem cell. Totipotency, or 'can do all' is a term reserved for a fertilized egg cell and its immediate progeny, which can form both embryonic and extra-embryonic cells such as placental cells (Figure 1).

In this thesis, we are interested in PSCs. Examples of PSCs are Embryonic Stem Cells (ESCs), derived from a very early, blastocyst-stage embryo, or induced pluripotent stem cells (iPSCs), derived by 'reprogramming' a more differentiated somatic cell with defined transcription factors, in order to become pluripotent. The latter are indistinguishable from ESCs when compared under the right conditions[3-6]; we discuss the concept of reprogramming below.

There has been enormous scientific progress in the pluripotent stem cell field in the past 20 years, such as the derivation of human embryonic stem cells[7] and the discovery of four-factor reprogramming into induced pluripotent stem cells[8]. Also, there are new uses for stem cells in combination with new gene-editing methods and in vitro tissue generation techniques; these are applied to design cures for single-gene congenital diseases and degenerative disease.

However, the word stem cell is nearly 150 years old[9]. The term was popularized by Edmund Wilson in 1896, but can be attributed to German biologist Ernst Haeckel, who in 1868 referred to the fertilized egg as a "Stammzelle". He later described a Stammzelle as the ancestor of all cells in multicellular organisms, the base of a tree of offspring like a common ancestor in a phylogenetic tree he called "Stammbäume", hence the name. The first person who described ESCs and demonstrated their ability to form any embryonic tissue, was a Maine researcher called Leroy Stevensμ. He studied the teratomas that one of his strains

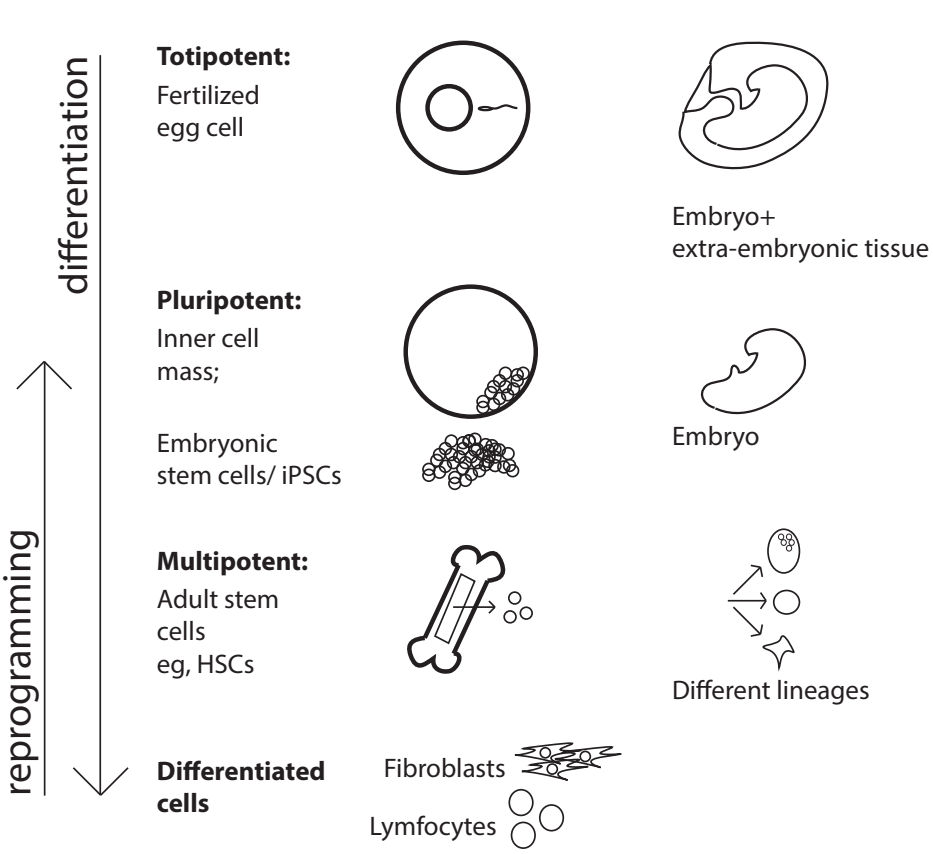


Figure 1. Different stages of differentiation and corresponding developmental potential.

of mice kept suffering from. Teratomas are tumors that contain an unorganized mix of adult tissues, including bone, muscle and hair; tissues representative of all embryonic germ layers. We know now that teratomas can also be derived experimentally. (Figure 2). He found that certain undifferentiated cells in a teratoma were transplantable from mouse to mouse and produced new teratomas indefinitely. He also noted that early embryonic progenitors to germ cells, primordial germ cells, were responsible for the teratomas in his mouse strain. Perhaps most importantly, he realized that primordial germ cells look like cells from the inner cell mass (ICM) of a few days-old embryo. In 1970 he showed that these cells were also capable of forming teratomas, calling them 'pluripotent embryonic cells' that 'appear to give rise to both rapidly differentiating cells and others which like themselves remain undifferentiated'[10]. This is the definition of an embryonic stem cell. Due to their ability to also give rise to cancerous cells however, they became known as embryonic carcinoma cells (ECCs)[10,11]. Five years later, Mintz and Illmensee showed that these cells indeed were capable of giving rise to a whole new mouse[12]. In 1981 several groups managed to isolate similar cells from mouse blastocysts, which we now call embryonic stem cells (ESCs)[13,14]. Since then, a body of work has defined what genes are key to pluripotency, such as Oct4, Sox2 and Nanog[15-18], how embryonic stem cells are best grown to maintain a pluripotent state[19] or to differentiate into defined lineages[20].

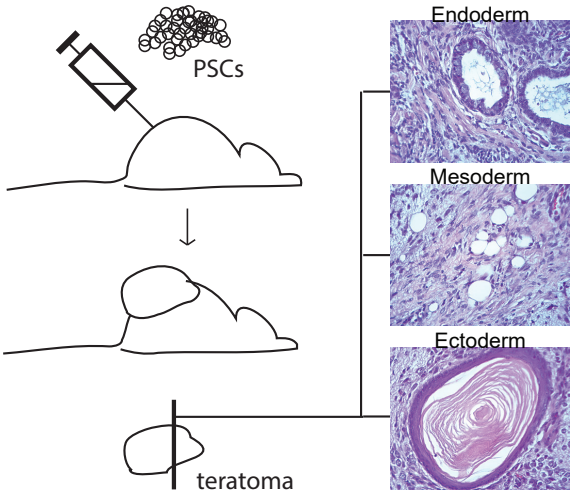


Figure 2. Schematic presentation of a teratoma formation assay.

Factors required to maintain embryonic stem cells

A combination of transcription factors, signaling factors and extracellular signals can maintain PSCs in a pluripotent state. Three transcription factors have been found to form the core regulatory network of pluripotency. These transcription factors are influenced by three main signaling cascades.

Pluripotency gene regulatory network

Transcription factors OCT4, SOX2 and NANOG form what has been called the 'pluripotency gene regulatory network (PGRN)'[21]. They are highly interconnected and maintain the pluripotent state through other transcription factors, in response to signaling cascades, epigenetic mechanisms and metabolic cues.

OCT4 is essential to pluripotency both in vivo and in vitro. OCT4 prevents trophectoderm differentiation by interacting with CDX2. *Cdx2* and *Oct4* expression are mutually exclusive, directing cells either to trophectoderm or ICM formation[17,22]. SOX2 is a key regulator of OCT4 and is, similar to OCT4, required to prevent trophectoderm differentiation[18,22]. SOX2 is also required for the formation of the epiblast, the structure within an implantation stage embryo which contains pluripotent cells[23]. Depending on the culture conditions, transcription factor NANOG can prevent differentiation into specific lineages. Formation of primitive endoderm, which forms the extraembryonic membranes, can be induced by removing Leukemia Inhibitory Factor (LIF) or when embryoid bodies are formed (a differentiation test in culture, mimicking early embryogenesis). Artificial *Nanog* expression blocks primitive endoderm formation in these settings[24,25]. Also, neuronal differentiation, induced by removal of LIF or of bone morphogenetic protein (BMP) from the culture media, can be prevented by *Nanog* overexpression[26]. When mesoderm is formed, NANOG can reverse this specification by repressing Brachyury[27].

According to Chromatin immunoprecipitation (ChIP) studies, these three factors co-occupy hundreds of regulatory elements including their own[21]. Thus, through positive feedback towards each other and downstream effectors, they stabilize the pluripotent state; however, disruption of the PGRN through (lack of) external signals quickly sends the cell to a path of differentiation. External signals are:

LIF

ESCs can differentiate into three cell types: primitive ecto- or endoderm, and under certain conditions, trophectoderm[28]. Extracellular signaling by LIF, a member of the interleukin-6 cytokine family is a key factor in preventing differentiation. LIF binding to its receptor activates the JAK/STAT pathway. Activation of STAT3 upregulates targets including *Klf4* and *Tfcp2l1* and represses non-neural differentiation. STAT3 expression is essential to maintenance of pluripotency[29].

BMP

BMP exerts its signal through the SMAD transcription factors. SMAD transcription factors

suppress differentiation by upregulating inhibitor of differentiation (Id) genes. Together with LIF, BMP suffices to prevent differentiation mouse ESCs[26].

WNT

In the absence of canonical WNT/β-catenin signaling, WNT effector and transcriptional repressor TCF3 co-localizes with OCT4 and SOX2 pluripotency-associated genes, antagonizing their function. GSK3-β marks β-catenin for degradation, and as a result GSK-β inhibition promotes pluripotency. However, GSK3-β inhibition is not sufficient to maintain a pluripotent state.

Manipulation of these pathways is applied to maintain pluripotent stem cells in culture and promote reprogramming towards pluripotency. In this thesis, LIF is added to all reprogramming, iPSC or ESC cultures. Inhibition of GSK3β and the FGF4-ERK pathway is used to stimulate reprogramming efficiency.

History and recent discoveries of cellular reprogramming

Eleven years ago, culturing ESCs and human ESCs had been well established. The discovery of four-factor reprogramming by the Yamanaka lab in 2006 allowed this knowledge to be applied to these new induced pluripotent stem cells. This boosted interest in the field because of the promise of biological insight into pluripotency through reprogramming. It also inspired clinical applications by means of disease modeling, and patient- (not embryo-) derived pluripotent stem cells for transplantable-tissue generation[8].

The history of reprogramming starts with the realization that there is a program. The first step was demonstrating that all cells from an early embryo have the information to form a complete embryo. In 1902, Hans Spemann succeeded split the relatively large cells from salamander embryos in two with a baby hair, without killing the cells or their potential to form an enmbryo as long as the nucleus was preserved[30]. He went on to temporarily separate the cytoplasm from the nucleus to show that only the half of the cell with the nucleus would continue to divide and form an embryo, but that giving back the nucleus to the temporarily enucleated half of the cell would permit this cell to generate an embryo too[31]. Thus, the nucleus was shown to be programmed to- and essential to form an embryo. Based on this experiment, he suggested that a nucleus should be transplantable into an enucleated receiving cell and carry out its programmed destiny[9]. This prediction was confirmed by Briggs and King in 1952, when they performed the first ‘nuclear transfer’ experiment with tadpole embryos[32]. During their experiments, they made the interesting observation that such “cloning” attempts were more successful with cells from earlier embryos compared to later, more differentiated cells. However, future Nobel prize winner

John Gurdon made the important observation in 1958 that the nucleus from an adult cell can also form an embryo when transferred into an enucleated egg cell, that this capability is not lost upon cellular differentiation and adult cells thus retain all information required to make an embryo[33]. Somehow factors in the cytoplasm of the host egg cell can instruct a

Figure 3. A historical overview of major events in stem cell- and reprogramming fields.

differentiated cell to acquire an omnipotent state.

During the 1970's, 80's and 90's, several steps were made to expand nuclear transfer to mammalian animals, adult cell nuclei, and fully grown mammals including primates[34-36]. The first live mammal cloned from an uncultured adult breast cell, the sheep named 'Dolly', helped bring this technique under the attention of a wider public[37]. Definitive proof that a terminally differentiated cell could be reprogrammed to pluripotency, was supplied by performing nuclear transfer on T- and B-lymphocytes. As a mark of their terminal differentiation these cells had undergone DNA rearrangements of their antibody encoding sequences[38] unlike their less differentiated precursors. The mice that were cloned from lymphocytes, contained the same DNA rearrangement in each of their cells.

The ability of cytoplasm factors to reprogram a nucleus of a differentiated cell was further demonstrated by alternative reprogramming methods. In 1976, Miller and Ruddle studied pluripotent embryonic carcinoma cells (ECCs), which, as discussed above, are derived from teratocarcinomas and can form tissues from all germ layers when injected in an early embryo[39]. They fused these ECCs with somatic cells and showed that the resulting cells were also pluripotent. Thus, the ECCs had reprogrammed the thymus cells. This experiment was also demonstrated with murine and human ESCs and another pluripotent cell line called embryonic germ cells (EGCs) which are derived from embryonic primordial germ cells[40-42].

Next, the field looked to find which factors in the enucleated egg's cytoplasm or pluripotent stem cell were responsible for the reprogramming of differentiated cells - factors dominant enough to overwrite the program of a differentiated nucleus. In Japan, Kazutoshi Takahashi and Shinya Yamanaka reasoned that factors in the cytoplasm of ESCs or egg cells should be capable of reprogramming a somatic cell's nucleus as demonstrated by nuclear transfer and cell fusion reprogramming methods[8]. Therefore, they looked for factors that were known to be important for maintaining ESC fate, that were exclusively expressed in ESCs, and were known to contribute to ESC proliferation in culture. They found that a group of 24 candidates were capable of reprogramming somatic cells and systematically narrowed down the list to four transcription factors. The four 'Yamanaka' factors; OCT4, SOX2, KLF4 and c-MYC, were capable of reprogramming adult murine cells. The resulting pluripotent cells were dubbed 'induced Pluripotent Stem Cells' ('iPSCs')[8]. His colleagues around the world, including our research group, quickly confirmed and expanded his results to different mammals including humans, and different cell types including terminally differentiated T- and B-lymphocytes[43]. Many candidates beyond the selected 'reprogramming factors' OCT4, SOX2, KLF4 and c-MYC, were later found to enhance reprogramming even though they are not essential. In this thesis, we describe the Yamanaka candidates in a screen for enhancers of reprogramming (chapter 3). It was found that the state of differentiation impacts the reprogramming efficiency; hematopoietic stem cells generated more iPSCs than blood

precursors who were in turn more easily reprogrammed than B-lymphocytes[43].

To summarize, reprogramming towards a pluripotent stem cell is possible in sea urchins, but also in any mammal; from any kind of (un)differentiated somatic cell, and needs just overexpression of four transcription factors. Figure 3 gives an overview of the above historic events. However, an important question remains: why is it harder to reprogram adult cells? What are the molecular differences between differentiated somatic cells and pluripotent stem cells, and what factors control the pluripotent state? What roadblocks exist to prevent transformation to a pluripotent state? Further below we will discuss what changes take place inside a cell during reprogramming; changes that are generally more dramatic if the cell of origin has gone further down the path of differentiation. Understanding this would help us to understand why cells hold on to their cell fate normally, an important factor in optimizing iPSC generation and in our understanding of how they alter their fate in oncogenic transformation.

Impact of research into stem cells and reprogramming

Naturally, the fast developments involving direct reprogramming to pluripotency or to different somatic cells has fueled interest in stem cell-based therapies. On one hand, proof-of-principle experiments, iPSC banks, drug testing experiments and even clinical trials have moved us closer to treating patients. On the other hand dubious clinics advertise stem cell treatments that are not supported by science and potentially dangerous; some serious concerns need to be overcome before pluripotent- or somatic stem cells can be applied that way[44]. In this thesis, we describe efforts to reproduce the promising and simply spectacular method of reprogramming by stressors to generate 'STAP' cells and -stem cells. Unfortunately, faults in the original data were discovered and independent laboratories were unable to reproduce the results. This thesis aims to better understand the mechanism of reprogramming, particularly the roadblocks to forming iPSCs, for several reasons:

To understand cell fate changes

The interest in cell fate changes occurring in vivo and that are induced in vitro seem to be stimulated by breakthroughs in reprogramming to pluripotent stem cells. Transdifferentiation entails an experimentally induced switch from one somatic cell type to another somatic cell type, without having to pass through a pluripotent state. Experiments of transdifferentiation were first described in 1987 when overexpression of the muscle transcription factor MyoD was found to turn fibroblasts directly into myoblasts[45]. Recently this line of research was revived, when the demonstrated plasticity induced by a handful of transcription factors used in iPSC derivation inspired transdifferentiation protocols using lineage specific factors.

Some of these protocols do require cells to go through a transient pluripotent state or can produce transiently pluripotent cells, possibly affecting safe application in the clinic[46]. There is also increased interest and understanding of plasticity in somatic tissues (in-vivo transdifferentiation), with labs showing remarkable ability of somatic cells to replace lost cells in their niche through transdifferentiation in liver, lung, intestine and even the inner ear. Mouse studies showed that inhibition of NOTCH signaling, an important factor in embryonic development, enables transdifferentiation of cochlear supporting cells to outer hair cells[41]. The latter are lost as a result from acoustic overexposure and age which causes irreversible hearing loss. In vivo transdifferentiation to replace lost hair cells could be the first curative approach to sensory hearing loss and is currently being tested in humans. Several molecules that were found to enhance reprogramming to iPSCs also enhance transdifferentiation, such as the four reprogramming factors[47], inhibition of CAF-1, TGF-β signaling or GSK3b and supplementation of ascorbic acid[48].

To understand a process that has many commonalities with oncogenesis

As detailed below, reprogramming to pluripotent stem cells shares molecular characteristics with oncogenic transformation, such as the switch in glucose metabolism, higher proliferation rate, reliance on oncogenes and a germline switch. However, resulting iPSCs do not share the same propensity to accumulate mutations and loss of the ability to differentiate; meaning that they could be seen as healthy counterparts to cancer cells. They thus serve as a useful model to study dramatic cell fate changes and as a reference for how non-cancerous cells conduct themselves when undergoing cellular transformation.

To improve speed and efficacy of patient-derived PSC generation

Three clinic-oriented efforts can and to some degree have already benefitted from the iPSC technology: (i) regenerative medicine, aiming to replace damaged tissue in the patient with stem-cell derived transplants; (ii) disease modelling, using patient-derived cells to model and examine the development of diseases with poorly understood genetic predispositions; and (iii) drug testing, using these iPSC-derived disease models to predict drug response in a model close, or equal to the target patient.

An example is the idea of gene therapy; repairing genes that cause serious disease in a specific organ and giving the patient's own, fixed cells back. A trial with HSC transplantation to cure the incurable severe combined immunodeficiency syndrome (SCID) showed its potential but also resulted in leukemia in some patients, frustrating further attempts at such therapy[49]. In recent years however, a series of increasingly powerful gene-editing techniques have been developed, as well as ever more affordable and thorough DNA sequencing

techniques and awareness of culture induced mutations. Together with better availability of patient tissues through reprogramming into iPSCs, very interesting efforts to cure diseased animals have been performed. This suggests that gene therapy for severe genetic disease is back on the table[50]. The next section discusses other applications for the clinic.

To facilitate therapeutic applications

Disease modeling: Using patient-derived iPSC lines, differentiated cells can be derived at a scale unimaginable with primary patient tissue. For example, neurons can be generated from expanded iPSCs from a patient. These secondary differentiated cells can be used for drug testing of thousands or even millions of compounds, to search for existing drugs that combat the patient's disease[51]. iPSC banks hope to support research into patient-derived materials by deriving and maintaining up to 10,000 iPSC lines derived from patients for use on a not-for-profit basis[52]. Differentiation of iPSCs into neurons has been one of the most popular applications, giving access to a cell type that cannot normally be harvested through a biopsy. Another example is generation of organoids from iPSC to use for drug testing. A group treated organoids derived from iPSCs from cystic fibrosis (CF) patients with new CF drugs. Using the in vitro response of organoids on the drugs, they defined which subpopulation of patients would benefit most from which drug. They were able to link drug response to individual mutations of the CFTR gene[53]. The same group repaired the CF causing mutation with the CRISPR/CAS9 DNA editing tool to explore gene therapy for this single-mutation disease. Ideally, modern DNA editing techniques such as the CRISPR/CAS9 system will proof reliable enough to generate healthy, repaired cells to give back to the patient they were derived from.

Quality standards to be applied to research on induced pluripotent stem cells.

With the burst of publications on direct reprogramming following Yamanaka's work, some concern exists about the quality of papers published under the pressure of the competition. One of the main concerns is to guarantee true pluripotency of PSCs that are derived experimentally. Unlike most medical and psychological fields, this field is depending on data quality more than measures of quantity and proper statistics. A guideline was published in 2008 for the evaluation of pluripotent stem cells[54]. The guideline was widely supported by other stem cell groups[55] and still is valuable to determine what scrutiny murine or human PSCs should be subjected to prior to paper submission. An important notion from this guideline is the hierarchy of pluripotency tests, from easy, scalable tests that are limited in specificity, to very time-consuming and expensive tests that represent the ultimate gold standard of pluripotency (Figure 4).

For day-to-day monitoring, morphology is very helpful. Early changes in reprogramming mesenchymal cells such as fibroblasts, indicative of the mesenchymal-to-epithelial transition (MET) taking place, can be seen by light microscopy as soon as 2 days after OKSM induction and reassure that the experiment is working[56]; cells reorganizing themselves into tightly packed colonies, and stably expanding in colonies even when OKSM expression is withdrawn are good visual indicators of pluripotent stem cells[54].

The next step is a simple macroscopic stain for alkaline phosphatase, a stem cell marker that is not very specific for pluripotent stem cells but allows for macroscopic quantification of the iPSC clone yield. Strong staining of well-defined colonies after OKSM withdrawal correlates to endogenous pluripotency reporter expression, although this does not mean true pluripotency of each AP positive colony is guaranteed. A further step in demonstrating pluripotency is expression of an endogenous pluripotent stem cell specific transcription factor such as *Oct4, Sox2* or *Nanog*, conveniently with a knocked-in fluorescent reporter such as GFP. Expression of these genes is very specific to PSCs.

Further proof of pluripotency requires transcriptional analysis by RNAseq or micro-array, to establish a wider pluripotency-related transcription profile. Also, further functional testing such as in vitro directed differentiation[54] can be tested along with the ability to maintain a PSC like growth pattern in appropriate PSC media, without exogenous factors. So-called 'partially reprogrammed' iPSC can also be isolated after a reprogramming experiment and are dependent on residual expression of exogenous factors, either by continued expression of dox-inducible constructs, or upon incomplete silencing of constitutive vectors expressing OKSM. They can be converted to true iPSC by treatment with DNA demethylating small molecules[57]. Also, analysis of the transcriptional state of iPSC right after colony formation and withdrawal of exogenous factors suggests retention of some epigenetic memory of the cell-of-origin. However, this memory is erased when passaging and culturing the iPSC longer in the absence of exogenous factors[4].

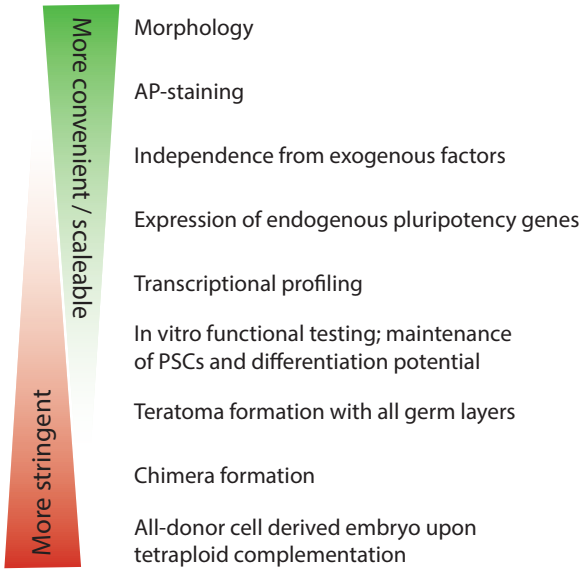


Figure 4. Testing for pluripotency, a balance between stringency and convenience.

Formation of teratomas

upon subcutaneous injection can show differentiation into cell types from all germ layers[58]. More stringently, cells can be challenged to contribute to a chimeric mouse upon blastocyst injection, for which a contrasting coat color to the host embryo is most convenient to visualize chimerism. Chimera contribution is a very convincing test for pluripotency. One test is more stringent still, but technically even more challenging: tetraploid embryo complementation. This test uses tetraploid blastocysts generated by 2-cell fusion, which are incapable of forming other than extra-embryonic tissues. PSCs injected into such an embryo should give rise to the entire mouse[59,60]. iPSC clones that easily pass all other tests can still fail at tetraploid complementation. This deficiency has been linked to aberrant silencing of a single genetic locus, overcome by adding ascorbic acid to the reprogramming media[61]. Implications of failing this ultimate test are unclear as this test is usually not required for publication.

In human cells, these high stringency tests are nearly impossible. Chimera-forming ability can be tested by generating interspecies chimeras, thus injecting human iPSC in mouse blastocysts, culturing the embryos shortly in vitro and assessing their contribution pattern and -efficiency in the very primitive embryo[62]. This procedure was found to be too inefficient for regular pluripotency testing and does not assess ability to form a viable neonate[63,64].

Events during reprogramming of somatic cells to iPSCs

With the promise and pitfalls of research into reprogramming to pluripotent stem cells in mind, we will now discuss what is known about the molecular and cellular alterations during reprogramming. In this thesis, we mainly work with murine embryonic fibroblasts (MEFs); connective-tissue cells derived from the mesoderm lineage and harvested around day 14 of embryonic development. Though embryonic, these cells are very far removed from the pluripotent state less than two weeks earlier. They differ at the levels of transcription, translation, and protein production, affecting their rate of cell growth and survival, their mode of metabolism, and their morphology (Figure 5). Here we zoom in on molecular and cellular changes that occur during OKSM induced reprogramming of MEFs to iPSCs.

Molecularly:

Transcription factors

Lineage-specific transcription factors are determinants of cell fate, depending on the context of a transcriptional network. Naturally, the transcriptional networks of MEFs and iPSCs differ strongly. Also, some transcription factors have no role in MEF or iPSC identity but affect the transition between the two cell types.

Several transcription factors were found to affect reprogramming based on their known function. When Takahashi et al realized that in the human system, early reprogramming cells at some point resemble cells from the primitive streak, they found that FOXH1, a transcription factor required for primitive streak specification, enhances reprogramming efficiency[65]. Many of the candidate reprogramming factors in their initial iPSC paper were transcription factors that, even though not essential, turned out to be upregulated during reprogramming and stimulated the induction of pluripotency. These include NANOG[66] and UTF1[67].

Reversily, the Tucker group reasoned that the transcription factor Bright/Arid3A might block reprogramming because its depletion led to increased developmental plasticity and pluripotency factor expression; indeed depletion in MEFs enhanced conventional reprogramming and even induced spontaneous reprogramming[68].

Another interesting group of transcription factors known to help induce reprogramming, and capable of replacing OCT4 or SOX2 from the 4 factor mix, consists of lineage specifiers such as GATA3, CEBPα and PAX1[69]. Notably, these factors individually steer PSCs towards differentiation to a mesendodermal or ectodermal lineage. When combined they synergistically facilitate a pluripotent state and replace reprogramming factors OCT4 and SOX2. The authors suggest that the balance between the differentiation factors results in a neutral, pluripotency-prone state. Although these factors are not upregulated in pluripotent

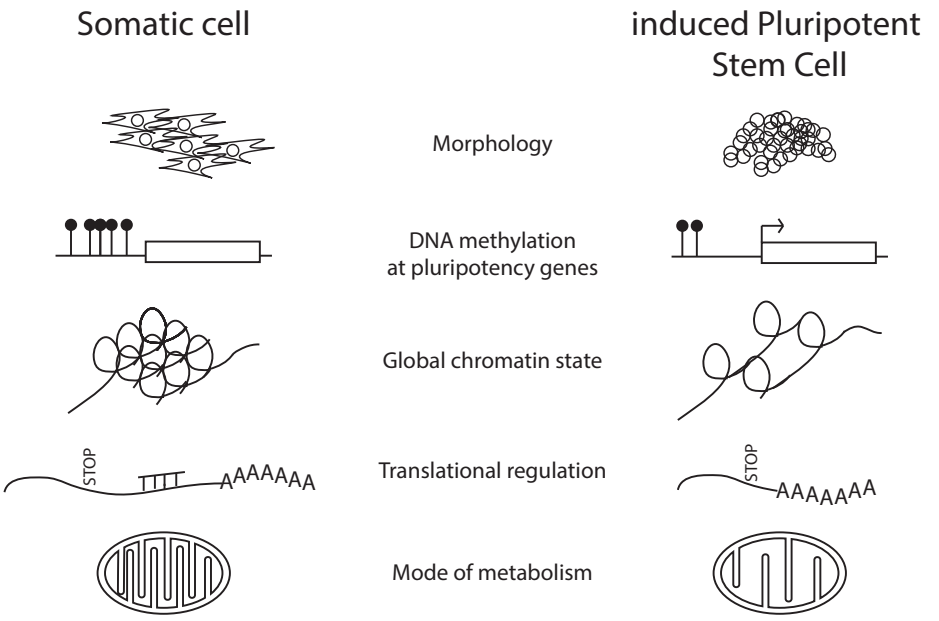


Figure 5. Comparison of somatic cells and induced pluripotent stem cells on several aspects.

stem cells, their temporary, combined overexpression steers cells towards pluripotency[69].

Transcriptional regulation: DNA methylation and chromatin structure

The epigenetic program of MEFs changes during reprogramming, meaning that transcriptional regulation is altered in an inheritable manner without altering the DNA. This epigenetic program is enacted by DNA methylation, chemical modification and composition of the chromatin around the DNA, and binding of transcription factors and non-coding RNA strands to the DNA and chromatin. A particular gene may be regulated by all different modes, affecting transcription differently: DNA methylation is highly stable but slow to change while switching out of different histone subunits by histone chaperones is more dynamic[70,71]. As a result of epigenetic regulation, target genes may be up- or downregulated, or transiently altered during intermediate stages of reprogramming[70,72,73].

In general, pluripotent cells maintain their genome in a more transcription factor accessible state, with 'open chromatin', to facilitate differentiation in many directions when prompted. On a molecular level, this is established by overall lower levels of DNA methylation at gene promotors of key developmental regulatory genes[62]. DNA methylation of gene promotors is a potent inhibitory mark to transcription of the target gene. It can be removed promptly during reprogramming via nuclear transfer, but during four-factor reprogramming is removed only partially. Marks reminiscent of the cell-of-origin can be detected in the iPSC, that only fade away after continued passaging of the iPSCs[4,74]. Remnant DNA methylation influences the differentiation potential of iPSC; there is a preference for forming tissues similar to the cell-of-origin and they contribute poorly to chimeras[3,4]. DNA methylation also influences the ability to be reprogrammed; higher overall levels of DNA methylation reduced nuclear transfer efficiency[75].

Induced pluripotent stem cells have been reported to aberrantly gain DNA methylation, for example hypermethylation of the Dlk1-Dio3 locus[6]. The resulting silencing of this locus prevents iPSCs from fulfilling the most stringent test of pluripotency: the formation of a complete mouse embryo upon injection in a tetraploid blastocyst (which itself can only form extraembryonic tissues). This aberrant silencing was overcome by culturing reprogramming cells with ascorbic acid, a cofactor to DNA and histone demethylases[6,61]. Ascorbic acid was already known to enhance reprogramming efficiency[48]. Moreover, passaged iPSC derived using ascorbic acid, a co-factor to various chromatin- and DNA methylation modifying enzymes, are indistinguishable from ESCs when compared within the same genetic background[6,61].

On a gene-by-gene level, specific loci are activated or inactivated during reprogramming. Upon the start of reprogramming, so-called 'euchromatic' regions

with somatic genes contain chromatin that is accessible to transcription but need to be deactivated. This is achieved by factors such as the polycomb repressive complex (PRC2) that deposits repressive H3K27me3 (Histone 3 Lysine 27 trimethylation) marks on the chromatin. Also, inhibition of Dot1l- which keeps these loci active through deposition of H3K79me2- facilitates somatic gene repression[76,77]. In this thesis, inhibition of DOT1l amongst other factors is shown to dramatically reduce reprogramming time (chapter 4).

Activation of heterochromatin-containing genes essential to pluripotency is facilitated by histone demethylases that remove repressive H3K9me3, or by inhibition of histone methyl transferases that deposit these repressive marks such as *G9a*, *Suv39h1* and *Suv39h2*[78-80].

Besides enzymes that deposit or remove these chemical modifications of histones, non-enzymatic DNA binding factors can also act as chromatin modulators by binding DNA and recruiting activating histone modifications. An example is Thrithorax complex member WDR5 which correlates with global levels of the activating H3K4me3 mark and is upregulated during reprogramming. *Wdr5* knockdown-fibroblasts reprogram less efficiently[81]. In chapter 5 we find that elevated WDR5 levels could explain more efficient reprogramming by knockdown of polyadenylation factor *Nudt21*.

In PSCs, one can identify three groups of loci depending on their DNA methylation- and chromatin state and corresponding accessibility for reprogramming factors[82]. For example, MEF specific and mesenchymal genes are immediately accessible by the transcription factors, and are inactivated first by chromatin modifications, their inactivation made more permanent later in the reprogramming process by DNA methylation[70]. Promotors and enhancers of early stem cell genes can be bound by some of the reprogramming factors that serve as pioneering factors to open up the chromatin. As a late step, heterochromatic regions with pluripotency-specific genes are demethylated to permit OKSM-binding and achieve stable expression of the pluripotency program. Some genes are not activated but brought into a rapid-inducible, 'bivalent' state; these are genes not expressed in pluripotent cells but can be rapidly activated as a first step in differentiation. They contain both active and repressive chromatin marks (H3K4me and H3K27me, respectively).

RNA translation; alternative polyadenylation

There is an increasing understanding of the role of translational regulation of mRNAs, influenced by secondary structure of transcripts but also by binding of micro-RNAs (miRNAs) to untranslated regions beyond the stop codons (3'UTRs). MiRNAs have been found to have profound roles during development. When the miRNA biogenesis machinery is impaired, pluripotent stem cells have reduced differentiation potential and reprogramming to iPSCs

is impaired[83,84]. Specifically, the miR290/295 and miR302/367 clusters of miRNAs are highly expressed in pluripotent cells, important for human iPSC generation and ESC proliferation, and downregulated during early embryonic development[72]. MiRNAs affect their specific target mRNA by binding the 3'UTR, thereby preventing translation or marking the transcript for degradation. Binding sites for miRNAs and regulatory RNA binding proteins (RBPs) along the 3'UTR are essential for the mRNA to be regulated in this manner.

Many 3'UTRs have multiple sites at which the UTR can be cleaved and polyadenylated by a process called alternative polyadenylation (APA)[85] (Figure 6). Comparing differentiated cells with stem cells or iPSCs; or healthy cells with cancer cells, a general tendency for shorter 3'UTRs of transcripts has been described in less differentiated and more proliferative cells. This evidence suggests that the cleavage- and polyadenylation processes are altered during reprogramming and oncogenesis[86-89]. Transcripts that are polyadenylated harbor

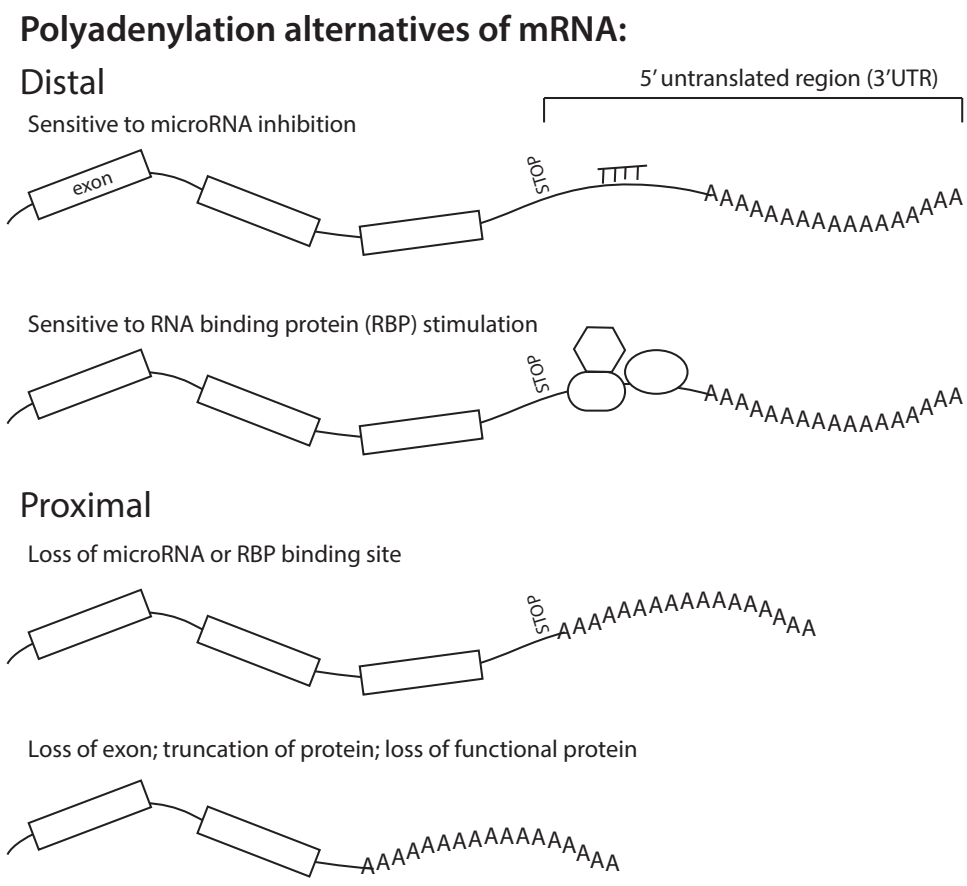


Figure 6. Alternative polyadenylation of a single mRNA molecule and regulatory consequences.

multiple polyadenylation sites (PAS), either within in the same or different exons. How the choice between multiple PAS is made is largely unknown; different models suggest a host of different influences, such as RNA binding regulatory proteins, the affinity of the APA machinery for the PAS, and transcription factors that recruit APA machinery components.

Concentration of components of the cleavage- and polyadenylation complex can influence PAS choice: this has been shown previously for cleavage factor CFIm68[135]. Depletion of CFIm68 leads to a widespread shift to more proximal PAS. In this thesis, we find that downregulation of a component of the APA machinery, CPSF5/CFIm25, encoded by *Nudt21*, enhances reprogramming efficiency of MEFs into iPSCs (chapter 5). Transcriptome-wide analysis of PAS site choice (PAS specific RNAsequencing; 'PAS-Seq') reveals APA occurs upon *Nudt21* suppression. This occurs for many transcripts, including transcripts involved in reprogramming or pluripotency.

Proteome

The overall proteome of MEFs, reprogramming- and pluripotent cells is harder to study than the genome or transcriptome due to the limitations of proteomic analysis. Recent advances in multiplexing proteomics samples and improved data analysis for mass spectrometry promise to change this[90]. It is very important nonetheless as RNA levels may have no relation to resulting protein levels due to translational regulation, post-translational chemical modifications, protein tagging with ubiquitin or sumo peptides and protein degradation. A study directly comparing protein and RNA abundance found that protein abundance was primarily determined at the level of translation[91]. Moreover, organization of proteins in compartments of the cell may define the impact of a protein on the cell's identity, as exemplified by B-catenin localization in gut stem cells[92,93].

Proteomic analysis of reprogramming intermediates, similar to the molecular characterization of intermediates in chapter 3, has shown that pathways and associated proteins are regulated during different phases of reprogramming and that these proteins are important inhibitors or facilitators of reprogramming. An example is the nucleoporin NUP210 which is essential for reprogramming into iPSCs and was identified by proteomic analysis of reprogramming cells[73].

Sumoylation

One particularly interesting protein modification is the tagging by the ubiquitin-like sumo peptide, a process called 'sumoylation'. One or multiple SUMO peptides can be attached to a lysine, usually within a consensus sequence of the target protein. This posttranslational modification is referred to as 'sumoylation'. This pathway is highly conserved; SUMO precursor

proteins are expressed in all eukaryotes that have been tested so far[84]. Sumoylation can affect target protein trafficking, stability, interaction with other proteins or DNA; depending on the multiplicity and paralog (SUMO1, SUMO2/3, SUMO4) of the SUMO tag[85-87] (Figure 7). The SENP group of enzymes can remove SUMO from its target. The SUMO paralogs are functionally redundant to some degree; however, in specific contexts, different peptides may have an opposite effect on the target protein; see examples below.

In this thesis, we find that depletion of SUMO2 increases reprogramming efficiency and speed. However, the abundance of SUMO2 target proteins makes narrowing down the mechanism difficult. SUMO2 was previously described to be essential for embryonic development, unlike SUMO1 or SUMO3, suggesting a regulatory role for SUMO2 sumoylation in the early embryo[88]. Another study in HeLa cells found that active sumoylation inhibits expression of RNAs required for polymerases, histones, and protein biogenesis genes. Also, the SUMO machinery seemed to play a role in repressive chromatin found in senescent cells; suggesting that sumoylation of chromatin associated factors inhibits cell growth and proliferation[89]. However, no difference in cell proliferation was found upon downregulation of SUMO2 during reprogramming of embryonic fibroblasts (Chapter 4, this thesis). One report describes that preventing sumoylation by all SUMO variants through downregulation of the common sumoylation enzyme UBC9 inhibits reprogramming. However, different SUMO paralogs may have opposing effects, and loss of all sumoylation may be toxic[90,91], explaining why the effect is opposite to specific SUMO2 downregulation. Also, the same study

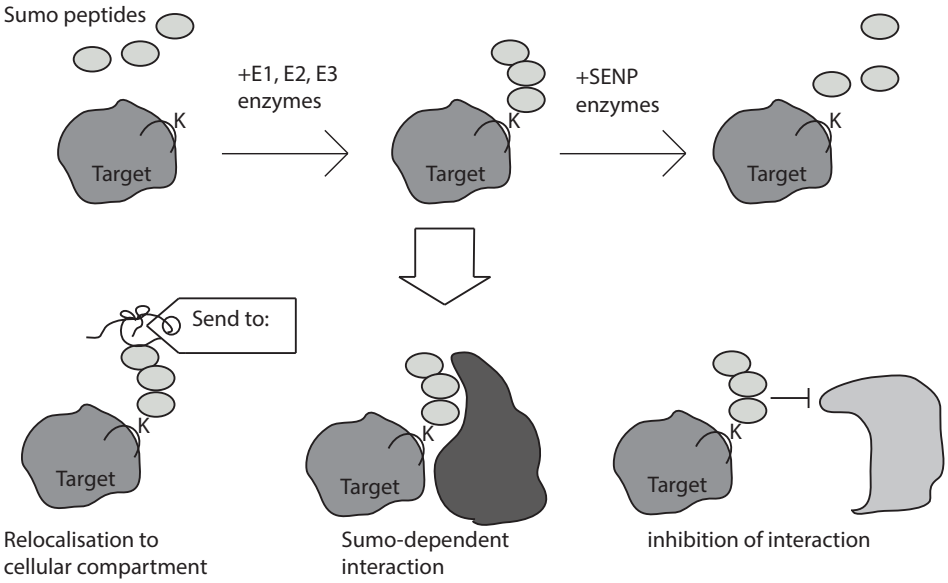


Figure 7. Protein-level regulation by sumoylation.

describes a loss of pluripotency in UBC9 depleted ESCs, which may reduce reprogramming yield upon reaching the iPSC stage. We found that iPSC derived upon SUMO2 depletion had no apparent problem in maintaining pluripotency or giving rise to healthy high grade chimeric mice. This points to the importance to regard sumoylation, not as a single mode of regulation. Instead one can specify the effects of the different Sumo paralogues, the site of sumoylation and the multitude of sumo peptides that are conjugated to the target protein. For example, many studies use antibodies or use genetic models that do not distinguish between SUMO paralogs SUMO1 and SUMO2.

Notably, sumoylation has been implicated in different methods of repression of some genes and of activation of transcription of other genes through DNA methylation or chromatin modification[92]. Some observations that point to a possible mechanism in reprogramming are:

- Inhibition of transcriptional synergy of C/EBPα, a transcription factor previously implied in facilitating reprogramming of B-cells into iPSCs[94];
- Increasing stability of GSK3b, a signaling molecule whose inhibition enhances reprogramming[95,96];
- Inhibition of c-MYB activity and downregulation of its targets such as reprogramming factor c-MYC upon stress; downregulation of SUMO-2 might enable higher expression of c-MYB targets[97].
- Stimulation of SMAD4 transcriptional activity, therefor stimulating TGF-β signaling; inhibition of TGF-β is known to enhance reprogramming[95];
- Roles in both transcriptional activation and repression could favor a differentiated state, whereby inhibition of SUMO-2 could facilitate transcriptional repression and activation required for reprogramming to iPSCs[100].

Cellular signaling

Cell fate determining transcription factors and chromatin modulators have been the focus of many studies looking into the mechanisms behind reprogramming. Regardless, signaling pathways responding to external signals play a major role in facilitating or inhibiting reprogramming of MEFs to iPSCs. One of the earliest examples is the TGF-β pathway, inhibition of which by a small molecule inhibitor or by the antagonistic factors BMP4 or SMAD7, boosts reprogramming[96,97]. Of course, this fits with the necessity to undergo mesenchymal-to-epithelial transition, since TGF-β signaling is upregulated in mesenchymal cells.

Another important signaling pathway implied in reprogramming is the WNT pathway. WNT is an important factor both in stem cell self-renewal and somatic cell differentiation.

According to a study which dissected WNT's role during early and late phases of reprogramming, WNT inhibits early reprogramming steps but enhances the later steps[98].

In this thesis (addendum to chapter 3) we describe a role for embryonic stem cell-specific Ras (*ERas*), a constitutively active RAS family member expressed in pluripotent stem cells and upregulated during reprogramming. Overexpression of *ERas* boosts reprogramming efficiency, independent of the growth factor presence needed by somatic RAS members *KRas* and *NRas*.

Metabolism

As mentioned above, altering the mode of metabolism is essential to facilitate the rate of cell proliferation. An important metabolic switch that occurs during reprogramming of MEF, is from predominant oxidative phosphorylation of glucose, which yields more cellular energy and just CO2 and water as end products, to predominantly aerobic glycolysis, which yields less energy but produces building blocks for fatty acids, amino acids and nucleotides[99]. Important glycolysis regulating factors such as hexokinase are upregulated during reprogramming; and inhibition of glycolysis inhibits reprogramming[99]. Importantly the pentose phosphate pathway (PPP), which leads to the production of nucleotides, is activated during reprogramming; the PPP is fueled by aerobic glycolysis. Threonine-glycine metabolism is also activated to fuel the 1-carbon pathway to give rise to S-Adenyl-methionine, a universal methyl donor necessary for chromatin- and DNA methylation. This in turn is implicated in facilitating activating chromatin modifications (H3K4 trimethylation) for a pluripotent state[100].

In this thesis, we study the oncogenic mutations in the important metabolic enzymes Isocitrate Dehydrogenase 1 and -2 (IDH1 and IDH2). These mutations alter the product of the enzyme and as a consequence the 'oncometabolite' 2-HG is produced at high levels. Mutant IDH (mIDH) has been implied in impaired differentiation in leukemic cells, possibly through interfering with establishment of a differentiated chromatin state by epigenetic regulators that are inhibited by 2-HG. Other effects of mutant IDH, such as disturbance of metabolic pathways, cannot be ruled out. Our analyses suggest that early steps of reprogramming are facilitated by mIDH expression; later steps of establishing pluripotency appear mostly unaffected. Loss of MEF identity as monitored by *Thy1* expression is enhanced in mIDH expressing reprogramming cells. This could be due to aberration of metabolic pathways as well as interference with 2-HG sensitive epigenetic regulators (chapter 6).

Genomic alterations during reprogramming

There have been various concerns regarding the genomic alterations brought on by reprogramming somatic cells into iPSCs. The original reprogramming protocols required infection of MEFs with viral vectors carrying the four reprogramming factors8. This results in random viral insertions in the genome, potentially disturbing endogenous genes or regulatory sequences. This raised the suspicion that specific insertions were required to form iPSC. Our lab did find that some insertion sites were common amongst multiple independent iPSC lines, such as one that perturbs TGF-β signaling, leading to the discovery of its negative role in reprogramming[95]. However, no insertion sites were common to all iPSC lines excluding the requirement for particular DNA mutations during iPSCs derivation[101]. Later, successful iPSC derivation without viral vectors, using mRNA, protein, or episomal vectors to deliver the four factors, further strengthened this point[102-105].

To study general genomic damage in iPSC and iPSC-derived transplantable tissues, reprogramming cells and resulting iPSC have been examined for genomic stability and random mutations during the prolonged time in culture required for deriving and expanding iPSCs. Of a considerable amount of mutations, in iPSC lines, a large proportion can be traced back to the parent cell line[106,107]. The remaining mutations however seem to be caused specifically by the reprogramming process[108]. Thus, careful analysis of iPSC clones seems to be required to ensure safety of iPSC-derived transplants for patients and to avoid bias in experiments with iPSC-derived cells.

On the cellular level:

Morphology

By far the most obvious change during reprogramming is the transition from large, spread-out, highly structured fibroblasts to small, round, dense colony-forming pluripotent cells. Under the light microscope, unstained iPSC can no longer be observed individually; the clonally expanding cells form three-dimensional clumps, densely packed together. Macroscopic morphology is not a specific identifier of their pluripotent state, but facilitates monitoring and picking iPSC colonies for expansion. However, one can study morphology at the nanoscale using electron microscopy (EM); this has been done for murine[109] and human[110] pluripotent stem cells. These groups found iPSCs to be indistinguishable from ESCs. When slicing through an iPSC colony with EM, they describe a more regular, smaller and round cell compared to the fibroblast of origin. Also they note that like ESCs, iPSCs form stronger adhesions to each other, and less to their surrounding fibroblasts or cell culture surface. MEFs on the other hand primarily attach to the culture surface, not each other.

They also noted that mitochondria in ESCs and iPSCs are more rounded in shape with less internal membranes compared to MEF, possibly linked to their altered metabolic state and overall reduced mitochondrial activity. To illustrate the distinct mitochondrial function in stem cells: mitochondrial staining can be used to distinguish hematopoietic stem cells from their progeny in cord blood[111]. However, this has not been shown to help characterizing pluripotent stem cells.

Cell growth and survival

Although embryonic fibroblasts still readily divide; they gradually slow down their rate of proliferation, to retreat to a senescent state eventually (depending on derivation and culture conditions, after 8 passages or later). Senescence is an irreversible quiescent state possibly followed by apoptosis and is an important mechanism for somatic cells to prevent malignant transformation upon cellular aging[112]. Immortalization by downregulation of the apoptosis machinery through knockdown of *Trp53* is a major roadblock of reprogramming[113]. Moreover, it has been suggested that any cell will reprogram eventually if it goes through enough cell divisions, when using cells that will not senesce[114].

Pluripotent cells continue to proliferate when cultured in conditions inhibitory to their differentiation. During reprogramming, induced pluripotent cells reactivate their telomerase to maintain telomere length[115], downregulate pro-apoptotic factors[113], and adjust their mode of metabolism to facilitate cell growth[116]. Their rate of proliferation is increased so much, that by the time the reprogramming process is completed, the resulting iPSCs have started to outnumber the reprogramming-resistant fibroblasts by far, despite low reprogramming efficiencies.

Germline switch; mesenchymal-to-epithelial transition (MET)

At several stages during development, starting with gastrulation, embryonic cells go through an epithelial-to-mesenchymal transition (EMT) or mesenchymal-to-epithelial transition (MET). A mesenchymal-to-epithelial transition is required for reprogramming of mesenchymal fibroblasts to the epithelial-like pluripotent stem cells[56,117,118]. *Cdh1*, which encodes E-cadherin, is a major regulator of the epithelial state. *Cdh1* is downregulated during EMT and involved in regulation of pluripotency[119]. Mesenchymal transcription factors such as SNAIL, TWIST and ZEB as well as mesenchymal specific surface markers THY1 and CD44 are downregulated during reprogramming[118,120]. EMT inducer TGF-β blocks reprogramming and inhibition of its receptor enhances reprogramming[95]. Moreover, EMT-related miRNAs are downregulated and MET-related miRNAs upregulated[56]. In conclusion, early essential steps of reprogramming fibroblasts to pluripotent stem cells closely resemble the MET

required for normal development.

Epithelial cells do not require MET to form iPSCs. Indeed, several studies comparing reprogramming efficiency of cells of epithelial origin (e.g., keratinocytes) to cells of mesenchymal origin (e.g., fibroblasts) found that epithelial cells reprogram with a much greater efficiency[121,122]. However, others found no difference[54]. Also, a direct epithelial to mesenchymal cell comparison is difficult since these cell types differ in more aspects than just their origin. Also, reprogramming factor *Klf4* is highly expressed in epithelial cells, explaining why epithelial cell reprogramming does not require exogenous KLF4 expression[56] and hypothetically contributing to a higher reprogramming efficiency of epithelial cells in general. To our knowledge, no study was performed comparing epithelial and mesenchymal cell types on a molecular level during reprogramming towards iPSC.

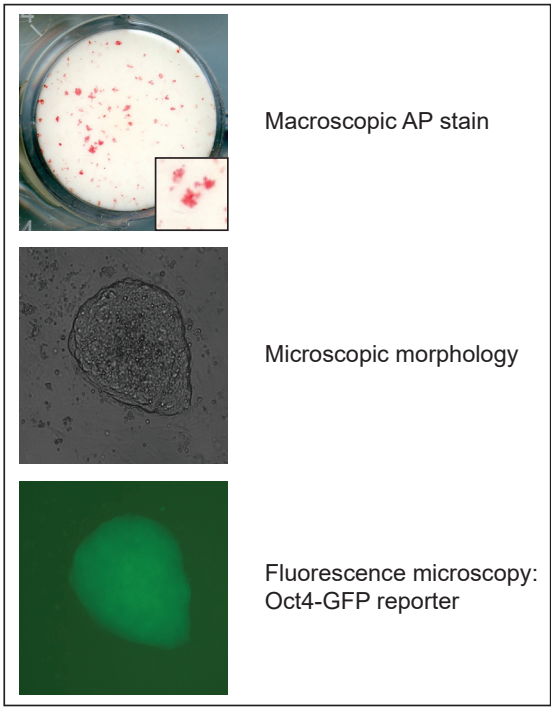


Figure 8. Characterisation methods of pluripotent stem cell colonies in culture.

Experimental parameters of reprogramming: Markers to follow reprogramming of MEFs into iPSCs

Certain proteins are of importance not just to understand the biology of reprogramming, but as reference points for studying it. We use several proteins as markers for different stages of reprogramming or as pluripotency markers. Cell surface markers are useful in combination with live cell staining and flow cytometry. Defined reprogramming intermediates have been isolated based on a combination of cell-surface markers[70,123]. The first intermediates lose expression of MEF-marker THY1, then they start expressing stem cell markers Alkaline Phosphatase (AP), SSEA1 and later EPCAM and PECAM. Finally, we use a pluripotency-specific transcription factor such as *Oct4* through a knock-in of a GFP reporter at the *Oct4* locus, as a sign of reaching pluripotency. Other pluripotency reporter systems use *Nanog* or *Sox2*[124]. For macroscopic quantification of reprogramming efficiency, the stem cell specific expression of Alkaline Phosphatase (AP) can be used in a

substrate-based assay, giving strong red staining to iPSCs colonies so they can be counted without a microscope (Figure 8; see section on quality standards above). This assay is not as stringent as pluripotency- transcription factor-reporters as it is expressed very early during reprogramming, well before establishment of a pluripotent state. However, in combination with the macroscopic colony morphology and persistence beyond withdrawal of ectopic reprogramming factors, it correlates well with expression of pluripotency reporters such as *Oct4*-GFP.

Experimental parameters of reprogramming: Independence from exogenous factors

When doxycycline is withdrawn from the culture, only bona fide iPSCs that have fully activated their endogenous pluripotency network, can maintain stem cell phenotype, whereas any partially reprogrammed cells or reprogramming intermediates regress back to a MEF-like state within 48 hours (Mikkelsen et al[57] and own observations). Thus, it is important to note that the transition is stable and independent from expression of the exogenous factors used to induce reprogramming. For example, exogenous *Oct4* expression could activate the endogenous locus by itself through a positive feedback loop[64], but cells are not considered fully reprogrammed iPSC until they keep expressing endogenous *Oct4* after withdrawal of exogenous *Oct4*. The molecular and cellular characteristics of iPSC described above, are not dependent on exogenous signals other than those present in ESC growth conditions, once they are established during reprogramming.

A view of oncogenic transformation from a reprogramming perspective

Similarities between somatic cell reprogramming and oncogenic transformation: germline switches

Cancer cells use mechanisms of transformation normally only seen during development- or during somatic cell reprogramming. Somatic cell reprogramming to pluripotency results in embryonic stem cells capable of forming healthy adult animals; oncogenic transformation results in cancer cells that form unorganized tumors and acquire many cellular aberrations in the process. We study stem cells to generate healthy new tissues from them and cancer cells to learn how to kill them. Nevertheless, these cells and the processes that form them share important molecular characteristics.

Here we will shortly discuss similarities with oncogenic transformation. A well-known similarity is the acquisition immortality; the ability to divide without senescing; along with

a heightened proliferation rate. This leads to expansion of the cell population, regardless of growth signals from the environment or endogenous signals of DNA instability in the case of cancer. Molecularly there are many parallels as well:

DNA methylation and modulation of chromatin structure

Although cancer is still considered a disease of the DNA, caused by DNA mutations and progressive genomic instability, the role of epigenetic reprogramming during transformation has been recognized as well. In a proof-of-principle experiment, a Japanese group reprogrammed cells in vivo using the Yamanaka factors in an inducible reprogrammable mouse model similar to our model. They found that by withdrawing the reprogramming factors before iPSC had formed in the mice, they instead got tumors in various tissues with widespread DNA methylation changes; however cells from the same experiment could still be reprogrammed into iPSC and generate normal mouse tissue suggesting no oncogenic DNA mutations had occurred[125]. These results made the point that epigenetic aberrations alone may suffice to turn somatic cells into cancer cells. Widespread epigenetic alterations found in cancer cells as well as mutations in epigenetic regulators support this point[126]. For example, cancer-associated hypermethylation was found to occur very frequently at target sites of the repressive polycomb complex 2 (PC2), which is an essential repressor of differentiation associated genes in pluripotent stem cells[127].

At the same time, global hypomethylation, is a hallmark of cancer as well as pluripotent stem cells. DNA methyltransferases (DNMT1, DNMT3A) which maintain or establish DNA methylation are lost in some cancers or experimentally found to support malignancy when inhibited[128-130].Inhibition of DNMT1 enhances reprogramming to iPSCs[131]. Finally, cancer cells, in this case melanoma, have been found to use epigenetic reprogramming to temporarily gain independence from pathways inhibited by cancer drugs; this therapy resistance is lost when treatment is paused due to the growth advantage inferred by the target pathways. This way, cancer cells use epigenetic programming to adapt to the environment[132].

Translational regulation

Regulation of protein production at the RNA level is also similar between stem cells and cancer cells in some cases. For example, alternative polyadenylation which determines protein production by manipulating miRNA binding sites or truncating proteins, occurs also in cancer cells, upregulating oncogenes by shortening their 3'UTR[86].

Metabolic switches

Transforming and reprogramming cells share an important switch of glucose metabolism: from oxidative phosphorylation to aerobic glycolysis. Cancer cells make this switch to facilitate cell division, called the 'Warburg effect' after its discoverer Otto Warburg[133]. Glycolysis increases glucose uptake in tumors to such an extent that can be used for diagnostics. To monitor cancer metastasis, imaging where radioactive labeled glucose is highly used provides a clear image (PET scan).

Scope of this thesis

The introduction discusses progress in the field of induced pluripotency in the 10 years since the discovery of iPSCs, about general stem cell and reprogramming fields 100 years prior, and what we have learned about the biology of pluripotent stem cells and possible applications. This thesis aims to expand our understanding of reprogramming towards a stem cell state and use this as a model for oncogenesis.

A model for reprogramming murine somatic cells into pluripotent stem cells

In chapter 2 we describe a model system that enabled studying reprogramming to pluripotency in a more reproducible and standardized way than was possible before. This model system, named the 'reprogrammable mouse' was used in all following projects to generate murine iPSCs. During the development of this model system I supported the first author with verification experiments and cell culture.

Defining transitions during reprogramming; Screening for enhancers of reprogramming

Chapter 3 describes efforts from our lab to characterize intermediate stages of reprogramming into iPSCs on an RNA and chromatin level. The addendum describes a more in-depth, systematic candidate-based screen for enhancers of reprogramming. This screen identified the embryonic specific signaling protein ERas as a potent enhancer of reprogramming, validating the finding that this factor is upregulated in reprogramming intermediates in chapter 3.

Screening for inhibitors of reprogramming

In chapter 4 and 5 we sought to perform an unbiased screen for protein-coding genes involved in reprogramming. We applied a genome-wide, RNA interference screens, and

design a screening method to be able to apply this during reprogramming into iPSCs. The addendum specifies the dilemmas and considerations required to perform a large scale screen. The screen identified the sumoylation peptide SUMO2 and the alternative polyadenylation factor CPSF5 (*Nudt21)* as important inhibitors of reprogramming. Reduction of SUMO2 levels allowed for more efficient and faster reprogramming in mouse and human model systems, independent of known reprogramming-enhancing pathways, suggesting an important role for SUMO2-sumoylation in maintaining cell fate. CPSF5 is a regulator of 3'UTR length by determining the site of cleavage and polyadenylation of 3' UTRs of many messenger RNAs, thereby modulating sensitivity to microRNAs and protein length in the case of intronic polyadenylation sites.

Modeling cancer during reprogramming

In chapter 6 we apply the 'reprogrammable mouse' model system to study an oncogenic mutation in a metabolic enzyme. Mutations in isocitrate dehydrogenases (IDH) 1 and 2 occur frequently in glioblastomas, leukemias and cholangiocarcinomas. Mutations in IDH1 or IDH2 are dominant and neomorphic; wildtype IDH generates the metabolite α-KG, but mutant IDH generates the highly similar oncometabolite 2-HG. Here we characterize early reprogramming cells as a model for oncogenic transformation, when they express mutant IDH1 and IDH2. We look both at transcriptional and metabolic perturbations as well as functional outcome. The results suggest that mutant IDH derails reprogramming, facilitating loss of cell identity but not promoting gain of pluripotency. Ascorbic acid is a co-factor to α-KG dependent enzymes including DNA demethylating TET enzymes and Jumanji domain-containing histone demethylases[134], Partial reversal of the phenotype by ascorbic acid strengthens the hypothesis that specific epigenetic regulators are affected by mutant IDH. Finally, we specify which metabolic pathways are affected transcriptionally and metabolically by the mutant enzymes.

Assessing a role for cellular stress during reprogramming

In chapter 7 we revisit a new approach to establishing pluripotency, as was described to occur by applying mechanical or chemical stress to somatic cells. Unfortunately, several stem cell laboratories including our own showed that this method was irreproducible and the original data a result of artifact or sample mix-ups. This chapter highlights a vulnerability in the way researchers and journals assess new discoveries, which is a consequence of the high interest in the stem cell field combined with pressure to publish high impact manuscripts. Moreover, it documents a large and swift collaboration between normally competing researchers to set the record straight.

References

1. De Los Angeles, A. et al. Hallmarks of pluripotency. Nature 525, 469–478 (2015).
2. Flores, I., Benetti, R. & Blasco, M. A. Telomerase regulation and stem cell behaviour. Curr. Opin. Cell Biol. 18, 254–260 (2006).
3. Kim, K. et al. Epigenetic memory in induced pluripotent stem cells. Nature 467, 285–290 (2010).
4. Polo, J. M. et al. Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. Nat Biotechnol 28, 848–855 (2010).
5. Choi, J. et al. A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. Nat Biotechnol 33, 1173–1181 (2015).
6. Stadtfeld, M. et al. Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. Nature 465, 175–181 (2010).
7. Thomson, J. A. Embryonic Stem Cell Lines Derived from Human Blastocysts. Science 282, 1145–1147 (1998).
8. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell 126, 663–676 (2006).
9. Ramalho-Santos, M. & Willenbring, H. On the Origin of the Term 'Stem Cell'. Cell Stem Cell 1, 35–38 (2007).
10. Stevens, L. C. The development of transplantable teratocarcinomas from intratesticular grafts of pre- and postimplantation mouse embryos. Dev. Biol. 21, 364–382 (1970).
11. Stevens, L. C. Embryology of testicular teratomas in strain 129 mice. J. Natl. Cancer Inst. 23, 1249–1295 (1959).
12. Mintz, B. & Illmensee, K. Normal genetically mosaic mice produced from malignant teratocarcinoma cells. Proc. Natl. Acad. Sci. U.S.A. 72, 3585–3589 (1975).
13. Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. Nature 292, 154–156 (1981).
14. Martin, G. R. 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 (1981).
15. Chambers, I. & Tomlinson, S. R. The transcriptional foundation of pluripotency. Development 136, 2311–2322 (2009).
16. Silva, J. et al. Nanog Is the Gateway to the Pluripotent Ground State. Cell 138, 722–737 (2009).
17. Niwa, H., Smith, A. G. & Miyazaki, J.-I. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nature Genetics 24, 372–376 (2000).

18. Masui, S. et al. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. Nat Cell Biol 9, 625–635 (2007).
19. Suda, Y., Suzuki, M., Ikawa, Y. & Aizawa, S. Mouse embryonic stem cells exhibit indefinite proliferative potential. J. Cell. Physiol. 133, 197–201 (1987).
20. Murry, C. E. & Keller, G. Differentiation of Embryonic Stem Cells to Clinically Relevant Populations: Lessons from Embryonic Development. Cell 132, 661–680 (2008).
21. Li, M. & Belmonte, J. C. I. Ground rules of the pluripotency gene regulatory network. Nat Rev Genet 18, 180–191 (2017).
22. Thomson, M. et al. Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell 145, 875–889 (2011).
23. Avilion, A. A. et al. Multipotent cell lineages in early mouse development depend on SOX2 function. Genes Dev. 17, 126–140 (2003).
24. Hamazaki, T., Oka, M., Yamanaka, S. & Terada, N. Aggregation of embryonic stem cells induces Nanog repression and primitive endoderm differentiation. J. Cell. Sci. 117, 5681–5686 (2004).
25. Chambers, I. et al. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 113, 643–655 (2003).
26. Ying, Q. L., Nichols, J., Chambers, I. & Smith, A. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. Cell 115, 281–292 (2003).
27. Suzuki, A. et al. Maintenance of embryonic stem cell pluripotency by Nanog-mediated reversal of mesoderm specification. Nat Clin Pract Cardiovasc Med 3 Suppl 1, S114–22 (2006).
28. Niwa, H. et al. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. Cell 123, 917–929 (2005).
29. Matsuda, T. et al. STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells. EMBO J. 18, 4261–4269 (1999).
30. H, S. Entwickelungsphysiologische Studien am Tritonei II. (1902).
31. Fässler, P. E. 40, 49–57 (1996).
32. Briggs, R. & King, T. J. Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. Proc. Natl. Acad. Sci. U.S.A. 38, 455–463 (1952).
33. Gurdon, J. B., Elsdale, T. R. & Fischberg, M. Sexually mature individuals of Xenopus laevis from the transplantation of single somatic nuclei. Nature 182, 64–65 (1958).
34. Bromhall, J. D. Nuclear transplantation in the rabbit egg. Nature 258, 719–722 (1975).
35. Willadsen, S. M. & Godke, R. A. A simple procedure for the production of identical sheep twins. Vet. Rec. 114, 240–243 (1984).
36. Meng, L., Ely, J. J., Stouffer, R. L. & Wolf, D. P. Rhesus monkeys produced by nuclear

transfer. Biol. Reprod. 57, 454–459 (1997).

37. Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. S. Viable offspring derived from fetal and adult mammalian cells. Nature 385, 810–813 (1997).
38. Hochedlinger, K. & Jaenisch, R. Monoclonal mice generated by nuclear transfer from mature B and T donor cells. Nature 415, 1035–1038 (2002).
39. Miller, R. A. & Ruddle, F. H. Pluripotent teratocarcinoma-thymus somatic cell hybrids. Cell 9, 45–55 (1976).
40. Cowan, C. A. Nuclear Reprogramming of Somatic Cells After Fusion with Human Embryonic Stem Cells. Science 309, 1369–1373 (2005).
41. Tada, M. Embryonic germ cells induce epigenetic reprogramming of somatic nucleus in hybrid cells. EMBO J. 16, 6510–6520 (1997).
42. Tada, M., Takahama, Y., Abe, K., Nakatsuji, N. & Tada, T. Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. Current Biology 11, 1553–1558 (2001).
43. Eminli, S. et al. Differentiation stage determines potential of hematopoietic cells for reprogramming into induced pluripotent stem cells. Nature Genetics 41, 968–976 (2009).
44. Barrilleaux, B. & Knoepfler, P. S. Inducing iPSCs to Escape the Dish. Cell Stem Cell 9, 103–111 (2011).
45. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell 51, 987–1000 (1987).
46. Bar-Nur, O. et al. Lineage conversion induced by pluripotency factors involves transient passage through an iPSC stage. Nat Biotechnol 33, 761–768 (2015).
47. Efe, J. A. et al. Conversion of mouse fibroblasts into cardiomyocytes using a direct reprogramming strategy. Nat Cell Biol 13, 215–222 (2011).
48. Esteban, M. A. et al. Vitamin C Enhances the Generation of Mouse and Human Induced Pluripotent Stem Cells. Cell Stem Cell 6, 71–79 (2010).
49. Howe, S. J. et al. Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J. Clin. Invest. 118, 3143–3150 (2008).
50. Daley, G. Q. The promise and perils of stem cell therapeutics. Cell Stem Cell 10, 740–749 (2012).
51. Engle, S. J. & Vincent, F. Small Molecule Screening in Human Induced Pluripotent Stem Cell-derived Terminal Cell Types. Journal of Biological Chemistry 289, 4562–4570 (2014).
52. De Sousa, P. A. et al. Rapid establishment of the European Bank for induced Pluripotent Stem Cells (EBiSC) - the Hot Start experience. Stem Cell Res 20, 105–114 (2017).
53. Dekkers, J. F. et al. A functional CFTR assay using primary cystic fibrosis intestinal organoids. Nat. Med. 19, 939–945 (2013).

54. Maherali, N. & Hochedlinger, K. Guidelines and Techniques for the Generation of Induced Pluripotent Stem Cells. Cell Stem Cell 3, 595–605 (2008).
55. Daley, G. Q. et al. Broader Implications of Defining Standards for the Pluripotency of iPSCs. Cell Stem Cell 4, 200–201 (2009).
56. Li, R. et al. A Mesenchymal-to-Epithelial Transition Initiates and Is Required for the Nuclear Reprogramming of Mouse Fibroblasts. Cell Stem Cell 7, 51–63 (2010).
57. Mikkelsen, T. S. et al. Dissecting direct reprogramming through integrative genomic analysis. Nature 454, 49–55 (2008).
58. Gertow, K. et al. Isolation of human embryonic stem cell-derived teratomas for the assessment of pluripotency. Curr Protoc Stem Cell Biol Chapter 1, Unit1B.4–1B.4.29 (2007).
59. Jaenisch, R. & Young, R. Stem Cells, the Molecular Circuitry of Pluripotency and Nuclear Reprogramming. Cell 132, 567–582 (2008).
60. Wernig, M. et al. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. Nature 448, 318–324 (2007).
61. Stadtfeld, M. et al. Ascorbic acid prevents loss of Dlk1-Dio3 imprinting and facilitates generation of all–iPS cell mice from terminally differentiated B cells. Nature Genetics 44, 398–405 (2012).
62. Mascetti, V. L. & Pedersen, R. A. Human-Mouse Chimerism Validates Human Stem Cell Pluripotency. Cell Stem Cell 18, 67–72 (2016).
63. Masaki, H. et al. Interspecific in vitro assay for the chimera-forming ability of human pluripotent stem cells. Development 142, 3222–3230 (2015).
64. Theunissen, T. W. & Jaenisch, R. Molecular control of induced pluripotency. Cell Stem Cell 14, 720–734 (2014).
65. Takahashi, K. et al. Induction of pluripotency in human somatic cells via a transient state resembling primitive streak-like mesendoderm. Nature Communications 5, 3678 (2014).
66. Theunissen, T. W. et al. Nanog Overcomes Reprogramming Barriers and Induces Pluripotency in Minimal Conditions. Current Biology 21, 65–71 (2011).
67. Yang, C.-S., Chang, K.-Y. & Rana, T. M. Genome-wide Functional Analysis Reveals Factors Needed at the Transition Steps of Induced Reprogramming. Cell Reports 8, 327–337 (2014).
68. Popowski, M. et al. Bright/Arid3A Acts as a Barrier to Somatic Cell Reprogramming through Direct Regulation of Oct4, Sox2, and Nanog. Stem Cell Reports 2, 26–35 (2014).
69. Shu, J. et al. Induction of Pluripotency in Mouse Somatic Cells with Lineage Specifiers. Cell 153, 963–975 (2013).
70. Polo, J. M. et al. A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. Cell 151, 1617–1632 (2012).

71. De Koning, L., Corpet, A., Haber, J. E. & Almouzni, G. Histone chaperones: an escort network regulating histone traffic. Nat Struct Mol Biol 14, 997–1007 (2007).
72. Takahashi, K. & Yamanaka, S. A developmental framework for induced pluripotency. Development 142, 3274–3285 (2015).
73. Hansson, J. et al. Highly Coordinated Proteome Dynamics during Reprogramming of Somatic Cells to Pluripotency. Cell Reports 2, 1579–1592 (2012).
74. Kim, K. et al. Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nat Biotechnol 29, 1117–1119 (2011).
75. Blelloch, R. et al. Reprogramming Efficiency Following Somatic Cell Nuclear Transfer Is Influenced by the Differentiation and Methylation State of the Donor Nucleus. Stem Cells 24, 2007–2013 (2006).
76. Fragola, G. et al. Cell Reprogramming Requires Silencing of a Core Subset of Polycomb Targets. PLoS Genet. 9, e1003292 (2013).
77. Onder, T. T. et al. Chromatin-modifying enzymes as modulators of reprogramming. Nature 483, 598–602 (2012).
78. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome. Cell 151, 994–1004 (2012).
79. Sridharan, R. et al. Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1γ in reprogramming to pluripotency. Nat Cell Biol 15, 872–882 (2013).
80. Chen, J. et al. H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. Nature Genetics 45, 34–42 (2012).
81. Ang, Y.-S. et al. Wdr5 Mediates Self-Renewal and Reprogramming via the Embryonic Stem Cell Core Transcriptional Network. Cell 145, 183–197 (2011).
82. Apostolou, E. & Hochedlinger, K. Chromatin dynamics during cellular reprogramming. Nature 502, 462–471 (2013).
83. Kanellopoulou, C. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes Dev. 19, 489–501 (2005).
84. Kim, B.-M. et al. MicroRNAs Are Indispensable for Reprogramming Mouse Embryonic Fibroblasts into Induced Stem Cell-Like Cells. PLoS ONE 7, e39239 (2012).
85. Shi, Y. & Manley, J. L. The end of the message: multiple protein–RNA interactions define the mRNA polyadenylation site. Genes Dev. 29, 889–897 (2015).
86. Mayr, C. & Bartel, D. P. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. Cell 138, 673–684 (2009).
87. Masamha, C. P. et al. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. Nature 510, 412–416 (2014).

88. Ji, Z., Lee, J. Y., Pan, Z., Jiang, B. & Tian, B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proceedings of the National Academy of Sciences 106, 7028–7033 (2009).
89. Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. Science 320, 1643–1647 (2008).
90. Hansson, J. & Krijgsveld, J. Proteomic analysis of cell fate decision. Curr. Opin. Genet. Dev. 23, 540–547 (2013).
91. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. Nature 473, 337–342 (2011).
92. Nelson, W. J. Convergence of Wnt, -Catenin, and Cadherin Pathways. Science 303, 1483–1487 (2004).
93. Le, N. H., Franken, P. & Fodde, R. Tumour–stroma interactions in colorectal cancer: converging on β-catenin activation and cancer stemness. Br. J. Cancer 98, 1886–1893 (2008).
94. Subramanian, L. A Synergy Control Motif within the Attenuator Domain of CCAAT/Enhancer-binding Protein alpha Inhibits Transcriptional Synergy through Its PIASy-enhanced Modification by SUMO-1 or SUMO-3. Journal of Biological Chemistry 278, 9134–9141 (2003).
95. Maherali, N. & Hochedlinger, K. Tgfβ Signal Inhibition Cooperates in the Induction of iPSCs and Replaces Sox2 and cMyc. Current Biology 19, 1718–1723 (2009).
96. Maherali, N. https://commerce.invitrogen.com/index.cfm?fuseaction=orderComplete.receipt&orderGUID=IDB5DF3A7FF680D9C79ABB9B6F336C1AAE. 1–2 (2010).
97. Chen, J. et al. BMPs functionally replace Klf4 and support efficient reprogramming of mouse fibroblasts by Oct4 alone. Cell Res 21, 205–212 (2010).
98. Ho, R., Papp, B., Hoffman, J. A., Merrill, B. J. & Plath, K. Stage-Specific Regulation of Reprogramming to Induced Pluripotent Stem Cells by Wnt Signaling and T Cell Factor Proteins. Cell Reports 3, 2113–2126 (2013).
99. Folmes, C. D. L. et al. Somatic Oxidative Bioenergetics Transitions into Pluripotency-Dependent Glycolysis to Facilitate Nuclear Reprogramming. Cell Metabolism 14 (2011).
100. Shyh-Chang, N. et al. Influence of Threonine Metabolism on S-Adenosylmethionine and Histone Methylation. Science 339, 222–226 (2013).
101. Varas, F. et al. Fibroblast-Derived Induced Pluripotent Stem Cells Show No Common Retroviral Vector Insertions. Stem Cells 27, 300–306 (2009).
102. Warren, L. et al. Highly Efficient Reprogramming to Pluripotency and Directed Differentiation of Human Cells with Synthetic Modified mRNA. Cell Stem Cell 7, 618–630 (2010).

103. Okita, K., Nakagawa, M., Hyenjong, H., Ichisaka, T. & Yamanaka, S. Generation of Mouse Induced Pluripotent Stem Cells Without Viral Vectors. Science 322, 949–953 (2008).
104. Stadtfeld, M., Nagaya, M., Utikal, J., Weir, G. & Hochedlinger, K. Induced Pluripotent Stem Cells Generated Without Viral Integration. Science 322, 945–949 (2008).
105. Zhou, H. et al. Generation of Induced Pluripotent Stem Cells Using Recombinant Proteins. Cell Stem Cell 4, 381–384 (2009).
106. Gore, A. et al. Somatic coding mutations in human induced pluripotent stem cells. Nature 471, 63–67 (2011).
107. Young, M. A. et al. Background Mutations in Parental Cells Account for Most of the Genetic Heterogeneity of Induced Pluripotent Stem Cells. Cell Stem Cell 10, 570–582 (2012).
108. Sugiura, M. et al. Induced Pluripotent Stem Cell Generation-Associated Point Mutations Arise during the Initial Stages of the Conversion of These Cells. Stem Cell Reports 2, 52–63 (2014).
109. Zeuschner, D., Mildner, K., Zaehres, H. & SchOler, H. R. Induced Pluripotent Stem Cells at Nanoscale. Stem Cells and Development 19, 615–620 (2010).
110. Courtot, A.-M. et al. Morphological Analysis of Human Induced Pluripotent Stem Cells During Induced Differentiation and Reverse Programming. Biores Open Access 3, 206–216 (2014).
111. Romero-Moya, D. et al. Cord blood-derived CD34+ hematopoietic cells with low mitochondrial mass are enriched in hematopoietic repopulating stem cell function. Haematologica 98, 1022–1029 (2013).
112. Collado, M., Blasco, M. A. & Serrano, M. Cellular Senescence in Cancer and Aging. Cell 130, 223–233 (2007).
113. Utikal, J. et al. Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. Nature 460, 1145–1148 (2009).
114. Hanna, J. et al. Direct cell reprogramming is a stochastic process amenable to acceleration. Nature 462, 595–601 (2009).
115. Marion, R. M. et al. Telomeres Acquire Embryonic Stem Cell Characteristics in Induced Pluripotent Stem Cells. Cell Stem Cell 4, 141–154 (2009).
116. Xu, X. et al. Mitochondrial Regulation in Pluripotent Stem Cells. Cell Metabolism 18, 325–332 (2013).
117. Golipour, A. et al. A Late Transition in Somatic Cell Reprogramming Requires Regulators Distinct from the Pluripotency Network. Cell Stem Cell 11, 769–782 (2012).
118. Samavarchi-Tehrani, P. et al. Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming. Cell Stem Cell 7, 64–77 (2010).

119. Chou, Y.-F. et al. The Growth Factor Environment Defines Distinct Pluripotent Ground States in Novel Blastocyst-Derived Stem Cells. Cell 135, 449–461 (2008).
120. O'Malley, J. et al. High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. Nature 499, 88–91 (2013).
121. Aasen, T. et al. Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. Nat Biotechnol 26, 1276–1284 (2008).
122. Drozd, A. M. et al. Generation of human iPSCs from cells of fibroblastic and epithelial origin by means of the oriP/EBNA-1 episomal reprogramming system. Stem Cell Res Ther 6, 122 (2015).
123. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining Molecular Cornerstones during Fibroblast to iPS Cell Reprogramming in Mouse. Cell Stem Cell 2, 230–240 (2008).
124. Brambrink, T. et al. Sequential Expression of Pluripotency Markers during Direct Reprogramming of Mouse Somatic Cells. Cell Stem Cell 2, 151–159 (2008).
125. Ohnishi, K. et al. Premature Termination of Reprogramming In Vivo Leads to Cancer Development through Altered Epigenetic Regulation. Cell 156, 663–677 (2014).
126. Tung, P.-Y. & Knoepfler, P. S. Epigenetic mechanisms of tumorigenicity manifesting in stem cells. Oncogene 34, 2288–2296 (2014).
127. Widschwendter, M. et al. Epigenetic stem cell signature in cancer. Nature Genetics 39, 157–158 (2006).
128. Ley, T. J. et al. DNMT3AMutations in Acute Myeloid Leukemia. N. Engl. J. Med. 363, 2424–2433 (2010).
129. Gao, Q. et al. Deletion of the de novo DNA methyltransferase Dnmt3a promotes lung tumor progression. Proceedings of the National Academy of Sciences 108 (2011).
130. Robert, M.-F. et al. DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. Nature Genetics 33, 61–65 (2002).
131. Mikkelsen, T. S. et al. Dissecting direct reprogramming through integrative genomic analysis. Nature 454, 49–55 (2008).
132. Sun, C. et al. Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma. Nature 508, 118–122 (2014).
133. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation. Science 324, 1029–1033 (2009).
134. Young, J. I., Züchner, S. & Wang, G. Regulation of the Epigenome by Vitamin C. Annu. Rev. Nutr. 35, 545–564 (2015).
135 Martin G., Gruber AR, Keller W, Zavolan M. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length Cell Reprogramming 2012

Chapter 2

A Reprogrammable Mouse Strain from Gene-Targeted Embryonic Stem Cells

Matthias Stadtfeld, Nimet Maherali, Marti Borkent & Konrad Hochedlinger

Nature Methods Brief Communications December 2009

Abstract

The derivation of induced pluripotent stem cells (iPSCs) usually involves the viral introduction of reprogramming factors into somatic cells. here we used gene targeting to generate a mouse strain with a single copy of an inducible, polycistronic reprogramming cassette, allowing for the induction of pluripotency in various somatic cell types. as these 'reprogrammable mice' can be easily bred, they are a useful tool to study the mechanisms underlying cellular reprogramming.

Introduction

Somatic cells can be reprogrammed into induced pluripotent stem cells (iPSCs) by transcription factors c-MYC, KLF4, OCT4 and SOX2[1]. The study of the molecular mechanisms underlying this conversion is hampered by the low efficiency of reprogramming and by the requirement to introduce the reprogramming factors into target cells. To circumvent these constraints, 'secondary systems' have been developed, which are based on iPSCs made with doxycycline-inducible lentiviruses[2–6].

Secondary systems based on viral vectors have several limitations. First, many iPSC lines show viral transgene silencing, leading to heterogeneous expression and thus variable results. Second, viral integrations at random genomic locations can alter endogenous gene expression. Third, the breeding of mice derived from such iPSCs is complicated by the need to screen their offspring for multiple viral transgenes.

Results

To develop an improved reprogramming system, we placed a doxycycline-inducible polycistronic cassette encoding the four reprogramming factors OCT4, KLF4, SOX2 and c-MYC (OKSM)[7] in the 3' untranslated region of the *Collagen* type I, alpha 1 gene (*Col1a1*, herein referred to as *Collagen*)[8] (Figure 1a). This gene targeting was mediated by Flp recombinase and was performed in embryonic stem cells (ESCs) that express an optimized reverse tetracycline-dependent trans-activator (*M2-rtTA*) from the ubiquitously expressed *Gt(ROSA)26Sor* (*Rosa26*) locus[8]. We refer to this allele as *Rosa26-rtTA*.

The OKSM cassette integrated into the *Collagen* locus in 12 of 12 picked ESC clones named *Collagen*-OKSM ESCs (Supplementary Figure 1). To test the function of this transgenic system in comparison to a viral secondary system, we isolated mouse embryonic fibroblasts (MEFs) derived from either *Collagen*-OKSM ESCs (called *Collagen*-OKSM MEFs) or from iPSCs originating from *Rosa26-rtTA* MEFs carrying two copies of a doxycycline-inducible lentiviral vector expressing the same polycistronic OKSM (Figure 1b). Upon culture in the presence

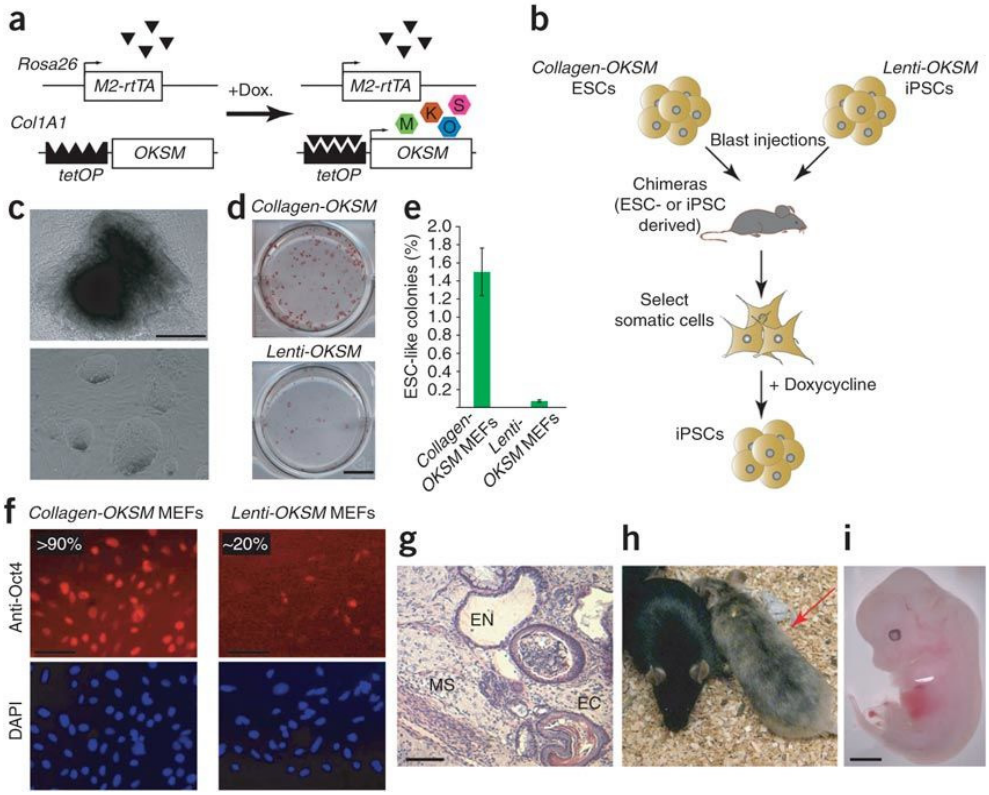


Figure 1. An ESC-based reprogramming system. (A) In *Collagen*-OKSM MEFs, cDNA encoding the optimized reverse tetracycline-dependent transactivator (M2-rtTA; triangles) was targeted to the *Rosa26* locus and a polycistronic cassette encoding OCT4 (O), KLF4 (K), SOX2 (S) and c-MYC (M) (OKSM) was targeted to the *Col1a1* locus under control of a tetracycline-dependent promoter (tetOP). In the presence of doxycycline (dox.), M2-rtTA binds to tetOP, inducing expression of the four reprogramming factors. (B) Outline for a secondary reprogramming system based on *Collagen*-OKSM ESCs or Lenti-OKSM iPSCs. (C) Brightfield images of a chimera-derived MEF iPSC clone at passage 0 (top, dark appearance due to dying or differentiating cells at the centre of the colony) and after three passages (bottom). (D) Alkaline phosphatase staining of colonies derived from *Collagen*-OKSM MEFs or MEFs carrying two copies of a doxycycline-inducible lentiviral vector encoding the same polycistronic cassette (Lenti-OKSM). (E) Efficiency of iPSC formation as a percentage of input cells (n = 3). (F) Fluorescence images of *Collagen*-OKSM MEFs and Lenti-OKSM MEFs cultured for 2 days in the presence of doxycycline and stained with an antibody to Oct4 (anti-Oct4). Percentages of Oct4+ cells are indicated. (G) Histological section through a teratoma derived from a *Collagen*-OKSM iPSC clone showing differentiation into ectodermal (EC; keratinized epithelium), endodermal (EN; glandular structures) and mesodermal (MS; muscle) derivatives. (H) Image of an adult coat color chimera (arrow) from MEF-derived iPSCs. Also shown is a nonchimeric BDF1 mouse. (I) Image of an embryonic day (E)13.5 embryo derived after injection of MEF iPSCs into tetraploid blastocysts ('all iPSC' fetus). Scale bars, 500 µm (C), 10 mm (D), 100 µm (F), 50 µm (G) and 2 mm (I).

of doxycycline, both *Collagen*-OKSM MEFs and Lenti-OKSM MEFs formed colonies with ESC-like morphology that stained positive for pluripotency-associated markers such as alkaline phosphatase, NANOG and endogenous OCT4 (Figure 1c and Supplementary Figure 2). Such colonies could be expanded apparently indefinitely (>25 passages when submitting this manuscript) in the absence of exogenous factor expression, and we will refer to them as iPSCs. *Collagen*-OKSM MEFs produced iPSC colonies approximately tenfold more efficiently than Lenti-OKSM MEFs (Figure 1d,e), consistent with the observation that *Collagen*-OKSM MEFs homogenously expressed exogenous Oct4 after addition of doxycycline to the cultures, whereas we observed only patchy reactivation in Lenti-OKSM MEFs (Figure 1f). *Collagen*-MEF iPSCs formed highly differentiated teratomas in severe combined immunodeficient (SCID) mice (Figure 1g) and contributed to adult coat color chimeras (Figure 1h). These cells supported the development of 'all-iPSC' midgestation fetuses after injection into tetraploid blastocysts (Figure 1i), further demonstrating pluripotency.

To track cells undergoing reprogramming, we labeled *Collagen*-OKSM MEFs with a constitutively active lentiviral vector expressing the reportergene *tdTomato*. We established MEF cultures from chimeric embryos and plated brightly red-fluorescent cells at single-cell density onto irradiated feeder cells. MEFs expressing the four reprogramming factors were smaller and proliferated faster within the first 48 h of doxycycline administration compared to uninduced cells. This resulted in clusters of coalescent cells between days 4 and 6, from which iPSC colonies emerged between days 8 and 14 of culture (Supplementary Figure 3).

Withdrawal of doxycycline at different time points showed that the first stably reprogrammed cells appeared by day 7, consistent with previous reports using lentiviral vectors[2,3] (Supplementary Figure 4a). When we withdrew doxycycline before day 7, MEFs that had undergone morphological changes reverted back to a fibroblast-like shape and ceased proliferation (Supplementary Figure 4a). Re-exposure of these cells to doxycycline after 1 week of withdrawal did not yield stable iPSC colonies from more than 100 regressed colonies assayed, suggesting that these cells had become refractory to reprogramming, possibly owing to the exhaustion of their proliferative capacity[9].

The observation that cells undergoing reprogramming revert to a fibroblast-like state upon premature withdrawal of doxycycline suggests that previously reported 'partially reprogrammed cells'[10,11] are not a common phenotype during factor-mediated reprogramming but rather may be the consequence of incomplete retroviral transgene silencing, which drives continuous cell proliferation without acquisition of pluripotency.

We next established a mouse strain carrying the inducible reprogramming cassette in all tissues. We bred *Collagen*-OKSM chimeras (Figure 2a) with mice that carried the EGFP gene in the *Pou5f1* (also known as *Oct4*) locus, which serves as a faithful indicator for the acquisition of pluripotency[12]. We will refer to the resulting mice as 'reprogrammable

mice' (Figure 2b). Notably, some reprogrammable mice developed a single aggressively growing tumor that histologically presented as a largely undifferentiated teratoma (Supplementary Figure 5a,b). This phenotype is likely due to leaky transgene expression in an undefined cell type and raises the possibility that somatic cells can be reprogrammed into pluripotent cells not only in vitro but also in vivo. To validate the *Collagen*-OKSM–Oct4-EGFP reprogramming system, we grew postnatal tail-tip fibroblasts (TTFs) in the presence of doxycycline, which resulted in the formation of green fluorescent ESC-like colonies at a frequency of approximately 0.1% (Figure 2c,d). The lower reprogramming efficiency of TTFs compared to that of MEFs (Figure 1e) was not due to incomplete reactivation of the reprogramming cassette (data not shown), indicating that fetal fibroblasts are more amenable to reprogramming than postnatal cells, as recently reported[13]. The efficiency of TTF reprogramming increased approximately twofold when we grew the cells in medium containing serum replacement (Figure 2d). These iPSCs adopted an ESC-like morphology faster than cells in serum-containing medium, consistent with previous observations[14]. In the absence of feeder layers, TTFs grown in serum-replacement medium had to be seeded at a high density (~10,000 cells cm^{-2}) to reach optimal reprogramming efficiencies compared to cells grown in serum-containing medium, which required a sparse starting culture (~1,000 cells cm^{-2}) (Figure 2e). Together, these results suggest that reprogramming is not entirely cell-autonomous but is influenced by both inhibitory and supportive signals from the culture environment.

Collagen-OKSM TTFs homozygous for the *Rosa26-rtTA* allele reached reprogramming efficiencies of 0.5–1% (Figure 2d) and reprogrammed faster than heterozygous cells, as demonstrated by the earlier appearance of doxycycline-independent green fluorescent colonies (Figure 2f). This shows that higher levels of transgene expression are favorable for both reprogramming kinetics and efficiency.

To reprogram a defined somatic cell lineage, we isolated different hematopoietic cell types from adult reprogrammable mice (Supplementary **Figure** 6). We consistently did not obtain iPSC colonies from mature T cells, B cells and granulocytes heterozygous for both Rosa26-rtTA and Collagen-OKSM, although we occasionally observed colonies from monocytes with this genotype (**Figure** 2g). In contrast, Collagen-OKSM heterozygous cells that were homozygous for Rosa26-rtTA reproducibly gave rise to green fluorescent iPSCs from all tested cell types at frequencies ranging from 0.01% (B cells) to 0.3% (monocytes) (Figure 2g and Supplementary Figure 7). This suggests that a critical level of factor expression is required for reprogramming of terminally differentiated blood cells. In contrast, hematopoietic stem cells and granulocyte-macrophage progenitors from Rosa26-rtTA heterozygous mice generated iPSCs at high (5–10%) efficiencies (Figure 2h), consistent with our previous finding that reprogramming potential inversely correlates with differentiation stage[15]. Notably, we

derived iPSCs from hematopoietic stem cells homozygous for Rosa26-rtTA at frequencies of up to 40% (Figure 2h), which is to our knowledge the highest reprogramming efficiency for an adult somatic cell type.

Reprogramming of blood cells was equally successful in the presence or absence of cell type–specific cytokines supporting hematopoietic cell survival and proliferation, with two exceptions. First, granulocytes did not form iPSCs in the presence of the growth factor granulocyte-macrophage colony stimulating factor, and second, *Rosa26-rtTA* heterozygous hematopoietic stem cells and granulocyte-macrophage progenitors reprogrammed 3–4 times less efficiently without cytokines (Figs. 2g,h). *Rosa26* homozygous hematopoietic stem cells and granulocyte-macrophage progenitors did not show this effect (Figure 2h), suggesting that higher expression levels of reprogramming factors can substitute for the signals provided by cytokines.

The mouse strain described here overcomes several of the limitations of viral secondary reprogramming systems and should be a useful tool to study the mechanisms underlying the conversion of somatic cell types into iPSCs. It may be especially attractive for dissecting the biochemistry of induced pluripotency in the context of adult stem and progenitor cells that reprogram at high efficiency. Another crucial question that can be addressed with this system is whether ESCs and iPSCs are molecularly and functionally equivalent. Specifically, the possibility to compare ESCs to genetically matched iPSCs from different tissues will allow this question to be resolved.

Methods

Mice

Derivation, handling and genotyping of *Oct4-EGFP12* and *Rosa26-rtTA*8 mice were described previously. *Collagen*-OKSM reprogrammable mice will be deposited with The Jackson Laboratories once the colony has been expanded. All animal studies were carried following approved guidelines of the animal protocol of the Massachusetts General Hospital Cancer Center.

Cell culture

ESCs and iPSCs were cultured in ESC medium (DMEM with 15% FBS, L-glutamine, penicillin-streptomycin, nonessential amino acids, β-mercaptoethanol and 1,000 U ml−1 LIF) on irradiated feeder cells. ESC-derived MEFs were isolated by trypsin digestion of midgestation (E14.5) chimeric embryos followed by culture in fibroblast medium (DMEM with 10% FBS, l-glutamine, penicillin, streptomycin, nonessential amino acids and β-mercaptoethanol)

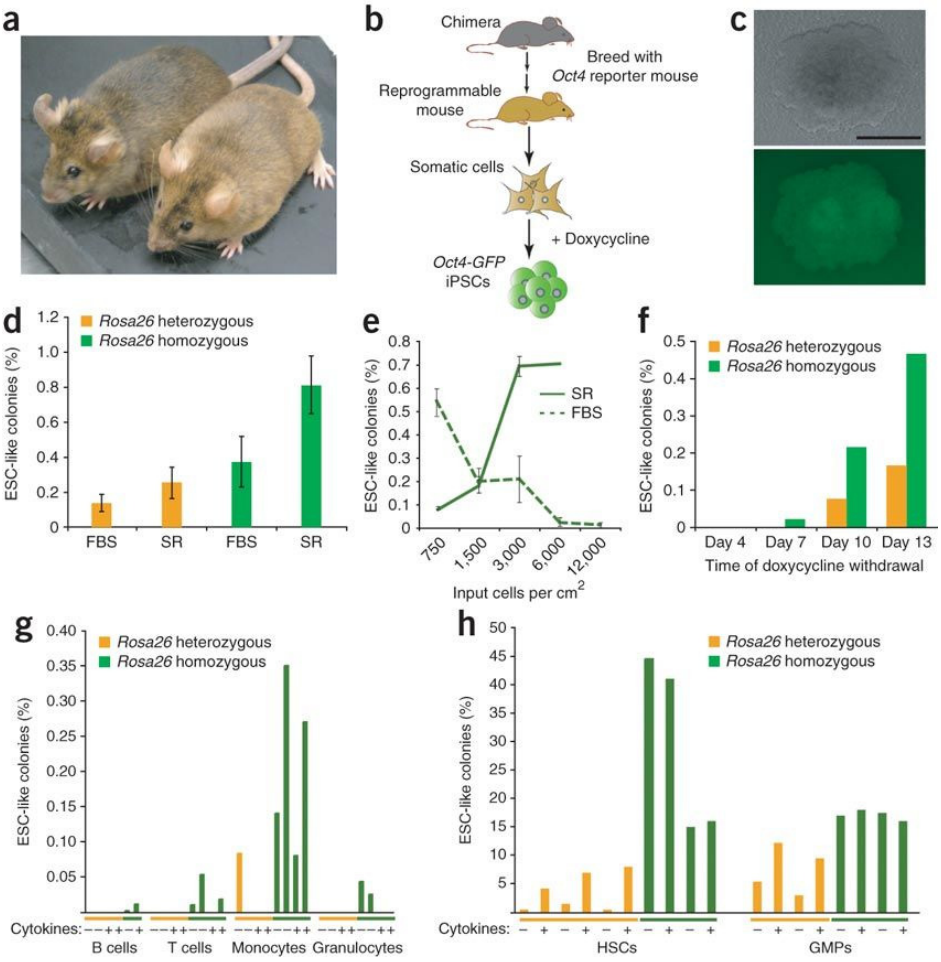


Figure 2. Derivation and characterization of reprogrammable mice. (A) Representative chimeras derived from *Collagen*-OKSM MEFs. (B) Outline for the establishment of reprogrammable mice. (C) Brightfield (top) and green fluorescence (bottom) images of an iPSC colony derived from postnatal tail-tip fibroblasts (TTFs) from a reprogrammable mouse. Scale bar, 500 µm. (D) Effect of serum (FBS) and serum replacement (SR) on reprogramming efficiency from *Rosa26-rtTA* heterozygous or homozygous TTFs. Reprogramming efficiencies were determined by counting green-fluorescent transgene-independent colonies with ESC- like morphology per number of input cells. Error bars, s.d. (n = 3). (E) Effect of cell density on the reprogramming efficiency of *Rosa26-rtTA* homozygous TTFs grown in FBS or SR. (F) Reprogramming efficiencies after culture of TTFs heterozygous or homozygous for *Rosa26-rtTA* in doxycycline-containing medium. Doxycycline was withdrawn on the indicated days and colonies were scored on day 20. (G) Reprogramming efficiencies of mature hematopoietic cell types from the reprogrammable mice heterozygous or homozygous for *Rosa26-rtTA*. Cells were cultured either in the presence (+) or absence (−) of cytokines. (H) Reprogramming efficiencies of hematopoietic stem cells (HSCs) and granulocyte-macrophage progenitor cells (GMPs) from reprogrammable mice.

containing puromycin. TTF cultures were established from tail-tip biopsies of newborn (3–8 d) mice and maintained in fibroblast medium. Hematopoietic cells were isolated from peripheral blood and bone marrow as previously described[15].

Gene targeting of ESCs

A polycistronic cassette encoding OCT4, KLF4, SOX2 and c-MYC was cloned into the shuttle plasmid pBS31 using NotI and ClaI digestion. The resulting plasmid was electroporated into KH2 ESCs together with a plasmid driving expression of Flp recombinase[8]. Correct targeted clones were screened for by hygromycin selection and confirmed by Southern blotting as previously described[8].

Lentiviral vectors

To produce infectious viral particles, 293T cells cultured on T75 flasks were transfected with a lentiviral vector expressing tdTomato from the ubiquitin promoter (11 μg) together with the packaging plasmids VSV-G (5.5 μg) and Δ8.9 (8.25 μg) using Fugene (Roche) transfection reagent. Viral supernatants were collected on 3 consecutive days starting 24 h after transfection, yielding a total of ~50 ml of supernatant per virus. Viral supernatant was concentrated ~200-fold by ultracentrifugation at 60,000g for 1.5 hour at 4 °C and resuspension in 250 μl of PBS. Viral concentrates were stored at −80 °C. iPSCs were transduced in 0.2 ml ES medium containing 5 μg ml^{-1} polybrene using 20–40 μl of viral concentrate on a 48-well plate. Regular ESC medium was added 24 h after infection and red-fluorescent iPSCs isolated by flow cytometry.

Immunofluorescence

iPSCs were cultured on pretreated cover slips, fixed with 4% PFA and permeabilized with 0.5% Triton X-100. The cells were then stained with primary antibodies to murine OCT4 (Santa Cruz, sc-8628) and to mNANOG (Abcam, ab21603), followed by staining with the respective secondary antibodies conjugated to either Alexa Fluor 488 or 546 (Invitrogen). Nuclei were counterstained with DAPI (Invitrogen). Cells were imaged using a Leica DMI4000B inverted fluorescence microscope equipped with a Leica DFC350FX camera. Images were processed and analyzed using Adobe Photoshop software.

Teratoma formation

iPSCs were collected by trypsinization and injected into the flanks of nonobese diabetic (NOD)-SCID mice, using ~5 million cells per injection. Mice were killed 3 weeks later, and teratomas were isolated and processed for histological analysis.

Blastocyst injections

Female BDF1 mice were superovulated by intraperitoneal injection of pregnant mare serum (PMS) and human chorionic gonadotropin (hCG) and mated to BDF1 stud males. Zygotes were isolated from females with a vaginal plug 24 h after hCG injection. Zygotes for diploid (2n) blastocyst injections were cultured for 3 days in EmbryoMax KSOM medium (Millipore); blastocysts were identified, injected with ESCs or iPSCs and transferred into pseudopregnant recipient females. For tetraploid (4n) blastocyst injections, zygotes were cultured overnight until they reached the 2-cell stage, at which point they were electrofused. One hour later, 1-cell embryos were carefully identified and separated from embryos that did not fuse, cultured in KSOM for another 2 days and then injected.

Reprogramming

MEFs and TTFs were counted and seeded in fibroblast medium at the desired density on either gelatin-coated plates or plates that also contained a layer of irradiated feeder cells. The next day, ESC medium containing 2 μg ml^{-1} doxycycline was added and replenished every 3 days. Upon doxycycline withdrawal, cultures were washed twice with PBS and then continued in standard ESC medium until colonies were either picked, scored by fluorescence microscopy or stained for alkaline phosphatase activity (Vector Red substrate kit; Vector Labs). Freshly isolated hematopoietic cells were isolated by flow cytometry and immediately plated on top of irradiated feeder layers in ESC medium containing doxycycline on either 35-mm or 100-mm plates. Cell densities were 200 cells per well for hematopoietic stem cells (HSCs) and granulocyte-macrophage progenitors (GMPs), and 5,000–100,000 cells per well for all other cell types. The following cytokines were used: IL-15 (2 ng μl^{-1}) and IL-7 (10 ng μl^{-1}) for B cells, ConA (2.5 μg ml–1) and IL-2 (10 ng μl^{-1}) for T cells, M-CSF (5 ng μl^{-1}) for monocytes, G-CSF (10 ng μl^{-1}) for granulocytes and Flt3-ligand (10 ng μl^{-1}), SCF (10 ng μl^{-1}) and TPO (10 ng μl^{-1}) for HSCs and GMPs. All cytokines were purchased from Peprotech. Doxycycline was withdrawn from HSCs and GMPs after 2 weeks and from the slower reprogramming peripheral blood cells after 4–6 weeks.

Acknowledgements

We thank G. Mostoslavsky (Boston University School of Medicine) for the STEMCCA construct and secondary viral MEFs, A. Foudi and S. Eminli for help with hematopoietic cell isolation and culture, A. Khalil for help in cloning the *Collagen*-OKSM targeting construct, J. Polo for help with cell culture as well as for inspiring discussions, and C. Konrad for advice on tumor histology. M.S. was supported by a postdoctoral fellowship from the Schering Foundation,

and K.H. was supported by a US National Institutes of Health Director's Innovator Award as well as by funds provided by the Harvard Stem Cell Institute and Massachusetts General Hospital.

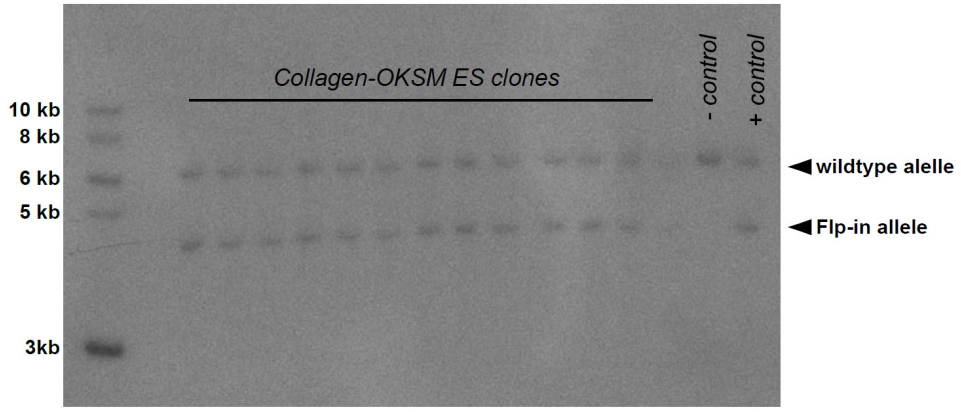
Author contributions

M.S. and K.H. conceived the study; N.M. performed blastocyst injections; M.S. performed all other experiments with help from M.B.; M.S. and K.H. analyzed the data and wrote the manuscript.

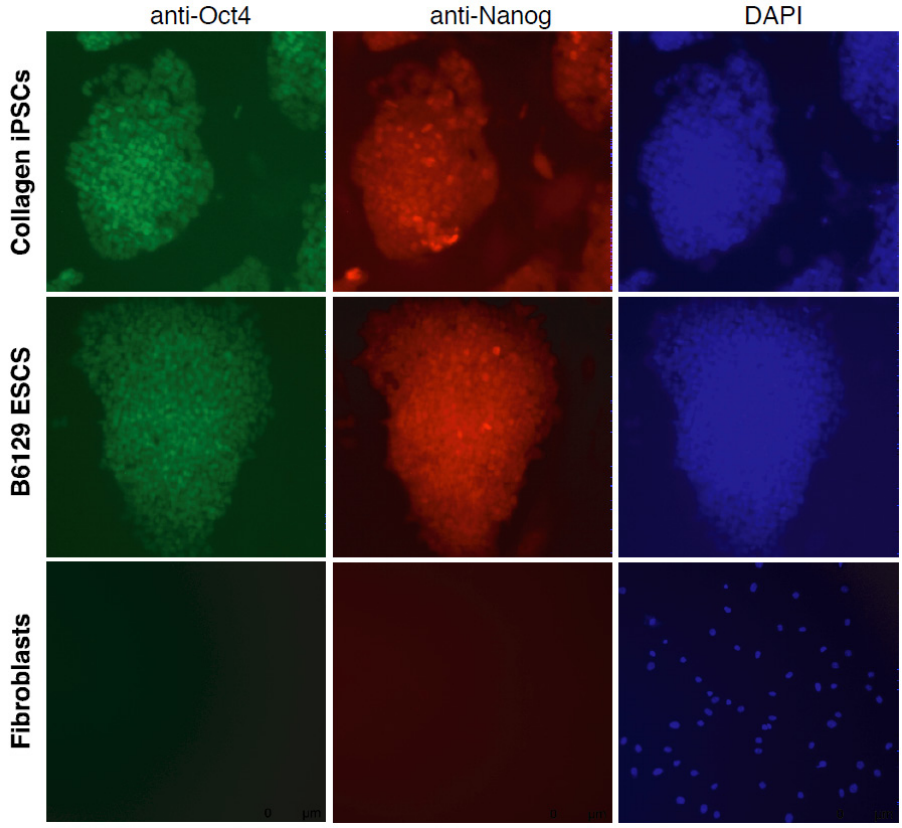
References

1. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663–676 (2006).
2. Brambrink, T. et al. Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. Cell Stem Cell 2, 151–159 (2008).
3. Stadtfeld, M., Maherali, N., Breault, D.T. & Hochedlinger, K. Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. Cell Stem Cell 2, (2008).
4. Maherali, N. et al. A high-efficiency system for the generation and study of human induced pluripotent stem cells. Cell Stem Cell 3, 340–345 (2008).
5. Wernig, M. et al. A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. Nat. Biotechnol. 26, 916–924 (2008).
6. Hockemeyer, D. et al. A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. Cell Stem Cell 3, 346–353 (2008).
7. Sommer, C.A. et al. Induced pluripotent stem cell generation using a single lentiviral stem cell cassette. Stem Cells 27, 543–549 (2009).
8. Beard, C., Hochedlinger, K., Plath, K., Wutz, A. & Jaenisch, R. Efficient method to generate single-copy transgenic mice by site-specific integration in embryonic stem cells. Genesis 44, 23–28 (2006).
9. Utikal, J. et al. Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. Nature 460, 1145–1148 (2009).
10. Mikkelsen, T.S. et al. Dissecting direct reprogramming through integrative genomic analysis. Nature 454, 49–55 (2008).
11. Sridharan, R. et al. Role of the murine reprogramming factors in the induction of pluripotency. Cell 136, 364–377 (2009).
12. Lengner, C.J. et al. Oct4 expression is not required for mouse somatic stem cell self-renewal. Cell Stem Cell 1, 403–415 (2007).

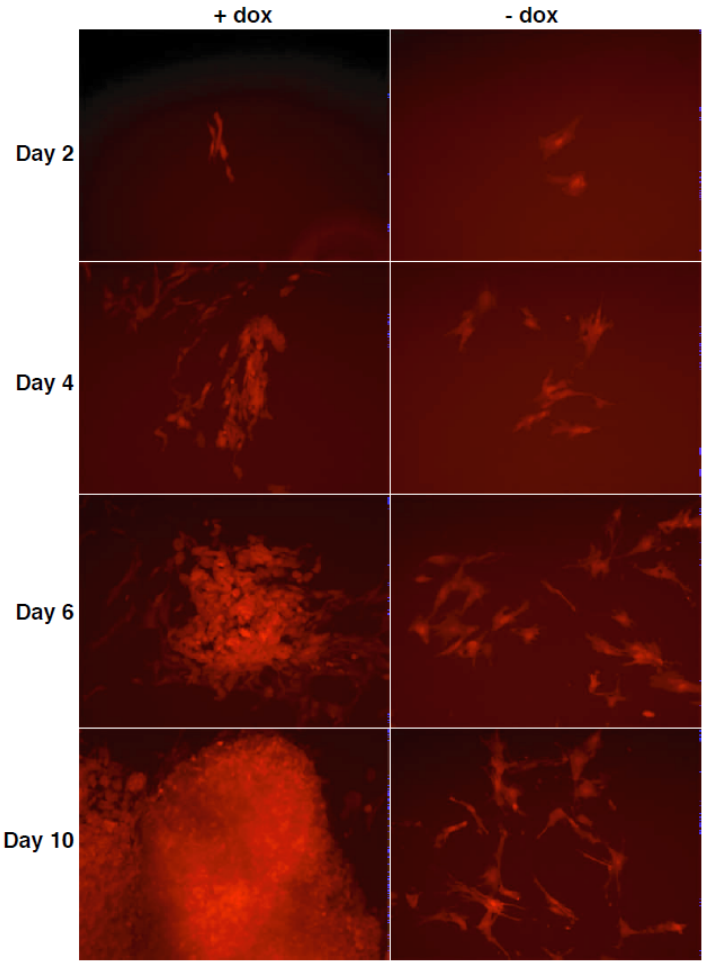
13. Markoulaki, S. et al. Transgenic mice with defined combinations of drug-inducible reprogramming factors. Nat. Biotechnol. 27, 169–171 (2009).
14. Blelloch, R., Venere, M., Yen, J. & Ramalho-Santos, M. Generation of induced pluripotent stem cells in the absence of drug selection. Cell Stem Cell 1, 245–247 (2007).
15. Eminli, S. et al. Differentiation stage determines potential of hematopoietic cells for reprogramming into induced pluripotent stem cells. Nat. Genet. 41, 968–976 (2009).



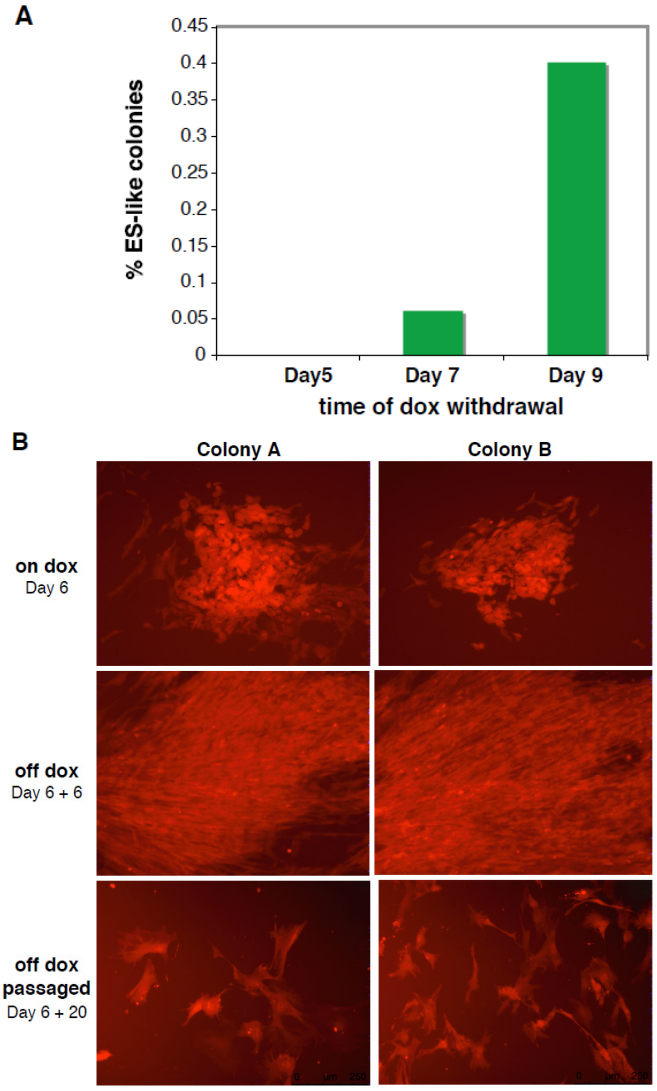
Supplementary Figure 1. Confirmation of correct targeting of *Col1a1*. Southern blot using genomic DNA extracted from 12 individual *Collagen*-OKSM ESC clone as well as a negative (iPSC clone with *Col1a1* locus in wildtype configuration) and positive (iPSC clone with *Oct4* cDNA in *Col1a1*) control DNA.



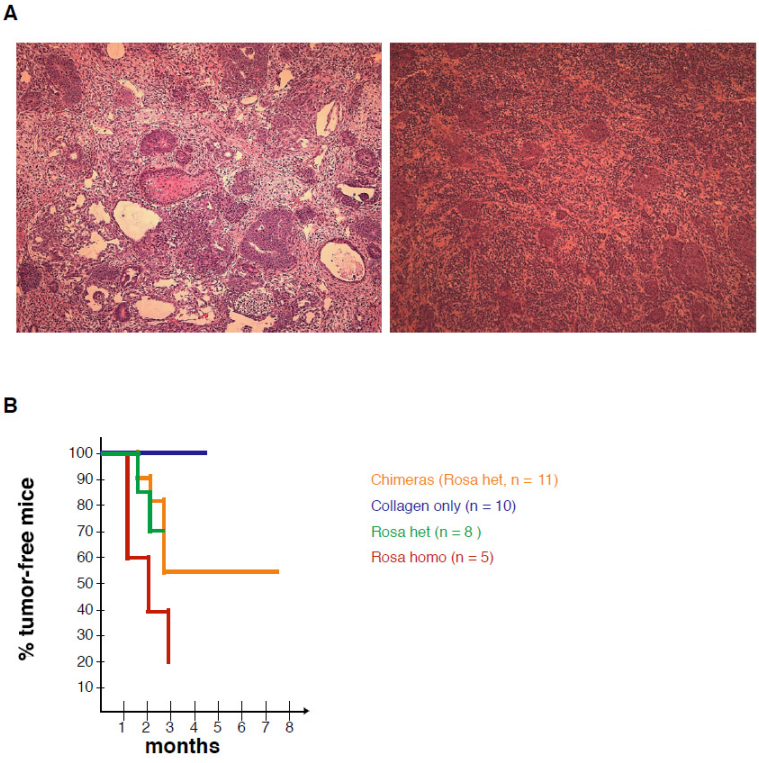
Supplementary Figure 2. Expression of pluripotency markers in *Collagen*-OKSM iPSCs. Immunofluorescence images after staining iPSCs cultured for several weeks in the absence of doxycycline (upper panel), ESCs (middle panel) and *Collagen*-OKSM fibroblasts cultured in the absence of doxycycline (lower panel) with antibodies against OCT4 and NANOG. Nuclei were counterstained with DAPI.



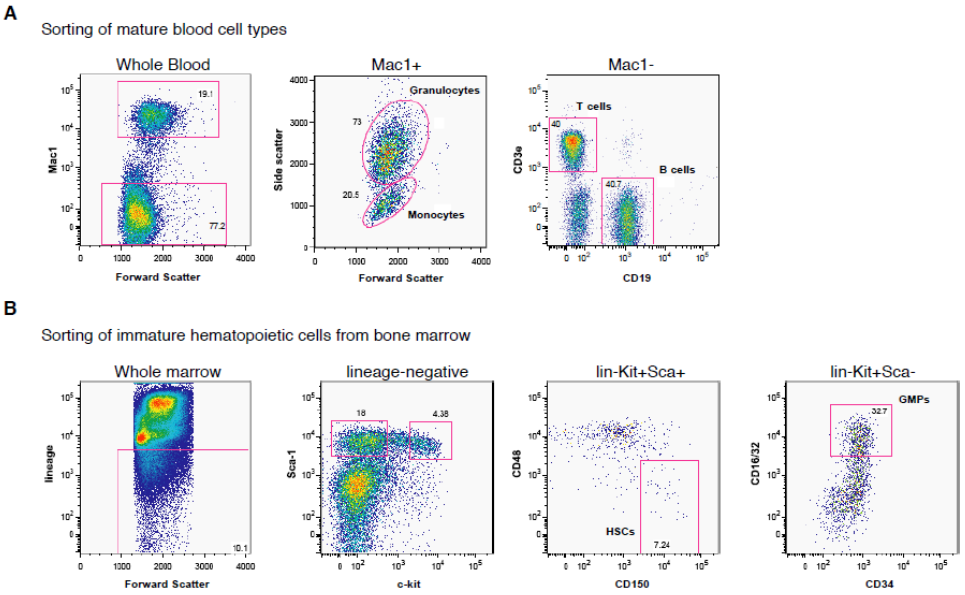
Supplementary Figure 3. Morphological changes of reprogramming fibroblasts. RFP+ MEFs isolated from midgestation chimeras derived after blastocyst injections with *Collagen*-OKSM ESCs transduced with a lentivirus expressing tdTomato were cultured either in the presence (+ dox, left panel) or absence (no dox, right panel) of doxycycline. Images were taken after the indicated days of culture.



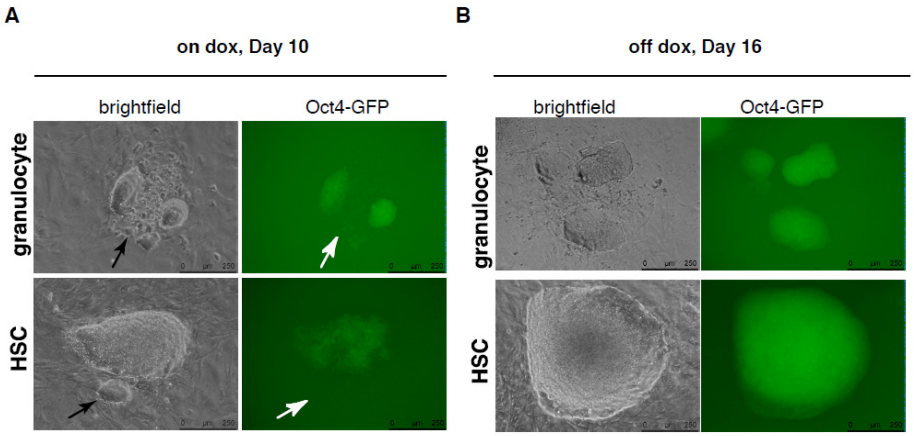
Supplementary Figure 4. Effect of premature dox withdrawal. (A) Reprogramming efficiencies of *Collagen*-OKSM MEFs after culture for 5, 7 and 9 days in the presence of doxycycline. (B) RFP fluorescence images of two representative intermediate colonies formed after culturing MEFs for 6 days in the presence of doxycycline (upper panel). After withdrawal of doxycycline cells (Day 6 + 6, middle panel) cells stopped reprogramming, reverted to an elongated fibroblast-like morphology (compare to Supplementary **Figure** 3) and upon passaging stopped dividing (Day 6 + 20, bottom panel).



Supplementary Figure 5. Tumor formation in *Collagen*-OKSM mice. (A) Histological sections through two tumors found in adult chimeras derived after blastocyst injection of *Collagen*-OKSM ESCs. (B) Curve showing percentage of tumor-free mice at the indicated timepoints.



Supplementary Figure 6. Isolation of hematopoietic cells. Gating logic for the isolation of four different mature hematopoietic cell types from blood (upper panel) as well as two immature cell types from bone marrow (lower panel). Cells were defined by the following criteria: Monocytes (MAC1+ SSClow), Granulocytes (MAC1+ SSChigh), B cells (CD19+ CD3- MAC1-), T cells (CD3+ CD19- MAC1-), GMPs (LinKit+ SCA1- CD34+ FcγR+) and HSCs (LinKit+ SCA1+ CD150+ CD48-).



Supplementary Figure 7. Hematopoietic iPSC colonies. (A) Brightfield and GFP fluorescence images of nascent iPSC colonies forming from granulocytes and HSCs cultured in the presence of doxycycline for 10 days. Note that *Oct4*-GFP fluorescence is patchy and that GFP- cells with ES morphology can be identified (highlighted by arrows), indicated incomplete reprogramming at this timepoint. (B) Images of the same colonies after 6 additional days of culture in the absence of doxycycline. GFPcells are absent and GFP fluorescence has become homogenous.

Chapter 3

A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells

Jose M. Polo*, Endre Anderssen*, Ryan M. Walsh, Benjamin A. Schwarz, Christian M. Nefzger, Sue Mei Lim, Marti Borkent, Effie Apostolou, Sara Alaei, Jennifer Cloutier, Ori Bar-Nur, Sihem Cheloufi, Matthias Stadtfeld, Maria Eugenia Figueroa, Daisy Robinton, Sridaran Natesan, Ari Melnick, Jinfang Zhu, Sridhar Ramaswamy#, and Konrad Hochedlinger#

*equally contributing authors
corresponding authors
Cell December 2012

Abstract

Factor-induced reprogramming of somatic cells into induced pluripotent stem cells (iPSCs) is inefficient, complicating mechanistic studies. Here, we examined defined intermediate cell populations poised to becoming iPSCs by genome-wide analyses. We show that induced pluripotency elicits two transcriptional waves, which are driven by c-MYC/KLF4 (first wave) and OCT4/SOX2/KLF4 (second wave). Cells that become refractory to reprogramming activate the first but fail to initiate the second transcriptional wave and can be rescued by elevated expression of all four factors. The establishment of bivalent domains occurs gradually after the first wave, whereas changes in DNA methylation take place after the second wave when cells acquire stable pluripotency. This integrative analysis allowed us to identify genes that act as roadblocks during reprogramming and surface markers that further enrich for cells prone to forming iPSCs. Collectively, our data offer new mechanistic insights into the nature and sequence of molecular events inherent to cellular reprogramming.

Introduction

Induced pluripotent stem cells (iPSCs) have been generated from a number of mouse and human cell types upon enforced expression of transcription factors such as OCT4, KLF4, SOX2, and c-MYC (OKSM)[1,2]. iPSCs provide a valuable source of patient-specific cells for the study and potential treatment of human diseases[3]. In addition, iPSC technology offers a unique tool to dissect the principles of cell fate determination during normal development and its dysregulation in disease[4].

In general, less than 3% of somatic cells expressing OKSM give rise to iPSC colonies, complicating efforts to dissect the mechanisms of reprogramming. Owing to this limitation, most previous studies focused on the immediate response of somatic cells to factor expression. For example, fibroblasts were shown to go through a process that was reminiscent of a mesenchymalto-epithelial transition (MET) within a few days of OKSM expression[5,6]. At the epigenetic level, widespread remodeling of certain histone modifications, but not of DNA methylation patterns, was seen within the first few cell divisions of iPSC induction[7]. However, intermediate and late stages of reprogramming have remained inaccessible for more detailed molecular analyses.

We and others have documented that fibroblasts undergoing reprogramming pass through a number of defined intermediates[8,9]. Briefly, cells expressing OKSM from doxycycline (dox)-inducible lentiviral vectors initially downregulate the fibroblast-associated marker THY1 (day 1–2), then activate the SSEA1 antigen (day 3–5) and eventually upregulate an *Oct4-GFP* reporter (day 8–10) before forming stable iPSC colonies at approximately 1.5

weeks.

Importantly, isolation of these rare cell populations with the aforementioned markers allowed us to significantly enrich for cells that are poised to becoming iPSCs. Here, we have utilized this approach, in combination with a transgenic system that enables homogeneous dox-inducible OKSM expression in somatic cells (chapter 2, [10]), to purify intermediate stages of iPSC formation with the goal to elucidate the nature and sequence of molecular changes specific to cellular reprogramming.

Results

Experimental Approach to Studying Rare Reprogramming Intermediates

We first determined whether the reprogramming of fibroblasts with a recently reported dox-inducible transgenic system (''reprogrammable system'', chapter 2)[10] generates the same subpopulations of cells that we have previously described by using direct lentiviral infection. As shown in Figure 1A, murine embryonic fibroblasts (MEFs) carrying the *Col1a1-tetO-OKSM* transgene, the *ROSA26-M2-rtTA* allele, and an *Oct4*-GFP knockin reporter gave rise to THY1 cells, SSEA1+ cells, and *Oct4*-GFP+ cells with the expected kinetics. To verify that these intermediate populations were indeed enriched for cells that would form iPSCs, we sorted cells on feeders based on THY1, SSEA1, and GFP expression and treated them with dox for an equal number of days (see Extended Experimental Procedures). Consistent with our previous report, intermediate cells with the potential to give rise to iPSCs were initially present within both the THY1 and SSEA1+ populations, then progressed to SSEA1+ cells, and ultimately transited to the SSEA1+, *Oct4*-GFP+ population (Figures 1B and 1C). Importantly, sorting of THY1+ cells after day 3 and of THY1 cells after day 6 consistently failed to yield iPSC colonies, indicating that these cell populations had become refractory to reprogramming.

To examine the phenotypic progression of reprogramming intermediates, we sorted THY1+, SSEA1+, and *Oct4*-GFP+ cells after 3, 6, 9, and 12 days of dox induction, followed by culture in dox for another 3 days before reassessing their surface phenotype (Figure 1C). This analysis, combined with the abovementioned reprogramming results (Figure 1B), documents that (1) cells undergoing successful reprogramming with the Col1a1tetO-OKSM transgenic system transit in a linear fashion from a THY1+ to a THY1 to a SSEA1+ state in the first 6 days and eventually to a SSEA1+, *Oct4*-GFP+ state by days 9–12 (see Figure 1D for graphic summary; red arrows connect intermediates progressing toward iPSCs); (2) SSEA1+ cells are phenotypically still plastic until days 9–12 when they undergo commitment to a stable pluripotent cell fate; and (3) THY1+ cells lose their ability to progress toward a THY1 and SSEA1+ state as early as day 3. Note that progressing intermediates account for only 5%–10% of cells in regular reprogramming cultures compared with 90%–95% of THY1+ cells.

We generated gene expression profiles for these intermediate cell populations as well as for noninduced THY1+ MEFs (day 0) and dox-independent iPSCs (see Table S1, available online, for searchable database). Examination of candidate genes confirmed that THY1 was downregulated and SSEA1 (synthesized by the gene product of *Fut9*[11]) was upregulated in the THY1 and SSEA1+ populations, respectively (Figure 1E; note that red line depicts progressive intermediates as defined in Figure 1D by red solid arrows, whereas black line shows refractory THY1+ cells). In addition, we noticed that *Snai1* became downregulated, whereas *E-Cadherin* was upregulated at day 3, consistent with the occurrence of a MET. Alkaline phosphatase (*Alpl*) and *Fbxo15*, early markers of pluripotent cells, gradually increased their expression, whereas endogenous OCT4 and SOX2 transcripts were detectable only late during iPSC generation. Lastly, *Cyclin B1* became upregulated and the CDK inhibitor *Cdkn2b* (encoding for p15) was downregulated early in reprogramming. It is worth mentioning that THY1+ cells mirrored the gene expression changes of progressing cells until day 3 but then failed to sustain this trend at later time points, which correlated with their inability to produce iPSCs after day 3 (Figure 1B). We conclude that our sorting strategy allows us to analyze gene expression patterns of progressive intermediate cell populations transitioning toward iPSCs (Figure 1C) and to distinguish these from patterns in the bulk population of cells that are refractory to reprogramming.

Major Gene Expression Changes Occur in Two Discernible Phases during iPSC Formation

Principle component analysis (PCA) of the cell populations revealed a molecular connectivity reflecting their progression from the initial THY1+ cells toward THY1 cells and ultimately SSEA1+ cells, *Oct4*-GFP+ cells and iPSCs as depicted by the dashed red line (Figure 2A). PCA analysis further showed that all intermediates at day 3 clustered together, indicating that cells responded homogeneously to OKSM activation within the first few days. After day 3, however, SSEA1+ cells progressed toward *Oct4*-GFP cells, which were most closely related to established iPSCs, demonstrating that the SSEA1+ population gradually evolved toward a bona fide pluripotent state with time. Unsupervised clustering confirmed the similarities of SSEA1+ cells at days 3–9 and of *Oct4*-GFP+ cells at day 12 and iPSCs (Figure 2B). An examination of the number of differentially expressed genes between progressing (SSEA1+)

(continued from right) (D) Scheme illustrating the different subpopulations throughout reprogramming. Solid red arrows connect cell populations progressing toward iPSCs as inferred from data in (B). (E) Expression analyses of indicated genes at day 0, 3, 6, 9, and 12 of reprogramming and in established iPSCs (black lines depict THY1+ populations; red lines depict cells undergoing successful reprogramming as defined by red arrows in **Figure** 1D).

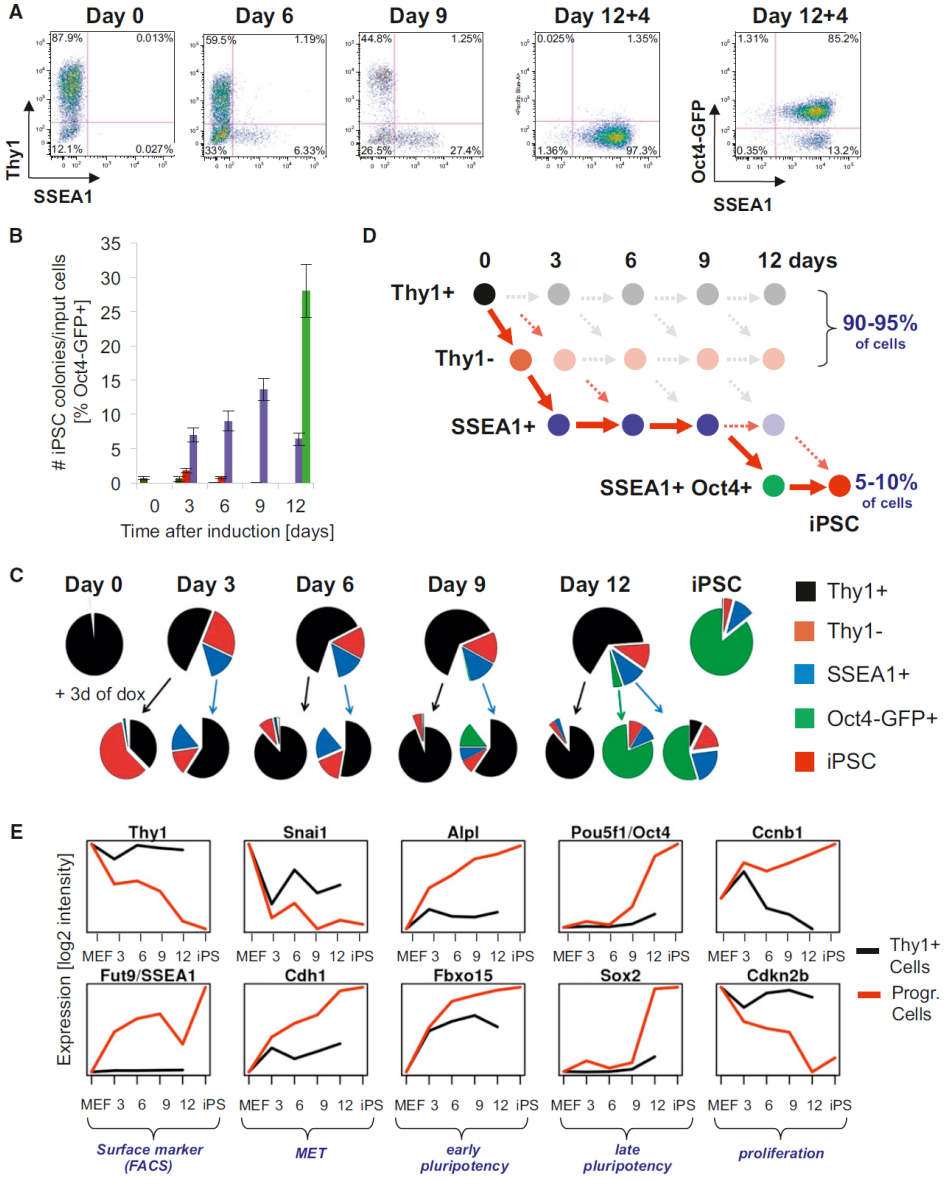


Figure 1. Strategy for Isolating Reprogramming Intermediates. (A) FACS analysis of reprogrammable MEFs at indicated time points. 12+4 denotes transgene-independent growth for 4 days. (B) Comparison of reprogramming efficiencies of intermediates purified at indicated time points. Note that established iPSCs have a colony formation efficiency of ~30%[9]. Data are represented as mean ± SEM (n = 3). (C) Pie charts summarizing FACS analysis of reprogrammable cells at indicated time points (top row). Bottom row shows FACS analysis for THY1, SSEA1 and *Oct4*-GFP 3 days after sorting and plating of the above cell populations in the presence of doxycycline. (continued on left)

and refractory (THY1+) cell populations at each time point showed a gradual increase, which culminated at ~1,500 genes by day 12 (Figure 2C). These observations document that the isolation of subpopulations with experimentally proven distinct reprogramming potentials are distinguishable by global gene expression patterns.

Remarkably, a comparison of relative gene expression changes among pairs of progressing cell populations at successive time points revealed two distinct waves of major gene activity (Figure 2D, left). The first wave occurred between days 0 and 3, whereas a second wave was detectable toward the end of reprogramming, after day 9. Refractory THY1+ cells initiated the first wave but failed to undergo the second wave (Figure 2D, right). Gene ontology (GO) analysis showed that expression changes within the first phase involved activation of processes related to cell proliferation, metabolism, cytoskeleton organization, and downregulation of genes associated with development. Genes upregulated during the second phase were associated with embryonic development and stem cell maintenance. A parallel study, which applied proteomics to study the same intermediates of reprogramming, concurs with our findings and further shows that molecular changes are highly coordinated during both phases[12]. Together, these data demonstrate that cells undergoing reprogramming into iPSCs, as defined by populations upregulating SSEA1+ at early time points and SSEA1+/*Oct4*-GFP at late time points, undergo a biphasic process at the transcriptional level that is separated by a period of less pronounced transcriptional change.

Defining Reprogramming-Specific Gene Expression Patterns

To gain further insights into the mechanisms of iPSC induction, we next determined categories of genes that changed their expression in characteristic patterns (Figure 2E and Table S1). A large number of genes became abruptly upregulated (cluster I; ~750 genes) or downregulated (cluster VI; ~1,200 genes) early in reprogramming and then remained largely unchanged until the iPSC state. Genes in these two categories were mainly involved in controlling DNA replication and cell division processes (upregulated genes) as well as cell adhesion and cell-cell contacts (downregulated genes) and account for the first transcriptional wave during reprogramming. Another category was comprised of ~400 genes that were gradually upregulated, such as the pluripotency-associated genes *Alpl*, *Fbx15,*

(continued from right) red trendlines depict gene expression patterns in cells undergoing successful reprogramming as defined in Figure 1D). Each gene is only represented once per category. (F) Expression analysis of candidate genes selected from (E) for overexpression or knockdown experiments shown in (G). Gene expression categories are shown in brackets. (G) Reprogramming potential of OKSM transgenic MEFs infected with lentiviral vectors expressing the indicated candidate genes or hairpins. Data are represented as mean ± SEM (n = 3). See also Figure S1.

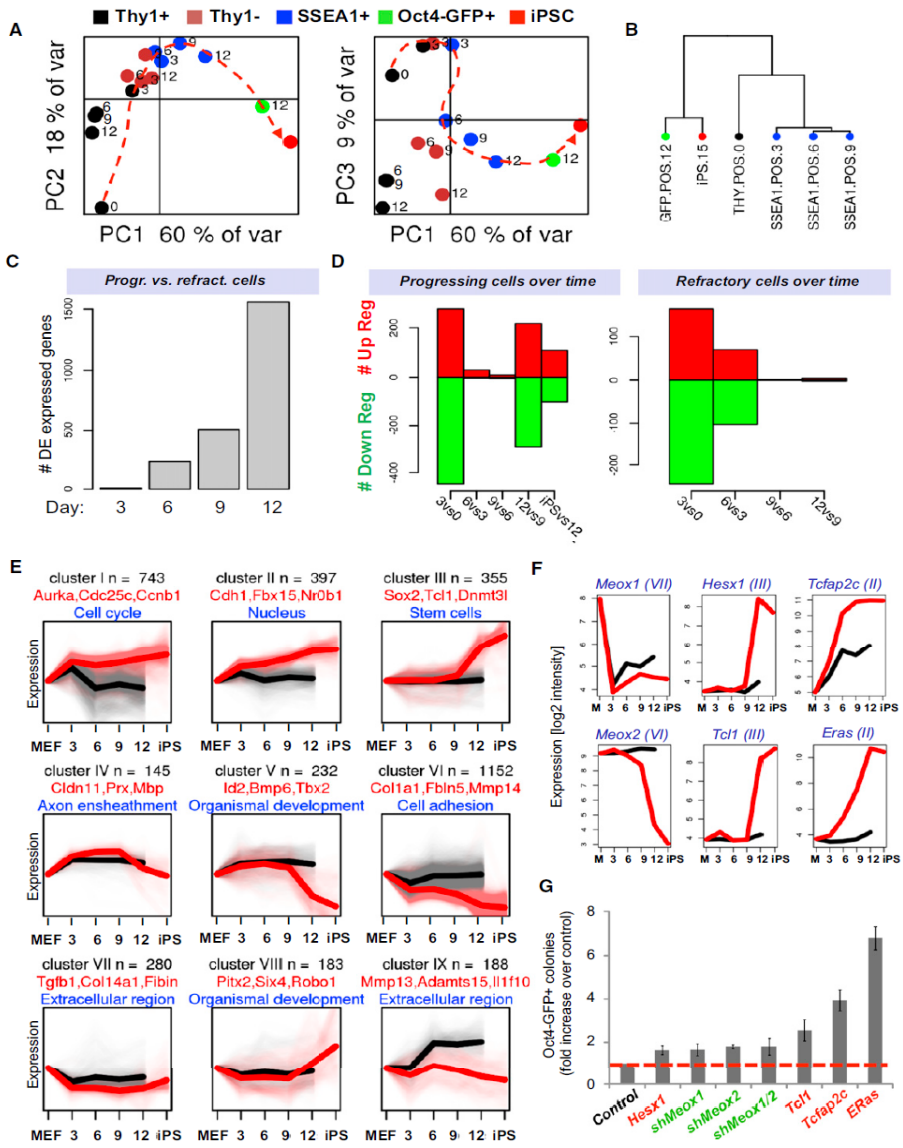


Figure 2. Gene Expression Dynamics during iPSC Formation. (A) Principal component analyses of global gene expression data of FACS-sorted subpopulations at indicated time points. (B) Unsupervised hierarchical clustering of gene expression profiles of indicated cell populations. (C) Number of differentially expressed genes between THY1+ and SSEA1+ cells at indicated time points. (D) Number of differentially expressed (DE) genes in progressing SSEA1+ cells at successive time points. Right panel shows gene expression changes in refractory THY1+ cells. (E) Gene expression categories (I to IX) clustered by common expression changes during reprogramming (black trendlines depict gene expression patterns in THY1+ population; (continued on left)

Nr0b1, Tcfcp2l1, and *Sall1* (cluster II), whereas roughly 350 genes were induced late during reprogramming and contained genes enriched for the categories stem cells and DNA binding (cluster III). The latter group, which contained well-known core pluripotency factors such as *Nanog*, *Oct4*, and Sox2 as well as *Esrrb*, *Dnmt3L*, *Tcl1*, and *Nr5a2*, is in part responsible for the second transcriptional wave and marks the acquisition of a stable pluripotent state.

Intriguingly, we also identified categories of genes that were either transiently upor downregulated during iPSC formation (cluster IV and VIII, respectively) or upregulated early and downregulated late (cluster V). Genes within those categories included a number of developmental and cell-type-specific regulators such as *Bcl11a*, *Prx*, and *Tbx21* among the transiently upregulated genes and *Spp1*, *Pitx2*, and *Six4* among the transiently downregulated genes (Figure 2E).

Lastly, we hypothesized that the manipulation of dynamically regulated genes from these categories might enhance reprogramming. We selected the *Akt* coactivator *Tcl1*, the transcription factors *Tcfap2c* and *Hesx1* and the ESC-specific *Ras* isoform *ERas* for overexpression experiments and the fibroblast-enriched genes *Meox1* and *Meox2* for knockdown experiments (Figure 2F). Accordingly, upregulation or downregulation of theses genes gave rise to up to seven times more *Oct4*-GFP+ colonies compared with control cells (Figure 2G). Collectively, these experiments prove that our gene expression categories facilitate the identification of molecules that positively or negatively influence the reprogramming process.

Gene Expression Patterns of Refractory Cells

Another category of genes (cluster IX) contained about 200 genes that were aberrantly activated in refractory THY1+ cells (Figure 2E and Table S1). Genes within this class were related to extracellular space/matrix, plasma membrane, retinoic acid binding, and immune response processes (e.g., *Mmp13, Rarres2, Fgf18, Fndc1, Aqp1* and *-4*, *Il1f10, Hsd11b1*, and Figure S1A) and likely contributed to the failure of THY1+ cells to reprogram.

To further understand the molecular reasons for the inability of THY1+ intermediates to reprogram, we analyzed other genes that were differentially expressed between SSEA1+ and THY1+ cells. This analysis revealed that mesenchymal genes were not properly downregulated, whereas epithelial genes failed to be upregulated in THY1+ cells compared to SSEA1+ cells after day 3 (Figure S1B). We also searched for differentially expressed genes between THY1+ and SSEA1+ cells at day 3, when overall gene expression patterns were still highly similar among all populations (Figure 2A). This analysis yielded a small number of significantly up and downregulated genes (e.g., *Il6, Nup210*, and *Bex1*) that might serve as valuable early discriminators between cells that succeed or fail in reprogramming

(Figure S1C). We conclude that THY1+ cells become refractory to reprogramming for a variety of reasons that include (1) an inability to undergo a MET, (2) aberrant activation of differentiation and immune-response-associated genes, and (3) a failure to maintain global gene expression trends beyond day 3.

Impact of Cellular Heterogeneity on Molecular Dissection of Reprogramming

We employed different strategies to determine the degree of heterogeneity among SSEA1+ cells. First, we used Fluidigm technology to perform single-cell expression analysis for 26 genes (see Figure S2A) in FACS-purified SSEA1+ intermediates at days 3, 6, and 9 as well as in day 0 THY1+ MEFs and established iPSCs. Correspondence analysis (COA) of all 26 genes across the different cellular groups showed that the three intermediate populations formed separate clusters that partially overlapped and gradually progressed from MEFs to iPSCs (Figure 3A). COA confirmed the early and late transcriptional waves and illustrated an increased degree of variation within SSEA1+ intermediates compared with MEFs and iPSCs (Figure 3B). Biplot analysis suggested that activation of *Nr5a2, ERas, Zfp42, Esrrb, Dnmt3l,* and *PECAM* was most informative for predicting the iPSC state, followed by activation of *Nanog, Lin28*, and *EpCAM* (Figure 3C). A comparison of the expression dynamics of individual genes between SSEA1+ bulk populations and single SSEA1+ cells revealed a similar overall kinetics and allowed us to differentiate between MEF- and ESC-associated genes that were either downregulated or upregulated immediately, gradually or late upon reprogramming factor expression (Figure 3D).

We next examined the shape of the violin plots (Figure 3D) in order to deduce whether gene expression changes took place in a minority or majority of SSEA1+ intermediates; whereas unimodal plots are consistent with uniform gene expression in a majority of cells, bimodal plots are indicative of distinct expression patterns and thus heterogeneous cell populations. We could distinguish between three characteristic patterns of gene expression change (Figure 3D and Figure S2B): (1) exclusively unimodal expression patterns, which were mostly characterized by genes that changed in one of the two transcriptional waves and included MEF genes that were silenced early (*Fibin, Snai1*) or gradually (*Fbn1*) as well as all examined pluripotency genes that were activated late (*Zfp42, Esrrb, Nr5a2, ERas, Lin28, PECAM, Tcl1, Dnmt3l*); (2) unimodal expression early and late with bimodal expression at intermediate stages of reprogramming, which contained the MEF gene *Zfpm* and the early iPSC marker *EpCAM, Nanog* and *Tcfap2c*; and (3) bimodal expression patterns at all time points. Examples included the MEF gene Hoxa10 and the intermediate-specific genes *Cldn11, Tbx21*, and *Six4*. Coexpression analysis of representative genes from each category showed that they were indeed activated within the same cells (Figure S2C).

To determine whether intermediate-specific genes were always expressed heterogeneously within SSEA1+ cells (Figure 3D), we performed immunohistochemistry (IHC) for PRX (Figure 3E) in reprogrammable MEFs induced with dox for 9 days. Costaining with antibodies recognizing SSEA1 and PRX revealed that all SSEA1+ cells also expressed PRX (28/28 examined cells) (Figure 3F). This result suggested that *Prx* and probably other intermediate-specific genes are expressed homogeneously among SSEA1+ cells and thus mark cells poised to becoming iPSCs. In contrast, other intermediate-specific genes such as *Tbx21* are expressed in more rare subsets of SSEA1+ cells (Figure 3D) whose fate remains unclear. To test whether activation of the latter group of genes correlated with their ability to form iPSCs, we infected tail fibroblasts isolated from *Tbx21-ZsGreen* mice with a polycistronic viral vector expressing OKSM. Flow cytometric analysis of SSEA1+ intermediates at days 6 and 9 confirmed the heterogeneous expression pattern (Figure 3G). Plating of equal numbers of SSEA1+ *Tbx21*-ZsGreen+ and of SSEA1+ *Tbx21*-ZsGreencells from day 6 on feeders gave rise to roughly equal numbers of iPSC colonies, indicating that *Tbx21* upregulation at this time point was neither necessary nor inhibitory for reprogramming (Figure 3H). However, SSEA1+ *Tbx21*-ZsGreen+ cells isolated at day 9 of reprogramming almost entirely lost their ability to form iPSC colonies, suggesting that a failure to downregulate this marker at later stages of reprogramming prohibited iPSC formation (Figures 3G and 3H). We conclude that our analysis of SSEA1+ bulk populations correlates well with expression patterns in individual SSEA1+ cells, thus validating our approach to study FACS-enriched intermediates of reprogramming. However, our observation that SSEA1+ cells exhibited some degree of heterogeneity warrants a search for markers that allow for further purification of reprogramming intermediates destined to form iPSCs (see Identification of Molecules to Enrich for Cells Poised to Becoming iPSCs).

Comparison with piPSCs and Bulk Populations Expressing OKSM

Partially reprogrammed iPSCs (piPSCs) are assumed to represent intermediate stages of reprogramming[13,14]. piPSC lines are stable cell lines that have silenced the somatic program

(Continued from right) Size of ovals indicates degree of variation. (C) Biplot displaying overlay of COA with genes associated with individual groups. (D) Comparison of Affymetrix (left) and single-cell (right) expression data for nine selected genes. Gene expression categories, as defined in Figure 2E, of selected candidates are shown in brackets. (E) Immunofluorescence for OCT4, SSEA1, and PRX on reprogrammable MEFs treated with dox for 9 days. (F) Quantification of data shown in (E). (G) FACS analysis of *Tbx21-ZsGreen* tail fibroblasts infected with dox-inducible lentivirus expressing OKSM at indicated time points. (H) Reprogramming potentials of indicated cell populations at days 6 and 9. Data are represented as mean ± SEM (n = 3). See also Figure S2.

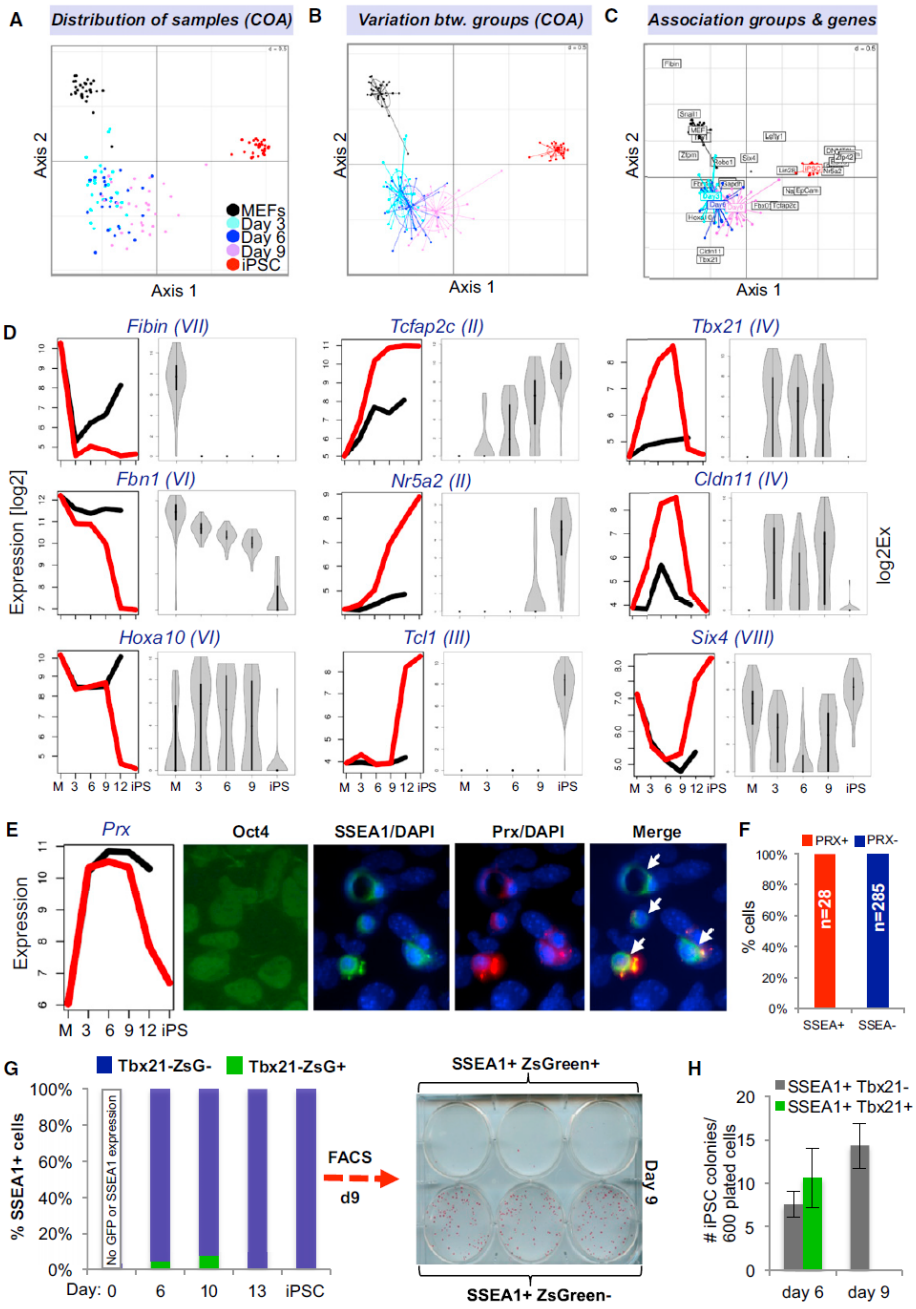


Figure 3. Characterization of Cellular Heterogeneity within SSEA1+ Cells. (A) Correspondence analysis (COA) of single-cell expression data obtained with Fluidigm technology for 26 genes in indicated cell populations. (B) COA of same groups as shown in (A) illustrates variation in gene expression. (Continued on left)

but failed to activate the pluripotency program and depend on continuous expression of viral transgenes. To assess whether their overall gene expression signature closely resembled any of our profiled cell populations, we performed PCA analysis between our data sets and published results from six different piPSC lines. Consistent with previous observations, piPSC lines derived from distinct cell types and produced in different laboratories clustered together, suggesting a similar molecular makeup (Figure S1D, gray symbols). Unexpectedly, these cell lines were quite distinct from any of our profiled intermediate populations along the depicted PC axes, showing essentially no overlap with one notable exception; piPSC line BIV1[13] that was generated with dox-inducible lentiviruses clustered together with retrovirally induced piPSCs in the presence of dox (''dox+'' marked triangle) but grouped with our late SSEA1+ intermediates after ~10 days of dox withdrawal (''dox-'' triangle). We conclude that piPSCs originate from cells that have exited the normal reprogramming route at an early time point and became immortalized, hence showing little overlap with progressing intermediates.

A comparison of gene expression profiles from SSEA1+ intermediates with those obtained from reprogrammable ''secondary'' cells exposed as bulk populations to dox for 0, 4, 8, 12, or 16 days[13] further showed that the latter samples clustered most closely with THY1+ and THY1 cells around day 3 but not with SSEA1+ intermediates at comparable later time points (Figure S1D, turquois triangles). This finding indicated that this previous study of bulk populations predominantly captured expression changes of cells that failed to reprogram after day 3 and underscores the importance of enriching for the rare subsets of cells that are prone to generating iPSCs, particularly at later stages of reprogramming.

Transcription Dynamics Predicts Distinct Reprogramming Factor Activities

We next wondered whether the biphasic transcriptional pattern could be explained by the activity of any individual or combinations of transcription factors. To this end, we compared our gene expression data with published genome-wide occupancy studies for OCT4, SOX2, KLF4, c-MYC, and NANOG in pluripotent stem cells (see Table S1 and Methods). Whereas a similar number of targets of OCT4, SOX2, and KLF4 were upand downregulated during both transcriptional waves, targets of c-MYC were mostly upregulated (~80%) with a bias for the first wave (Figure S3A).

Given that many pluripotency-associated genes are targets of OCT4, SOX2, KLF4, and c-MYC in ESCs/iPSCs, we next distinguished between expression changes of individual and combinatorial OKSM targets during both waves. This analysis confirmed that c-MYC alone or in combination with other factors is a dominant force behind early gene induction (shown for c-MYC and KLF4 targets in Figure S3B). To assess the contributions from individual

factors to reprogramming, we applied a mathematical approach that models and predicts transcription factor activities by using network component analysis[15]. As expected, c-MYC targets showed a striking upregulation during the first few days of OKSM expression, with no major changes detectable until the end of reprogramming (Figure 4A, left).

Notably, an analysis of transcriptional activity for genes that are bound by pairs of factors in ESCs showed a gradual change, as exemplified for OCT4/SOX2(OS) targets, suggesting that the combined activity of certain pluripotency factors is more likely to modulate targets than individual factors (Figure 4A, middle). To experimentally verify this mathematical prediction, we picked 3 OS targets that became upregulated early (*Fut9*, day 3) or late (*Nanog and Lefty1*, days 9–12) during reprogramming and performed chromatin immunoprecipitation (ChIP) on SSEA1+ intermediates with OCT4 and SOX2-specific antibodies, combined with real time PCR. Indeed, we found that *Fut9* was occupied by both OCT4 and SOX2 as early as day 3, whereas *Nanog and Lefty1* were occupied by OCT4 alone at early time points and by both OCT4 and SOX2 by day 12, consistent with their robust transcriptional activation (Figure 4B and Figure S3C). The different susceptibilities of OS targets to be transcriptionally activated correlated well with an underlying permissive or repressive chromatin structure (Figure S3D).

KLF4 was exceptional in that its targets changed their expression early and late in reprogramming with a phase of less activity change during intermediate stages (days 3 to 9), supporting a possible dual role of KLF4 in early somatic gene repression and subsequent pluripotency gene activation (Figure 4A, right). Accordingly, regulated KLF4 targets were comprised of factors associated with differentiation, such as *Tgfb1, Pdgfra, and Col6a1,* at early time points and of pluripotency-associated genes including *Pou5f1 (Oct4), Tdgf1, and Klf5* at late time points of reprogramming. Altogether, these results suggest that the first transcriptional wave is mostly mediated by c-MYC and occurs in both progressing and nonprogressing cells, whereas the second wave is the consequence of a gradual upregulation of OS targets, ultimately leading to the activation of other pluripotency genes, including Nanog, to consolidate the pluripotent transcription factor network. KLF4 seems to support both phases by suppressing genes during the first phase and enhancing pluripotency gene expression during the second phase.

MicroRNA Expression Follows Biphasic Pattern and Inversely Correlates with Known and Predicted Target mRNAs

Similar to the expression analysis for coding genes, miRNA expression analysis allowed us to cluster cell populations into different groups based on their phenotype by using PCA and unsupervised clustering (Figures 4C–4E and Table S2). Pairwise comparisons of progressing

SSEA1+ populations at successive time points again revealed two transcriptional waves, which both showed an over representation of downregulated versus upregulated miRNAs (Figure 4F). miRNAs changed their expression in similar patterns to mRNAs over the course of reprogramming (see Figure 4G for representative examples). Moreover, miRNAs that have previously been documented to inhibit (e.g., *let-7*, miR34c) or promote (e.g., *miR-294*, miR-106a)[16] iPSC formation, showed the expected downregulation and upregulation, respectively, in progressing intermediates (Figure 4G and Figure S3E). We conclude that forced expression of OKSM controls the expression of both coding and noncoding loci in a similar fashion.

A comparison of many miRNAs and their known targets indicated an inverse correlation (Figure S3E). This is exemplified for *miR-294*, which targets *TgfbR2*, and for *let-7*, which targets *Lin28*[17]. To extend this analysis beyond well-established miRNA-mRNA pairs, we built a table that links the expression changes during reprogramming of all differentially expressed miRNAs to their putative mRNA targets (Figure 4H, Figure S3F and Extended Experimental Procedures). Indeed, the previously validated *let-7*c targets *Lin28, N-Myc, and Sall4* and the *miR-294* targets *Lats2, TgfbR2, and Akt1* exhibited high negative correlation scores (Figure S3F). This analysis further suggested that the pluripotency factor *Esrrb*, the histone methyltransferases *Suv39h1/2*, and the coactivator *Ncoa3*, all of which are implicated in cellular reprogramming, are likely targets of *let-7*c, whereas the documented iPSC-inhibitory factor *Prrx1*[18] is predicted to be targeted by *miR-294*.

Unexpectedly, *Mir-302a*, whose forced expression was also shown to enhance iPSC generation in the context of the Yamanaka factors[17], exhibited transient activation specifically in *Oct4*-GFP+ cells at day 12 but remained otherwise unchanged (Figures S3G and S3H). *Mir-302a* is normally expressed in mouse epiblast stem cells[16] but barely detectable in mouse ESCs, suggesting that iPSC induction might entail a transient passage through an epiblast-like state before reaching naive pluripotency. In agreement with this idea, we detected

(Continued from right) (C) Principle component analysis of microRNA expression data of FACS-sorted subpopulations at the indicated time points. (D) Unsupervised hierarchical clustering of indicated microRNA expression profiles. (E) Number of differentially expressed (DE) microRNAs between THY1+ and SSEA1+ cells at indicated time points. (F) Number of differentially expressed microRNAs between progressing SSEA1+ cell populations at successive time points. (G) Examples of microRNA profiles that change in dynamic patterns. (H) Predicted target genes of *let-7*c (Targetscan database) are shown based on inverse expression patterns with *let-7*c. Examples of putative targets with an inverse expression score of 0.8 or lower are shown. Targets marked by red asterisks have previously been validated. (I and J) iPSC formation efficiencies (I) and *Oct4*-GFP FACS quantification of reprogrammable MEFs treated with mimics for miR-182 or miR-214 (J). Data are represented as mean ± SEM (n = 3). See also Figure S3.

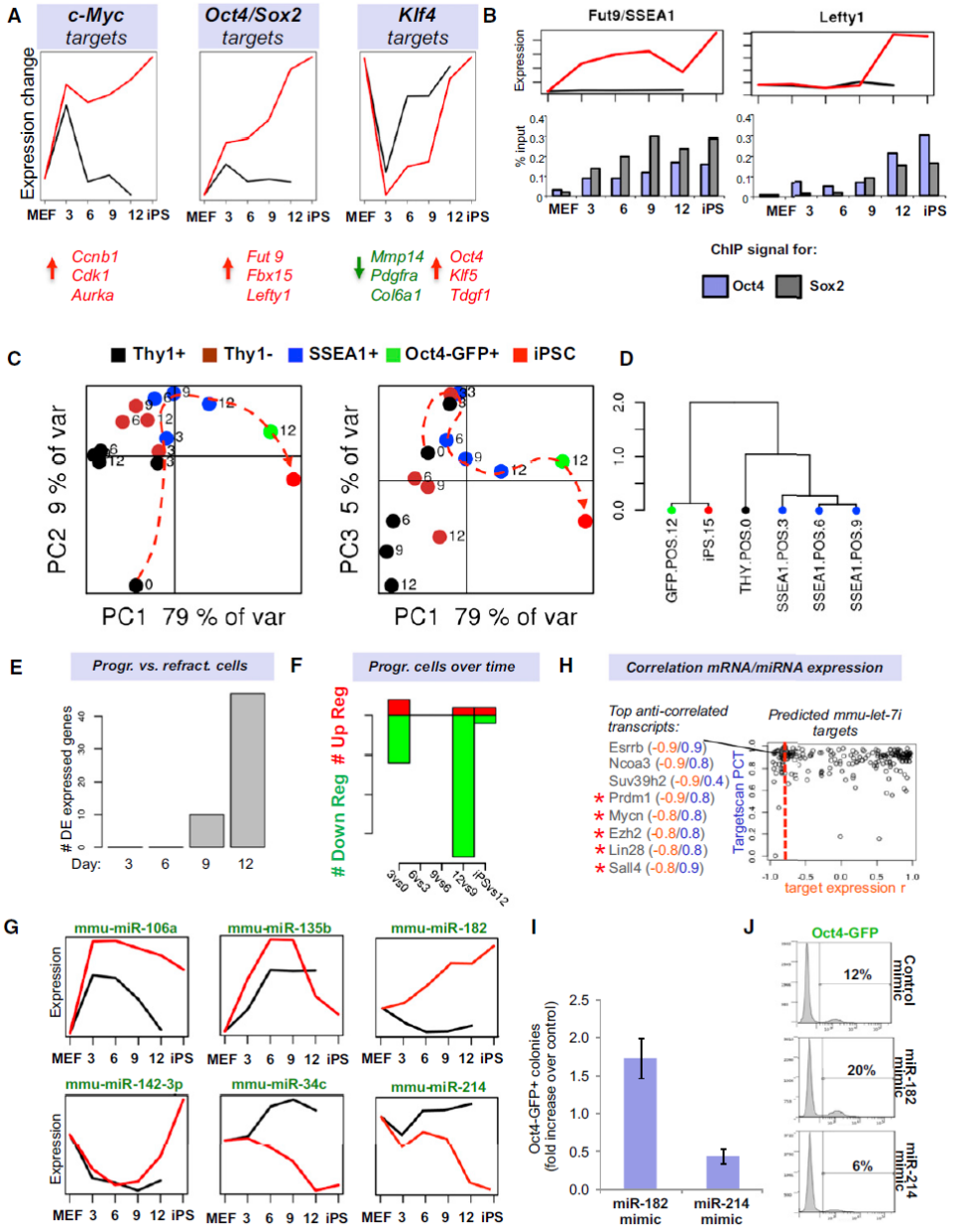


Figure 4. Predicted Reprogramming Factor Activities and MicroRNA Expression Dynamics (A) Transcription factor (TF) activities for c-MYC, KLF4 and the SOX2-OCT4 dimer based on network component analysis. Shown below are examples of activated (red) or repressed (green) targets. (B) Expression dynamics of an early (*Fut9*; left) and late (*Lefty1*; right) OCT4/SOX2 target during reprogramming. Shown below is promoter ChIP analysis for OCT4 and SOX2. (Continued on left)

transient downregulation of a number of putative *Mir-302a* targets in *Oct4*-GFP+ cells such as *Rbl2, Rab11fip5, Rbbp6*, and transient, albeit modest, upregulation of epiblast stem cell-associated markers including *Brachyury (T), Cer1, Foxa2, Eomes,* and *Fgf5* (Figures S3G and S3I). It remains to be tested whether activation of *Mir-302a* and associated transcripts takes place in all or only a subset of *Oct4*-expressing intermediate cells. Also, we cannot rule out the possibility that these transient expression changes are the result of aberrant targeting of exogenous OKSM to the respective genes, resulting in their activation or repression.

We finally wondered whether our miRNA expression data would allow us to identify novel modulators of reprogramming. Indeed, gain-of-function of *miR-182* by mimics increased, whereas that of *miR-214* decreased iPSC formation, consistent with their transcriptional changes during reprogramming (Figures 4G, 4I, and 4J).

Differential Chromatin States Provide an Epigenetic Logic for Early, Gradual, and Late Gene Regulation

We and others have previously shown that both transcriptional and DNA/histone methylation patterns are reset from a somatic state to a pluripotent state upon reprogramming into iPSCs[13,19]. However, it is unknown when these epigenetic changes occur during reprogramming and whether there is a hierarchy in their establishment and erasure, respectively. In an attempt to resolve these questions, we analyzed active and repressive histone methylation marks (histone H3 lysine 4 and lysine 27 trimethylation H3K4me3/H3K27me3) and DNA methylation patterns at a genome-wide scale in SSEA1+ cells throughout reprogramming. Analysis of H3K4me3 and H3K27me3 ChIP-seq patterns in progressing intermediates showed two waves (Figures 5A and 5B), which coincided with the observed mRNA and miRNA phases (Figures 5C and 5D). Kinetic analysis of bivalency formation (H3K4me3/H3K27me3 enriched promoters)[20] at genes that changed expression during reprogramming showed an initial burst of ~110 targets by day 3, which gradually increased to ~130 at day 9, ~160 by day 12, and ~180 in established iPSCs (Figure 5F, Figure S4A). Thus, activating and repressive histone marks individually exhibit a biphasic pattern akin to coding and noncoding genes, whereas the establishment of differentially expressed bivalent promoters is a more gradual process.

An examination of histone marks at genes that changed their transcription in characteristic patterns allowed us to study their underlying chromatin dynamics (Figure 5G and Table S1). Consistent with the different gene expression categories (Figure 2E), we were able to distinguish between genes that changed their H3K4me3 and H3K27me3 status early (day 3), at intermediate stages (day 6–9), or late (day 12). For example, the fibroblast-associated gene *Pdgfrb* was downregulated by day 3, which coincided with the early loss

of H3K4me3 and subsequent acquisition of H3K27me3 marks, suggesting efficient access of this locus by chromatin silencers. In contrast, the MEF expressed gene *Zfpm2* showed a gradual decrease in H3K4me3 marks and a concomitant increase in H3K27me3 marks, resulting in a bivalent state and transcriptional silencing around day 9, whereas the *Lats2* gene became decorated by H3K27me3 and was transcriptionally silenced only by day 12.

Similar to the deposition of H3K27me3 marks and the concomitant silencing of MEF-specific genes, we observed distinct classes of pluripotency genes that gained H3K4me3 and lost H3K27me3 at different time points (e.g., *Fgf4, Sall4, and Lin28*) (Figure 5G). Lastly, genes that changed their expression transiently acquired H3K4me3 or H3K27me3 in a temporally restricted manner (e.g., *Prx, Klf2*) (Figure 5G). We deduce from these results that the kinetics of silencing of MEF genes and activation of ESC genes is determined by a combination of parameters, including the type and complexity of underlying histone modifications as well as the availability and accessibility of transcription factors to regulate a given target (see Figure 4B). Indeed, the vast majority of genes (~90%) that were activated early or gradually (categories I and II; see Figure 2E) already carried activating H3K4me3 marks in MEFs (Figure 5E and Figure S4B), thus complementing observations made in a previous study (Koche et al., 2011). In contrast, genes that were activated late (category III, e.g., *Oct4, Nanog*) are often unmarked (~15% of genes) or bivalent (~15%) in MEFs, suggesting that loci associated with these chromatin patterns are more resistant to transcriptional activation.

DNA Methylation Patterns Are Reset Late in Reprogramming

In contrast to gene expression and histone modification patterns, genome-wide promoter DNA methylation changes occurred predominantly late in reprogramming as determined by HELP analysis (Figure 5H). Equal numbers of methylated restriction sites were gained and lost after day 9, indicating that a comparable number of loci became methylated and demethylated, respectively. In agreement, we found that enzymes implicated in DNA methylation and demethylation, such as *Dnmt3a, Dnmt3L, Apobec2, and Tet1* were transcriptionally upregulated late and specifically in SSEA1+ cells (Figure 5I). To confirm these global changes of DNA methylation at single base-resolution, we investigated promoter methylation levels at a number of candidate loci by mass array EpiTYPEr on genomic DNA isolated from SSEA1+ intermediates. We found that pluripotency-associated genes, such as *Nanog, OCT4, and Zfp42 (Rex1*) became demethylated very late during reprogramming (~days 9–12) (Figure 5J and Table S3). Similarly, genes that are normally methylated in pluripotent cells but demethylated in fibroblasts, including *HoxA10 and Gja8*, became de novo methylated late.

Rescue of Refractory Cells by Increased OKSM Expression

A comparison of gene expression intensities among the different subpopulations showed that THY1 and THY1+ cells generally failed to regulate ESC-enriched and MEF-enriched mRNAs and miRNAs to the same extent as SSEA1+ cells (Figures 6A and 6B). Although this observation cannot be explained by differential transcription of the OKSM transgene in these subpopulations (Figure 6C), we were surprised to detect substantially increased protein levels for OCT4 in SSEA1+ cells compared with THY1+ cells (Figures 6D and 6E). Consistently, we observed a 20%– 25% reduction in the number of THY1+ and THY1 cells and a concomitant 400% increase in the number of SSEA1+ cells when inducing reprogrammable MEFs carrying two copies of the OKSM cassette and *Rosa26-M2rtTA* allele (Ho/Ho), respectively, with dox compared with MEFs that only contained one copy of each transgene (Het/Het) (Figure 6F).

To test whether elevated OKSM protein levels could rescue refractory THY1+ cells at different stages of reprogramming, we infected THY1+ cells isolated at days 3, 6, 9, and 12 of dox induction from Het/Het reprogrammable MEFs with viral vectors expressing additional copies of OKSM (Figure 6G). Remarkably, THY1+ cells receiving extra copies of OKSM, but not untreated control cells or cells infected with c-MYC vector alone, gave rise to a substantial number of *Oct4*-GFP+ iPSC colonies (Figures 6H and 6I). This result thus documents that the inability to sustain OKSM protein expression in THY1+ cells on or after day 3 contributes to their failure to form iPSCs.

Identification of Molecules to Enrich for Cells Poised to Becoming iPSCs

In a last set of experiments, we aimed to identify new surface markers that would allow further enrichment for subpopulations of cells undergoing reprogramming in comparison with THY1, SSEA1, and *Oct4*-GFP expression. We focused on the molecules c-KIT, EpCAM, and PECAM1 because of their expression patterns specifically in SSEA1+ intermediates (EpCAM, early gene; c-Kit, intermediate gene; PECAM1, late gene) (Figure 7A and Figure

(continued from right) in relation to their chromatin status in MEFs. (F) Number of differentially expressed genes that become bivalent (H3K27me3 and H3K4me3 enriched) during reprogramming (red dots = bivalent promoters) and quantification.
(G) Integration of gene expression and histone modification data define subsets of genes with characteristic expression changes. Shown are examples of fibroblast-associated (top), pluripotency-associated (center), and transiently changing genes (bottom). (H) Number of genes, which change DNA methylation status in progressing cell populations during reprogramming as determined by genome-wide methylation analysis (HELP). (I) Expression dynamics of candidate genes associated with DNA methylation and demethylation. (J) Heatmap of DNA methylation patterns of specific CpGs (boxes) in the promoter regions of indicated genes during reprogramming with EpiTYPER DNA methylation analyses. Yellow indicates 0% methylation, and blue represents 100% methylation. See also Figure S4.

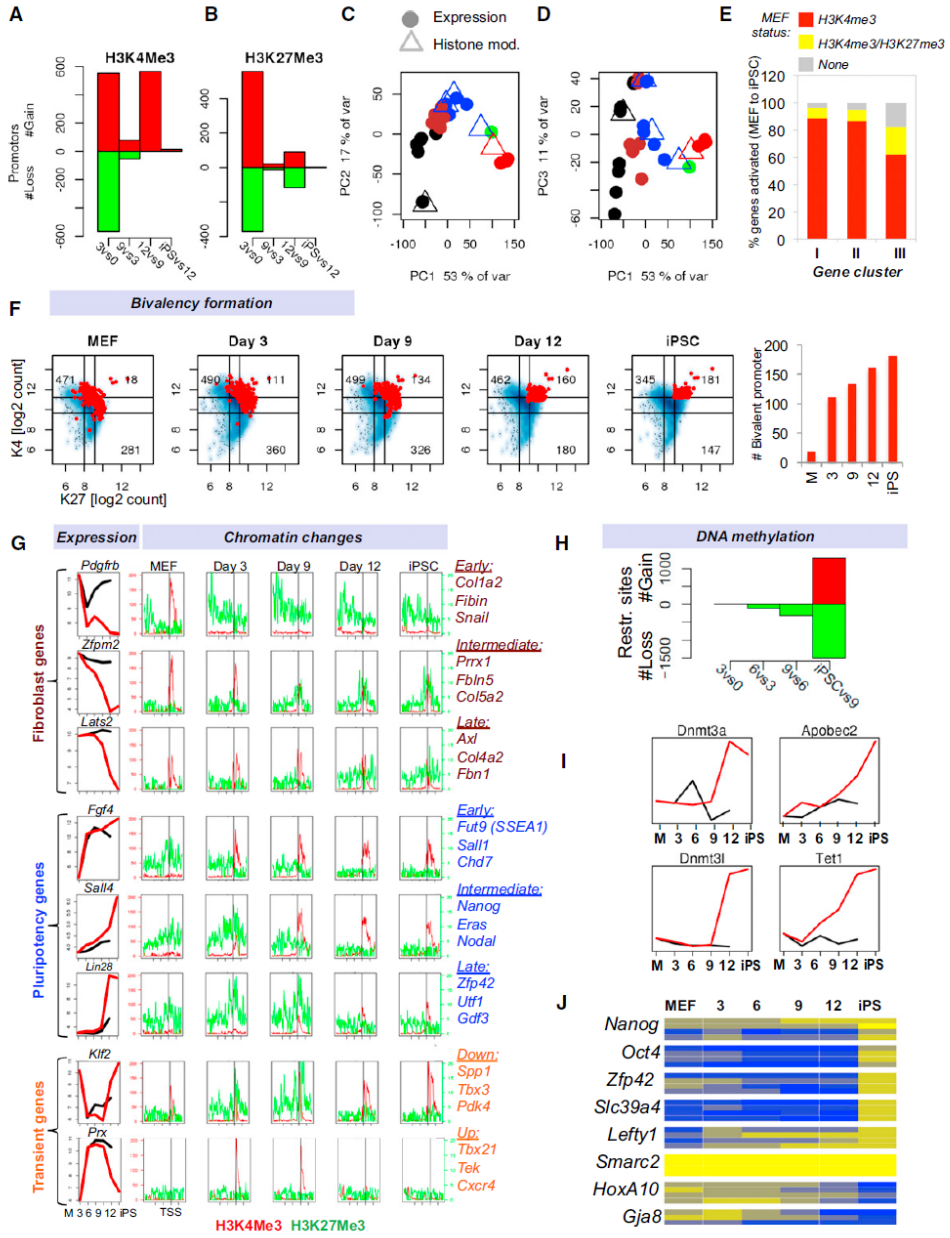


Figure 5. Histone and DNA Methylation Dynamics during Cellular Reprogramming (A and B) Enrichment for H3K4me3 or H3K27me3 at promoters of differentially expressed genes in progressing intermediates. (C and D) Superimposition of principal component analyses for genes enriched in H3K4me3 (C) or H3K27me3 (D) (triangles) with gene expression data (circles) of the same cell populations (see Figure 2A for color coding). (E) Display of activated genes from gene expression categories I, II, and III (see Figure 2E) (Continued on left)

S5A). Notably, EpCAM+ cells were first detectable at day 6 in a fraction (~25%) of SSEA1+ cells. In contrast, c-Kit became upregulated only by day 9 in ~25% of SSEA1+ cells, whereas PECAM1 was detectable exclusively in SSEA1+, *Oct4*-GFP+ cells at day 9. Altogether, these results show that EpCAM, c-Kit, and PECAM1 become activated at successive time points in subsets of SSEA1+ cells. This experiment further documents that SSEA1+ intermediates that activated the endogenous *Oct4(-GFP)* locus are generally more homogeneous for these three markers than are SSEA1+ *Oct4*-GFP+ cells at day 6 and day 9.

To assess the functional value of these markers, we sorted SSEA1+ EpCAM and SSEA1+ EpCAM+ cells 6 days after dox induction and plated equal numbers on feeders for another 8 days in the presence of dox. Counting of dox-independent alkaline phosphatase-positive colonies 5 days later showed a modest but significant increase in reprogramming efficiency among SSEA1+ EpCAM+ intermediates (Figure 7B). In agreement, expression analysis of both subpopulations revealed subtle differences with 68 genes being upregulated and 48 genes being downregulated more than 2-fold (Figure 7C). Upregulated genes included *Nanog* (11-fold), *ERas* (7-fold), *Sox2, Nr0b1, Sall4, Nr5a2,* and *Tdgf1* (3to 5-fold). Surprisingly, the core transcription factor *Oct4* was not differentially expressed, which is consistent with the absence of *Oct4*-GFP signal at this time point. At the epigenetic level, *Nanog* promoter methylation levels were reduced by 50%, whereas OCT4 promoter methylation levels decreased only mildly (Figure 7D and Figure S5B). These findings thus suggest that there is a hierarchy in the activation of core pluripotency factors within SSEA1+ cells with *Nanog* and associated transcripts being activated before.

Discussion

Our results constitute a comprehensive analysis of transcriptional and epigenetic changes in phenotypically defined intermediates of iPSC induction. These data have elucidated the identity and order of molecular changes inherent to transcription factor-induced reprogramming (see Figure 7 for summary of observations) and provide a rich resource of data to further dissect the mechanisms of induced cell fate transitions. Our findings suggest that the reprogramming of somatic cells follows a similar sequence of epigenetic changes as is seen during normal somatic cell differentiation; differentiating cells are thought to undergo transcriptional and histone modification changes before DNA methylation changes[21]. It will be interesting to assess whether the rather abrupt loss of methylation after day 9 is solely the consequence of a replication-dependent passive process or also involves active demethylation. Notably, methylation changes coincided with the acquisition of a stably reprogrammed state and are in line with the interpretation that methylation patterns stably lock in the reprogrammed state.

Our molecular analysis allowed us to define nine categories of dynamically expressed genes, which characterize distinct stages of reprogramming and whose overexpression or knockdown enhanced iPSC formation. We surmise that a failure to activate these (Figure 2F) and related genes constitute roadblocks of reprogramming and is part of the reason why iPSC formation is inefficient and takes relatively long. Our observation of two major transcriptional waves and an intermediate period of less transcriptional change is reminiscent of a previous study that has identified an ''initiation,'' ''maturation,'' and ''stabilization'' phase when examining bulk populations of cells expressing OKSM[6]. The observed transient activation/repression of developmental regulators may indicate that the OKSM proteins aberrantly activate/repress these targets or, alternatively, that (some) reprogramming intermediates undergo a transient phase of transdifferentiation or dedifferentiation as part of the reprogramming process. Our results might thus explain recent successes in deriving epiblast stem cells, neural progenitors, or cardiomyocytes directly from fibroblasts upon brief expression of OKSM and exposure to culture conditions conducive of the respective cell type[22]. This finding could be potentially exploited to generate other desired cell fates directly from fibroblasts. We have applied a mathematical model that faithfully predicts activation of OKSM targets in the course of reprogramming. Our results provide a transcriptional logic for the previously seen early requirement for c-MYC[14] and the late requirement for SOX2[23] during reprogramming and suggest an unanticipated dual function for KLF4 by predominantly repressing somatic targets early and activating pluripotency targets late in iPSC formation. Our finding that THY1+ refractory cells produce less OKSM protein and thus fail to properly regulate target gene expression compared with SSEA1+ cells warrants further examination. One plausible molecular explanation is that the OKSM factors are prone to more ubiquitination-mediated degradation in THY1+ cells[24].

We recognize the fact that SSEA1+ cells, although enriched for cells poised to forming iPSCs, still exhibit some degree of heterogeneity. Single-cell analysis of 26 genes as well as FACS analysis of three additional genes (*EpCAM, c-Kit,* and *PECAM*) documented that gene expression changes occur more homogeneously at early (0–3 days) and late time points (day 9 onward), whereas they are more heterogeneous at intermediate stages (days 6–9). There is some debate as to whether reprogramming entails a hierarchic/deterministic or probabilistic/stochastic process[25]. A previous study identified an early deterministic phase of reprogramming[26], whereas another recent report concluded that reprogramming involves an early stochastic and a late deterministic phase[27]. Our data may explain both observations, and we therefore suggest that iPSC formation follows an early and late deterministic phase, which is separated by a more probabilistic phase. With the aid of new surface marker, such as EpCAM, PECAM, and c-KIT, or novel reporter alleles, it may be possible to identify rare intermediates that progress toward iPSCs in a purely deterministic manner.

Methods

Reprogramming Experiments

Murine Embryonic fibroblast (MEF) cultures were established from E13.5 embryos from reprogrammable mice carrying one (het/het) or two (ho/ho) copies of the OKSM cassette and ROSA26-M2rtTA allele[10]. Reprogramming was performed in ESC medium (KODMEM with 15% FBS, L-Glutamin, penicillin-streptomycin, nonessential amino acids, b-mercaptoethanol and 1000 U/ml LIF) in the presence of doxycycline. Tail fibroblasts were established from 3cm tail snips after removing skin and hair and chopping up the tail with scalpels in trypsin. Reprogramming was initiated by direct infection of cells with two separate lentiviral vectors delivering *tetOP-OKSM and M2rtTA*, respectively as described previously[9,28]. All procedures, including maintenance of animals, were performed according to a mouse protocol approved by the MGH Subcommittee on Research Animal Care. *Tbx21-ZsGreen* mice were generated as previously described[29]; these mice were bred and maintained in the NIAID-specific pathogen free animal facility and the experiments were done under a protocol approved by the NIAID Animal Care and Use Committee.

Viral Vectors and Production

293T cells were transfected with lentiviral vectors, VSV-G and Δ8.9 plasmids and Fugene. Viral supernatant was collected after 36, 48 and 60 hr. postinfection and filtered through a 40um filter. Supernatant was concentrated approximately 100-fold by ultracentrifugation at 20,000rpm for 1.5 hr. at 4C, resuspended in 300 µl 1× PBS, and stored at −80°C. Infections were carried out in 1 ml medium with polybrene diluted 1:2,000 and 20 µl of each viral concentrate were used per well of a 6-well plate for 12 hr. Infected cells were trypsinized and replated in triplicate at densities of 3,000-5,000 cells per well of a 6-well on irradiated feeder cells for reprogramming experiments.

Calculation of Reprogramming Efficiencies

Equal numbers of cells were plated in the absence or presence of doxycycline on 100 mm dishes coated with gelatin and containing a layer of irradiated MEF feeders. Efficiencies

(Continued from right) (G) Experimental outline to rescue refractory THY1+ cells by supplying viral copies of OKSM. (H) Alkaline phosphatase stained colonies obtained after infecting THY1+ reprogrammable cells at indicated days with a dox-inducible vector expressing OKSM. Controls were uninfected, dox-treated THY1+ cells (''Dox'') and THY1+ cells infected with c-MYC virus alone. Representative *Oct4*-GFP+ colonies are shown at the bottom. (I) Quantification of results in (G). Data are represented as mean ± S.E.M. (n = 3)

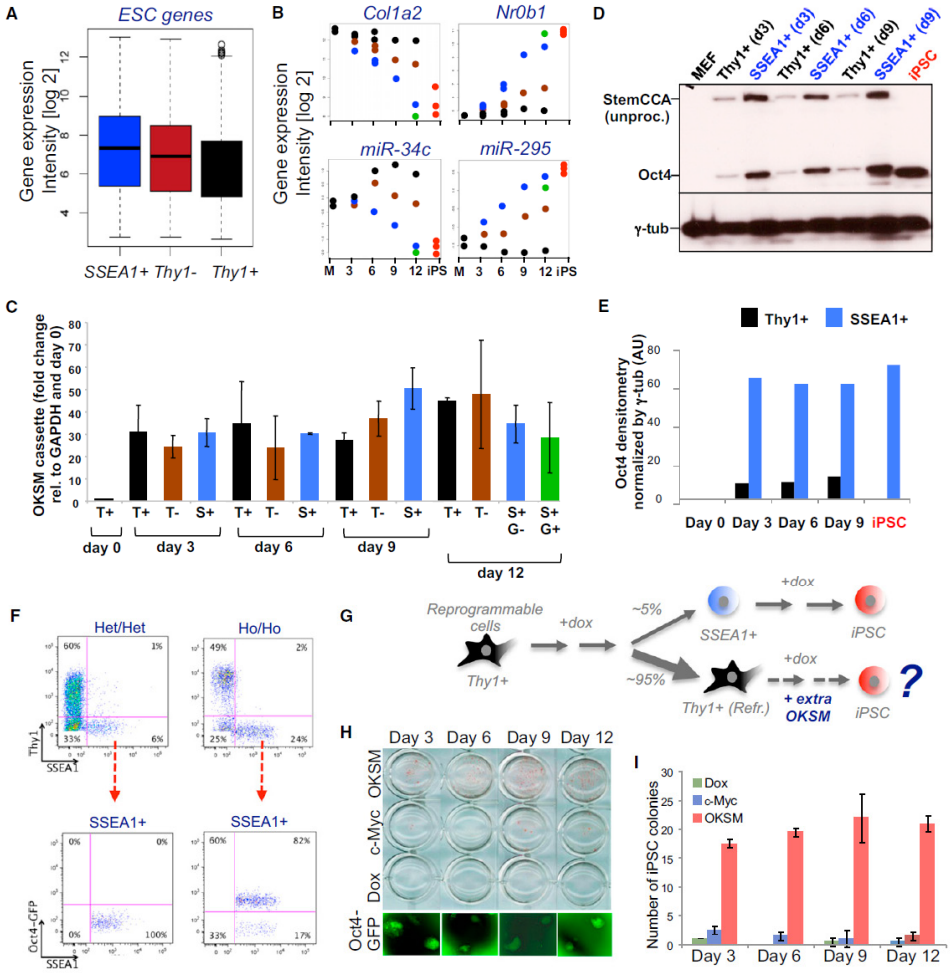


Figure 6. Rescue of Refractory THY1+ Cells by Increased OKSM Expression. (A) Box plot depicting average expression levels in THY1+ cells, THY1 cells and SSEA1+ cells of genes that are significantly upregulated between MEFs and iPSCs. (B) Expression dynamics of indicated MEF-associated and ESC-associated transcripts in THY1+, THY1 and SSEA1+ cells (see **Figure** 2A for color coding). (C) Exogenous OCT4 expression levels, normalized to GAPDH for the indicated cell populations and time points.

(D) Western blot analysis for OCT4 and gamma-tubulin (g-tub) for the indicated cell populations and time points. Higher molecular weight band for exogenous OCT4 compared with endogenous OCT4 (iPSC) reflects unprocessed protein originating from the polycistronic construct as described previously by Carey et al. (2011)[35]. (E) Densitometric quantification of western blot analysis shown in (D). AU is an abbreviation for arbitrary units.

(F) FACS analysis of reprogrammable fibroblasts carrying one (Het/Het) or two (Ho/Ho) copies each of the OKSM cassette and Rosa26-M2rtTA allele (top row). Bottom row shows FACS analysis for SSEA1 and *Oct4*-GFP of the same samples. (Continued on left)

were determined ~20 days later by dividing the number of Alkaline Phosphatase positive or *Oct-4*-GFP positive iPSC colonies that grew after withdrawal of doxycycline by the number of seeded cells. For determining reprogramming efficiencies of FACS-sorted intermediates at different time points (Figure 1), isolated cells were plated on feeders and grown in the presence of dox for an additional number of days, so that every cell population was exposed to dox for a total of 12 days, followed by withdrawal of dox for 5 days before colonies were scored.

Alkaline Phosphatase Staining

Alkaline phosphatase (AP) staining was performed with an Alkaline Phosphatase substate kit (Vector laboratories) according to manufacturer's recommendations.

RNA Analysis, mRNA profiling and microRNA profiling

iPSCs were harvested when they reached about 50% confluency and preplated on nongelatinized T25 flasks for 45 min to remove feeder cells. Intermediate cell types were harvested after FACS sorting. Cells were spun down and the pellet used for isolation of total RNA with the miRNeasy Mini Kit (QIAGEN) without DNase digestion. RNA was eluted from the columns with 50 ml RNase-free water or TE buffer, pH 7.5 (10 mM Tris-HCl and 0.1 mM EDTA) and quantified with a Nanodrop machine (Nanodrop Technologies).
Total RNA samples (RIN > 8) were subjected to transcriptomal analyses with Affymetrix HTMG- 430A mRNA expression microarrays as previously described. All microarray data will be available from the GEO repository.
For miRNA, RNA was isolated with the miRNeasy Mini Kit (QIAGEN) and submitted to Exiqon for micro-RNA profiling with the miRNAmiCURY LNA microRNA Profiling platform. Each sample was labeled with the miRCURY Hy3/Hy5 power labeling kit and hybridized on Exiqon BWA 0.5.9 arrays against a pool of all samples. All microarray data will be available from the GEO repository.

Quantitative PCR Detection of mRNA and miRNA

cDNA was produced with the first Strand cDNA Synthesis Kit (Roche) with 1 mg of total RNA input. Real-time quantitative PCR reactions were set up in triplicate with 5 ml of cDNA (1:100 dilution) with the Brilliant II SYBR Green QPCR Master Mix (Stratagene) and run on a Mx3000P QPCR System (Stratagene). Primer sequences are listed in Table S1.
miRNAs were extracted from MEFs, iPSCs and reprogramming intermediates (SSEA1+ cells at day 3, SSEA1+ cells at day 6, SSEA1+ cells at day 9, SSEA1+/*Oct4*-GFP+ cells at day 12) with the mirVana miRNA isolation kit (Life Technologies). Using the first Strand III kit (Life

Technologies), we reverse transcribed 10 ng of RNA into cDNA by using specific primers for miRNA-302a or housekeeper snoRNA-135 (Life Technologies). Subsequently Taqman assays (LifeTechnologies) against miRNA-302a or snoRNA-135 were used for quantification on a LightCycler 480 instrument (Roche). The housekeeping snoRNA-135 was used for standardization.

HELP DNA Methylation Analysis

High-molecular-weight DNA was isolated from iPSCs with the PureGene kit from QIAGEN (Valencia, CA) and the HELP (HpaII tiny fragment enrichment by ligation mediated PCR) assay was carried out as previously described[30]. Briefly, One microgram of genomic DNA was digested overnight with either HpaII or MspI (NEB, Ipswich, MA). On the following day the reactions were extracted once with phenol-chloroform and resuspended in 11 µl of 10 mM Tris-HCl pH 8.0 and the digested DNA was used to set up an overnight ligation of the JHpaII adaptor with T4 DNA ligase. The adaptor-ligated DNA was used to carry out the PCR amplification of the HpaII and MspI-digested DNA as previously described[30]. All samples for microarray hybridization were processed at the Roche-NimbleGen Service Laboratory. Samples were labeled with Cy-labeled random primers (9mers) and then hybridized onto a mouse custom-designed oligonucleotide array (50-mers) covering 25,720 HpaII amplifiable fragments (HAF) (>50,000 CpGs), annotated to 15,465 unique gene symbols (Roche NimbleGen, Design name: 2006-10-26_MM5_HELP_Promoter Design ID = 4803). HpaII amplifiable fragments are defined as genomic sequences contained between two flanking HpaII sites found within 200-2,000 bp from each other and is represented on the array by 15 individual probes, randomly distributed across the microarray slide. HAF were first realigned to the MM9 July 2007 build of the mouse genome and then annotated to the nearest transcription start site (TSS), allowing for a maximum distance of 5 kb from the TSS. Scanning was performed with a GenePix 4000B scanner (Axon Instruments) as previously described (Figueroa et al., 2009). Quality control and data analysis of HELP microarrays were performed as described[30].

Signal intensities at each HpaII amplifiable fragment were calculated as a robust (25% trimmed) mean of their component probe-level signal intensities. Any fragments found within the level of background MspI signal intensity, measured as 2.5 mean-absolute-differences (MAD) above the median of random probe signals, were categorized as “failed.” These “failed” loci therefore represent the population of fragments that did not amplify by PCR, whatever the biological (e.g., genomic deletions and other sequence errors) or experimental cause. On the other hand, “Methylated” loci were so designated when the level of HpaII signal intensity was similarly indistinguishable from background. PCR-amplifying fragments (those not flagged as either “methylated” or “failed”) were normalized

with an intra-array quantile approach wherein HpaII/MspI ratios are aligned across density-dependent sliding windows of fragment size-sorted data. DNA methylation was therefore measured as the log2(HpaII/MspI) ratio, where HpaII reflects the hypomethylated fraction of the genome and MspI represents the whole-genome reference. Analysis of normalized data revealed the presence of a bimodal distribution. For each sample a cutoff was selected at the point that more clearly separated these two populations and the data were centered around this point. Each fragment was then categorized as either methylated, if the centered log HpaII/MspI ratio was less than zero, or hypomethylated if on the other hand the log ratio was greater than zero. All microarray data will be available from the GEO repository.

Quantitative DNA Methylation Analysis by MassARRAY EpiTyping

Single locus methylation analysis was performed by MALDI-TOF mass spectrometry with EpiTyper by MassARRAY (Sequenom, CA) on bisulfite-converted DNA following manufacturer's instructions46 but with the Fast Start High Fidelity Taq polymerase from Roche for the PCR amplification of the bisulfite converted DNA. MassArray primers were designed to cover the promoter regions of the indicated genes and sequences are given in Table S1.

DNA Methylation Analysis by Pyrosequencing

Frozen cell pellets were sent to EpigenDx (Hopkinton, MA) for DNA isolation, bisulfite treatment, and pyrosequencing. For Nanog, CpGs 1-7 correspond to positions −541, −502, −491, −434, −412, −302, and −280 from the transcription start site, respectively. For OCT4, CpGs 1-6 correspond to positions −165, −140, −127, −103, −99, and −47 from the transcription start site, respectively.

Flow Cytometry and Immunofluorescence

For flow cytometry, harvested cells were incubated with antibodies against THY1.2 (PE or eFluor450, 53-2.1, eBioscience) and SSEA-1 (purified, MC-480, Developmental Hybridoma Bank; biotin, MC-480, eBioscience; or AlexaFluor647, MC-480, BioLegend) for 20 min. Cells were washed in PBS and then incubated for 20 min with APC conjugated anti mouse IgM (eBioscience), streptavidin PerCP (BD PharMingen), or streptavidin Pacific Blue (Invitrogen). The cells were washed in PBS, resuspended in propidium iodide or 4',6-diamidino-2-

(continued from right) (B) Potential of EpCAM subpopulations at day 6 to form iPSC colonies. (C) Affymetrix expression analysis of EpCAM subpopulations.
(D) Methylation analysis of *Nanog* promoter by bisulfite sequencing of ESCs, MEFs and intermediates shown in (B). The following abbreviations are used: S, SSEA1; Ep, EpCAM.

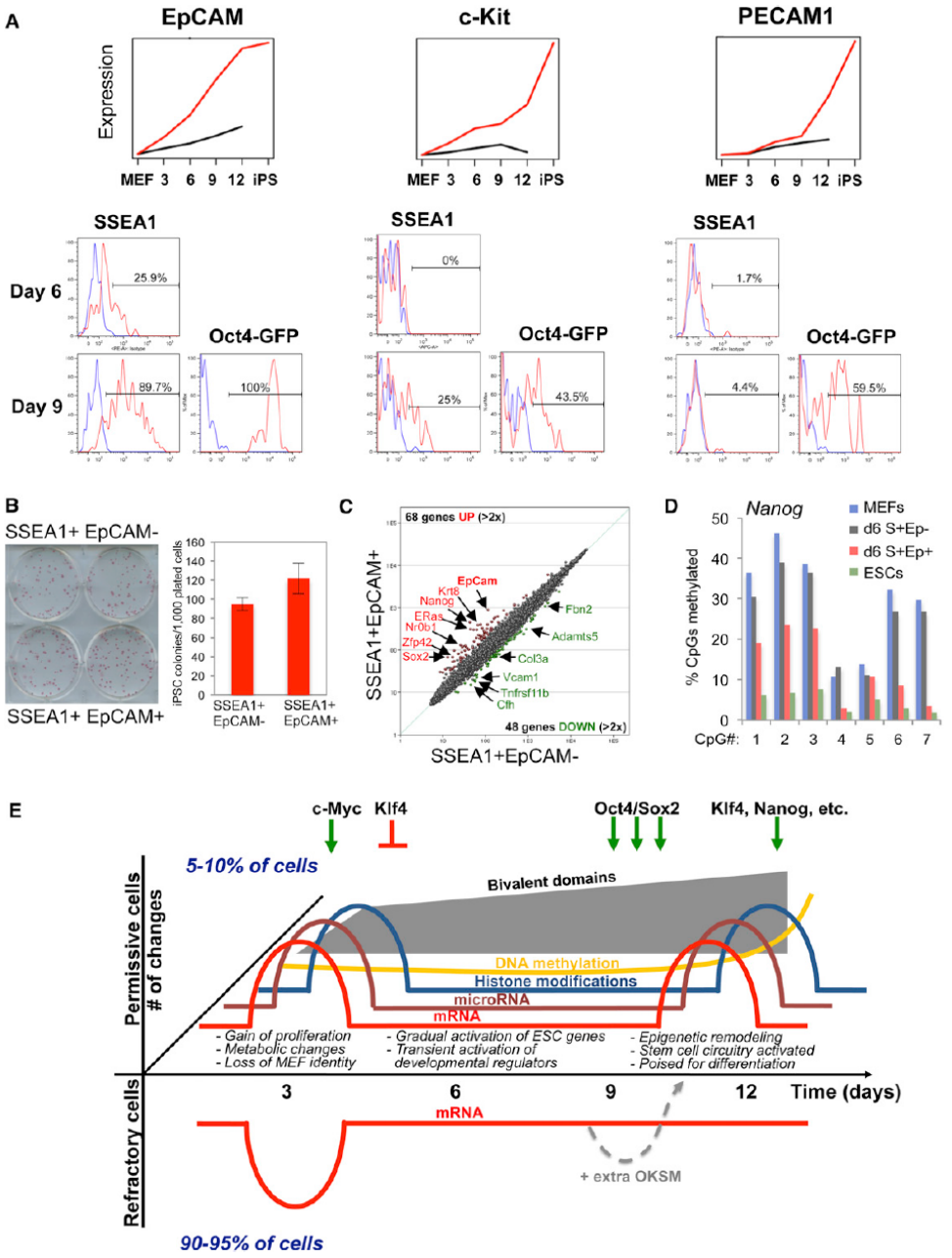


Figure 7. Identification of Surface Markers to Enrich for Reprogramming Intermediates, and Model. (A) Expression data and histogram plots of FACS analysis for c-Kit, EpCAM, and PECAM1 in SSEA1+ and *Oct4*-GFP+ populations at the indicated days of reprogramming. Red lines depict antibody-specific signal, blue lines show signal obtained with isotype control. No expression was seen before day 6. (continued on left)

phenylindole 5% FBS/PBS solution and passed through a 40mm cell strainer to achieve single-cell suspension. Cells were sorted on a FACSAria (BD Biosciences), Influx cell sorter instrument (BD Biosciences) and/or analyzed in a LSRII (BD Bioscience). For analysis and/or sorting of intermediates, cells were stained with THY1.2 and SSEA1 antibodies and sorted or analyzed as indicated. Additional antibodies include anti-EpCAM (PE, G8.8, eBioscience), anti-PECAM-1 (PE, 390, eBioscience), Rat IgG2a κ Isotype Control (PE, eBR2a, eBioscience), and anti-c-Kit (APC, 2B8, eBioscience). For the c-Kit stain, an APC unstained control was used as the negative staining control.

MicroRNA Mimics Experiments

Reprogrammable MEFs were counted and seeded in fibroblast medium at a density of 20,000 cells/well on 12-well gelatin-coated plates. The following day, MEFs were transfected in 12-well plates, with Lipofectamine RNAiMAX Reagent (Life Technologies) according to the manufacturer's instructions. MiR-182 and hsa-miR-214 overexpression and inhibition were obtained from mirVana miRNA Mimics and Inhibitors (Life Technologies). Briefly, miRNA mimics and inhibitors (30 pmol) in 100 µl of serum and antibiotic-free medium Opti-MEM (Life Technologies), were mixed with 2 µl of Lipofectamine RNAiMAX Reagent (Life Technologies), dissolved in 100 µl of Opti-MEM (Life Technologies), and left to incubate at room temperature for 20 min. Both transfection solutions (200 µl volume) were added to each well containing 1 ml of ES cell medium containing LIF and doxycycline (minus antibiotic). Four hours later, the cultures were replaced with 1 ml of fresh ES cell medium supplemented with antibiotics. All test and control experiments were performed in triplicate wells. After 15 days of culture in the presence of doxycycline, followed by an additional 5 days in the absence of doxycycline, the presence of iPS cell colonies was scored according to morphology and GFP fluorescence (indicative of Oct-4 expression) on an Olympus 1x71 inverted fluorescence microscope and processed with the analySIS image capture software.

For flow cytometry analysis, day 20 transfected cultures were dissociated into single cells with 0.25% Trypsin-EDTA (Life Technologies) and filtered through a 40-µm cell strainer (Falcon). Live cells were gated according to side-scatter, forward-scatter and propidium iodide exclusion. Flow cytometric gates were designed with control ES cells with the inclusion of propidium iodide. Flow cytometric analysis of GFP was used as a reporter for monitoring endogenous Oct-4 expression. Analysis was performed with the BD LSR II flow cytometer (BD Biosciences). Flow cytometric analysis was processed with Flowlogic software (Flowlogic).

Immunofluorescence

For immunofluorescence, cells were washed twice with PBS, fixed with 10% formalin solution (Sigma-Aldrich) for 15 min, washed twice with PBS and blocked for 1 hr. at room temperature (RT) with PBS containing 2% BSA (Fisher Scientific) and 0.1% Triton X-100 (Fisher Scientific). The cells were then stained with primary antibodies diluted in blocking solution against mouse OCT4 (IgG, 1:200, Santa Cruz sc-8628), mouse SSEA1 (IGM, 1:200, eBioscience MC-480) and mouse PRX (IgG, 1:100, Santa Cruz sc-240773). The cells were washed twice with PBS and incubated with the appropriate secondary antibodies for 1 hr. in RT. The secondary antibodies used were Alexa Fluor 546 donkey anti goat IgG, Alexa Fluor 488 goat anti mouse IgG and Alexa Fluor 488 goat anti mouse IgM. DNA was counterstained with DAPI (Invitrogen). Images were taken with Leica DMI4000B inverted fluorescence microscope equipped with a Leica DFC350FX camera.

Chromatin Immunoprecipitation and ChIP-Seq analysis

Cells were fixed in 1% formaldehyde for 10 min, quenched with glycine and washed 3 times with PBS. Cells were then resuspended in lysis buffer and sonicated 10 × 30 s in a Bioruptor (Diagenode, Philadelphia, PA) to shear the chromatin to an average length of 600 bp. Supernatants were precleared with Protein-A/G Dynabeads (Invitrogen) and 10% input was collected. Immunoprecipitations were performed with antibodies to H3K4trimethylated, H3K27trimethylated (Millipore) and OCT4 and SOX2(Santa Cruz). DNA-protein complexes were pulled down using Protein-A/G Dynabeads (Invitrogen) washed. DNA was recovered by overnight incubation at 65°C to reverse crosslinks and purified using QIAquick PCR purification columns (QIAGEN, Maryland). Enrichment of the modified histones at different genes was validated by quantitative real time PCR.
For ChIP-Seq, Libraries were prepared from ChIP DNA experiments with the ChIP-seq Sample Prep Kit (Illumina) following manufacturer's instructions46. The libraries were then sequenced with single end for 50 bp in GAII Illumina sequencer at the MGH next-Generation Sequencing Core. Raw reads were filtered with standard quality control measures and reads that passed this prefiltering step were aligned with the BWA algorithm (version 0.5.9) to the mouse genome (Mm9 build).

Single-Cell Expression Analysis

Template for single-cell PCR was produced with Life Technologies' Single cell to Ct kit according to manufacturer's instructions. Briefly, single cells for each time point (MEFs; SSEA1+ cells at day 3, day 6, day 9; and iPSCs) were directly sorted (Influx cell sorter instrument, BD-Biosciences) into separate wells of qPCR 96-well plates containing cell

lysis solution. Following a cDNA synthesis step, target sequences were preamplified for 18 cycles according to instructions. Single-cell PCR and data collection were performed with a Biomark instrument (Fluidgm). For both preamplification and single-cell PCR, Taqman probes (Life Technologies) were used (see Table S1). Results are expressed as Log2Ex = LOD (Limit of Detection) Cq – Cq [Gene]. The limit of detection was set to 30. If Log2Ex value is negative, Log2Ex = 0.

Statistical and Bioinformatic Analyses

All statistical and bioinformatics analysis was performed with the R / Bioconductor software suite. mRNA and miRNA data were preprocessed and quantile normalized with the RMA method [31]. HELP DNA methylation data were normalized with an intra-array quantile approach. Differentially expressed or methylated genes were identified for different comparisons with a t test with Benjamini and Hochberg correction ($p \leq 0.05$)[32]. Hierarchical clustering was performed with correlation distances and the Ward method. For transcription factor meta-analysis, we identified three ChIP-Chip data sets from the published literature for *Oct4, Sox2, Klf4, Myc,* and *Nanog*[14,33,34]. Of our differentially expressed mRNA targets, we identified those that were bound by each of these factors in at least two published experiments. We then used this information and our differentially expressed mRNA profiles to calculate transcription factor activity profiles with network component analysis[15]. ChIP-seq coverage was plotted by centering a −10Kb and +5Kb window on the transcriptional start site for each gene. Differential histone methylation changes were determined with a 2-fold cutoff. Gene ontology analysis was performed with the GOstats package. To integrate mRNA and miRNA expression data, we listed potential mRNA targets based on a scoring algorithm that combined the presence of seed sequences (with “targetscan” and “mirSVR” databases) and inverse mRNA-miRNA expression patterns. Genes whose expression inversely correlated with that of their predicted microRNA (high negative correlation value) are expected to be high confidence targets.

Supplemental Information includes three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.11.039.

Acknowledgements

We thank members of the Hochedlinger and Ramaswamy labs for helpful suggestions and critical reading of the manuscript. We thank Laura Prickett and Kat Folz-Donahue at the MGH/HSCI flow cytometry core and Flowcore Monash for expert cell sorting; Paul Lacaze, Danielle Evans, and Michelle Garred at Millenium Science for support with the Fluidigm

experiments; and Lucy Dagostino from Lifetech Australia for help in miRNA assays. We are grateful to Ben Wittner for help with formatting and uploading genome-wide data. Support to J.M.P. was from an ECOR postdoctoral fellowship, Monash Larkins Program, and a NHMRC CDF. J.Z. was supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, NIH, USA. K.H. was supported by the NIH (DP2OD003266 and R01HD058013). S.R. was supported by an HHMI Physican-Scientist Early Career Award.

References

1 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.

2 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.

3 Wu, S.M., and Hochedlinger, K. (2011). Harnessing the potential of induced pluripotent stem cells for regenerative medicine. Nat. Cell Biol. 13, 497–505.

4 Stadtfeld, M., and Hochedlinger, K. (2010). Induced pluripotency: history, mechanisms, and applications. Genes Dev. 24, 2239–2263.

5 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.

6 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.

7 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.

8 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.

9 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.

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

11 Kudo, T., Kaneko, M., Iwasaki, H., Togayachi, A., Nishihara, S., Abe, K., and Narimatsu,

H. (2004). Normal embryonic and germ cell development in mice lacking alpha 1,3-fucosyltransferase IX (Fut9) which show disappearance of stage-specific embryonic antigen 1. Mol. Cell. Biol. 24, 4221–4228.

12 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 Rep., in press.

13 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.

14 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.

15 Chang, C., Ding, Z., Hung, Y.S., and Fung, P.C. (2008). Fast network component analysis (FastNCA) for gene regulatory network reconstruction from microarray data. Bioinformatics 24, 1349–1358.

16 Huo, J.S., and Zambidis, E.T. (2012). Pivots of pluripotency: The roles of non-coding RNA in regulating embryonic and induced pluripotent stem cells. Biochim. Biophys. Acta. Published online October 24, 2012.

17 Subramanyam, D., and Blelloch, R. (2011). From microRNAs to targets: pathway discovery in cell fate transitions. Curr. Opin. Genet. Dev. 21, 498–503.

18 Yang, C.S., Lopez, C.G., and Rana, T.M. (2011). Discovery of nonsteroidal anti-inflammatory drug and anticancer drug enhancing reprogramming and induced pluripotent stem cell generation. Stem Cells 29, 1528–1536.

19 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.

20 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.

21 Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13, 484–492.

22 Orkin, S.H., and Hochedlinger, K. (2011). Chromatin connections to pluripotency and cellular reprogramming. Cell 145, 835–850.

23 Chen, J., Liu, J., Yang, J., Chen, Y., Chen, J., Ni, S., Song, H., Zeng, L., Ding, K., and Pei, D. (2011). BMPs functionally replace KLF4 and support efficient reprogramming of mouse fibroblasts by OCT4 alone. Cell Res. 21, 205–212.

24 Buckley, S.M., Aranda-Orgilles, B., Strikoudis, A., Apostolou, E., Loizou, E., Moran-Crusio, K., Farnsworth, C.L., Koller, A.A., Dasgupta, R., Silva, J.C., et al. (2012). Regulation of Pluripotency and Cellular Reprogramming by the Ubiquitin-Proteasome System. Cell Stem Cell. Published online October 24, 2012.

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

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

27 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.

28 Sommer, C.A., Stadtfeld, M., Murphy, G.J., Hochedlinger, K., Kotton, D.N., and Mostoslavsky, G. (2008). iPS Cell Generation Using a Single Lentiviral Stem Cell Cassette. Stem cells (Dayton, Ohio).

29 Zhu, J., Jankovic, D., Oler, A.J., Wei, G., Sharma, S., Hu, G., Guo, L., Yagi, R., Yamane, H., Punkosdy, G., et al. (2012). The Transcription Factor T-bet Is Induced by Multiple Pathways and Prevents an Endogenous Th2 Cell Program during Th1 Cell Responses. Immunity 37, 660-673.

30 Figueroa, M.E., Melnick, A., and Greally, J.M. (2009). Genome-wide determination of DNA methylation by Hpa II tiny fragment enrichment by ligation-mediated PCR (HELP) for the study of acute leukemias. Methods Mol Biol 538, 395-407.

31 Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31, e15.

32 Hochberg, Y., and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. Stat Med 9, 811-818.

33 Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell 132, 1049-1061.

34 Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J., et al. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 134, 521-533.

35 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.

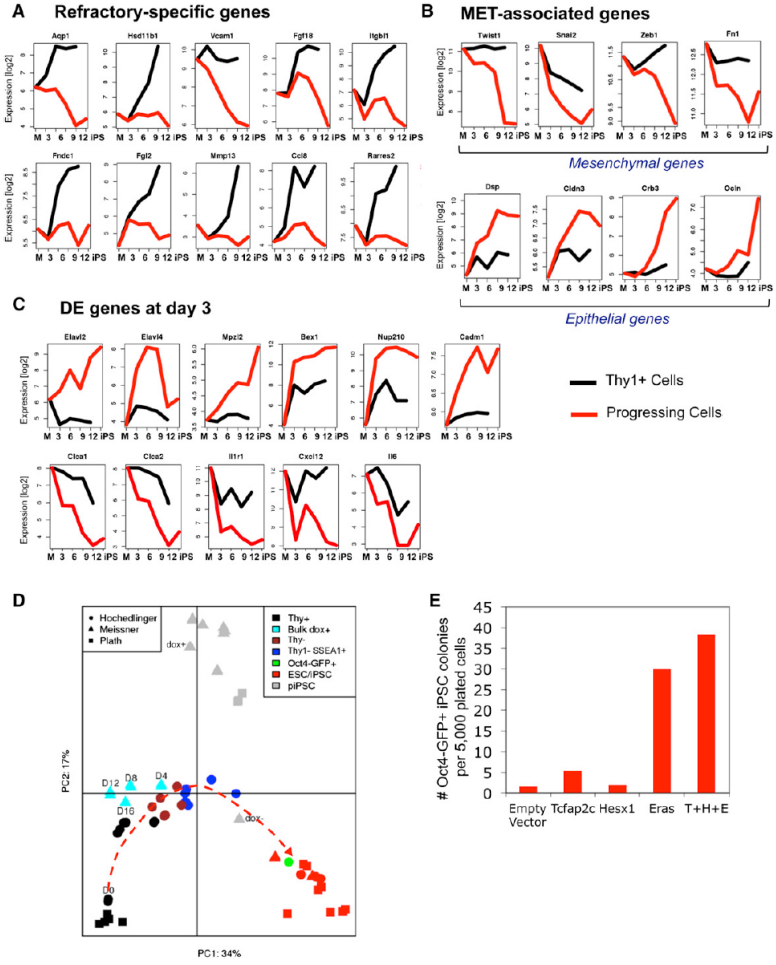


Figure S1. Comparison of Gene Expression Data from Progressing Intermediates with Those of Refractory Cells, Partially Reprogrammed iPSCs and Bulk Populations Expressing OKSM, Related to Figure 2 (A) Ten of the <?>200 genes that are aberrantly activated in refractory THY1+ cells (see Figure 2E). (B) Failure to regulate mesenchymal-to-epithelial (MET)-associated genes in refractory cells. (C) Potential biomarker that might distinguish refractory (THY1+) cells from progressing (SSEA1+) cells at day 3 when overall gene expression is very similar. Shown are all genes that are differentially expressed (DE) more than two-fold. (D) Principal component analysis of gene expression profiles of FACS-sorted populations (circles) in comparison with published data sets for partially reprogrammed iPSCs (piPSCs) (gray symbols) and bulk populations expressing *Oct4, Sox2, Klf4 and c-Myc* (turquois triangles). Dashed red line depicts reprogramming path. ''dox+,'' dox-dependent piPSC line on dox; ''dox-,'' same cell line <?>10 days after dox withdrawal. (E) iPSC formation frequencies upon overexpression of individual or combination of genes shown in Figure 2G.

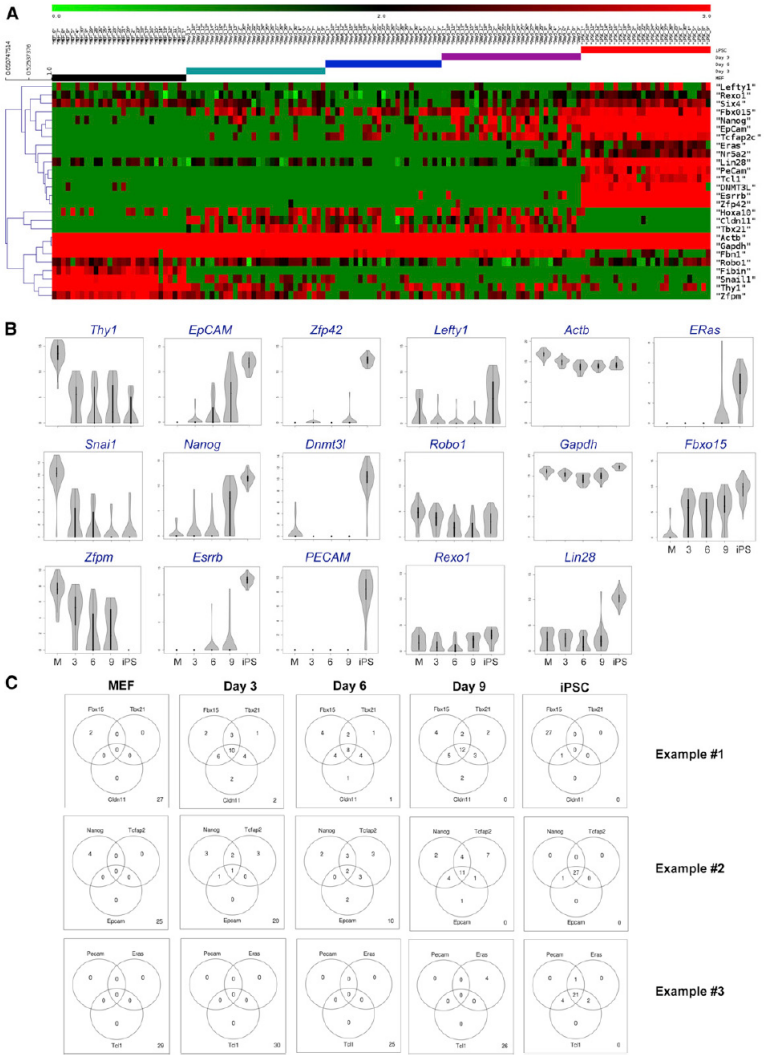


Figure S2. Fluidigm Single-Cell Gene Expression Analysis of Reprogramming Intermediates, Related to Figure 3 (A) Heatmap of Fluidigm single-cell expression data and hierarchical clustering. Note that related genes cluster together within MEFs (black bar), intermediates at day 3 (green bar), day 6 (blue bar), day 9 (purple bar) and iPSCs (red bar). (B) Violin plots of gene expression data that are not shown in Figure 3. (C) Three examples of genes that show coexpression in single cells, either transiently (example #1), gradually (example #2) or late (example #3).

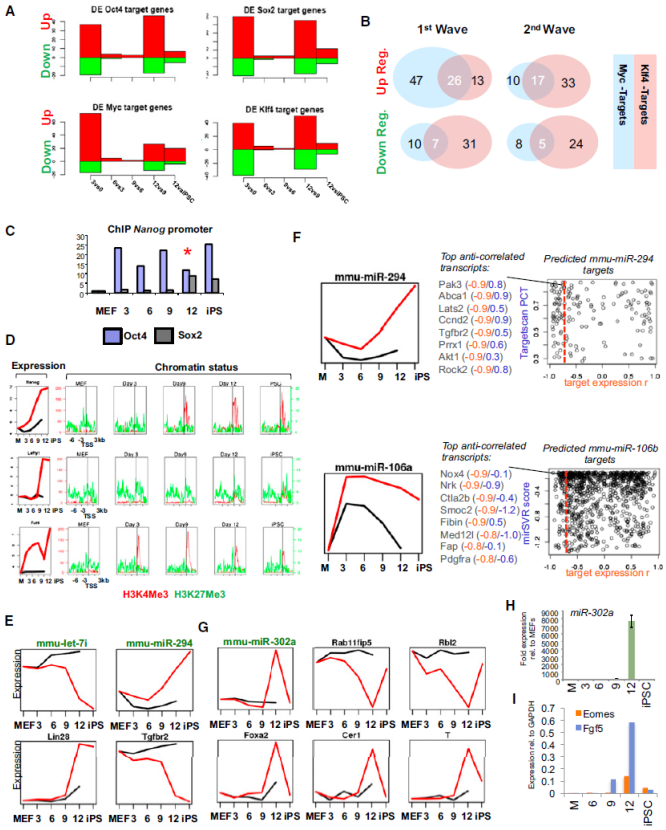


Figure S3. OKSM Target Gene Analysis, Predicted miRNA/mRNA Pairs, and Transient Upregulation of Epiblast Stem Cell-Related Genes in Reprogramming Intermediates, Related to Figure 4 (A) Number of differentially expressed (DE) genes that are targets of OCT4, SOX2, c-MYC or KLF4 when comparing progressive SSEA1+ cell population at successive time points (d3 and d0, d6 and d3, d9 and d6, d12 *Oct4*-GFP and d9, and iPSCs and d12 *Oct4*-GFP). DE target genes of OCT4, SOX2and KLF4 represent both single targets and multi-targets. Only single targets are shown for c-MYC. Note dominant effect of c-MYC on gene activation specifically in the first wave. (B) Number of single and shared KLF4 and c-MYC target genes that are either upregulated or downregulated during the two gene expression waves. Note overrepresentation of c-MYC targets among upregulated genes and of KLF4 targets among downregulated genes during first wave. Similar observations were made for other combinations of reprogramming factors and motivated us to generate a mathematical model of ‘‘transcription factor activities’’ that predicts the contribution of individual factors to gene expression by controlling for the respective other factors. (Continued on next page)

(Continued from previous page) (C) Chromatin immunoprecipitation for OCT4 and SOX2at the Nanog promoter in SSEA1+ intermediates at indicated time points of reprogramming. (D) H3K4me3 and H3K27me3 maps of *Nanog, Lefty1 and Fut9* promoter regions in SSEA1+ intermediates during reprogramming. (E) Expression dynamics of proven miRNA-mRNA pairs *let-7*i/Lin28 and *miR-294*/Tgfbr2. (F) Expression dynamics of *miR-294* and miR-106a (left panels) and their predicted mRNA targets ranked by targetscan or mirSVR score and inverse expression score (r). Shown are examples of top anticorrelated putative targets with r = 0.8 or higher. (G) Transient upregulation of epiblast stem cell-associated mir-302 and concomitant downregulation of two known targets (*Rab11fip5 and Rbl2*) at day 12. Epiblast marker *Foxa2, Cer1 and T* (*brachyury*) show similar transient activation at day 12. (H) Verification of transient *Mir-302a* upregulation by quantitative PCR in indicated reprogramming intermediates. (I) Quantitative PCR for additional epiblast stem cell-associated marker *Fgf5 and Eomes* at indicated time points of reprogramming.

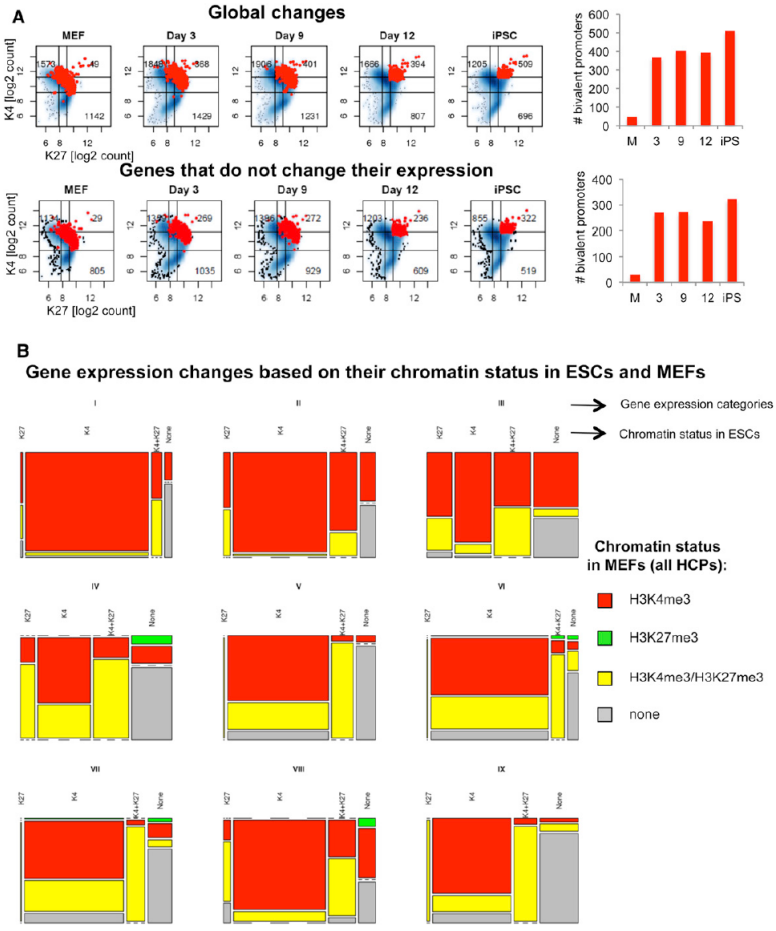


Figure S4. Bivalency Formation during iPSC Induction, and Correlation of Gene Expression Categories, I-IX, with Chromatin Marks in ESCs and MEFs, Related to Figure 5 (A) Number of bivalent domains (H3K27me3 and H3K4me3) at promoters depending on expression changes of associated genes (red dots = bivalent domains). Quantification is shown on the right. (B) Correlation of gene expression categories (I-IX) presented in Figure 2E with chromatin status in MEFs and ESCs/iPSCs (ESC/iPSC chromatin status is shown along the x axis, MEF status is shown by color coding). Note that genes that are induced early or gradually (categories I and II) show the most striking correlation with H3K4me3 enrichment status in MEFs.

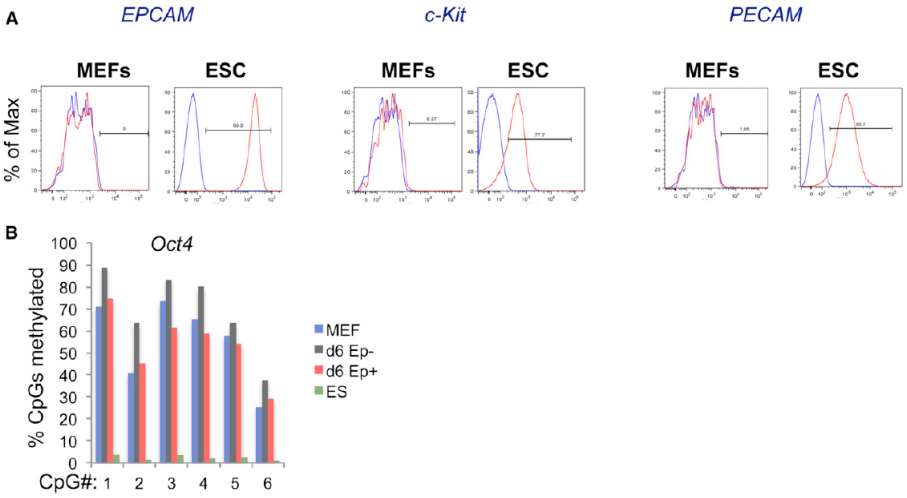


Figure S5. Flow Cytometry Analysis of EpCAM, PECAM1 and c-KIT in MEFs and ESCs, and *Oct4* Promoter Methylation Analysis in EpCAM Subpopulations, Related to Figure 7 (A) Differential expression of *EpCAM, c-Kit* and *PECAM* on MEFs and ESCs, as determined by flow cytometry. Red lines depict antibody-specific signal, blue lines show signal obtained with isotype control. (B) *Oct4* promoter methylation in day 6 SSEA1+ EpCAM+ and SSEA1+ EpCAM- intermediates.


a3

Addendum chapter 3

A candidate screen for facilitators of reprogramming to pluripotency identifies *ERas* and *Tcl1*

Marti Borkent, Matthias Stadtfeld, Jose Polo and Konrad Hochedlinger

Abstract

Somatic cell reprogramming holds great promise to study cell plasticity and epigenetic regulation, to model diseases and to potentially regenerate tissues and organs. To further our understanding and usability of the technique, we sought to define factors that facilitate reprogramming to pluripotency. To this end, we selected candidates from the literature related to pluripotency, epigenetics, cell proliferation and apoptosis in order to overexpress during *Oct4*, *Sox2*, *Klf4* and *c-Myc* (OKSM)-mediated reprogramming of mouse embryonic fibroblasts (MEF). Notably, out of the 24 candidates that were screened, Embryonic Ras (*ERas*) increases reprogramming efficiency up to 9-fold by colony count and reduces OKSM expression requirement by two days, while *ERas* knockdown reduces reprogramming efficiency by 60%. T Cell Leukemia 1 (*Tcl1*) overexpression boosts reprogramming up to 4 fold and like *ERas*, is known to activate the PI3K/AKT signaling pathway. Accordingly, we find that PI3K inhibition, but not MAPK inhibition, rescues the effect of *ERas* overexpression on reprogramming. In conclusion, our data suggest that *ERas* and *Tcl1* are involved in reprogramming, most likely through activation of the PI3K-AKT signaling pathway. Results from our group and other groups validate our approach of a candidate screen to study the poorly understood process of reprogramming cells to pluripotency based on educated guesses.

Introduction

Pluripotency is the ability of a cell to form any tissue in the body. One kind of pluripotent cell is the embryonic stem cell (ESC), derived from the inner cell mass of a blastocyst stage embryo. Because ESCs can self-renew and be maintained in culture, they are of great promise to generate new tissue to treat degenerative disease by transplantation. Ethical and practical constraints limit the supply of human ESCs however, and ESC derived transplants are at risk of being rejected by the patient's immune system. Induced pluripotent stem cells (iPSCs) derived from somatic cells can be used as a patient-specific alternative to embryonic stem cells (ESCs) for generating transplantable cells and tissues[1]. Moreover, the patient specific iPSCs can be used for modeling genetic diseases in vitro[2]. Hence, a better knowledge of the factors that are involved in the reprogramming process is key to predicting cellular behavior and the quality of the iPSCs if they are to be used for cell replacement therapy or for modeling diseases[3,4 5]. In addition, reprogramming into iPSC provides a platform to study cell fate changes, cellular plasticity and establishment of pluripotency on a molecular level[6,7]. Therefore, molecular and genetic approaches to determine key players in the reprogramming process are important for the understanding and usability of this technology[8].

Facilitating factors in iPSC generation

Before the discovery of iPSCs, somatic cell reprogramming could be achieved by alternative methods such as somatic cell nuclear transfer (SCNT, or 'cloning') or fusion between a somatic cell and ESC. Transcription factor-mediated reprogramming into iPSCs does not require existing oocytes or ESCs, making it much more practical and applicable, but it is also much slower. SCNT or fusion-induced pluripotency is established within 24-48 hours, and is thought to be 'deterministic' because cells reprogram with a short latency. In contrast, iPSC derivation takes about two weeks at a minimum, because expression of the four reprogramming factors alone is not enough for the somatic cell to become pluripotent. Thus, these factors need to induce expression of other factors. Work in a defined homogenous cell population has shown that every somatic cell has the potential to form an iPSC, after a sufficient number of cell divisions, suggesting that transcription factor-mediated reprogramming is a stochastic process[9].

To understand which factors are essential to reprogramming, we reasoned that overexpression of such 'facilitators', along with OKSM would increase the speed and efficiency of iPSC formation and eventually make it a more deterministic process. Facilitators could be known pluripotency regulators present in ESCs or oocytes, but absent from the transcription factor cocktail. They could also help cells upregulate the necessary regulators by modifying epigenetic marks or preventing senescence and apoptosis. We used this rationale to select candidate facilitators from the existing literature and test the effect of overexpression on iPSC formation from mouse embryonic fibroblasts.

Results

Selecting candidate factors to enhance direct reprogramming

We hypothesized that genes involved in pluripotency, ESC biology, epigenetic regulation, growth and survival regulation, could facilitate reprogramming. In 2007 Surani et al reviewed known regulators of pluripotency, that were shown to help the establishment of pluripotent cell lines[10]. Also, in his 2006 landmark paper, Yamanaka selected 24 factors to screen for reprogramming induction resulting in the discovery of *Oct4*, *Klf4*, *Sox2* and *c-Myc* as reprogramming factors to generate iPSCs[11]. Most of these 24 factors had not yet been described for their role in reprogramming and were added to our screening list. Furthermore, individual candidates were selected from different publications linking them to cellular reprogramming into pluripotent cells. For example, we selected the putative DNA demethylation factor AID, as it is required for reprogramming by cell fusion between mouse ESCs and human fibroblasts[12]. *ERas* and *HRas* were selected because somatic Ras family

members are well known oncogenes in a range of cancer types, and *ERas* was described to stimulate tumor-like growth properties of ESCs. Aiding growth and preventing OKSM-triggered senescent or cell death could be essential for MEFs to transform into iPSCs.

Moreover, publications about the role of the p53 apoptosis pathway and the role of cell division rate during reprogramming suggested that cellular survival and cell cycling should be promoted, possibly to prevent apoptosis or senescence induced by transgene expression and in vitro cell culture[9,13,14]. Finally, we screened molecules that could potentially establish and maintain the new expression patterns obtained by iPSCs. These molecules comprised epigenetic regulators form the third group of candidates, such as chromatin modifiers and DNA-methylation related factors. Chemical inhibitors blocking methylation have already been shown to increase the efficiency of reprogramming. The DNA-demethylating drug 5-aza-deoxycytidine enhances reprogramming by the four factors[15]. It was also shown that reprogramming without c-MYC or KLF4 occurs efficiently when the histone deacetylase (HDAC) inhibitor valproic acid is administered[16]. Lastly, reprogramming without c-Myc, Oct4 or Sox2 can be facilitated by the histone methyltransferase inhibitor G9a[17]. In total, 75 candidate genes were selected for screening, as shown in Figure 1.

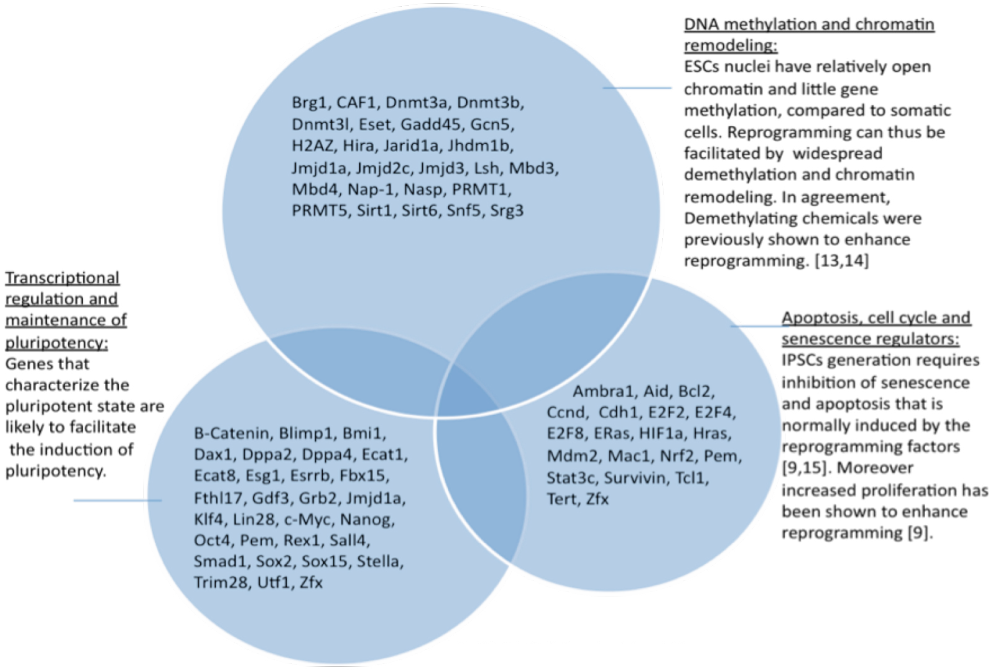


Figure 1. Selection of candidate facilitators of reprogramming based on literature.

ERas and Tcl1, but not HRas overexpression increases reprogramming efficiency

Of all candidates, 24 were tested so far after being cloned into a doxycycline-inducible overexpression vector and validated by sequencing and quantitative PCR (qPCR) for total cDNA levels upon infection and doxycycline induction. Next, we tested each individual candidate by overexpression in mouse embryonic fibroblasts (MEF) derived from a 'reprogrammable mouse', containing a doxycycline-inducible polycistronic cassette with the four Yamanaka reprogramming factors (*Oct4*, *Sox2*, *Klf4* and *c-Myc*)(Col1A-OKSM), a Tet-on advanced transactivator (*Rosa26*-M2rtTA) and a fluorescent pluripotency reporter (*Oct4*-GFP). *Oct4* is expressed in ESCs and EpiSCs, but not in somatic tissues or somatic stem cells[18]. These cells were then tested for their ability to form iPSC colonies upon 12 days of doxycycline induction, allowing the reprogramming factors and the candidate cDNA to be expressed (Figure 2). After 12 days, doxycycline was removed to select for true iPSC colonies that could maintain pluripotency without exogenous OKSM expression. The resulting iPSC colonies were counted by Alkaline Phosphatase (AP) staining, which is a surrogate assay to detect pluripotent stem cell colonies (Figure 2a-b). Repeated experiments showed that overexpression of Embryonic stem cell specific Ras (*ERas*) and T-cell Leukemia 1 (*Tcl1*) increased reprogramming efficiency by up to 9 and 4 fold by colony count, respectively. *ERas* is specifically expressed in ESCs and is 43 to 47% identical to the family members *KRas*, *NRas* and *HRas*[19]. ERAS is a small GTPase that stimulates growth of teratomas derived from ESCs, suggesting it helps ESCs expand through maintenance of pluripotency, proliferation or prevention of cell death. In contrast to other family members like *HRas* and *KRas*, *ERas* is not expressed in MEFs and is constitutively active in ESCs[19]. Conversely, although *HRas* belongs to the same family and is known to drive tumor growth, *HRas* overexpression nearly abolishes iPSC formation. To validate the overexpression on a protein level, we confirmed the presence of ERAS protein in *ERas* overexpressing MEFs by Western Blot (Figure 2c). *Tcl1* is an oncogene activated by regulatory elements of T cell receptors after translocations and inversions in T cell leukemia's[20] and stimulates cell growth and survival in mature T cells and cancer cells[21,22].

ERas overexpressing MEFs reprogram faster

To assess if *ERas* could also accelerate reprogramming when overexpressed alongside OKSM, doxycycline was removed prematurely (after 6 days) to select for iPSCs, which had lost their dependency on exogenous OKSM expression. Both control cells and *ERas* overexpressing cells formed iPSCs when treated with doxycycline for 12 days (Figure 3). However, upon 6 days of doxycycline administration, only *ERas* overexpressing cells could form iPSCs.

Thus *ERas* overexpression reduces the time necessary for reprogramming cells to become independent of exogenous reprogramming factors.

MEFs that reprogram upregulate ERas and Tcl1, but not HRas

Earlier work in our lab showed that the MEF marker THY1 and the early stem cell marker SSEA1 predict which populations will progress towards fully reprogrammed, *Oct4* expressing iPSCs. Cells that silence *Thy1* and turn on *Fut9* (encoding SSEA1) at sequential time points, are intermediates of reprogramming; a transient population of cells that are increasingly likely to form iPSCs after 13 days and following dox withdrawal[23]. Our lab performed genome-wide microarray expression analysis, identifying genes that are upregulated in intermediate cell populations (chapter 3)[24]. Both *ERas* and *Tcl1* were upregulated during reprogramming (Figure 4a). So is *Nanog*, a core pluripotency transcription factor that is required for reprogramming[25]. *ERas* closely follows the pattern of upregulation of *Nanog*,

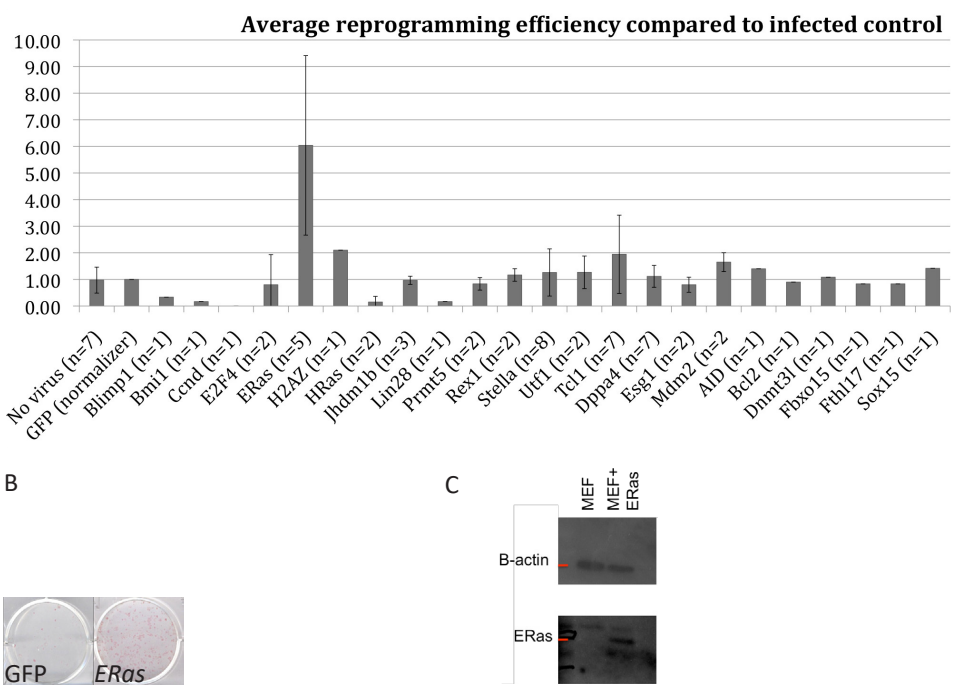


Figure 2. Overexpression of *ERas* and *Tcl1* enhances reprogramming efficiency. (A) Compiled data of independent experiments testing overexpression of individual candidates. Y axis represents fold increase of efficiency compared to GFP control. Error bars show the standard deviation. (B) Example of Alkaline Phosphatase (AP) staining to quantify iPSC colonies derived from control- (GFP) and candidate- (*ERas*) expressing MEFs. Each bright, well-defined red dot is one iPSC colony. (C) Validation of *ERas* overexpression by Western Blot.

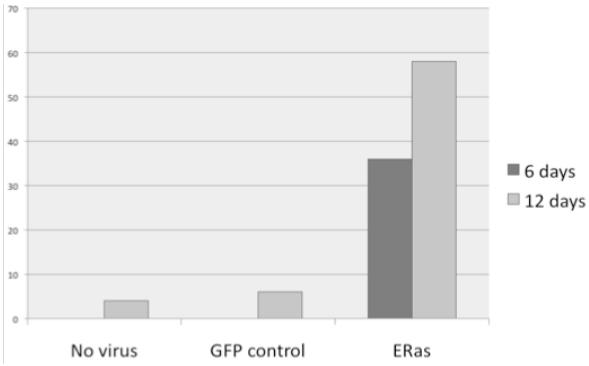


Figure 3 When *ERas* is overexpressed, MEFs need reprogramming factor expression for just 6 days. Y axis: Total iPSC colonies observed after two weeks, upon 6 or 12 days of doxycycline induction. Induction corresponds to expression of OKSM alone, or in combination with GFP or *ERas*.

starting in SSEA1+ populations at day 6 of OKSM expression. In contrast, *Tcl1* is upregulated late, when the Oct4-GFP reporter is activated and iPSC have become independent from exogenous factor expression. Cells that do not reprogram and maintain expression of the MEF marker THY1, maintain low or absent transcription of all three genes (Figure 4b). HRas expression stays constant in all cell populations. Thus, cells that are prone to form iPSCs upregulate *ERas* and *Tcl1*, supporting the notion that *ERas* and *Tcl1* may be important or even required for iPSC formation.

Requirement for ERas expression in iPSC formation

To test whether *ERas* upregulation is essential for somatic cell reprogramming, we depleted *ERas* transcript from reprogrammable MEFs by RNA Interference. To this end, we screened 6 short hairpin RNAs (shRNAs), delivered by a constitutive lentiviral vector, for their ability to knock down transcript levels of *ERas* individually or as a pool (Figure 5a). The 'sh*ERas*#1' exhibited an 85% reduction of *ERas* transcripts by quantitative PCR and was therefore used in the following experiments. The average loss of reprogramming efficiency after knock-down of *ERas* was 60% compared to a shRNA targeting Firefly luciferase (shFF) (Figure 5b-c). Thus, reductions in *ERas* transcript are correlated with impaired iPSC formation. Knockout experiments should be performed to determine whether *ERas* is required for reprogramming.

Mechanism by which ERas and Tcl1 facilitate reprogramming

To define the mechanism underlying the *ERas* and *Tcl1* phenotype, we sought to manipulate downstream signaling pathways and determine their effects on reprogramming. Based on published data, we hypothesized that *ERas* and *Tcl1* would enhance reprogramming through activation of the PI3K/Akt signaling cascade. To test this possibility, we treated reprogrammable MEFs with PI3K inhibitor LY294002 (PI3Ki). Furthermore, we sought to

exclude the possibility that during reprogramming ERAS does signal through the MAPK/ERK pathway, contrary to what is the case in ESCs, by treating with MAPK inhibitor PD98059 (MAPKi)[19]. We treated both GFP- and *ERas*- overexpressing MEFs during reprogramming to test the synergy with *ERas* overexpression (Figure 6). Both PI3Ki and MAPKi treatment sharply reduced iPSCs formation in control MEFs, compared to a treatment with solvent alone. Co-treatment with the two compounds abrogated reprogramming completely. However, MAPKi failed to affect iPSC formation in *ERas* overexpressing cells, suggesting that indeed, *ERas* overexpression does not enhance reprogramming through the MAPK pathway. Conversely, PI3K inhibition almost completely returned reprogramming efficiency to that of control cells, suggesting that the PI3K/AKT pathway is key to ERAS action. Co-treatment also inhibits reprogramming in *ERas* overexpressing cells, suggesting that MEFs need activity of at least one of the pathways to reprogram, regardless of *ERas* expression. Reproduction of this experiment and inclusion of *Tcl1* overexpressing cells can lead to a more conclusive result on the role of AKT and MAPK pathways during reprogramming.

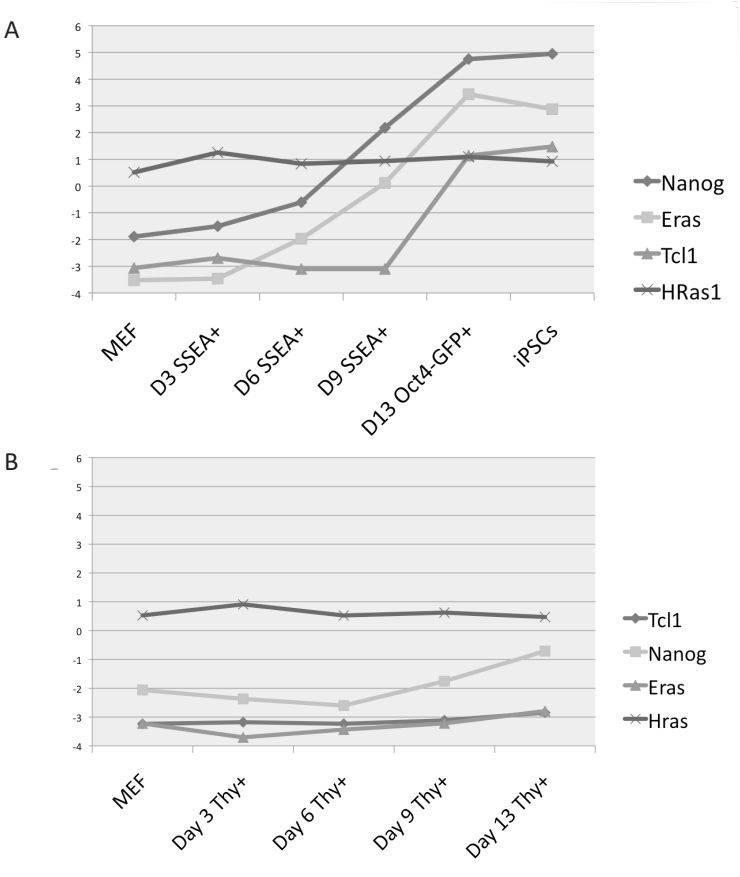


Figure 4.
ERas and *Tcl1* are upregulated during reprogramming in cells that are progressing towards pluripotency, based on global gene expression analysis. Relative transcript levels of *Nanog*, *ERas*, *HRas* and *Tcl1* in populations of (A) reprogramming intermediates, MEFs and iPSCs; (B) refractory populations at corresponding time points. Error bar: standard deviation.

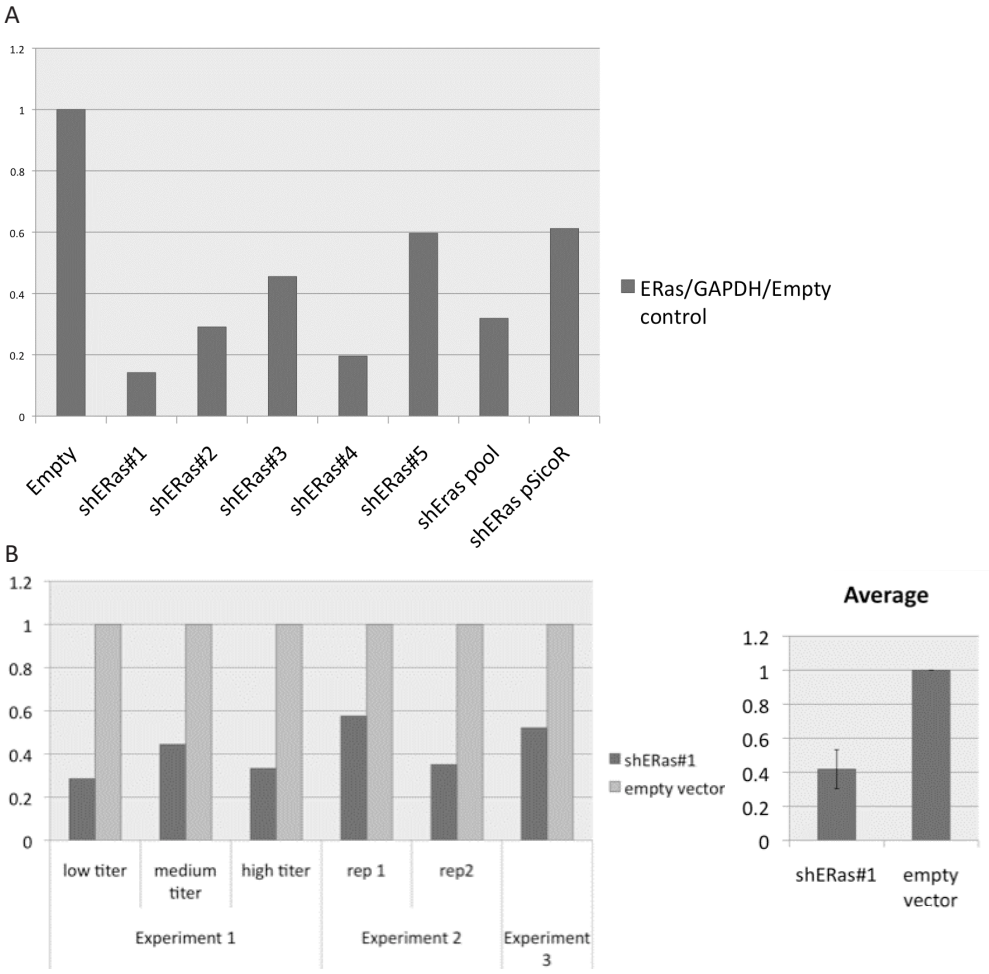


Figure 5. Downregulation of *ERas* reduces reprogramming efficiency. (A) qPCR analysis of *ERas* mRNA levels, normalised by GAPDH mRNA levels. Empty vector control =1. shRNA pool: combination of sh*ERas* 1-5. (B) reprogramming efficiency of sh*ERas*#1 infected MEFs compared to empty vector control, per experiment and on average.

Discussion

We selected candidates based on the literature in order to identify facilitators of transcription factor mediated reprogramming into iPSCs. After screening 24 candidates of our library, we found two interesting genes. The first, ESC-specific Ras or *ERas*, enhanced reprogramming efficiency up to 9-fold and gave rise to doxycycline-independent iPSC colonies after only 6 days of exogenous factor expression, compared to 8 days in control infected cells. The second gene, T Cell Leukemia 1 or *Tcl1*, enhanced reprogramming up to 4-fold. *ERas*

expression is absent in MEFs but upregulated during reprogramming in a pattern similar to *Nanog*, suggesting it could be required for reprogramming. *Tcl1* also is upregulated late in reprogramming nascent-formed iPSCs become transgene-independent. Next, we examined the importance of *ERas* for iPSC formation by shRNA-mediated knockdown, demonstrating a more than 2-fold reduction in reprogramming efficiency when 85% of transcript is depleted. Further studies should examine reprogramming ability of *ERas*-null cells. Also, *ERas*-overexpression-derived iPSC should be subjected to an analysis of differentiation potential and genomic integrity to confirm that true iPSC formation is facilitated by *ERas*.

To determine the mode of action of *ERas* and *Tcl1*, we noted that both factors are known to activate the PI3K/Akt1 signaling pathway. Notably, other forms of Ras also activate MAPK/ERK signaling. Preliminary data on the effect of a PI3K inhibitor or MAPK inhibitor in combination with *ERas* overexpression suggest that PI3K activity is indeed required for the *ERas* phenotype, whereas MAPK activity is dispensable. A recent publication supports our preliminary conclusions by showing that the ERAS-AKT-FOXO1 signaling axis stimulates reprogramming, and that AKT inhibition or knockdown suppresses reprogramming[26]. FOXO1 is degraded by active AKT and is found to inhibit reprogramming, which could be due to induction of a G0/1 arrest, apoptosis or unfavorable glucose metabolism regulation[27-29]. Chemical inhibition of FOXO1 may be an interesting approach to enhance reprogramming efficiency and study the role of AKT targets in iPSC formation.

Another explanation for the observed *ERas* overexpression phenotype is that *ERas* simple stimulates higher cell proliferation or prevents cell death, without changing the inherent reprogramming potential of cells. AKT is known to inhibit apoptosis through inhibition of MAD and CASP9, as well as increase proliferation by inhibiting cell cycle regulators and activation of MTORC signaling[30]. However, HRAS is also a potent activator of AKT but seems to diminish iPSC formation. This could mean that simultaneous MAPK activation by HRAS prevents reprogramming, or that ERAS promotes pluripotency independently from increasing proliferation and survival through AKT activation.

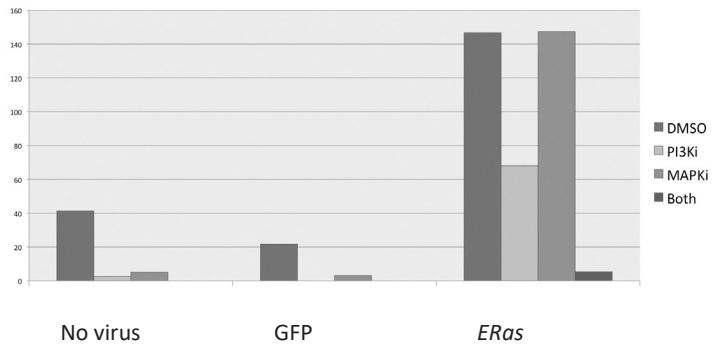


Figure 6. Effect of PI3K and MAPK inhibition on reprogramming of *ERas* or GFP overexpressing MEFs. Y axis: Total number of iPSC colonies, n=1. PI3Ki, PI3K inhibitor PD98059; MAPKi, MAPK inhibitor LY294002.

Next, we performed a preliminary experiment to test the requirement of both the AKT and MAPK pathway for iPSC formation. Interestingly, we find that combined inhibition of PI3K and MAPK, but not that of individual pathways, blocks reprogramming, suggesting that MEFs require either one of the pathways to transform, regardless of *ERas* expression. The interaction of PI3K and MAPK signaling has been extensively studied in human cancers. In these cancers, MAPK signaling is often erroneously activated by mutations in the upstream molecule BRAF. Mutated BRAF has been successfully used as a therapeutic target in solid tumors such as melanoma. Colorectal cancers frequently acquire mutant BRAF and aberrant MAPK signaling. However, it was shown that BRAF mutant colorectal cancers, in contrary to other tumors, are not responsive to BRAF inhibition by Vemurafenib. Instead, colonic RAS expression allows an alternative route to promote cell growth and survival via PI3K signaling[31]. Somatic RAS family members such as KRas require activation of the upstream EGF receptor (EGFR), and colon cancer without KRas mutations can be treated with EGFR antagonist Cetuximab. Prahallad and colleagues thus were able to demonstrate that the combination of BRAF and EGFR inhibition in BRAF mutant colon carcinoma is highly synergistic in attenuating tumor growth. This example supports the idea that proliferative cells, like cancer cells or nascent iPSCs, require either active MAPK- or PI3K-signalling at any time.

It has to be noted that the same MAPK inhibitor was previously shown to enhance reprogramming efficiency in human cells when combined with an ALK5 (TGF-β) inhibitor[32]. Our results suggest that MAPKi alone blocks reprogramming in murine cells or has no effect when *ERas* is overexpressed. Our experiment will have to be repeated before we can begin to understand this contradiction.

Finally, other candidate approaches to identify modulators of reprogramming have shed light on the reprogramming process. Since the start of this project, other candidates that we originally picked but have not yet tested were linked to reprogramming by other groups, suggesting that our candidate screen was a valid approach. For example, the transcription factors TbBX3, SALL4, UTF1 and NANOG were shown to enhance reprogramming when added to the 4 'Yamanaka factors' in mouse fibroblasts[14,33-35]. Chromatin modifiers JHDM1B, JMJD1A, BRG1[36-38] were also shown to enhance reprogramming. Other candidate screens proved successful too, like Yamanaka's original screen for reprogramming factors and a candidate screen for chromatin modifiers using just 20 candidates by RNAi, that found a chromatin modifier, DOT1L, to inhibit reprogramming[v].

In conclusion, we used a candidate screen to find that *ERas* and *Tcl1* are facilitators of murine somatic cell reprogramming, most likely through activation of AKT signaling. Our approach was independently validated by other groups. More work is required to identify potential targets of *ERas* and *Tcl1* during reprogramming.

References

1. Kiskinis, E. & Eggan, K. Progress toward the clinical application of patient-specific pluripotent stem cells. J. Clin. Invest. 120, 51–59 (2010).
2. Dimos, J. T. et al. Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons. Science 321, 1218–1221 (2008).
3. Daley, G. Q. The promise and perils of stem cell therapeutics. Cell Stem Cell 10, 740–749 (2012).
4. Liang, G. & Zhang, Y. Genetic and epigenetic variations in iPSCs: potential causes and implications for application. Cell Stem Cell 13, 149–159 (2013).
5. Wu, S. M. & Hochedlinger, K. Harnessing the potential of induced pluripotent stem cells for regenerative medicine. Nat Cell Biol 13, 497–505 (2011).
6. Hochedlinger, K. & Plath, K. Epigenetic reprogramming and induced pluripotency. Development 136, 509–523 (2009).
7. Stadtfeld, M. & Hochedlinger, K. Induced pluripotency: history, mechanisms, and applications. Genes Dev. 24, 2239–2263 (2010).
8. Hanna, J. H., Saha, K. & Jaenisch, R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. Cell 143, 508–525 (2010).
9. Hanna, J. et al. Direct cell reprogramming is a stochastic process amenable to acceleration. Nature 462, 595–601 (2009).
10. Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and Epigenetic Regulators of Pluripotency. Cell 128, 747–762 (2007).
11. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell 126, 663–676 (2006).
12. Bhutani, N. et al. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature 463, 1042–1047 (2010).
13. Utikal, J. et al. Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. Nature 460, 1145–1148 (2009).
14. Zhao, Y. et al. Two supporting factors greatly improve the efficiency of human iPSC generation. Cell Stem Cell 3, 475–479 (2008).
15. Mikkelsen, T. S. et al. Dissecting direct reprogramming through integrative genomic analysis. Nature 454, 49–55 (2008).
16. Huangfu, D. et al. Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. Nat Biotechnol 26, 795–797 (2008).
17. Shi, Y. et al. Induction of Pluripotent Stem Cells from Mouse Embryonic Fibroblasts by Oct4 and Klf4 with Small-Molecule Compounds. Cell Stem Cell 3, 568–574 (2008).
18. Lengner, C. J. et al. Oct4 expression is not required for mouse somatic stem cell self-

renewal. Cell Stem Cell 1, 403–415 (2007).
19. Takahashi, K., Mitsui, K. & Yamanaka, S. Role of *ERas* in promoting tumour-like properties in mouse embryonic stem cells. Nature 423, 541–545 (2003).
20. Rozovskaia, T. et al. Upregulation of Meis1 and HoxA9 in acute lymphocytic leukemias with the t(4 : 11) abnormality. Oncogene 20, 874–878 (2001).
21. Hoyer, K. K. et al. T cell leukemia-1 modulates TCR signal strength and IFN-gamma levels through phosphatidylinositol 3-kinase and protein kinase C pathway activation. J. Immunol. 175, 864–873 (2005).
22. Pekarsky, Y. et al. *Tcl1* enhances Akt kinase activity and mediates its nuclear translocation. Proceedings of the National Academy of Sciences 97, 3028–3033 (2000).
23. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining Molecular Cornerstones during Fibroblast to iPS Cell Reprogramming in Mouse. Cell Stem Cell 2, 230–240 (2008).
24. Polo, J. M. et al. A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. Cell 151, 1617–1632 (2012).
25. Silva, J. et al. *Nanog* Is the Gateway to the Pluripotent Ground State. Cell 138, 722–737 (2009).
26. Yu, Y. et al. Stimulation of somatic cell reprogramming by *ERas*-Akt-FoxO1 signaling axis. Stem Cells 32, 349–363 (2014).
27. Adachi, M. et al. The forkhead transcription factor FoxO1 regulates proliferation and transdifferentiation of hepatic stellate cells. Gastroenterology 132, 1434–1446 (2007).
28. Sangodkar, J. et al. Targeting the FOXO1/KLF6 axis regulates EGFR signaling and treatment response. J. Clin. Invest. 122, 2637–2651 (2012).
29. Nakae, J., Kitamura, T., Silver, D. L. & Accili, D. The forkhead transcription factor Foxo1 (Fkhr) confers insulin sensitivity onto glucose-6-phosphatase expression. J. Clin. Invest. 108, 1359–1367 (2001).
30. Manning, B. D. & Cantley, L. C. AKT/PKB signaling: navigating downstream. Cell 129, 1261–1274 (2007).
31. Sun, C. et al. Reversible and adaptive resistance to BRAF(V600E) inhibition in melanoma. Nature 508, 118–122 (2014).
32. Lin, T. et al. A chemical platform for improved induction of human iPSCs. Nat Meth 6, 805–808 (2009).
33. Silva, J. et al. *Nanog* Is the Gateway to the Pluripotent Ground State. Cell 138, 722–737 (2009).
34. Han, J. et al. Tbx3 improves the germ-line competency of induced pluripotent stem cells. Nature 463, 1096–1100 (2010).
35. Tsubooka, N. et al. Roles of Sall4 in the generation of pluripotent stem cells from

blastocysts and fibroblasts. Genes to Cells 14, 683–694 (2009).
36. Chen, J. et al. H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. Nature Genetics 45, 34–42 (2012).
37. Singhal, N. et al. Chromatin-Remodeling Components of the BAF Complex Facilitate Reprogramming. Cell 141, 943–955 (2010).
38. Wang, T. et al. The Histone Demethylases Jhdm1a/1b Enhance Somatic Cell Reprogramming in a Vitamin-C-Dependent Manner. Cell Stem Cell 9, 575–587 (2011).
39. Onder, T. T. et al. Chromatin-modifying enzymes as modulators of reprogramming. Nature 483, 598–602 (2012).

Chapter 4

A serial shRNA screen for roadblocks to reprogramming identifies the protein modifier SUMO2

Marti Borkent*, Brian D. Bennett*, Brad Lackford, Ori Bar-Nur, Justin Brumbaugh, Li Wang, Ying Du, David C. Fargo, Effie Apostolou, Sihem Cheloufi, Nimet Maherali, Stephen J. Elledge, Guang Hu#, Konrad Hochedlinger#

* equally contributing authors
# corresponding authors
Stem Cell Reports, April 2015

Abstract

The generation of induced pluripotent stem cells (iPSCs) from differentiated cells following forced expression of *Oct4*, *Klf4*, *Sox2* and *C-Myc* (OKSM) is slow and inefficient, suggesting that transcription factors have to overcome somatic barriers that resist cell fate change. Here, we performed an unbiased serial shRNA enrichment screen to identify potent repressors of somatic cell reprogramming into iPSCs. This effort uncovered the protein modifier SUMO2 as one of the strongest roadblocks to iPSC formation. Depletion of SUMO2 both enhances and accelerates reprogramming, yielding transgene-independent, chimera-competent iPSCs after as little as 38 hours of OKSM expression. We further show that the SUMO2 pathway acts independently of exogenous *c-Myc* expression and in parallel with small molecule enhancers of reprogramming. Importantly, suppression of SUMO2 also promotes the generation of human iPSCs. Together, our results reveal sumoylation as a crucial post-transcriptional mechanism that resists the acquisition of pluripotency from fibroblasts using defined factors.

Introduction

The reprogramming of somatic cells into pluripotent cells using the classical set of transcription factors, *Oct4*, *Klf4*, *Sox2* and *C-Myc* (OKSM) and conventional culture conditions (LIF, serum) usually takes several weeks and yields induced pluripotent stem cells (iPSCs) at extremely low frequencies (0.1-3%)[1]. This observation suggests that reprogramming factors need to overcome undefined barriers that have been established by somatic cells to preserve cell identity and resist cell fate change. Identifying roadblocks to iPSC generation thus provides a valuable platform to dissect general principles of cell identity and cell fate change[2].

Previously identified barriers to reprogramming include regulators of cell cycle progression and senescence (e.g., P53, INK4A/ARF)[3], histone and DNA modifications (e.g., DNMT1, KDM2B, MBD3)[4-6] as well as signaling pathways and epigenetic processes that can be targeted by small compounds (e.g., ascorbic acid, GSK3 inhibitor, DOT1L inhibitor)[7-9]. However, suppression of some of these barriers enhances iPSC formation only under specific culture conditions (e.g., MBD3)[5,10], potentially limiting its usefulness in different cellular contexts. Moreover, manipulation of certain barriers causes permanent aberrations of the epigenome (e.g. DNMT1)[11], complicating its applications in a therapeutic setting.

More recently, unbiased shRNA screens have been performed during iPSC formation, leading to the identification of novel roadblocks to reprogramming[12-14]. Surprisingly, individual suppression of hits that emerged from these screens showed rather modest effects (2-4-fold enhancement) compared to the simultaneous suppression of multiple hits (5-10-fold enhancement). Furthermore, there was little overlap among independent

screening efforts, suggesting that reprogramming may be restrained by additional, yet to be identified barriers.

The goal of this study was to identify potent roadblocks to reprogramming by performing a serial genome-wide shRNA enrichment screen in combination with a well-defined transgenic reprogramming system. Our screening strategy uncovered SUMO2 as a top-scoring hit, thus implicating protein sumoylation as a mechanism that effectively resists transcription factor-induced pluripotency.

Results

Serial shRNA screen for roadblocks to reprogramming

To identify roadblocks to iPSC formation in an unbiased manner, we combined a well-defined transgenic reprogramming system with a genome-wide shRNA library targeting 18,464 genes with 60,642 hairpins. We utilized murine embryonic fibroblasts (MEFs) carrying a doxycycline (dox)-inducible polycistronic cassette encompassing the open reading frames for *Oct4*, *Klf4*, *Sox2* and *c-Myc* (OKSM) in the *Col1a1* locus, the M2-rtTA transactivator in the Rosa26 locus and an EGFP reporter in the endogenous *Pou5f1* (*Oct4*) locus (chaper 2)[15]. We will refer to these transgenic MEFs as "reprogrammable cells" and the genotype as "*Col1a1*-tetOP-OKSM; *R26*-M2rtTA; *Oct4*-GFP".

The shRNA library was generated by cloning shRNAs into the pHAGE-Mir lentiviral vector carrying a puromycin resistance gene and a tRFP reporter[16,17](see experimental procedures for details). Transduction of reprogrammable MEFs with an identical empty vector gave rise to *Oct4*-GFP+, tRFP+ iPSC colonies upon exposure to dox, albeit at slightly lower frequencies than uninfected cells (Figure 1A,B and data not shown), demonstrating the feasibility of an shRNA screen using these cells and vector system.

To identify shRNAs that potently enhance reprogramming with low background signal from passenger shRNAs, we devised a pooled screening strategy using serial enrichment of hairpin libraries. Briefly, we infected reprogrammable MEFs with the pooled shRNA library 2 days before dox induction to ensure effective suppression of targets prior to initiation of reprogramming. After 10 days of OKSM expression, dox was withdrawn for 4 days to select for stably reprogrammed, transgene-independent colonies, followed by purification of emerging *Oct4*-GFP+ cells by flow cytometry (Figure 1C).

Enriched hairpins were amplified by PCR from genomic DNA and subsequently re-cloned into the original viral backbone before initiating another round of viral transduction and reprogramming. We performed 5 rounds of reprogramming and shRNA enrichment before focusing on candidates (Figure 1D,E). Parallel cultures of reprogrammable MEFs were exposed to dox alone or transduced with the viral library in the absence of dox before

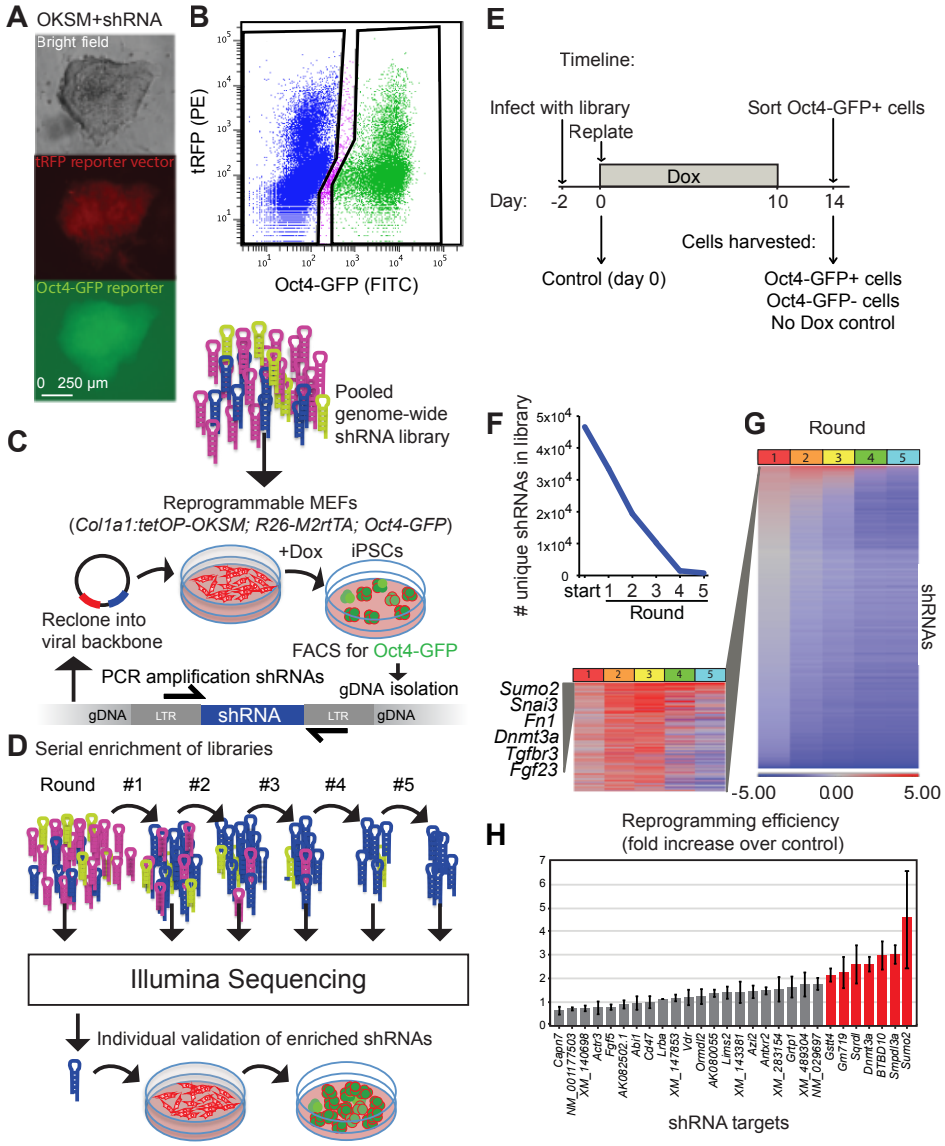


Figure 1. A genome-wide serial shRNA enrichment screen during iPSC generation. (A) Fluorescence microscopy image of a primary iPSC colony showing lentiviral tRFP expression and endogenous *Oct4*-GFP expression. (B) Gating strategy to purify *Oct4*-GFP+ cells from lentivirally transduced cultures undergoing reprogramming. (C) Schematic representation of one reprogramming/shRNA enrichment cycle. (D) Overview of serial enrichment screen and validation experiments. (E) Timeline of reprogramming experiments and strategy to collect control and experimental samples for subsequent analysis of shRNA library representation. (F) Change in shRNA library complexity during enrichment screen, i.e. number of unique shRNAs at the start of rounds 1-5. (G) Heatmap depicting fold-change enrichment of shRNAs during 5 rounds of reprogramming. (Continued on right)

extracting genomic DNA (Figure 1E and Figure S1); these samples served as controls for possible passenger hairpins that became passively enriched in expanding iPSC colonies or hairpins that merely affected the growth of uninduced reprogrammable MEFs. Library representation was then determined in all samples by deep (Illumina) sequencing of genomic DNA. Notably, we detected a gradual reduction of library complexity and a progressive accumulation of specific shRNA vectors during the 5 rounds of screening, suggesting enrichment of biologically meaningful hairpins (Figure 1F,G).

SUMO2 emerges as a top-scoring candidate barrier to reprogramming

We next applied stringent criteria to call hits based on the number of normalized reads and the cumulative fold-change of shRNAs across all rounds of reprogramming relative to controls (Figure S1). Analysis of shRNAs that were consistently enriched revealed several candidate barriers to reprogramming including FGF5, SMPDL3a and SUMO2 (Figure 1F-H). Of note, some of the top-scoring shRNAs affected pathways that were previously shown to block reprogramming including members of the FGF/FGFR and OLFR families[12,18]. A complete list of candidates with associated cumulative enrichment scores is shown in Supplemental Table 1. Of the 27 validated shRNAs, 7 showed a more than 2-fold increase in iPSC formation (*Gstt4, Gm719, Sqrdl, Dnmt3a, BTBD10, Smpdl3a, Sumo2*)(Figure 1H) with *Sumo2* shRNA exhibiting the strongest phenotype.

Given the prominent effects on reprogramming of shRNAs targeting Sumo2, we will focus on this gene for the remainder of the manuscript. SUMO2 (small ubiquitin-like modifier 2) is an ubiquitin-related protein that can be covalently attached to proteins as a monomer or lysine-linked polymer. This post-translational modification, lysine sumoylation, controls the stability, activity and localization of hundreds of proteins and has been implicated in a number of biological processes including DNA replication and repair, heterochromatic gene silencing and signal transduction[19].

Suppression of SUMO2 promotes reprogramming without compromising proliferation or pluripotency

Using more quantitative reprogramming assays, we found that suppression of SUMO2 increases the number of transgene-independent alkaline phosphatase-positive (AP+) iPSC-

(Continued from left) Blue bars represent lost shRNAs whereas red bars represent enriched shRNA relative to controls (see experimental procedures for details). (H) Validation of candidates identified in serial shRNA screen (mean value from 3 biological replicates +/- s.d.). Red bars depict shRNAs yielding a 2-fold or higher increase in reprogramming efficiency.

like colonies and the fraction of *Oct4*-GFP+ cells by 7-8 fold (16% *Oct4*-GFP+ cells with Sumo2 shRNA; 2% with control shRNA at day 8)(Figure 2A-D). We were able to recapitulate enhanced reprogramming with 6 independent shRNAs as well as siRNA pools targeting Sumo2, documenting the consistency of the observed phenotype using either permanent or transient knockdown approaches (Figure 2C,D and Figure S2A). Expression of *Sumo2* shRNAs led to reduced Sumo2 transcript levels as well as a decrease of free and conjugated SUMO2 protein levels, demonstrating specificity of the knockdown vector and diminution of global sumoylation levels (Figure 2E and Figure S2B).

Considering that sumoylated *Klf4* reportedly inhibits cellular reprogramming[20], we next determined whether SUMO2 suppression leads to reduced KLF4 sumoylation. However, we failed to detect differences in KLF4's sumoylation status in cells expressing OKSM and *Sumo2* shRNAs, suggesting that SUMO2's effect on iPSC formation may involve other targets (Figure S2C). Moreover, forced expression of SUMO2 during iPSC generation (Figure S2D) was insufficient to block reprogramming (data not shown), indicating that the high physiological SUMO2 levels are already saturated; enough to facilitate any sumoylation reaction.

iPSCs generated with *Sumo2* shRNAs could be stably propagated over many passages and gave rise to well-differentiated teratomas, demonstrating that transient suppression of SUMO2 does not compromise the self-renewal or pluripotency of iPSCs (Figure S3A). It is also noteworthy that endogenous *Sumo2* mRNA levels were comparable between MEFs and iPSCs and barely changed during the reprogramming process, suggesting that SUMO2 may play independent roles in somatic and pluripotent cell types (Figure 2F).

To complement the aforementioned marker-based assays of reprogramming with a functional readout, we determined whether suppression of SUMO2 could promote the formation of transgene-independent iPSC colonies after reduced pulses of OKSM expression (Figure 2G). Indeed, we found that knockdown of SUMO2 yielded transgene-independent iPSC colonies after only 4-5 days of OKSM expression, whereas stable iPSC colonies only emerged between days 6 and 8 in controls, consistent with previous findings[21] (Figure 2H). In agreement with this observation, we detected a significant upregulation of key ESC-associated transcripts (e.g., *Nanog*, *Nr5a2, Sall4*) and epigenetic regulators (e.g., *Dnmt3b* and *Tet1*) as well as a downregulation of MEF-associated transcripts (e.g., *Twist2, Fgf18, Meox2*) in cells expressing OKSM and *Sumo2* shRNAs at day 6 of reprogramming relative to a non-targeting shRNA control (Figure 2I,J).

Critically, knockdown of SUMO2 had no discernible effect on cell proliferation or apoptosis of bulk cultures, thus excluding the possibility that the observed phenotype is due to accelerated growth or reduced cell death (Figure S3B,C). Together, these results demonstrate that both transient and constitutive suppression of SUMO2 markedly enhances

and accelerates the formation of iPSCs from somatic cells.

SUMO2 suppression acts during early to mid stages of reprogramming

In order to understand how SUMO2 suppression influences the dynamics of iPSC formation, we utilized surface markers and a reporter allele to distinguish between early, mid and late stages of reprogramming. We previously showed that cells undergoing successful reprogramming initially upregulate SSEA1 (early stage), followed by sequential activation of EPCAM (mid stage) and *Oct4*-GFP (late stage)(Polo et al., 2012; Stadtfeld et al., 2008) (Figure 3A). SUMO2 depletion had no pronounced effect on the earliest intermediates of reprogramming, as shown by comparable fractions of SSEA1+ and EPCAM+ cells at day 3 relative to controls (5% vs. 6% for SSEA1; 4% vs. 1% for EPCAM)(Figure 3B,C and Figure S4). However, we noticed a 3-fold increase in the fraction of SSEA1+ cells and a 5-fold increase in the fraction of EPCAM+ cells by day 5 as well as an 8-fold increase in the fraction of *Oct4*-GFP+ cells by day 8 of reprogramming. These data indicate that SUMO2 suppression facilitates early to mid stages of iPSC formation based on surface marker expression in bulk cultures.

To corroborate the notion that SUMO2 suppression promotes early to mid stages of reprogramming, we determined iPSC colony formation efficiencies after transfecting reprogrammable MEFs with siRNAs against *Sumo2* either once (on day 0) or twice (on days 0 and 3)(Figure 3D,E). iPSC colony formation was essentially the same when SUMO2 was suppressed initially or continuously during a 6-day reprogramming period, suggesting that early SUMO2 suppression is sufficient to elicit enhanced reprogramming. Consistent with the acceleration of reprogramming upon suppression of SUMO2, we find that MEFs expressing OKSM and *Sumo2* shRNAs for only 6 days are molecularly most similar to advanced stages of reprogramming (day 9 and day 12 intermediates) using a previous expression time course (Polo et al., 2012), whereas MEFs expressing OKSM and a control shRNA are most similar to day 6 intermediates, as expected (Figure 3F).

Collectively, these phenotypic, functional and molecular data suggest that SUMO2 suppression accelerates early to mid stages of reprogramming, ultimately leading to a strong increase in the number of *Oct4*+ transgene-independent iPSCs.

SUMO2 suppression acts independently of C-MYc and in parallel with small molecules

We next investigated whether exogenous *C-Myc* expression is required for the enhancement of iPSC formation by *Sumo2* shRNAs. To this end, we derived reprogrammable MEFs from mice carrying the *Col1a1*-tetOP-OKS-mCherry allele (lacking the *c-Myc* transgene) in combination

with the *R26*-M2rtTA allele (Figure 4A). Exposure of these MEFs to dox alone gave rise to extremely few if any AP+ colonies after 9-21 days of OKS expression, and no *Oct4*-GFP positive cells could be detected by day 9 of reprogramming (Figure 4B-D). By contrast, depletion of SUMO2 in these cells using transient transfection of siRNA pools yielded iPSC colonies and stable *Oct4*-GFP+ cells by flow cytometry as early as 9 days after OKS induction. We conclude that suppression of SUMO2 enhances reprogramming independently of exogenous *C-Myc* expression, thus enabling iPSC formation under suboptimal reprogramming conditions and in the absence of this potent oncogene.

To determine whether SUMO2 suppression acts in parallel with small molecules that were previously shown to enhance reprogramming, we treated reprogrammable cells harboring shRNAs against *Sumo2* or *Firefly* luciferase with doxycycline in the presence or absence of ascorbic acid (AA)[7], a DOT1L inhibitor (DOT1Li)[8] and a GSK3β inhibitor (GSK3i)[9] (Figure 4E). Consistent with previous reports, we found that exposure of reprogrammable cells to each of these compounds significantly enhanced the generation of AP+ iPSC colonies, with combined AA/GSK3i treatment (AGi) exhibiting the strongest effect (Bar-Nur et al., 2014). Strikingly, SUMO2 depletion alone even surpassed the effect of AGi exposure based on AP+ colony numbers (Figure 4E). Moreover, suppression of SUMO2 further enhanced iPSC formation in the presence of either ascorbate, GSK3i or DOT1Li (Figure 4F). These results underscore the strong effects SUMO2 suppression alone has on the reprogramming process and suggest that the sumoylation pathway may act in parallel to previously described facilitators of iPSC formation including AA and inhibitors of H3K79 methylation and GSK3/WNT signaling.

Generation of iPSCs after as little as 38 hours of OKSM expression

Given the additive effect between SUMO2 suppression and small molecule enhancers of reprogramming, we asked whether this combination treatment would allow us to further reduce the minimal time period of OKSM expression required to produce stable transgene-independent iPSCs. We used early-passage reprogrammable MEFs (passage 2) carrying two copies of each the *Col1a1*-tetOP-OKSM and *R26*-M2rtTA alleles to achieve

(Continued from right) (H) Data obtained from experiments depicted in (G) using indicated shRNAs and 3 biological replicates (mean +/- s.d.). (I) Scatter plot comparing microarray data (log2 values) of day 6 reprogramming intermediates expressing indicated shRNAs. Representative upregulated pluripotency genes are shown in red whereas downregulated somatic genes are shown in green. (J) Normalized expression levels of representative pluripotency-associated and MEF-associated genes in indicated samples at day 6 of OKSM expression. Data obtained from one experiment. Rep-MEFs, reprogrammable MEFs, ESCs, embryonic stem cells.

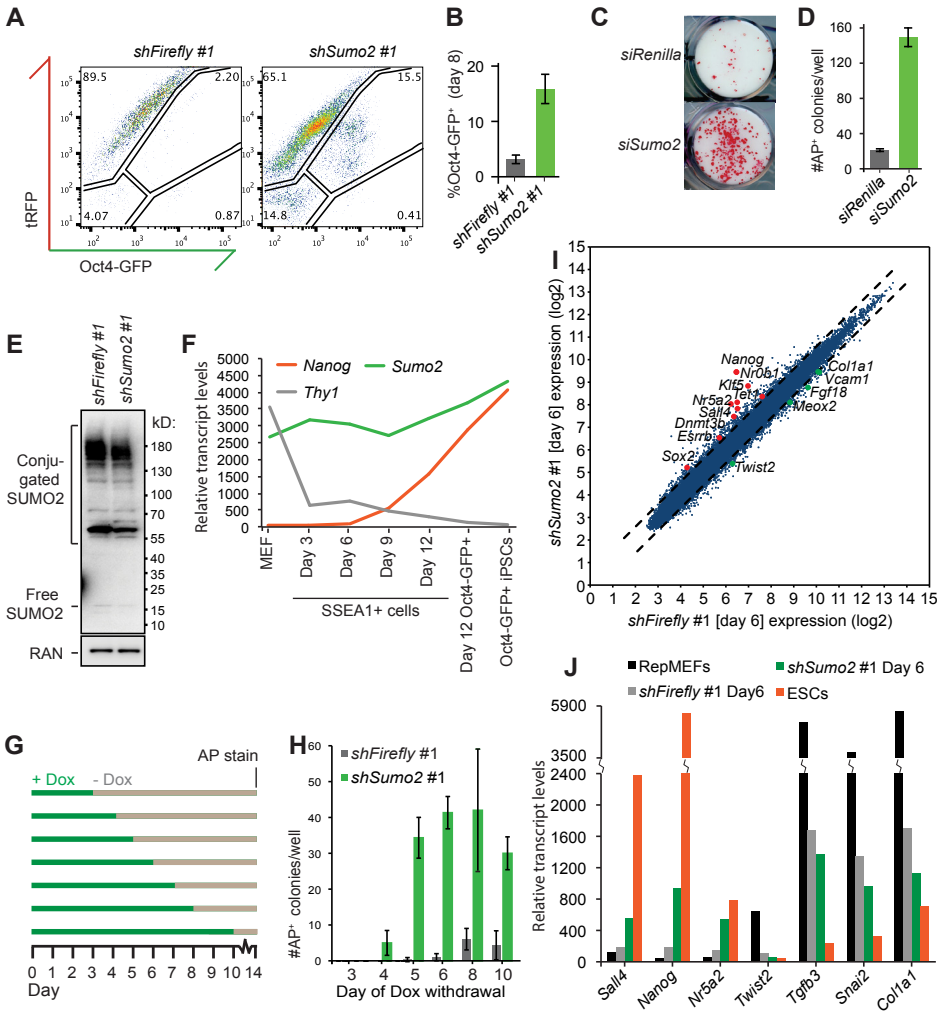


Figure 2. Suppression of SUMO2 robustly enhances and accelerates reprogramming. (A) Flow cytometric analysis of *Oct4*-GFP expression in reprogrammable MEFs after 8 days or OKSM expression in the presence of indicated shRNAs. (B) Quantification of data shown in (A); shown is percentage of *Oct4*-GFP+ cells per total number of cells using 3 biological replicates (mean +/- s.d.). (C) Alkaline Phosphatase (AP) staining of iPSC colonies derived from reprogrammable MEFs transfected with indicated siRNAs. (D) Quantification of data shown in (C); data represent mean from 3 biological replicates +/- s.d. (E) Western blot analysis for SUMO2 expression in reprogrammable MEFs infected with indicated shRNA vectors and treated with doxycycline (dox) for 3 days. (F) Expression dynamics of *Sumo2* mRNA in MEFs, iPSCs and intermediate stages of reprogramming using a previously published expression time course[22]. *Thy1*, fibroblast marker; *Nanog*, pluripotency marker. (G) Scheme to determine minimal duration of OKSM expression (in days) required to achieve transgene-independent iPSC colonies. (Continued on left)

optimal reprogramming efficiencies[15]. MEFs exposed to DOT1Li, GSK3i and AA required only 72 hours of OKSM expression to produce the first dox-independent AP+ iPSCs, which is faster than any previously reported protocol (Figure 5A,B). Remarkably, suppression of SUMO2 further reduced this time window to 38 hours. Emerging iPSC colonies activated the endogenous *Oct4*-GFP reporter, expressed endogenous *Nanog*, *Oct4* and *Sox2*, gave rise to well-differentiated teratomas and supported the formation of coat-color chimeras, indicating acquisition of an authentic iPSC state (Figure 5C-F).

Analysis of global transcriptional patterns of these iPSCs further revealed a remarkable similarity with either iPSCs generated after 10 days of OKSM expression or with an established ESC line, suggesting that abbreviated OKSM expression in the presence of chemicals and *Sumo2* siRNAs does not compromise the pluripotency program (Figure 5G-I and Figure S5). These results show that 1-2 days of OKSM expression are sufficient to produce stable, pluripotent iPSCs when SUMO2 expression is transiently suppressed under optimal culture conditions.

Finally, we tested whether suppression of SUMO2 also enhances human reprogramming. Towards this end, we infected human dermal fibroblasts with vectors expressing OKSM and either *Sumo2* or *Firefly* (FF) shRNAs and measured the formation of AP+ colonies at day 21 (Figure 5J, K). Consistent with our observations in the mouse system, we find that suppression of SUMO2 increases the formation of human iPSC-like colonies by 4-6 fold (Figure 5L). We also noticed that human iPSC-like colonies expressing *Sumo2* shRNAs formed earlier than controls, suggesting acceleration of reprogramming (data not shown). Thus, sumoylation is a conserved reprogramming barrier across murine and human cells.

Discussion

Here, we identified SUMO2 as a potent roadblock to iPSC generation by combining a well-defined transgenic reprogramming system with a genome-wide shRNA screening approach. In contrast to previous shRNA or siRNA screens conducted during iPSC formation, we employed a serial shRNA enrichment strategy, which may reduce the number of false positive hits and allow for selection of shRNAs with stronger phenotypes. Indeed, suppression of a top candidate, SUMO2, markedly enhanced and accelerated iPSC formation compared to individual hits that emerged from previous large-scale screens or candidates that were selected based on gene expression differences between somatic and pluripotent cells.

In agreement with the notion that the expression of barrier genes does not necessarily have to be different between MEFs and iPSCs, we found that *Sumo2* mRNA levels do not dramatically change during reprogramming. To our knowledge, iPSC formation after

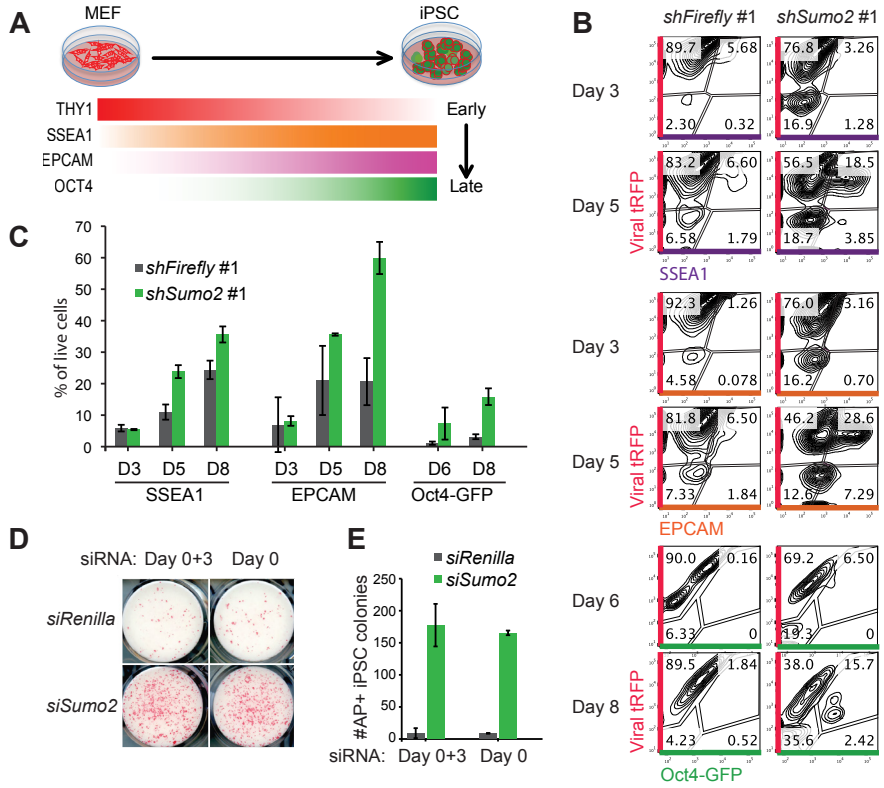


Figure 3. Effect of SUMO2 suppression on defined reprogramming intermediates. (A) Overview of surface markers and reporter alleles to distinguish between early, mid and late stages of reprogramming. (B) Flow cytometry analysis of these markers at intermediate stages of reprogramming in the presence of indicated shRNAs (tRFP+ cells). (C) Quantification of data shown in (B) and Figure S4 using 3 biological replicates +/- s.d. (D) AP+ transgene-independent iPSC colonies obtained after transfection of reprogrammable MEFs with indicated siRNAs either once (day 0) or twice (day 0 and day 3) in the presence of dox for 6 days; iPSC colonies were scored after 4 days of dox withdrawal to capture stable iPSCs. (E) Quantification of data shown in (D) using 3 biological replicates (mean +/- s.d.). (F) Correlation analysis between microarray data obtained in this study (day 6 reprogramming cultures expressing either *Firefly* or *Sumo2* shRNAs, single data points) and a previously reported reprogramming time course[22].

38 hours of OKSM expression represents the shortest time period that has been reported to obtain stably reprogrammed cells from fibroblasts. In addition to SUMO2, our screen uncovered a number of other candidate barriers to iPSC formation, which provide a useful resource for future mechanistic studies of the reprogramming process.

While SUMO2 has not previously been recognized as a roadblock to reprogramming, a recent report suggested that knockdown of the upstream SUMO-conjugating enzyme UBC9 blocks iPSC formation[23]. Several reasons may account for the apparent discrepancy between our studies. For example, UBC9 knockdown is expected to inhibit SUMO1, SUMO2 and SUMO3, which may be toxic to cells, whereas SUMO2 depletion may be tolerated by cells owing to compensation by SUMO1 and/or SUMO3. In support of this idea, *Ubc9* deletion in vivo causes a more severe phenotype compared to individual deletion of *Sumo1, Sumo2* or *Sumo3*[24]. Furthermore, knockdown of UBC9 in ESCs reportedly triggers differentiation[23]. Considering that the same constitutive knockdown vectors were used to assess UBC9's role in ESC self-renewal and reprogramming, it is possible that a potential increase in the number of iPSCs was overlooked due to their immediate loss through differentiation. Indeed, a recent shRNA screen conducted during iPSC formation with retroviral vectors, which become silenced in iPSCs, identified UBC9 as a potent reprogramming barrier[25].

Mechanistically, SUMO2 depletion may enhance iPSC formation by derepressing epigenetically silenced pluripotency loci[26,27] as well as histone and protein biosynthesis genes important for cellular growth and proliferation[28]. Consistently, we find that key pluripotency genes and epigenetic regulators are expressed more robustly in SUMO2-depeted reprogramming intermediates compared to controls. In addition, SUMO2 suppression may contribute to iPSC formation by directly modulating certain pluripotency factors during iPSC generation. In agreement with this view, sumoylation of SOX2 and KLF4 reportedly impairs transcriptional activity and compromises pluripotency[20,29].

Moreover, overexpression of sumoylation-deficient variants of these transcription factors slightly enhances reprogramming into iPSCs, although the reported effects were subtle compared to the phenotype reported here. Considering these observations and our preliminary finding that KLF4 is not differentially sumoylated upon suppression of SUMO2, we surmise that SUMO2 acts at multiple levels to resist the acquisition of pluripotency rather than to control the activity or stability of a single protein target. It should be informative to identify relevant SUMO2 targets in MEFs and iPSCs using recently developed proteomics and ChIP-Seq[28,30] approaches.

Our findings may have practical implications for basic science and cell therapy. The ease with which SUMO2 can be inhibited using transient siRNA delivery should facilitate the mechanistic dissection of the reprogramming process in more homogeneous cell cultures. The observation that SUMO2 depletion increases human iPSC generation, cooperates with

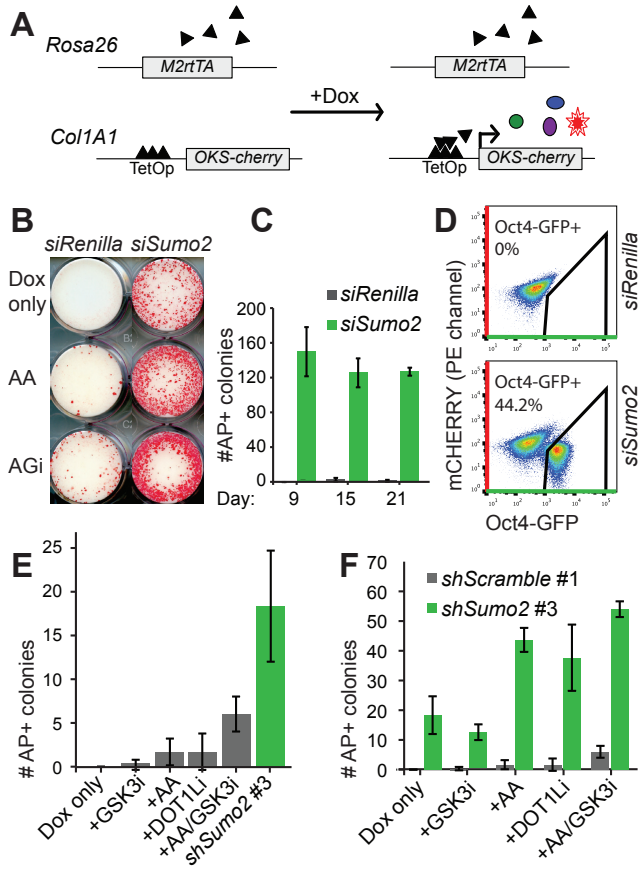


Figure 4. SUMO2 suppression acts independently of C-MYC and in parallel with small molecules. (A) Scheme depicting reprogrammable MEFs carrrying *Col1a1*-tetOP-OKS-mCherry and *Rosa26*-M2rtTA alleles. (B) Generation of AP+ transgene-independent iPSC colonies obtained from these reprogrammable MEFs transfected with indicated shRNAs after exposure to dox and small molecules for 9 days. (C) Quantification of data shown in (B), "dox only" samples and additional time points; data show mean from 3 biological replicates +/- s.d. (D) *Oct4*-GFP expression of OKS reprogrammable MEFs treated with indicated siRNAs and dox for 9 days, followed by 5 days of dox-independent growth. (E) Comparison of iPSC formation efficiencies from OKSM reprogrammable MEFs in the presence of either small molecules or siRNAs targeting *Sumo2*. Values show mean from 3 biological replicates +/- s.d. (F) Combination treatment of reprogrammable MEFs with siRNA targeting *Sumo2* and indicated small molecule. Values show mean from 3 biological replicates +/- s.d.

small molecule enhancers of reprogramming and obviates the need for exogenous *C-Myc* expression may further facilitate the efficient and safe generation of patient-specific iPSCs from rare donor cells.

Methods

Tissue culture and virus production

Reprogrammable Mouse Embryonic Fibroblasts (RepMEFs) were derived from E13.5-15.5 embryos carrying either the *Col1a1*-tetOP-OKSM or *Col1a1*-tetOP-OKS-mCherry alleles in combination with the *Rosa26*-M2rtTA alleles[31]. Reprogramming was initiated by adding

20ng/ml doxycyline (dox)(Sigma, cat.# D9891-25G) to RepMEFs and, where indicated, 25µg/ml L-Ascorbic acid (Sigma, cat.# A4544-25G), 3uM GSK3 inhibitor (CHIR99021, Stemgent, cat.# 04-0004), ALK5 inhibitor (Calbiochem, cat.# 616452), 1uM MEK inhibitor (PD0325901, Stemgent, cat.# 04-0006) or DOT1L inhibitor (generous gift from Dr. Peter Brown). Viral transductions, using pHage or pGIPZ vectors in combination with packaging plasmids psPax2 and pDM2.G, were performed by spin infection for 30 minutes at 2,150rpm at room temperature. The following shRNA seed sequences were used for mouse experiments: for pHage shRNA (shSumo2 #1): ATAAGAGCTGAATGAGCATGCC. For pGIPZ (shSumo2 #2-7; Dharmacon):

#2 (V2LMM_2701): TTCTGGAGTAAAGTAGCAG;
#3 (V2LMM_5114): TAAGAGCTGAATGAGCATG;
#4 (V3LMM_496391): TAGTAGACACCTCCAGTCT;
#5 (V3LMM_496392): AAACTGCACCACAGAACCA;
#6 (V3LMM_496393): TGTTCTCAGTCTTGACTCC;
#7 (V3LMM_496396): AATCTTAAACTGCACCACA.

Transfections of RepMEFs with 1.5ul siRNAs (esiRNA technology, Sigma; Sumo2: EMU095391, Renilla: EHURLUC) were performed in 12-well plates using Lipofectamine 2000 and Opti-MEM (Life Technologies) according to the manufacturer's instructions. For human reprogramming experiments, BJ fibroblasts were first infected with control or *Sumo2* pGIPZ shRNA (V3LHS_388696 and V3LHS412780, GE Dharmacon) viruses. Two days later, cells were infected with the human pHAGE-STEMCCA virus[32] to initiate reprogramming. Cells were re-plated in fibroblast medium (DMEM and 10% FBS) and cultured for four days. Medium was then switched to a 1:1 mixture of fibroblast medium and human ESC medium (E8, StemCell Techonolgies) for another two days, and finally to 100% human ESC medium. Cells were stained for alkaline phosphatase activity (Stemgent) and AP+ iPSCs were counted ~ 3 weeks after STEMCCA infection.

Flow cytometry

For isolation of *Oct4*-GFP+ iPSCs, SSEA1+ cells were first enriched by MACS sorting using SSEA1 antibody coated magnetic beads (Miltenyi). *Oct4*-GFP+ cells were then purified

(continued from right) and *Sumo2* siRNA transfection (38hr iPSCs) (G), 38hr iPSCs and ESCs (H) and iPSCs derived after 10 days of OKSM and *Sumo2* shRNA expression (10 day iPSCs) and ESCs (I). (J) Suppression of SUMO2 mRNA levels in foreskin fibroblasts using *Sumo2* or *Firefly* shRNA vectors as determined by qPCR. Values represent means from 3 technical replicates (+/- s.d.) (K) Reprogramming efficiency of cells characterized in (J) as measured by the number of AP+ colonies per 60,000 plated fibroblasts. Values represent means from 3 biological replicates (+/- s.d.) (L) Representative images of brightfield and AP+ colonies obtained with SUMO2 and FIREFLY shRNA vectors after 14 and 21 days of OKSM expression.

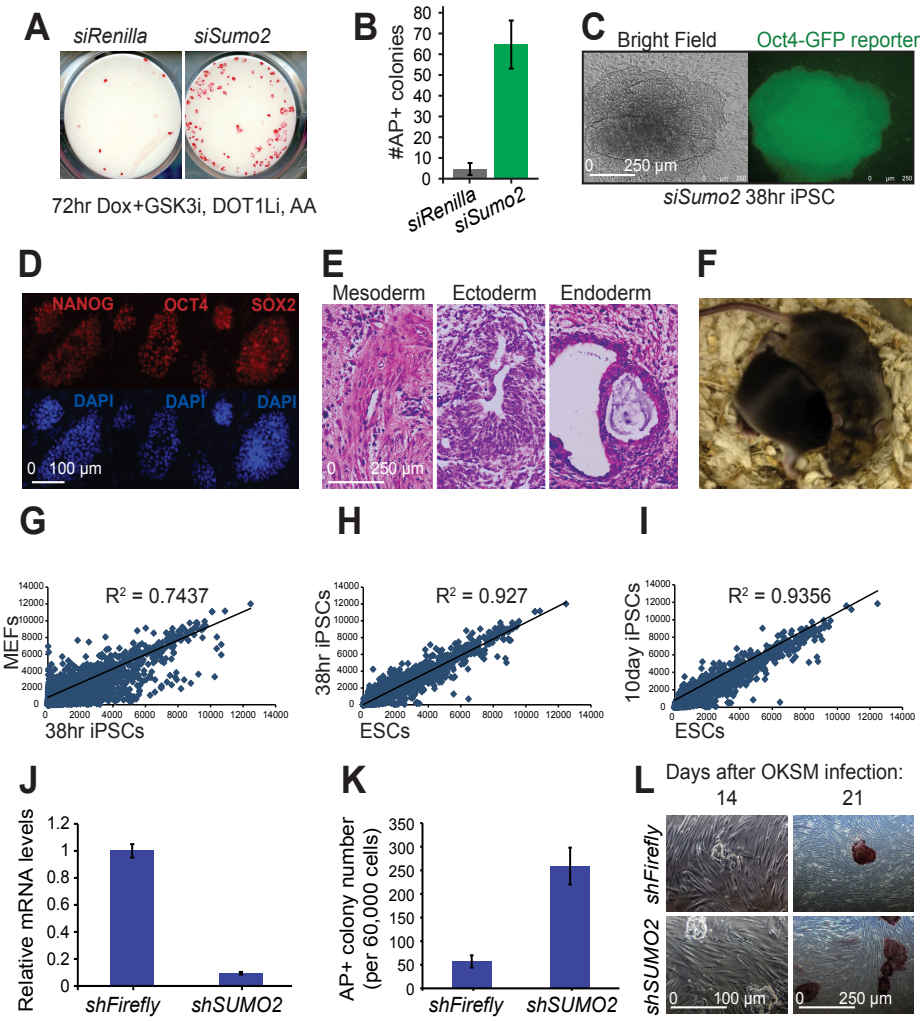


Figure 5. Generation of iPSCs after as little as 38 hours of OKSM expression. (A) Treatment with ascorbic acid (AA), DOT1L inhibitor (DOT1Li) and GSK3 inhibitor (GSK3i) facilitates the recovery of transgene-independent AP+ iPSC colonies from control and Sumo2 siRNA transfected MEFs after 72 hours of OKSM expression. (B) Quantification of data shown in (A) using 3 biological replicates (mean +/- s.d.). (C) Suppression of SUMO2 enables generation of *Oct4*-GFP+ transgene-independent iPSCs after 38 hours of OKSM expression in the presence of the indicated small molecules. (D) Representative immunofluorescence images showing expression of endogenous OCT4, NANOG and SOX2 in iPSCs generated after 38 hours of OKSM expression and *Sumo2* siRNA transfection. (E) iPSCs shown in (C) are pluripotent as determined by their potential to differentiate into the 3 germ layers in teratomas. (F) iPSCs produced after 38 hours of OKSM expression give rise to coat color chimeras. (G-I) Correlation analysis of microarray data (single data points) between MEFs and iPSCs derived after 38 hours of OKSM expression (Continued on left)

within the enriched SSEA1+ fraction by FACS. Gates were set based on uninfected MEFs, virally transduced MEFs (tRFP+) and *Oct4*-GFP+ iPSCs. Intermediates of reprogramming were analyzed by flow cytometry using the following antibodies: THY1-Viogreen (BD, cat.# 561616) or THY1-Pacific Blue (eBioscience, cat.# 48-0902-82), SSEA1-APC (Biolegend, cat.# 125608) or SSEA1-PE-Cy7 (Miltenyi Biotec, cat.# 130-100-426), and EPCAM-PE-Cy7 (eBioscience, cat.# 25-5791-80) (1:200 for 30min. at 4C). BD's Annexin V kit was used to measure apoptotic cells. All cytometry data were analyzed and plotted using FlowJo software.

Quantification of reprogramming efficiencies

For macroscopic detection of iPSC colonies, Alkaline Phosphatase (AP) staining was carried out according to manufacturer's instructions using the Vector Labs AP staining kit (cat.# 5100). AP staining was always performed 2 to 4 days after dox withdrawal to eliminate partially reprogrammed colonies and to score for transgene-independent iPSCs. Colonies were counted manually or by custom-made Nikon software (CL-Quant).

Teratoma and chimera formation

For teratoma generation, iPSC lines (passage 6 or higher) were harvested and resuspended in 600µl media per confluent 6-well. Mice were anesthetized with Avertin and injected with 150µl cell suspension subcutaneously. Tumours were harvested 3 to 4 weeks after injection and analyzed histologically. For chimera production, iPSC lines were injected as single cell suspension into day 3.5 blastocysts isolated from intercrosses of C57Bl/6xBDF1 mice. Blastocysts were transferred into pseudopregnant Swiss Webster recipient animals.

Immunofluorescence analysis

iPSC lines (passage 6 or higher) were seeded in a 24-well plate at a low density. Once small colonies emerged, wells were washed with 1xPBS and fixed by a 5-10 minute incubation in 10% formalin at room temperature. After washes in 1xPBS, cells were blocked in 1xPBS containing 2% BSA and 0.1% Triton-X 100. Primary and secondary antibodies were diluted in blocking solution at a concentration of 1:200 and added for 1 hour at RT or overnight at 4C. Primary antibodies were anti-*Nanog* (Abcam, cat.# AB80892), anti-*Sox2* (Santa Cruz, cat.# Sc-17320), anti-*Oct4* (Santa Cruz, cat.# Sc-8628); secondary antibodies were donkey anti-goat IgG or donkey anti-rabbit IgG Alexa Fluor 546-conjugated antibodies (Life Technologies). After 2 washes in 1xPBS, cells were immobilized on slides in mounting media containing DAPI (Vectashield, Vector Labs) and analyzed.

RNA expression analysis

For global gene expression analysis, total RNA was isolated from indicated samples using RNeasy Mini Kit (Qiagen) and analyzed by Affymetrix microarray chips. The raw microarray expression signals in the CEL files were normalized using the Affymetrix Expression Console software and RMA normalization. For quantitative PCR analysis, Brilliant III SYBR-green based master mix was used according to the manual (Agilent), following RNA isolation (RNeasy kit, Qiagen) and reverse transcription (Transcriptor First Strand cDNA Synthesis Kit, Roche) of Sumo2 or control knockdown samples two days after initiation of reprogramming. Samples were run in triplicate on the Lightcycler 480 (Roche). Primer sequences for qPCR were as follows: *Sumo2* (forward): AAGGAAGGAGTCAAGACTGAGAA, *Sumo2* (reverse): CGGAATCTGATCTGCCTCATTG, GAPDH (forward): AGGTCGGTGTGAACGGATTTG and GAPDH (reverse): TGTAGACCATGTAGTTGAGGTCA.

Scatter Plot

Microarray gene expression measurements for samples containing *Sumo2* or *Firefly* shRNA for 6 days were obtained. For genes with more than one probe, the average of all probes was used. These expression measurements were plotted (on a log2 scale), with Firefly expression on the x-axis, *Sumo2* knockdown expression on the y-axis, and each point representing a gene. Genes falling outside of the dashed lines have a fold change of greater than 1.5 between *Sumo2* knockdown and *Firefly*. Some key genes were highlighted, with those being upregulated in *Sumo2* knockdown colored red, and those downregulated colored green.

Correlation Matrix

Previously published microarray gene expression measurements representing stages of reprogramming were obtained from GEO (GSE42379). The following samples were used: KH2-MEF_m2, KH2-MEF_m3, Day3_SSEA1+ M2, Day3_SSEA1+ M3, Day6_SSEA1+ M2 and Day6_SSEA1+ M3, Day9_SSEA1+ M2, Day12_SSEA1+ M2, iPS_KH2-SC_MEF_1-5, iPS_KH2-SC_MEF_1-6, and iPS_KH2-SC_MEF_1-3 (GSM1038591, GSM1038592, GSM1038595, GSM1038598, GSM1038601, GSM1038604, GSM1038607, GSM1038611, GSM1038612, GSM1038613, GSM1038614), to represent day 0 of reprogramming (MEF), day 3, day 6, day 9, day 12, and fully reprogrammed (iPS). For timepoints with multiple samples, the average expression measurement for each probe was used. For genes with more than one probe, the average of all probes was used.

The gene expression measurements of the *Sumo2* knockdown and *Firefly* knockdown samples were compared to those of the samples representing stages of reprogramming. Only the genes that were present in both datasets were considered. To remove batch effect,

the “Remove Batch Effect” tool in the Partek Genomics Suite was used, which fits a linear model to the data which includes batch as a component, then subtracts out that component from the data.
A matrix was constructed with pairwise comparisons between the *Sumo2* knockdown and *Firefly* knockdown samples (as the rows in the matrix) to each of the samples representing stages of reprogramming (as the columns in the matrix). A scatterplot was generated and a Pearson’s R correlation coefficient was calculated for each pairwise comparison. Scatterplots that are shaped closer to a straight line and have a higher correlation coefficient indicate that the global expression patterns of the two samples more closely match.

Western Blot analysis

Protein lysates were run on 4-20% Mini Protean TGX gels (Biorad), blotted onto Immobilon-P membrane (EMD Millipore) and incubated with anti-SUMO2 antibody ab3742 (Abcam) and anti-RAN antibody 610341 (BD) for Western blot analysis. We used anti-KLF4 antibody AF3158 (R&D Systems), anti-OCT4 antibody 11263-1-AP (ProteinTech) and anti-SUMO2 antibody ab81371 (Abcam) for immunoprecipitation experiments.

Generation of pHAGE-Mir vector

The pHAGE lentiviral backbone was released from pHAGE-EF1a-eGFP-W vector (kindly provided by Dr. Richard Mulligan’s lab) using NotI + BamHI. Turbo RFP was PCR amplified from pTurboRFP-N (Evrogen) and blunt-end cloned into the above pHAGE backbone to generate the intermediate pHAGE-EF1a-turboRFP vector. The Mir30-shRNA cassette and PGK-Puro selection marker was digested from MSCV-PM using BglII + ClaI, and blunt-end cloned into the BamHI digested pHAGE-EF1a-turboRFP intermediate to generate the pHAGE-Mir vector.

shRNA screen and identification of hits

RepMEFs were expanded until passage 4 in 4% oxygen, switched to atmospheric oxygen and infected with the pooled shRNA library as described above. For each shRNA (621,000 shRNAs in total), 1-2 x103 cells were infected to achieve good coverage. Infected cells were passaged onto gelatinized 15cm cell culture dishes (Falcon) in reprogramming media (ESC media supplemented with ascorbic acid and doxycycline) for 10 days, and in doxycycline/ ascorbic acid-free ESC media for an additional 4 days. Cells were harvested, pooled, and purified with SSEA1-linked magnetic beads using an AutoMACS sorter (Miltenyi). SSEA1-enriched cells were then FACS-sorted for endogenous *Oct4*-GFP expression.
Genomic DNA was extracted from collected *Oct4*-GFP+ cells by lysing the cells in 10 mM Tris pH 8.0, 10 mM EDTA, 10 mM NaCl, 0.5% Sarkosyl. Lysates were treated with 0.1 mg/

ml Rnase A at 37˚C for 30 min, 0.5 mg/ml Proteinase K at 55˚C for 1-2 hr, and then phenol-chloroform extracted, ethanol precipitated, and resuspended in 10 mM Tris-HCl pH 8.0. For each sample, all the genomic DNA was used as template for shRNA PCR, usually in multiple PCR reactions. Each 50 µl PCR reaction contained: 2.5 µg genomic DNA template, 200 µM dNTPs, 400 nM of each PCR primer (pHAGE-Mir-PCR: 5'- GCAAACTGGGGCACAGATGATGCGG; BC1R-L: 5'- CGCCTCCCCTACCCGGTAGA), 1x Q5 reaction buffer, 1x Q5 high GC buffer, and 0.5 µl Q5 polymerase (NEB). PCR was performed with the following program: 94 ˚C 4 min, 35 cycles of (94˚C 30 sec, 60 ˚C 30 sec, 72 ˚C 45 sec), 72 ˚C 10 min. PCR products (~700 bp) for each sample were pooled, ethanol precipitated, resuspended, and gel-purified using the QIAquick Gel Extraction Kit (Qiagen). The purified shRNA PCR products were used to: 1) generate sub-libraries for the next round of shRNA library screens; 2) generate sequencing libraries for Illumina sequencing.

For sub-library generation, the purified PCR product was digested with NotI and MluI, and the ~400 bp fragment that contains the shRNAs was gel-purified. Separetely, the pHAGE-Mir plasmid was also digested with NotI and MluI to recover the ~ 9kb vector backbone. 25-50 ng of the purified shRNA fragment and 125-250 ng of the vector backbone were ligated in 5 µl ligation reaction using NEB T4 ligase. 1 µl ligation reaction was used to transform 20 µl Electromax competent cells DH10b (Life Technology) with electroporation. 1 µl of the transformation reaction was plated on one 10-cm LB-Amp (100 µg/ml) plate to estimate the total number of colonies, and the rest of the transformation reaction was plated on two 15-cm LB-Carbenicillin (100 µg/ml) plates and grown overnight at 37 ˚C. To maintain the representation of the library, at least 100 x coverage was needed (i.e., colony number = 100 x number of shRNAs in the library). When necessary, the entire ligation reaction may be used for transformation in multiple electroporation reactions to increase the number of colonies. The next day, lawn formed on the two 15-cm plates were scraped off and cultured in 300 ml LB-Carbenicillin (100 µg/ml) medium and grown at 30˚C for 2-3 hrs. The cloned sub-library DNA was extracted from collected bacteria by the Genelute Maxiprep kit (Sigma).

For Illumina sequencing, the purified shRNA PCR product was used as template for another round of PCR: 500 ng purified shRNA PCR product, 200 µM dNTPs, 2 µM of each PCR primer (p5 and p7), 1x Q5 reaction buffer, 1x Q5 high GC buffer, and 1 µl Q5 polymerase (NEB) in 100 µl PCR reaction. PCR was performed with the following program: 94 ˚C 4 min, 2 cycles of (94˚C 30 sec, 50 ˚C 20 sec, 72 ˚C 30 sec), 20 cycles of (94˚C 30 sec, 60 ˚C 20 sec, 72 ˚C 30 sec), 72 ˚C 10 min. PCR products (~120 bp) were and gel-purified using the QIAquick Gel Extraction Kit (Qiagen). Gel-purified products were submitted for Illumina sequencing on the Illumina MiSeq instrument, using a custom sequencing primer: mir30-EcoRI : 5'-TAGCCCCTTGAATTCCGAGGCAGTAGGCA. PCR primers:

p5-miSeq: 5'-ATGATACGGCGACCACCGAGATCTACACCTAAAGTAGCCCCTTGAATTC;

p7-miSeq-1: 5'-CAAGCAGAAGACGGCATACGAGACGATAGTGAAGCCACAGATGTA
p7-miSeq-2: 5'-CAAGCAGAAGACGGCATACGAGACACTAGTGAAGCCACAGATGTA
p7-miSeq-3: 5'-CAAGCAGAAGACGGCATACGAGACTATAGTGAAGCCACAGATGTA
p7-miSeq-4: 5'-CAAGCAGAAGACGGCATACGAGACCTTAGTGAAGCCACAGATGTA
Different p7 primers were used for multiplexing purpose.
Single-end 51 bp reads were obtained using the Illumina HiSeq or MiSeq instrument. The reads are expected to have an initial 22 nucleotides that identify the shRNA, followed by a constant region that is the same for all shRNAs and a 2 nucleotide barcode to identify the sample. Reads that contain perfect matches at the following 6 nucleotides were first extracted from the sequencing data: the 2 nucleotides adjacent to the initial 22 base sequence and the 2 nucleotides adjacent to the barcode on both sides. The shRNAs were then identified by requiring an exact match of the 22 nucleotides to the sequences in the shRNA library annotation file. The samples were identified by the 2 nucleotide barcodes.
The total number of reads that were identified for each shRNA, sample, and round were counted. The counts were normalized to be directly comparable between samples and rounds by first dividing by the total number of counts for that sample and round and then multiplying by the total number of shRNAs in the initial library. A pseudocount of 1 was added to each normalized count to downweight enrichment derived from low read counts and to avoid division by zero in calculating fold-changes.
The enrichment for each shRNA in each round was calculated as the log2 fold change of the *Oct4*-GFP+ normalized counts over the maximum of the normalized counts of the controls (T0, No-Dox and *Oct4*-GFP- cells). The cumulative enrichment for each shRNA in each round was calculated as the sum of the log2 fold changes for that round and all previous rounds. The overall enrichment of each shRNA was defined as the maximum of the cumulative enrichment scores among all rounds.
The heat map for Figure 1G was plotted using the cumulative enrichment scores. Only shRNAs that have at least one read in the *Oct4*-GFP+ sample in at least two rounds were used in the plot, resulting in a total of 23,853 shRNAs.

Accession numbers

Microarray expression data are available at the Gene Expression Omnibus database under accession number GSE76699.

Supplemental table 1 legend

(Excell file available at URL:
http://www.cell.com/cms/attachment/2055690886/2061109055/mmc2.xlsx)

List of identified shRNAs, ordered by overall (maximum) enrichment score with information on shRNA identifier, 22-mer target sequence, gene accession number, gene symbol/ description and enrichment scores for individual rounds.

Acknowledgements

We thank Laura Prickett, Kat Folz-Donahue and Meredith Weglarz of the CRM Flow Cytometry core facility and all K.H. and G.H. lab members for advice and discussions. We are grateful to Dr. Peter Brown for providing the DOT1L inhibitor, and Dr. Jianping Jin for providing the SUMO2 entry vector. We thank Dr. Mehrnaz Ghazvini for help with human reprogramming experiments. M.B. was supported by a Boehringer Ingelheim Fonds predoctoral fellowship. This work was supported by grants from the NIH (GM44664 and AG011085) to S.J.E.; S.J.E. is an Investigator with the Howard Hughes Medical Institute. G.H. was supported by the Intramural Research Program at the National Institute of Environmental Health Sciences (Z01ES102745). Support to K.H. was from the Gerald and Darlene Jordan Endowment, the NIH (HD058013) and an Early Career Award from the Howard Hughes Medical Institute.

Author contributions

M.B., B.D.B., N.M., S.J.E., G.H. and K.H. planned the experiments, M.B. and G.H. performed the serial enrichment screen, validated and characterized hits, B.D.B. conducted all bioinformatics analyses with assistance from Y.D. and D.C.F., B.L. performed sumoylation experiments, O.B.N. generated teratomas and helped with array analysis, J.B. generated chimeric mice, L.W. performed human reprogramming experiments, E.A. provided tetOP-OKS MEFs, S.C. shared unpublished information. M.B., B.D.B., G.H. and K.H. wrote the manuscript.

References

1 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.

2 Apostolou, E., and Hochedlinger, K. (2013). Chromatin dynamics during cellular reprogramming. Nature 502, 462-471.

3 Krizhanovsky, V., and Lowe, S.W. (2009). Stem cells: The promises and perils of p53. Nature 460, 1085-1086.

4 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.

5 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.

6 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.

7 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.

8 Onder, T.T., Kara, N., Cherry, A., Sinha, A.U., Zhu, N., Bernt, K.M., Cahan, P., Marcarci, B.O., Unternaehrer, J., Gupta, P.B., et al. (2012). Chromatin-modifying enzymes as modulators of reprogramming. Nature 483, 598-602.

9 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.

10 dos Santos, R.L., Tosti, L., Radzisheuskaya, A., Caballero, I.M., Kaji, K., Hendrich, B., and Silva, J.C. (2014). MBD3/NuRD facilitates induction of pluripotency in a context-dependent manner. Cell stem cell 15, 102-110.

11 Jackson-Grusby, L., Beard, C., Possemato, R., Tudor, M., Fambrough, D., Csankovszki, G., Dausman, J., Lee, P., Wilson, C., Lander, E., et al. (2001). Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. Nature genetics 27, 31-39.

12 Qin, H., Diaz, A., Blouin, L., Lebbink, R.J., Patena, W., Tanbun, P., LeProust, E.M., McManus, M.T., Song, J.S., and Ramalho-Santos, M. (2014). Systematic identification of barriers to human iPSC generation. Cell 158, 449-461.

13 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.

14 Yang, C.S., Chang, K.Y., and Rana, T.M. (2014). Genome-wide functional analysis reveals factors needed at the transition steps of induced reprogramming. Cell reports 8, 327-337.

15 Stadtfeld, M., Maherali, N., Borkent, M., and Hochedlinger, K. (2010). A reprogrammable mouse strain from gene-targeted embryonic stem cells. Nature methods 7, 53-55.

16 Meerbrey, K.L., Hu, G., Kessler, J.D., Roarty, K., Li, M.Z., Fang, J.E., Herschkowitz, J.I., Burrows, A.E., Ciccia, A., Sun, T., et al. (2011). The pINDUCER lentiviral toolkit for inducible RNA interference in vitro and in vivo. Proceedings of the National Academy of Sciences of the United States of America 108, 3665-3670.

17 Schlabach, M.R., Luo, J., Solimini, N.L., Hu, G., Xu, Q., Li, M.Z., Zhao, Z., Smogorzewska, A., Sowa, M.E., Ang, X.L., et al. (2008). Cancer proliferation gene discovery through functional genomics. Science 319, 620-624.

18 Dejosez, M., Ura, H., Brandt, V.L., and Zwaka, T.P. (2013). Safeguards for cell cooperation in mouse embryogenesis shown by genome-wide cheater screen. Science 341, 1511-1514.

19 Flotho, A., and Melchior, F. (2013). Sumoylation: a regulatory protein modification in health and disease. Annual review of biochemistry 82, 357-385.

20 Tahmasebi, S., Ghorbani, M., Savage, P., Yan, K., Gocevski, G., Xiao, L., You, L., and Yang, X.J. (2013). Sumoylation of Kruppel-like factor 4 inhibits pluripotency induction but promotes adipocyte differentiation. The Journal of biological chemistry 288, 12791-12804.

21 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.

22 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.

23 Tahmasebi, S., Ghorbani, M., Savage, P., Gocevski, G., and Yang, X.J. (2014). The SUMO conjugating enzyme Ubc9 is required for inducing and maintaining stem cell pluripotency. Stem Cells 32, 1012-1020.

24 Wang, L., Wansleeben, C., Zhao, S., Miao, P., Paschen, W., and Yang, W. (2014). SUMO2 is essential while SUMO3 is dispensable for mouse embryonic development. EMBO reports 15, 878-885.

25 Cheloufi, S., Elling, U., Hopfgartner, B., Jung, Y.L., Murn, J., Ninova, M., Hubmann, M., Badeaux, A.I., Euong Ang, C., Tenen, D., et al. (2015). The histone chaperone CAF-1 safeguards somatic cell identity. Nature 528, 218-224.

26 Poleshko, A., Kossenkov, A.V., Shalginskikh, N., Pecherskaya, A., Einarson, M.B., Marie Skalka, A., and Katz, R.A. (2014). Human factors and pathways essential for mediating epigenetic gene silencing. Epigenetics : official journal of the DNA Methylation Society 9, 1280-1289.

27 Yang, B.X., El Farran, C.A., Guo, H.C., Yu, T., Fang, H.T., Wang, H.F., Schlesinger, S., Seah, Y.F., Goh, G.Y., Neo, S.P., et al. (2015). Systematic identification of factors for provirus silencing in embryonic stem cells. Cell 163, 230-245.

28 Neyret-Kahn, H., Benhamed, M., Ye, T., Le Gras, S., Cossec, J.C., Lapaquette, P., Bischof, O., Ouspenskaia, M., Dasso, M., Seeler, J., et al. (2013). Sumoylation at chromatin governs coordinated repression of a transcriptional program essential for cell growth

and proliferation. Genome research 23, 1563-1579.

29 Wu, Y., Guo, Z., Wu, H., Wang, X., Yang, L., Shi, X., Du, J., Tang, B., Li, W., Yang, L., et al. (2012). SUMOylation represses Nanog expression via modulating transcription factors Oct4 and Sox2. PloS one 7, e39606.

30 Becker, J., Barysch, S.V., Karaca, S., Dittner, C., Hsiao, H.H., Berriel Diaz, M., Herzig, S., Urlaub, H., and Melchior, F. (2013). Detecting endogenous SUMO targets in mammalian cells and tissues. Nature structural & molecular biology 20, 525-531.

31 Bar-Nur, O., Brumbaugh, J., Verheul, C., Apostolou, E., Pruteanu-Malinici, I., Walsh, R.M., Ramaswamy, S., and Hochedlinger, K. (2014). Small molecules facilitate rapid and synchronous iPSC generation. Nature methods 11, 1170-1176.

32 Sommer, A.G., Rozelle, S.S., Sullivan, S., Mills, J.A., Park, S.M., Smith, B.W., Iyer, A.M., French, D.L., Kotton, D.N., Gadue, P., et al. (2012). Generation of human induced pluripotent stem cells from peripheral blood using the STEMCCA lentiviral vector. Journal of visualized experiments : JoVE.

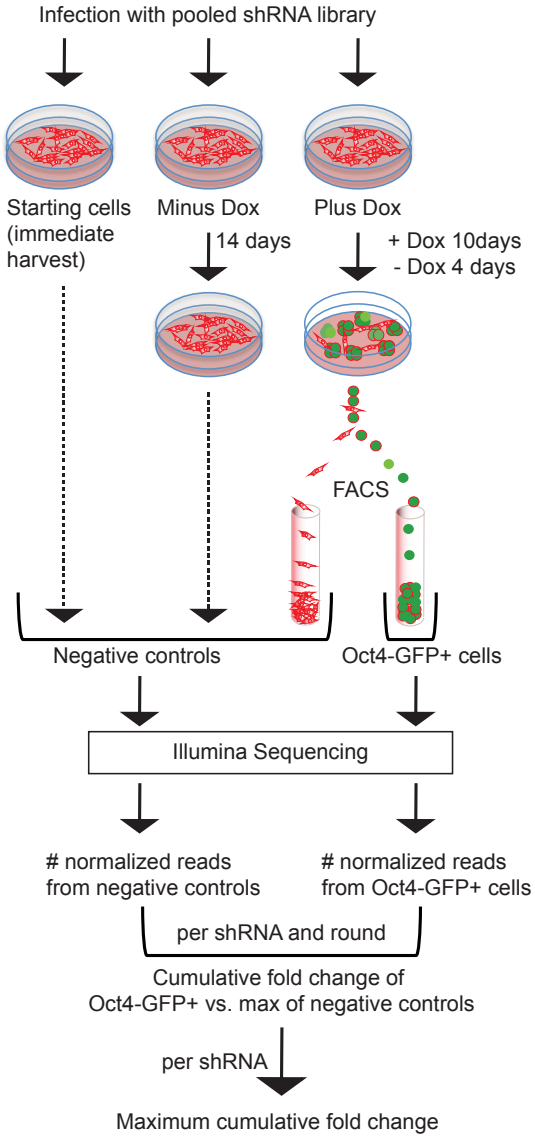


Figure S1. Workflow to obtain experimental and control samples for deep sequencing during serial shRNA screen. See text and methods section for details.

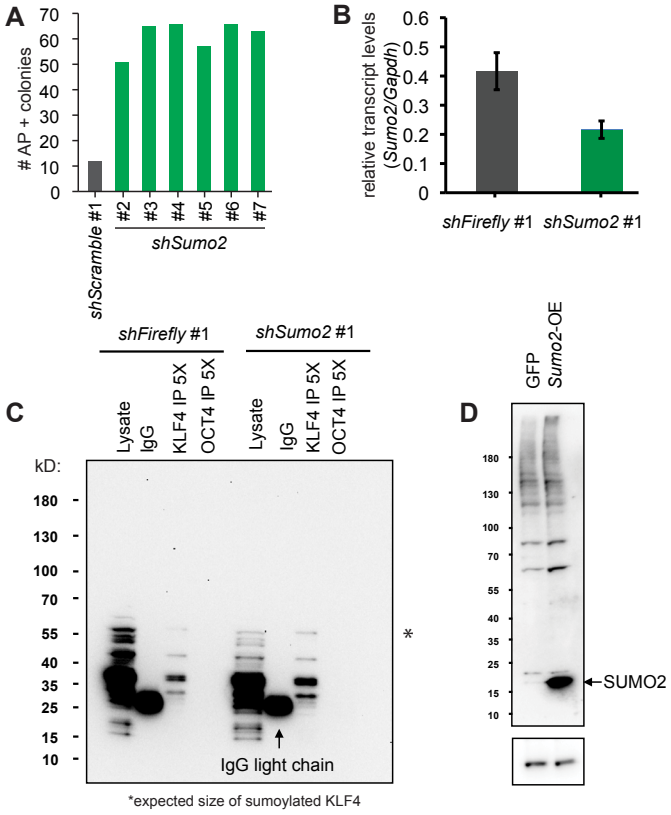


Figure S2. Confirmation of reprogramming phenotype with independent *Sumo2* shRNAs, and effect of SUMO2 suppression or overexpression on global sumoylation levels. (A) iPSC formation efficiencies after infecting reprogrammable MEFs with additional shRNAs targeting different seed sequences within *Sumo2* mRNAs. iPSC colonies were quantified after 10 days of dox treatment and 4 days of dox withdrawal. Data represent single experiments. (B) shRNAs against *Sumo2* lead to reduction of target mRNA levels. qPCR analysis for *Sumo2* transcripts normalized to *Gapdh* levels. Values are means from 3 technical replicates (+/- s.d.). (C) Analysis of KLF4 sumoylation by immunoprecipitation-Western blot analysis. Protein extracts from reprogrammable cells expressing OKSM and either *Firefly* (FF) or *Sumo2* shRNAs for 3 days were immunoprecipitated with either KLF4 or OCT4 specific antibodies, followed by Western blot analysis with a SUMO2 specific antibody. Amount of recovered *Oct4* immunoprecipitates was likely below the detection limit, explaining the lack of signal. Size of expected sumoylated KLF4 band is indicated with an asterisk. (D) Overexpression of SUMO2 in dox treated reprogrammable MEFs compared to a GFP-expressing control. Western blot analysis for SUMO2 in cells expressing OKSM and SUMO2 for 3 days.

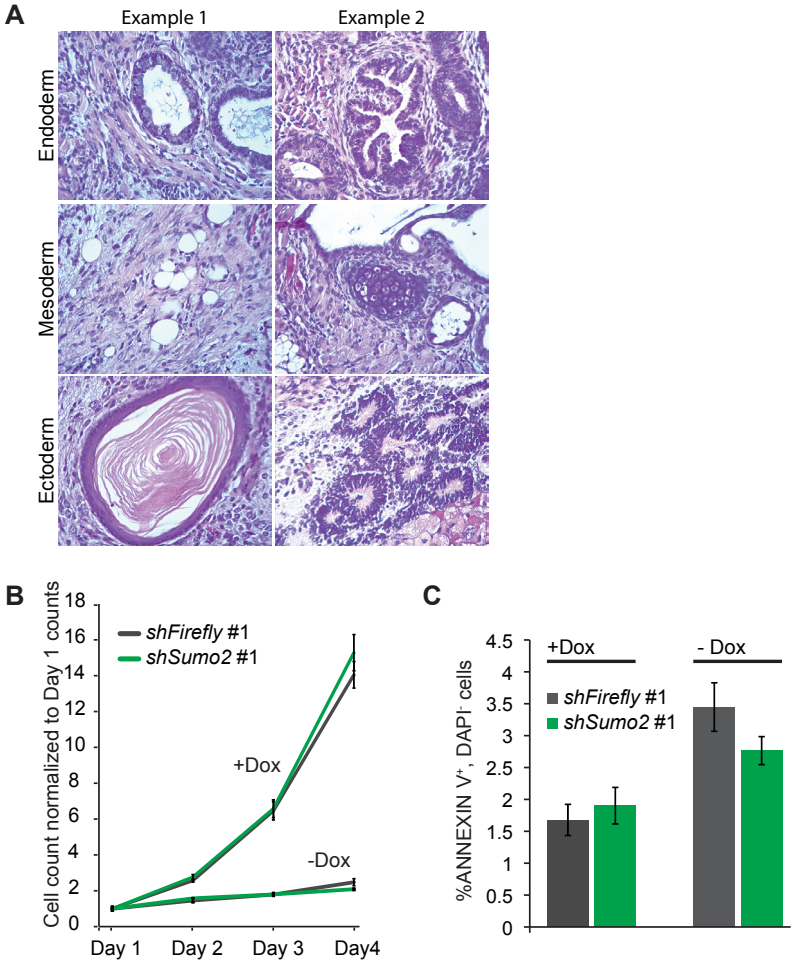


Figure S3. Effect of SUMO2 suppression during reprogramming on cellular proliferation, apoptosis and pluripotency. (A) iPSC clones obtained after *Sumo2* knockdown give rise to well-differentiated teratomas showing differentiation into ectodermal, endodermal and mesodermal derivatives. Shown are different regions of the same teratoma. (B) *Sumo2* knockdown does not affect cell proliferation or viability. Growth curves of reprogrammable MEFs expressing shRNAs targeting *Firefly* or *Sumo2* in the presence or absence of dox (i.e., OKSM expression). Cell counts were normalized to the cell number at day 1 (data are mean +/- s.d. from 3 biological replicates). (C) Parallel cultures as shown in (A) were stained for Annexin V (apoptotic cells) and DAPI (dead cells)(data are mean +/- s.d. from 3 biological replicates).

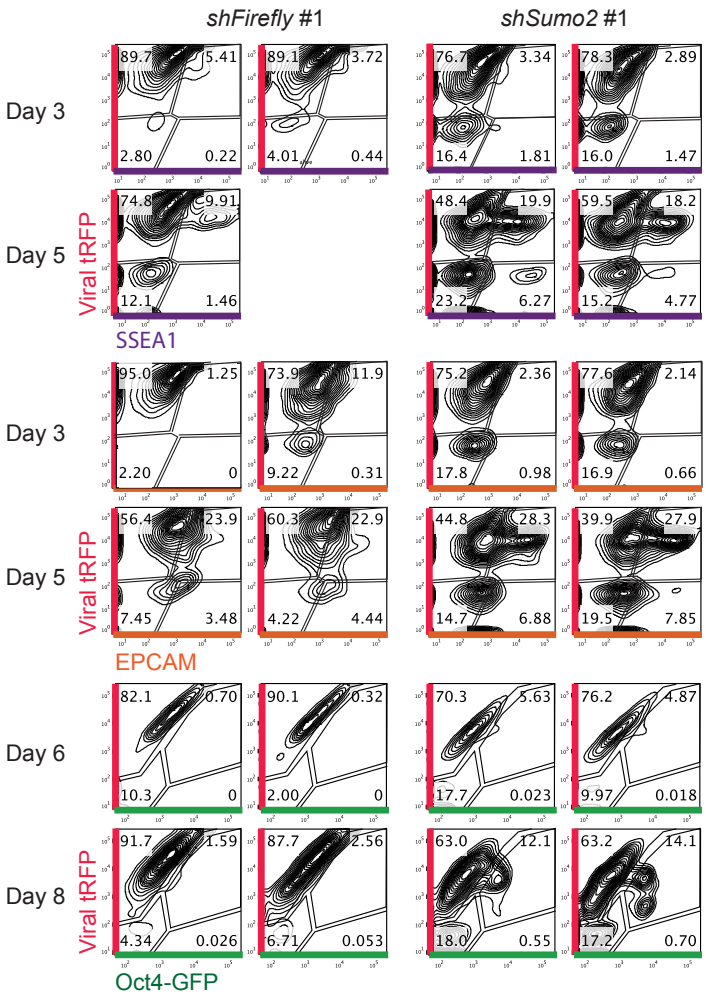


Figure S4. Flow cytometric analysis of SSEA1, EPCAM and *Oct4*-GFP expression in reprogramming intermediates carrying *Firefly* or *Sumo2* shRNAs. Additional biological replicates of assay shown in Figure 3B and quantified in Figure 3C.

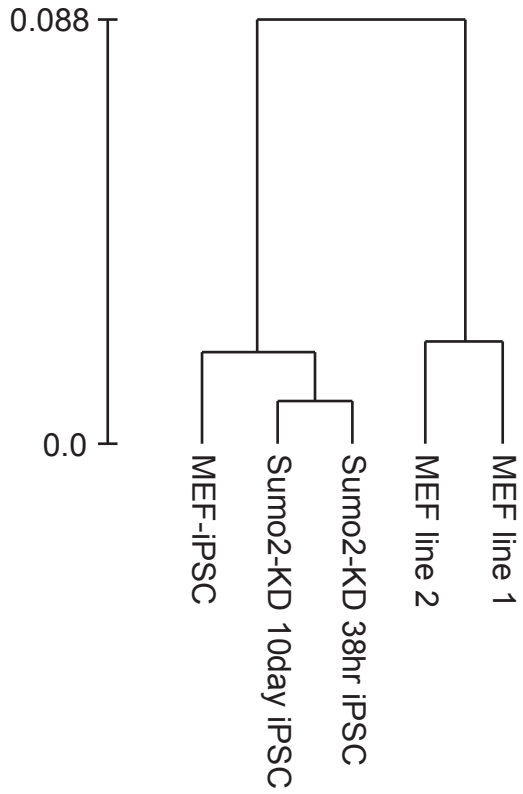


Figure S5. Unsupervised clustering of microarray expression data generated from MEFs and iPSCs derived under different conditions. MEF lines 1, 2 and MEF-iPSCs were previously published in Bar-Nur et al. (2014)[31].

Chapter 5

Alternative polyadenylation factor *Nudt21* inhibits cellular reprogramming

Marti Borkent*, Chengguo Yao*, Justin Brumbaugh*, Brian D. Bennett, Brad Lackford, Lingjie Weng, Ying Du, David C. Fargo, Sihem Cheloufi, Ori Bar-Nur, Effie Apostolou, Nimet Maherali, Stephen Elledge, Guang Hu, Yongsheng Shi#, Konrad Hochedlinger#

* equally contributing authors
# corresponding authors
Manuscript in preparation

Abstract

Differentiated, somatic cells can be reprogrammed back to a pluripotent state by the upregulation of a few transcription factors (OCT4, SOX2, KLF4 and c-MYC; OKSM). Resulting induced pluripotent stem cells (iPSCs) form at a low rate, indicating many intrinsic factors must prohibit this transformation. In a large scale RNA-inhibition screen, we identified suppression of *Nudt21* to stimulate iPSC formation. *Nudt21* suppression increases expression of pluripotency associated genes, facilitates formation of iPSCs upon just 2 days of exogenous OKSM expression and allows for omission of c-Myc from the reprogramming cocktail. *Nudt21* encodes for CPSF5; a factor involved in alternative polyadenylation (APA); the process that determines the length of the 3'UTR of many mRNAs. Preliminary analysis of RNA sequencing directed at the 3'UTR of all mRNAs reveals shifts in 3'UTR length upon *Nudt21* suppression during reprogramming, affecting genes including stem cell associated factor *Wdr5*. This study identifies an APA modifying factor as a roadblock to gaining pluripotency; further analysis of choice of polyadenylation site and resulting proteome upon *Nudt21* suppression during reprogramming can help clarify the role of APA during cell fate changes.

Introduction

Somatic cells can be reprogrammed into pluripotent stem cells (induced Pluripotent Stem Cells, iPSCs) using a set of transcription factors: *Oct4*, Klf4, Sox2 and c-Myc (OKSM). iPSCs have been generated in various model systems including human cells, and reproduced by many research groups in and outside the stem cell field. However, the efficiency of the classical protocol is limited to 0,01-3%, suggesting intrinsic factors prevent most cells from reprogramming to iPSCs[1,2]. We previously described a screening approach to identify genes that inhibit reprogramming. This screen originally identified the gene *Sumo2* (chapter 4), and in re-analysis of an earlier 3-round screen, we identified additional hits. The top hit from this group was *Nudt21*, which encodes for the alternative polyadenylation factor CPSF5. Here we validated the positive effect of *Nudt21* downregulation on reprogramming in different contexts; including with 3-factor (c-Myc-free) reprogramming. Reprogramming speed was optimized by combining *Nudt21* knockdown with small molecule enhancers. We characterized *Nudt21*-knockdown iPSC. Finally, we explore potential mechanisms by determining alternative polyadenylation profiles during reprogramming, using polyadenylation site-specific sequencing.

Results

Knockdown of Nudt21 increases reprogramming efficiency and speed

We used our model of a 'reprogrammable mouse'; a transgenic mouse with doxycycline-inducible expression of the four reprogramming factors OKSM3. We performed a genome-wide, serial RNAi screen for inhibitors of reprogramming to pluripotency (Figure 1A). We identified an shRNA targeting *Nudt21* to score as a strong hit. When we analyzed the composition of the shRNA pool before and after a reprogramming experiment. We noticed that during three screening rounds, the shRNA targeting *Nudt21* was relatively enriched, suggesting it has a beneficiary effect on forming iPSCs (Figure 1B). We were able to validate this result in several ways. The number of dox-independent (stable upon withdrawal of exogenous OKSM expression) upon *Nudt21* knockdown by several different shRNAs was increased 6-fold (Figure 1C). Also, the percentage of cells expressing the pluripotency reporter *Oct4*-GFP after a week of reprogramming was increased by around 16-fold (Figure D-E). Third, more dox-independent iPSC colonies seem to be formed from *Nudt21* knockdown reprogramming cells when decreasing the duration of OKSM expression (Figure 1F-G). Expression of the shRNA resulted in a strong decrease of *Nudt21*-encoded protein as shown by western blot (Figure 1H). When we used alternative knockdown strategies, by employing small inhibitory RNAs (siRNAs) targeting *Nudt21*, the increase in reprogramming efficiency was even bigger; up to 10 fold more iPSC colonies were formed when anti-*Nudt21* siRNAs were transfected once or twice during reprogramming (Figure 1I-J). Reprogramming with expression of the *Nudt21* shRNA resulted in iPSC that could be passaged and grown well under ESC conditions (not shown) and formed, upon subcutaneous injection, well-differentiated teratomas displaying lineages from all three germ layers (Figure 1K).

Effect of Nudt21 knockdown on cell proliferation and survival, as well as expression of stem cell genes

Proliferation and cell death can dictate reprogramming efficiency by determining the total number of cells undergoing reprogramming. Additionally, reprogramming has been shown to directly depend on the number of divisions undergone by an original cell[4]. The proliferation and cell death rate was determined for *Nudt21* knockdown cells undergoing reprogramming to dissect their contribution to the phenotype. Reprogrammable cells were counted daily during 4 days while cultured in media with or without doxycycline to induce reprogramming. Although dox-induced cells increased more in number than non-induced cells, no difference was noted between *Nudt21* knockdown cells or control (Figure 2A). Later time-points were not assessed to avoid bias of having more reprogrammed cells in

one culture which proliferate faster than non-reprogrammed cells. Rate of cell death was estimated by determining the number of Annexin V positive, DAPI negative cells; this is the fraction of cells that are in the early phase of cell death. At day 3 of the experiment, no significant difference could be noted in the percentage of Annexin V positive cells, between *Nudt21* knockdown- or control cells (Figure 2B).

Next, the effect of *Nudt21* knockdown on expression of reprogramming- or pluripotent stem cell related genes was assessed by RNA sequencing of messenger RNAs (mRNAs). At day 6 of reprogramming, transcript levels of epithelial markers *Epcam* and *Cdh1* were increased in *Nudt21* knockdown cells, as were of stem cell transcription factors *Nanog* and *Sall4*, and reprogramming associated epigenetic regulators *Dnmt3b* and *Tet1* (Figure 2C). A scatter plot depicts all genes that were differentially expressed between *Nudt21* knockdown and control cells (Figure 2D). Outliers include OCT4 encoding *Pou5f1*, and many factors that were not previously linked to reprogramming or pluripotency such as *Ly6g6c, Txnip* (involved in reduction of oxidative stress), *Fam25c* and *Morf4l1* (promotes cell proliferation of neural stem cells).

Some genes can be used to mark stages of reprogramming by presence of cell-surface proteins encoded by these genes, or by fluorescent reporter expression from the endogenous locus[6]. Intermediates of reprogramming can be defined by presence of protein markers SSEA1, EPCAM and of the reporter *Oct4*-GFP[5]. Reprogramming cells expressing the *Nudt21* shRNA or control were analyzed for expression of these markers at several timepoints (Figure 2E-F). At day 3 and 5, clear increases in cells expressing SSEA1 or EPCAM

(continued from right) X-axis: noting when, at the start or finish of screening round 1-3, representation was calculated from shRNA library sequencing results. Y-axis: abundance of shRNAs compared to total number of shRNAs identified from sequencing data.
(C) Reprogramming efficiency using alternative shRNAs targeting *Nudt21*; number of AP positive colonies counted after 10 days of OKSM expression.
(D) Efficiency of generating *Oct4*-GFP positive reprogramming intermediates at day 8 of OKSM expression (n=3, error bar=standard deviation) upon Nudt21 knockdown or control.
(E) FACS plot showing one replicate of data in (D). tRFP: reporter of shRNA vector expression.
(F) Overview of experiment in (G): timed withdrawal of OKSM expression through removal of doxycycline from the culture media.
(G) Reprogramming efficiency upon 3 to 10 days of OKSM expression, during *Nudt21* shRNA expression or control; n=3, error bar= standard deviation. Student's t-test: * p-value <0.10, ** p-value <0.05.
(H) Western blot stained for *Nudt21* encoded protein CPSF5 and loading control ACTIN.
(I) Reprogramming efficiency of cells transfected once or twice with siRNA targeting Nudt21 or control; AP staining of iPSC colonies (red dots).
(J) Average of three replicates of experiment shown in (I); error bar= standard deviation.
(K) Teratoma derived from *Nudt21* knockdown iPSCs; all pictures from same tumor; showing examples of differentiated structures from all three germ layers.

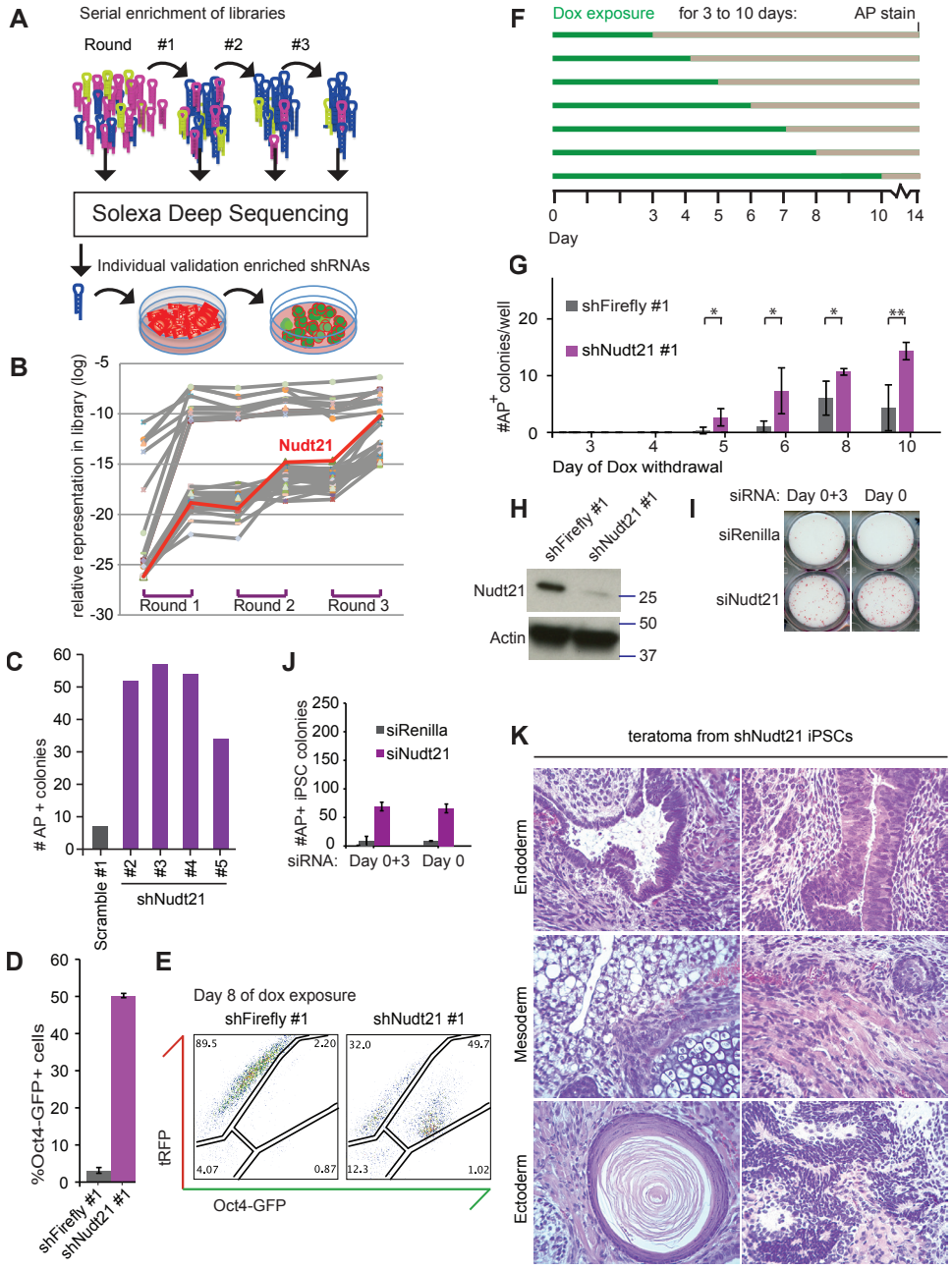


Figure 1. *Nudt21* knockdown facilitates somatic cell to iPSC reprogramming.(A) Overview of shRNA screen pilot identifying *Nudt21* shRNA as a hit. (B) Representation of *Nudt21* shRNA in libraries before and after 3 rounds of reprogramming experiments. Red line: *Nudt21* shRNA. Grey lines: other top scoring shRNAs. (continued on left)

were detected in *Nudt21* knockdown cells. Also, at day 6 a substantial group of *Oct4*-GFP positive cells arose that was absent in the control. Thus, reprogramming is enhanced within days of *Nudt21* knockdown and OKSM expression, giving *Nudt21* knockdown cells an almost immediate advantage in forming iPSCs.

Fast and efficient and c-Myc-free reprogramming with Nudt21 knockdown

Above experiments suggested that *Nudt21* knockdown affects reprogramming nearly immediately, pushing cells to express stem cell markers within days of OKSM expression. To optimize the reprogramming process to require minimal duration of OKSM expression and little handling, we tested other culture conditions recently found to enhance reprogramming efficiency6-9. GSK3 inhibitor (GSK3i), ascorbic acid (AA), DOT1l inhibitor (D1li), and the GSK3i/AA combination were tested in comparison and addition to *Nudt21* knockdown. Reprogramming duration (OKSM expression) was reduced to 8 days because these conditions yield enough iPSC to overgrow the culture plate when exposed to OKSM longer. *Nudt21* knockdown resulted in nearly as many iPSC colonies as the best culture condition (GSK3i/AA; 'AGi') (Figure 3A). When combined, *Nudt21* knockdown had a strong additional effect on reprogramming efficiency, on top of the effect of the various culture conditions (Figure 3B). Next, the best culture conditions were combined with *Nudt21* knockdown via siRNA transfection. Also, reprogrammable MEFs were used at earliest possible passage post-derivation to further raise efficiency. In this experiment, withdrawal of OKSM expression after 3 days resulted in iPSC colonies in both control siRNA and even more in *Nudt21* siRNA conditions (Figure 3C). Even after just 2 days of OKSM (48hr Dox), rare iPSC colonies emerged in *Nudt21* siRNA transfected cells, but not in the control (not shown; all colonies were picked for further analysis). These iPSCs were pluripotent as assessed by *Oct4*-GFP expression and teratoma formation (Figure 3D-E).

(continued from right) (C) Expression of epithelial- and stem cell markers (*Epcam, Cdh1, Sall4*) and epigenetic regulators (Dnmt3b, Tet1) in MEF, iPSC, and reprogramming cells with or without *Nudt21* knockdown, as determined by RNA sequencing. Error bars indicate standard deviation; n=3.
(D) Scatterplot of same dataset as in (C). Triangle shaped dots represent genes that were expressed at twofold lower (blue) or twofold higher (grey) levels in *Nudt21* knockdown reprogramming cells (day 6), when compared to *Firefly* knockdown control; average of three replicates. (E) Key results from flow cytometry analyses of reprogramming intermediates with *Nudt21* shRNA (purple) or control. X-axis designates what surface marker was analyzed at which day of OKSM expression (day 5 or day 8). Error bars indicate standard deviation; n=3. (F) Flow cytometry plots showing one replicate of experiment in (E). tRFP marks expression of the shRNA-carrying vector. Logarithmic scales; numbers in plots mark % of DAPI negative (live) single cells).

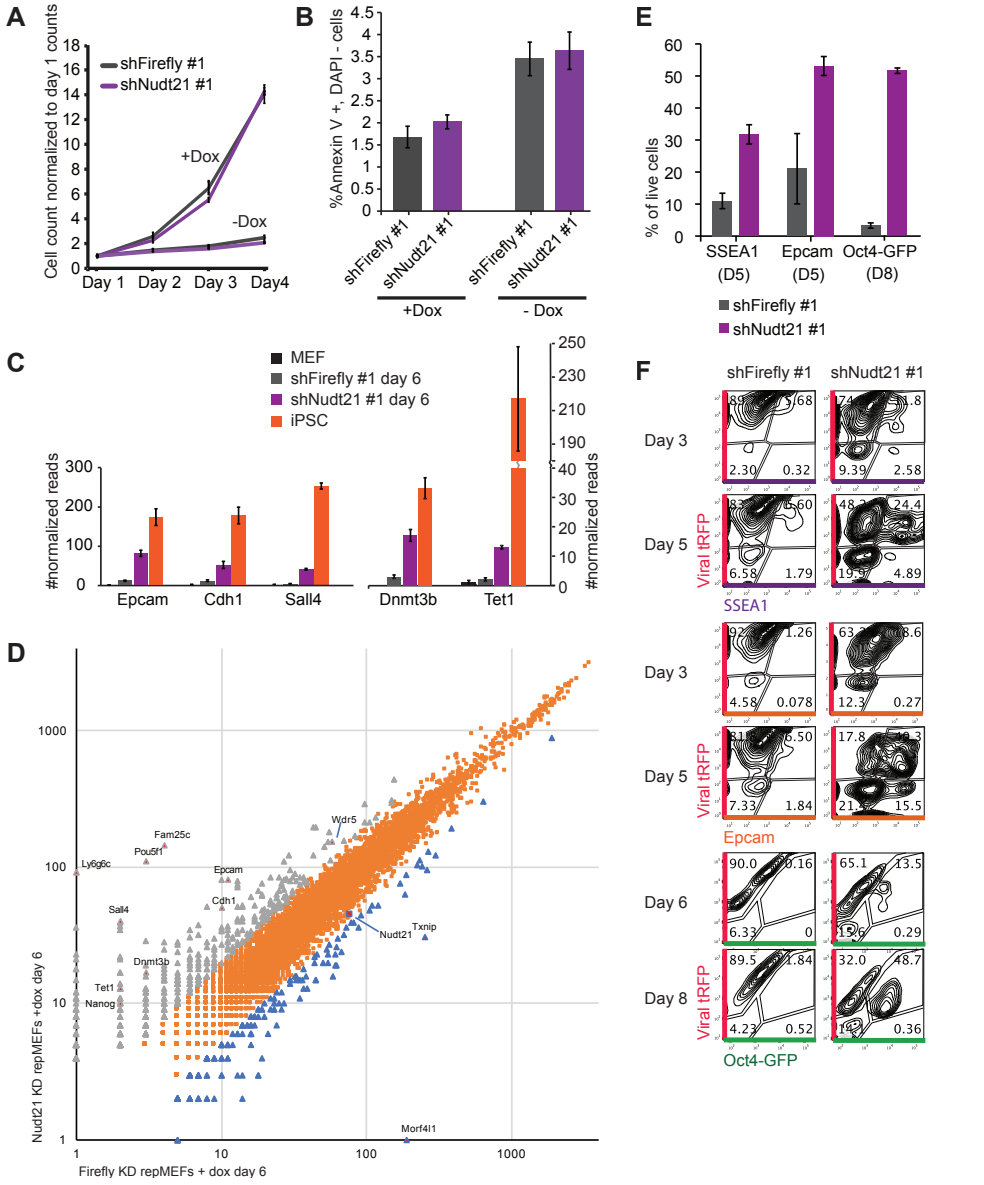


Figure 2. *Nudt21* knockdown does not affect cell proliferation or apoptosis, does affect stem cell/ reprogramming related genes. (A) Number of cells counted during 4 days, with *Nudt21* shRNA expression or control, and with (top lines) or without (bottom lines) dox induction of OKSM expression. Error bars indicate standard deviation; n=3. (B) Same cells as in (A), assessed for Annexin V expression; n=3, error bars: standard deviation. Y-axis: percentage of DAPI negative cells that were Annexin V positive. (continued on left)

Considering that in the future, iPSC may be derived with the goal of forming transplantable tissues for patients, avoiding the oncogene *c-Myc* in the reprogramming strategy has a preference. We tested whether *Nudt21* suppression could facilitate reprogramming without *c-Myc*. We used a mouse model identical to the OKSM reprogrammable mouse, but with a mCHERRY fluorescent reporter instead of the *c-Myc* cDNA. These OKS-mCherry reprogrammable cells require AA or AGi conditions to form iPSC efficiently in control conditions. However, *Nudt21* siRNA transfection resulted in some iPSC in dox only culture conditions, and increased reprogramming efficiency in AA or AGi conditions (Figure 3F). Dox-only reprogramming with *Nudt21* siRNA resulted in robust reprogramming efficiencies after 15 or 21 days of OKS expression (Figure 3G).

Nudt21¬knockdown affects the alternative polyadenylation (APA) profile of reprogramming-associated mRNAs

Upon choice of a polyadenylation site (PAS) and assembly of the required factors, the 3'UTR of an mRNA is cleaved and subsequently polyadenylated. In mRNAs with multiple PAS, choice of a more distal or proximal to the stop codon can strongly affect 3'UTR length and preservation or removal of binding sites for regulators such as RNA binding proteins (RBPs) and microRNAs (miRNAs) (Figure 4A). Sequencing of mRNA directed at PAS (PAS-Seq) reveals what PAS are used at what ratio10. Different biological conditions are associated with a shift to more distal or more proximal PAS usage, however it is important to note that a shift is not an on/off switch between PAS but rather a changing ratio. Comparing many mRNAs in different biological contexts to detect global shift to more proximal or more distal PAS is possible and has been done in the context of cancer versus healthy tissue and stem cells versus differentiated cells[11,12], showing that stem cells and cancer cells tend to have more proximal PAS usage.

PAS localization is unique to each mRNA, thus 'proximal' or 'distal' cannot be linked to an absolute distance from the stop codon. Nor can each PAS usage be compared; to compare two samples, only PAS sites used to some degree in both samples can be compared by their proximal/distal ratios. This prevents systematic comparison of more than two samples at once. One PAS site can thus represent proximal PAS usage in one comparison, or distal PAS usage in the next (Figure 4A). When comparing two samples, PAS sites used in both samples are identified and marked as distal or proximal sites in that particular context. A ratio of proximally polyadenylated reads divided by distally polyadenylated reads can be calculated to put a number to the use of proximal or distal PAS. This proximal/distal ratio can then be compared between the two samples.

When we compared PAS usage ratios between reprogramming cells (day 3 of

induction with OKSM) upon *Nudt21* knockdown or -control, many mRNAs had significantly altered their proximal/distal ratio, and most of those had shifted to more proximal PAS usage (Figure 4B). This effect was dependent on the presence of the reprogramming factors or ongoing reprogramming; non-induced *Nudt21* knockdown or control cells showed much less difference in APA profile. When we compared these early reprogramming cells to ESCs, we noted that *Nudt21* knockdown cells have less significantly different proximal/distal ratios than control cells (Figure 4C).

The consequences of these APA events are challenging to define purely based on

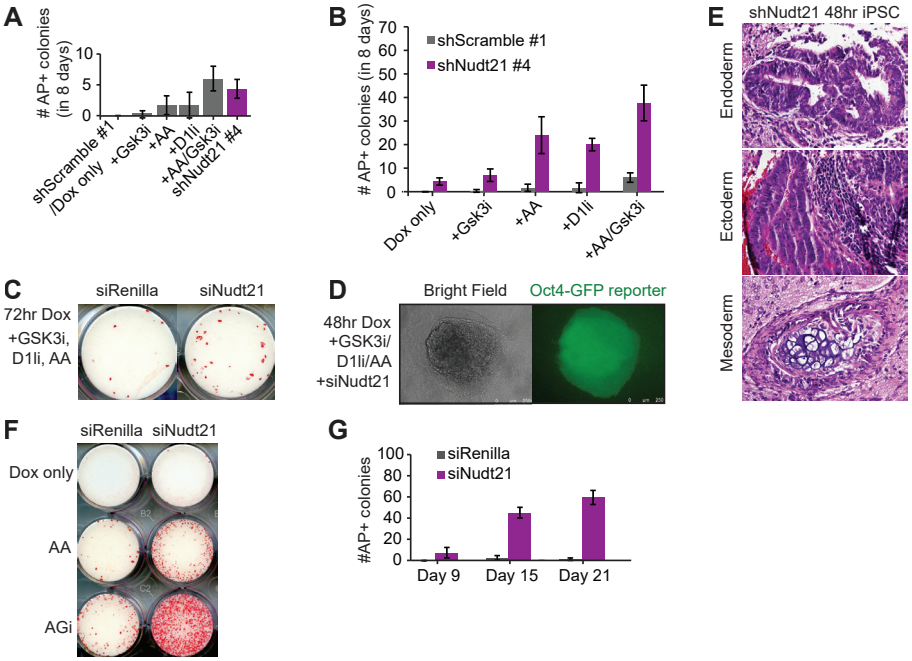


Figure 3. *Nudt21* knockdown phenotype is independent from known reprogramming pathways and c-Myc expression; enables ultra-fast reprogramming. (A) Reprogramming efficiency determined by number of AP positive colonies after 8 days of OKSM; under culture conditions with known reprogramming enhancing small molecules, or with *Nudt21* knockdown. Error bars indicate standard deviation; n=3. (B) Same experiment as in (A), this time combining *Nudt21* knockdown or control with culture conditions. (C) Reprogramming efficiency when combining all small molecules in (A, B) and *Nudt21* knockdown by an siRNA; after 72 hours of OKSM expression. (D) iPSC derived in same condition as in (C) but just 48hrs of OKSM expression. (E) Teratoma derived from iPSC in (D), containing differentiated structures from all three germ layers. (F) AP staining of iPSC colonies to determine reprogramming efficiency of OKS-mCherry MEFs (lacking *c-Myc*), in combination with AA, AGi conditions and/or *Nudt21* siRNA. (G) Quantification of 3 replicates of experiment shown in (F) after 9, 15 or 21 days of OKS expression; error bars indicate standard deviation.

transcriptional data. Thus, no generalizing statement of how *Nudt21* knockdown affects reprogramming through these APA events can be made. However, by looking at individual mRNAs, candidates for further study can be identified.

One possible effect of a proximal shift is higher stability of the transcript due to loss of miRNA binding sites or binding of RBPs that facilitate mRNA degradation. The PAS-Seq data also allowed for quantification of overall abundance of mRNAs. To find mRNAs that could be responsible for the phenotype of *Nudt21* knockdown, we selected mRNAs with a proximal shift in PAS usage and a higher abundance. Of those genes, nearly all are normally upregulated during reprogramming as determined by Polo et al. A few examples are shown in Figure 4D. Amongst them, *Wdr5* is a known regulator of pluripotency and thus a candidate effector of the *Nudt21* phenotype. Preliminary studies on the effect of *Wdr5* upregulation on reprogramming showed a modest increase in protein levels of WDR5 and also modest positive effects on reprogramming efficiency. Conclusively, APA of many transcripts together is likely required to recapitulate the *Nudt21* knockdown phenotype during reprogramming.

Discussion

Here we describe how we identified *Nudt21* as a roadblock to reprogramming. An shRNA targeting *Nudt21* was consistently enriched in a pool of shRNA that was transduced into reprogrammable fibroblasts, retrieved and re-transduced for consecutive reprogramming experiments. We first confirmed that reducing *Nudt21* expression was responsible for the reprogramming phenotype by using different methods of knockdown (alternative shRNAs and siRNAs) and confirming loss of *Nudt21* protein upon knockdown. *Nudt21* knockdown was found to decrease the duration of reprogramming factor expression (of *Oct4, Klf4, Sox2* and *c-Myc*; 'OKSM') needed to form pluripotency factor *Oct4*-expressing cells and later, OKSM- independent, stable iPSC colonies. *Nudt21*-knockdown iPSCs were confirmed to be pluripotent by deriving teratomas. *Nudt21* knockdown reprogramming cells start expressing stem cell- and pluripotency factors earlier as determined by looking at cell surface markers and RNA expression. The *Nudt21* knockdown phenotype could not be explained by alteration of the rate of proliferation or apoptosis. To determine synergy with inhibitors of specific cellular signaling pathways that are known to boost reprogramming, *Nudt21* knockdown was combined with these inhibitor containing culture conditions. *Nudt21* knockdown boosted reprogramming further in each of the conditions, suggesting the mechanism is independent of these pathways.

Reprogramming efficiency, speed, and dependence on oncogene *c-Myc* are limiting factors in the search to derive iPSC in experimental and therapeutic settings. Thus we took advantage of our finding by combining it with known reprogramming boosting culture

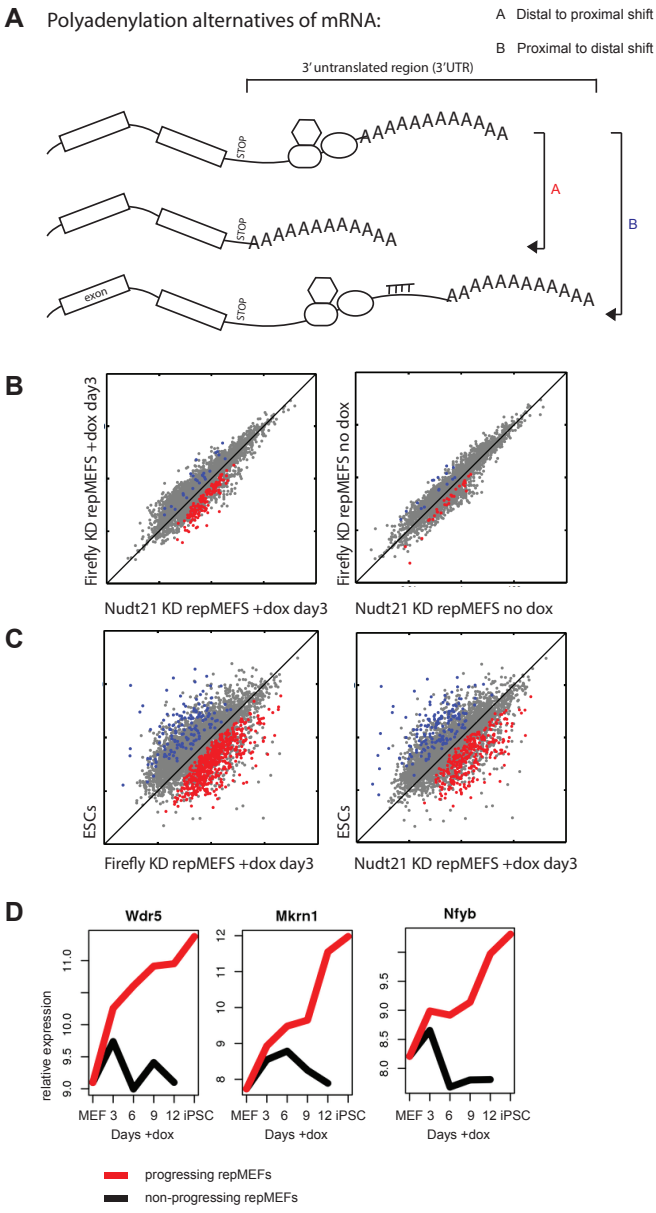


Figure 4. Polyadenylation site sequencing (PAS-Seq) analysis of reprogramming fibroblasts upon *Nudt21* knockdown.

(A) Schematic presentation of possible outcomes of pairwise comparison of samples following PAS-Seq: Distal to proximal shift or proximal to distal shift. Absolute distance from the stop codon of distal and proximal annotated polyadenylation sites (PAS) depends on the PAS used in the samples that are compared. Depending on the mRNA, shifts between proximal and distal PAS can affect binding of miRNAs and RNA binding proteins, which in turn can alter mRNA stability and translation.

(B) Ratio of proximal PAS/ distal PAS usage in reprogramming cells upon *Nudt21* KD (x axis) or control *Firefly* KD (y-axis). Each dot represents one mRNA; significantly different ratios are highlighted in blue (on average shorter 3'UTR in control cells) and red (on average shorter 3'UTR in *Nudt21* KD cells). Left graph: upon 3 days of dox induction of OKSM expression; right graph: no dox induction.

(C) Graphs as in (B), comparing ESCs (y-axis) to reprogramming cells upon Firefly KD (left graph) and *Nudt21* KD (right graph). (D) Expression of genes during reprogramming according to microarray analysis (Polo et all Cell 2010). Corresponding mRNAs experience a distal to proximal shift (thus acquire an on average shorter 3'UTR) in *Nudt21* KD reprogramming cells, compared to control reprogramming cells. Red line: expression in reprogramming intermediates; black line: expression in reprogramming-refractory cells.

conditions with small molecules[6,13,14]. Combination of ascorbic acid and GSK3 inhibitor ('AGi') with *Nudt21* knockdown enabled efficient reprogramming of reprogrammable MEFs lacking overexpression of *c-Myc*. Combination of AGi, a DOT1l inhibitor and *Nudt21* knockdown resulted in iPSC after just 48hrs of OKSM expression.

Combined, our results characterize the role of *Nudt21* in preventing reprogramming to pluripotency of murine somatic cells. It expands our toolbox of measures to optimize reprogramming efficiency and speed. It supports further research into posttranscriptional regulation and its role in cellular reprogramming. We are currently further exploring the mechanism behind the *Nudt21* knockdown phenotype. Our leading hypothesis is that alternative polyadenylation occurs and helps cells to become iPSCs, due to the known role of the *Nudt21* encoded protein CPSF5.

Alternative polyadenylation affects the processing of 3`UTRs of messenger RNAs (mRNAs)[15]. The choice between polyadenylation sites (PASs) are made by specific factors in the cleavage and polyadenylation complex, among which CPSF5, encoded by *Nudt21*. This affects availability of miRNA and RBP binding sites, influencing the rate of translation into protein. Global APA patterns have been found to be altered in cancer cells and pluripotent stem cells, compared to healthy, somatic cells[11,12]. Since PAS location, number and presence of regulatory sequences in relation to PAS is specific to each mRNA, it is challenging to translate global patterns to gene-by-gene level mechanisms. Also, since we show that *Nudt21* knockdown likely affects many transcripts during reprogramming (some effects may be secondary to the reprogramming phenotype), single target genes may not recapitulate the phenotype by themselves, complicating validation of possible mechanisms. To identify possible effectors, we compared APA affected genes to genes whose mRNA level was modified upon *Nudt21* suppression. We found that most mRNAs that acquire shorter 3'UTRs and gain in abundance, also are upregulated during reprogramming according to existing expression data. Further analysis of these individual genes, including known stem cell regulator Wdr5, is warranted. Expansion of the PAS-Seq dataset, using the *Nudt21* siRNAs for a stronger phenotype and analyzing more timepoints is ongoing. This effort should allow better understanding of how global APA is affected by *Nudt21* suppression and in turn affects the acquisition of a pluripotent state.

Materials and Methods

Tissue culture, virus production, siRNA transfections, flow cytometry, quantification of reprogramming efficiencies, teratomas formation and Western blot analysis were performed simultaneously with experiments described in chapter 4[16].

shRNA screen and identification of hits

The shRNA screen was performed as described in chapter 4[16], for a total of three screening rounds. Solexa sequencing data were retrieved as described. Identification of hits was performed using algorithms in Excel. Different algorithms were used based on enrichment of shRNAs in samples from GFP+ iPSCs each round, lack of enrichment of negative controls (GFP- samples and non-induced samples), consistency of enrichment during each round and total enrichment. Top scoring hits from different algorithms were pooled to create a list of candidates for further individual validation. *Nudt21* was identified as a hit based on enhanced reprogramming efficiency of *Nudt21* shRNA expressing reprogrammable MEFs.

PAS-Seq analysis

For PAS-Seq analysis, total RNA was extracted using the Qiagen RNeasy Mini Kit and sequencing libraries were prepared, processed and analyzed as described[10]. In short, extracted RNA was fragmentated, reverse transcribed to cDNA using a poly-A tail selective oligo-dT primer. A custom oligo-dT sequencing primer was used for single end sequencing. Reads were mapped to the genome after removal of T's. Reads in near vicinity of each other were clustered to assign a poly(A) site (PAS). Samples were compared in pairs for relative abundance of proximal versus distal PAS usage. mRNA abundance was determined from all reads.

References

1. Apostolou, E. & Hochedlinger, K. Chromatin dynamics during cellular reprogramming. Nature 502, 462–471 (2013).
2. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell 126, 663–676 (2006).
3. Stadtfeld, M., Maherali, N., Borkent, M. & Hochedlinger, K. A reprogrammable mouse strain from gene-targeted embryonic stem cells. Nat Meth 7, 53–55 (2009).
4. Hanna, J. et al. Direct cell reprogramming is a stochastic process amenable to acceleration. Nature 462, 595–601 (2009).

5. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining Molecular Cornerstones during Fibroblast to iPS Cell Reprogramming in Mouse. Cell Stem Cell 2, 230–240 (2008).
6. Bar-Nur, O. et al. Small molecules facilitate rapid and synchronous iPSC generation. Nat Meth 11, 1170–1176 (2014).
7. Onder, T. T. et al. Chromatin-modifying enzymes as modulators of reprogramming. Nature 483, 598–602 (2012).
8. Stadtfeld, M. et al. Ascorbic acid prevents loss of Dlk1-Dio3 imprinting and facilitates generation of all–iPS cell mice from terminally differentiated B cells. Nature Genetics 44, 398–405 (2012).
9. Esteban, M. A. et al. Vitamin C Enhances the Generation of Mouse and Human Induced Pluripotent Stem Cells. Cell Stem Cell 6, 71–79 (2010).
10. Shepard, P. J. et al. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 17, 761–772 (2011).
11. Mayr, C. & Bartel, D. P. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. Cell 138, 673–684 (2009).
12. Ji, Z. & Tian, B. Reprogramming of 3' Untranslated Regions of mRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types. PLoS ONE 4, e8419–13 (2009).
13. Esteban, M. A. et al. Vitamin C Enhances the Generation of Mouse and Human Induced Pluripotent Stem Cells. Cell Stem Cell 6, 71–79 (2010).
14. Onder, T. T. et al. Chromatin-modifying enzymes as modulators of reprogramming. Nature 483, 598–602 (2012).
15. Shi, Y. & Manley, J. L. The end of the message: multiple protein–RNA interactions define the mRNA polyadenylation site. Genes Dev. 29, 889–897 (2015).
16. Borkent, M. et al. A Serial shRNA Screen for Roadblocks to Reprogramming Identifies the Protein Modifier *Sumo2*. Stem Cell Reports 6, 704–716 (2016).


a4

Addendum chapters 4 and 5

Considerations when performing a serial, genome-wide RNA inhibition screen

Marti Borkent

Introduction

The main project in this thesis is a genome-wide screen for inhibitors of somatic cell reprogramming. With no previous experience with screening at this scale in the lab, we were grateful to collaborate with the screening library's creator Stephen Elledge, and mostly with his former postdoc Guang Hu. Together we optimized the screen for our platform of iPSC reprogramming. Since the resulting papers do not discuss more than our eventual experimental design and limited materials and methods, we will use this thesis chapter to highlight important lessons learned along the way.

One part will discuss important concepts considered in experimental design, such as the required coverage of shRNAs and how to achieve this at each step and why we looked at shRNAs that help reprogramming, not those that inhibit reprogramming.

The second part focuses on technicalities that did not make the materials and methods section of the published paper, but are helpful regardless for anyone considering a screen or other experiment with reprogramming cells at this scale.

But first, we will discuss the rationale behind using RNAi technology, on a genome-wide level, specifically with a pooled library.

Specific gene targeting with shRNAs

Long before the rise of DNA sequencing, or even our knowledge of DNA structure, genes were characterized by their loss through mating, recombination or mutation. Hence the naming of genes after properties opposite to their function (the 'White' gene in fruit fly is required for a dark eye color). RNA interference is a sequence-specific, conserved and highly customizable gene silencing tool that has modernized and hugely expanded loss-of-function genetic experiments.

RNA interference was first discovered in plants and worms. It requires a double strand RNA (dsRNA) that is broken down to 'small interfering RNAs' (siRNAs) that perfectly complement their target mRNA. Concurrent with the discovery of siRNA, several groups found endogenous and highly conserved microRNA (miRNA) which are processed by largely the same machinery as siRNAs, but target multiple genes with near perfect complementary sequencies. Most notably, miRNAs are derived from stem-loop structure forming RNA called 'short hairpin RNA (shRNA)' and are important endogenous regulators in mammalian cells. shRNAs were soon developed into a stably expressed, inducible gene silencing tool for mammalian cells (Figure 1).

Genome-wide versus gene-of-interest screening

Many genes involved in cell fate, stem cell- and cancer biology have been suspected and indeed shown to affect iPSC reprogramming. Therefore we preferred an unbiased approach to candidate genes. Where a small shRNA screen might include 100 candidates that need to be selected based on previous work, a genome-wide screen covers all transcripts annotated at that time including genes only known in a different context or not functionally described at all. Though probably no genome-wide screening library can truly test 100% of its candidates, and new annotations constantly alter what we consider 'all genes', a genome-wide approach gives a much broader reach and leads into unknown territories.

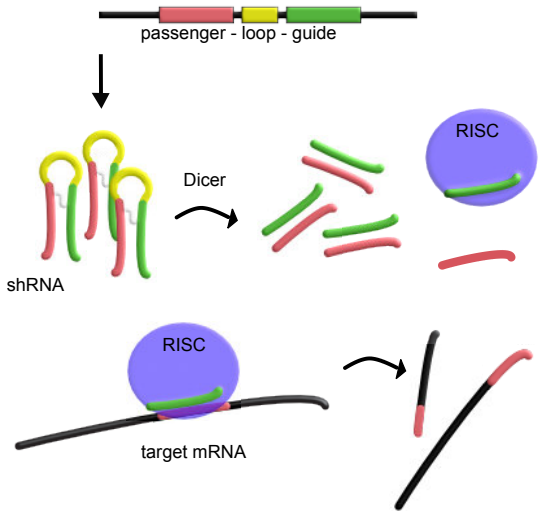


Figure 1. Schematic depiction of shRNA mediated RNAi. Top strand: transduced DNA encoding shRNA. Dicer digests shRNA to double strand RNA (dsRNA). RISC: RNA-induced Silencing complex, which finds the target mRNA and induces cleavage at the recognition sequence.

Pooled screening

A pooled library has all shRNAs for all genes mixed together for application to the target cells, as opposed to applying each shRNA to a different sample (e.g., well). Pooled libraries significantly ease access to genome-wide screening. Well-by-well testing requires cell culture automation; robots that can securely change media of 100,000+ wells. Moreover, the small wells complicate a quantitative assessment, such as of reprogramming efficiency, and are more amenable to qualitative questions such as whether or not cells remain viable in a given well. Lastly, there is possible well-to-well technical variation since wells are handled individually and have different locations on the plate. Pooled screening bypasses these requirements. And because the shRNAs are embedded in the DNA of their host, they can retrospectively be identified through a unique barcode or by sequencing of the short (19-21bp) unique miRNA sequence. The expression vector includes a turbo-RFP cDNA allowing detection of successfully infected cells.

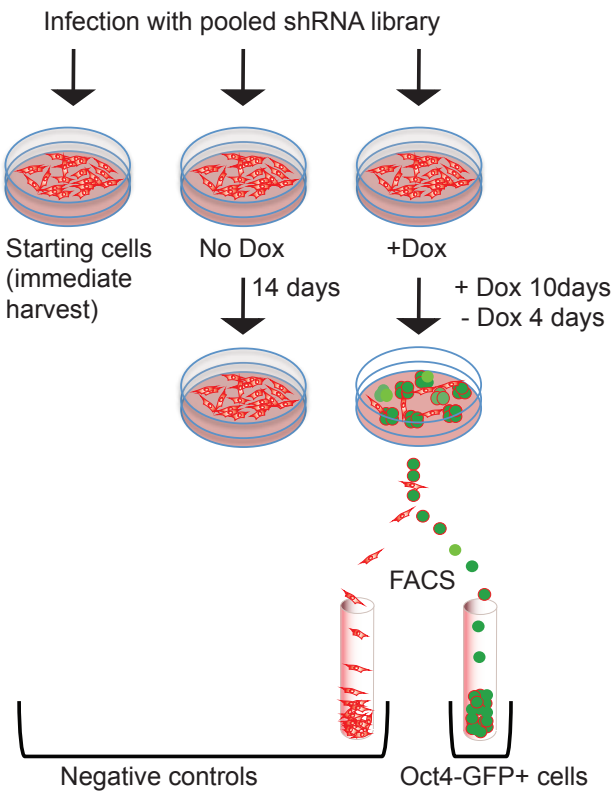


Figure 2. Experimental design of screening experiment. After infection of cells with the shRNA library, 4 samples are generated including 3 negative controls. One negative control is taken immediately to represent shRNA delivery to the cells. One negative control sample is cultured for the length of the experiment, but not exposed to doxycycline to reprogram. The largest sample is induced to reprogram with doxycycline for 10 days and withdrawn from doxycycline for 4 days, before separation into *Oct4*-GFP positive iPSCs, and *Oct4*-GFP negative unreprogrammed cells. FACS: Fluorescence activated cell sorting.

The Hannon-Elledge libraries

In a combined effort by the Greg Hannon and Stephen Elledge labs, genome-covering libraries of mRNA targeting shRNAs were developed, optimizing gene knockdown efficiency, viral delivery and shRNA expression[1]. Their libraries have enabled studies answering a range of different questions in human or murine cells. Examples are the identification of genes essential to the survival of mutant Ras cancer cells, cell migration and self-renewal[2-4].

Concepts

Coverage of the shRNA library.

A key requirement of a pooled screen is to allow each shRNA to exert its biological effect. In our case, this means that each shRNA should enter a cell that just needs knockdown of that particular gene to reprogram into a stable iPSC. To make sure this chance is high enough, each shRNA in a pooled screen is generally given cells to infect by a factor that is referred to

as 'coverage'. A coverage of 10 would mean that 10 cells are infected by the pooled library, per shRNA; so a 63,000 shRNA library would require 630,000 cells. If the coverage is too high, the experiment is simply bigger and using more resources then needed. However, a low coverage reduces the sensitivity of the entire screen; or the likelihood of an shRNA with the desired phenotype to be found. In that case not all shRNA can exert their biological effect on enough cells to be identified as hits. In reality, a large scale screen will never have 100% sensitivity and be completely reproducible. Ensuring sufficient coverage at each experimental step, not just the step of cell culture, will increase the chances of finding hits. These are considerations per step:

shRNA library production.

This library was created by printing oligo's on microarrays, cleaving them off, and mass-cloning them into viral vectors. Each of these steps may favor some shRNA sequencies over others (e.g. by different GC content affecting amplification). To account for this bias, the Elledge lab recommended a 10-fold coverage of each of the 63K shRNAs. Secondary library production starts with PCR amplification of shRNA sequencies from gDNA, which warrants the same coverage. During cloning of itself, an even higher coverage of 100-1000x per plasmid is recommended to account for possible bias during construct ligation, transformation of bacterial cells, and plasmid harvest.

Viral delivery.

Even after optimizing infection efficiency of the reprogrammable cells (see below), sometimes 50 million cells at a time, we would get infection efficiencies as low as 50% up to 90%. Thus we had to assume half the cells did not contribute to the screen and double coverage.

Reprogramming efficiency.

This is an issue very specific to reprogramming experiments. A screen for shRNAs that influence cell death of cancer cells, or differentiation of ES cells, can assume that each cell is capable of undergoing the process of interest such as apoptosis. However, the baseline reprogramming efficiency is very low in the conditions of this screen. We decided to assume that at most 5% of cells would be amenable to be pushed over the threshold towards pluripotency by an shRNA, and thus increased the coverage 20-fold.

Large scale reprogramming

Reprogramming experiments typically require a couple of cell culture plates with small wells to reprogram around 20 thousand cells per replicate and condition. In this case we were looking to reprogram one batch of shRNA library expressing cells, up to 70 million at a time. In chapter 2 we describe the reprogrammable mouse model which enabled us to generate large quantities of reprogrammable mouse embryonic fibroblasts (repMEFs) at once and reprogram these cells into iPSCs at a robust baseline efficiency.

At any step in the experiment, thorough calculation of scale is essential as usage of reagents, culture plates, incubators can easily get beyond feasible and requires more planning than usual. Another important consideration is representative sampling: the scale required for good coverage doesn't mean that all material harvested is used up for the following steps. Representative sampling is inevitable. Think about the soonest step you can fairly do sampling without compromising coverage, to save on handling of material you will not use down the line. In our case, we realized that the input for library sequencing is smaller than the absolute amount of DNA retrieved from the cells. Instead of sampling after cell purification, gDNA preparation from cells and generation of the sequencing library, we already sampled after cell harvest and thorough mixing of all cells.

Negative controls

As mentioned before, the readout of the screen is a quantification of each shRNA by deep sequencing, prior and after the experiment. A shift in abundance of an shRNA during the experiment points to a possible biological effect. Since we were interested in barriers to reprogramming, we looked for shRNAs that become more abundant; that enrich themselves in the library during reprogramming. These shRNAs must aide the repMEFs into forming iPSC; a more proliferative population that we then purified from the reprogramming culture. Some important controls for enrichment biases are (Figure 2):

- Non-reprogramming sample; a parallel, no-dox culture of repMEFs to control for enrichment independent from reprogramming, such as through stimulation of proliferation or inhibition of the normal cell death rate.
- Leftover cells from iPSC purification; cells that were exposed to doxycycline, possibly formed reprogramming intermediates or even iPSC, but no stable, reporter-expressing iPSC after dox withdrawal. This controls for impact on cell growth in the context of reprogramming, and negatively selects for aiding reprogramming without allowing establishment of pluripotency.
- Sample at the start of the experiment; to see at what relative abundance each shRNA reached the target cells for a fair comparison with the end library.

Serial enrichment strategy

In our model of heterogeneously reprogramming cells, we still needed be more stringent in our selection of candidates. Individual repMEFs have individual speeds and efficiencies of reprogramming, for reasons not completely understood. So in our case, we risked losing or selecting shRNAs depending on which cell they happened to reach. In particular, we needed to weed out 'passenger shRNAs' that happened to reach a cell that was efficiently reprogramming regardless. To this end we added consecutive screens to our design, every time starting with a shRNA library derived from the prior experiment (Figure 3). This way, an ever-smaller screening library was created, increasingly containing true pro-reprogramming shRNAs. Moreover, we could select for shRNAs that consistently increase their abundance each round. Luckily, each round required less cells for adequate coverage, because of the lower number of shRNAs present in the library.

Analysis: alternative uses of the data

An interesting question often asked, was whether this setup could be used not only for shRNAs that get enriched in reprogramming cells, but also to find biologically relevant dropouts; shRNAs that target genes that are beneficial to reprogramming. Initially we were very skeptic of this possibility because of the likelihood of shRNAs dropping out for merely technical reasons. Once an shRNA drops out, you don't know if your coverage was lacking, the shRNA was toxic or anti-proliferative, or primarily inhibitory to reprogramming. Plus, the serial screening approach does not apply.

However, we believe it is not entirely impossible to do. The negative controls are crucial for such a question as they can show how well the shRNA was represented in the fibroblasts. A dropout specific to only the iPSC fraction is interesting. When we selected for shRNAs that are well covered in the library and enrich in non-reprogramming cells but drop out of the reprogrammed cells, we did find an shRNA for *Brca*1, which was previously shown to be important for reprogramming, suggesting that this approach could be fruitful (Figure 4).

Another additional way the data can be mined, is to look at shRNAs that do influence cellular proliferation. These were filtered out of our analysis because rate of proliferation is a major and well-known bias to reprogramming efficiency, which we were not interested in. But our data should enable a look at shRNAs that enrich specifically in the non-reprogramming cells.

Validation (-rate)

After retrieving shRNA sequencing data, the identification of hits has only just started. An average screen of this kind will produce hits that validate upon individual scrutiny, at a rate of around 50%. We however were happy to see 10-20% of hits show any effect on reprogramming, let alone a very strong one. This influences a final important consideration: amount of hits to be further tested.

A low validation rate can suggest that testing many hits will lead to more real hits. However, depending on your assay, cloning and testing hits individually, with the required replicates, is cumbersome. We strived to test 20 to 30 hits only.

Knowing which genes to include in this list is impossible if the screen was not performed before, in the same model system. We tried different algorithms along the way, playing with the weight of different factors such as enrichment compared to one or all controls, fold of enrichment, multiple shRNAs targeting the same gene. It is key to set cut-off limits, or criteria for a hit such, that a workable number of shRNAs (whether it be 10 or 50) is selected for validation.

Technicalities

shRNA lentiviral vector

The vector in our library has a lentiviral backbone with a miRNA30 cassette with the shRNA coding sequence, as well as a tRFP reporter under individual, ubiquitous promotors. The Hannon and Elledge labs determined that not the half-processed pre-miRNA strand, but rather the unprocessed primary miRNA strand was most effective at producing miRNAs. They modeled their strands after the well-characterized *mir30* miRNA. The cassette is flanked with several PCR primer targets and restriction sites for easy sequencing and cloning.

Infection with library virus

Transduction of the virus into MEF introduces some toxicity and can be of varying efficiency. Therefor we infected cells right before passaging the MEF onto plates for reprogramming. During passaging, the number of surviving cells could be counted and plated at desired density, infection efficiency could be measured by viral reporter expression, and the infection could be done in just a couple of safe, enclosed cell culture flasks as opposed to the many reprogramming plates. We infected at a density of around 50% to allow cell proliferation during infection. The virus was produced as fresh, filtered supernatant from 293T HEK cells, to which we added fresh MEF media 1:1 along with 1:2000 Polybrene. Since the Polybrene is another potentially toxic compound to the cells, we strived to wash off the virus between

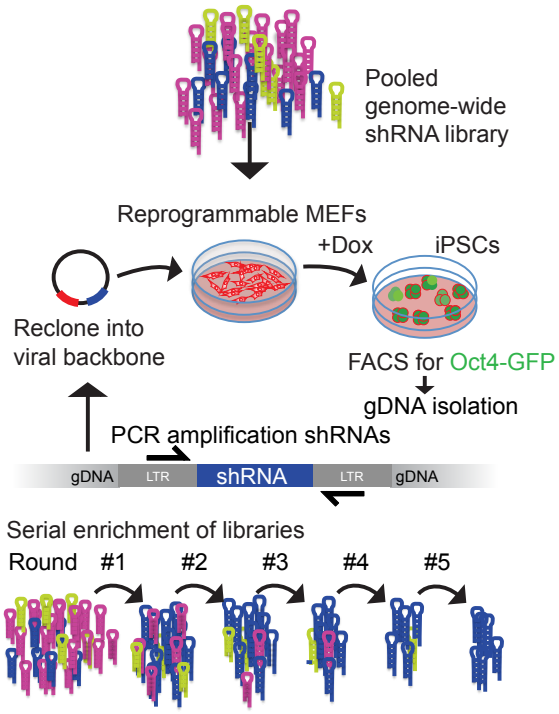


Figure 3. Experimental design of serial enrichment screening. Top panel: the experiment depicted in figure 2 formed one round of multiple rounds of serial enrichment screening. The *Oct4*-GFP positive iPSCs were used to generate a secondary shRNA library. This library was then screened during reprogramming again like in figure 2. Bottom panel: we continued this cycle until we had performed 5 rounds. All intermediate libraries including negative controls were analyzed. The library became less complex (containing less shRNAs) each step, while losing shRNAs that do not benefit reprogramming (depicted as green or pink shRNAs).

6 and 12 hours after the start of transduction. Next we waited 48-60 hours to allow for some recovery before the next harsh treatment of passaging cells onto reprogramming plates.

a4

Cell culture and harvest

In short, we experienced that likely due to handling larger quantities of cells, overall reprogramming efficiency was usually lower then the pilot experiments predicted. That said, large Petri dish-like, 15 cm culture plates served well to for a starting population of 300 thousand cells each. We seeded MEF directly into the reprogramming media containing doxycycline. When looking to reduce the use of reagents and time, we noted no reduction in efficiency upon culturing in a reduced volume of media of 25 ml versus 30 ml per plate, nor when delaying the first media change to day 3 of reprogramming.

The most practically challenging step was the harvest of many plates at a time. After reprogramming, non-reprogrammed MEF tend to create what seems to be a strong network of extracellular matrix proteins, causing the trypsinized culture to clump partially. We used a large pore (100 µm) to small pore (40 µm) filtering protocol to create a cell suspension for subsequent separation by flow cytometry. The expensive and time consuming FACS sort is a major bottleneck to the amount of gDNA that can be generated. Only a fraction of genomic DNA that can be harvested from all cells can be used to amplify the shRNAs by PCR. Thus

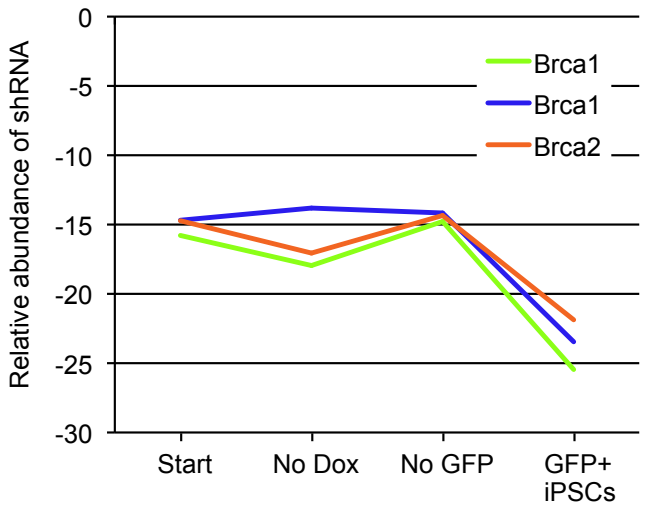


Figure 4. Relative abundance of 'dropout' shRNAs during the first round of screening. We selected for shRNAs well represented in the negative controls but diminished or lowered in the iPSC fraction (T1 Dox, GFPpos). Here, abundance of shRNAs targeting *Brca*1 and *Brca*2 are shown in reprogrammed cells (*Oct4*-GFP+ iPSCs) compared to starting population and controls (no dox, no GFP).

we took a representative fraction before the cell sort, after a very thorough mixing of all harvested and filtered cells.

Prior to FACS, we increased the yield of iPSC by enriching our sample for SSEA1 expressing stem cells by magnetic (MACS) sorting. We reached around 90% purity of *Oct4*-GFP positive cells depending on the stringency setting on the AutoMACS sorter (Figure 5). This SSEA1+ sorted cells were purified for *Oct4*-GFP expression by FACS which should give very near to 100% purity, although the limitation on sorting time prevented us from re-sorting to guarantee such total purity. The steps before FACS took around 12 hours for the largest screening rounds, during which cells were kept at 4 degrees Celsius. It was important to FACS right after to not lose cells to extensive cell death because of this long day.

Genomic DNA preparation from cell samples

During each round, four samples were generated from shRNA infected cells: the 'T0' sample of immediately harvested MEFs; the 'T1' sample of reprogrammed, sorted, *Oct4*-GFP+ cells; the non-reprogrammed, *Oct4*-GFP-negative cells and the MEF that were cultured in parallel but without doxycycline induction (Figure 2). All cells were pelleted and immediately frozen at -80C. Each of these samples is subject to a PCR to retrieve the shRNAs for sequencing (all) or re-cloning (only *Oct4*-GFP positive cells). This PCR is complicated by three factors: concentration of the shRNA template in total genomic DNA, secondary structure formation of the shRNA amplicon, and possible contaminants in the large gDNA preparation. To address the latter, we used a phenol-chloroform extraction with phase-lock tubes. Generally PCR steps needed more optimization and practicing than most PCR protocols, and resulted in a relatively low yield.

Recloning sequential libraries

The key of the experimental design of this screen is the serial re-screening of the shRNA library to select for reproducibly enriching shRNAs. We re-cloned the pooled shRNA library from genomic DNA of iPSC from the previous screening round to generate the screening library for the next round. For this purpose, all shRNAs were en masse amplified, purified, digested to create sticky ends and ligated into the original viral vector. These steps proofed difficult to do at the efficiency required for good coverage (100-1000x) of the library; count on optimizing this in your lab as well.

Sequencing

We used Solexa sequencing to assess the relative abundance of each shRNA. We added generic adaptors with a two-nucleotide library barcode to the amplified pool of shRNA encoding strands from each sample. This sequencing library was mixed with up to three other barcoded libraries to reduce the number of Solexa flow cells required. Each flow cell generated 100-200 million 50 base pair reads, resulting in 25-50 million reads per sample. Subsequent analysis allowed separation of reads from different samples prior to counting individual shRNA reads. Figure 6A gives an idea of our usual outcomes from the Solexa sequencing in terms of total reads, mixing of the barcoded samples, and reads per sample.

Our library required a 50 base pair readout to capture the unique 19 nucleotide shRNA sequence, along with the sample barcode and common shRNA loop sequence used to align the sequences during analysis of the raw data. A benefit of sequencing an shRNA library over, for example, genomic or mRNA derived libraries is the quality of data retrieved because the strands are homogenous, don't require high quality of the entire strand. Also, the sequencing primer is designed to amplify the most variable parts of the read first so the software can easily distinguish neighboring spots in the flow cell. In Figure 6B the average read is plotted against the data quality, showing how unique sequencies of the barcodes and shRNA sequence are sequenced well whereas the default sequencies are not because they are the same in all reads.

Analysis

We were able to rank shRNAs on many different criteria, such fold enrichment, consistency in enrichment, enrichment in negative controls, and many combinations. Initially we used an excel sheet with the data and tested different algorithms manually. Our manual algorithm delivered a list of 20 top hits from which we identified one of the strongest hits from the screen, an shRNA for polyadenylation modulator *Nudt21*.

When we repeated the screen, we obtained better coverage of the shRNA library and a more reliable, gradual loss of shRNAs after each round. The collaborating bio-informaticians were able to take these higher quality data, calculate a mean value from the negative controls to compare to the iPSC libraries, and rank all shRNAs based on how much they enrich in all rounds compared to the negative controls. (Figure 7) This list was topped by our strongest hit from the entire project, an shRNA targeting the sumoylation tag SUMO2.

Conclusion

A genome-wide screen is a powerful and relatively unbiased tool to find genes involved in various cellular processes. The pooled approach reduces the amount of specialized equipment needed and avoids using small wells for culture. In our case, this enabled a readout based on reprogramming efficiency of shRNA expressing cells. Furthermore, we piloted a serial screening approach to get reliable results in a process prone to create false-positives. Our approach allowed us to screen all protein-coding genes for a role during reprogramming. However, despite optimizations in design and execution, this is a demanding project, time-wise and material-wise. Smaller scale screening based on genes of

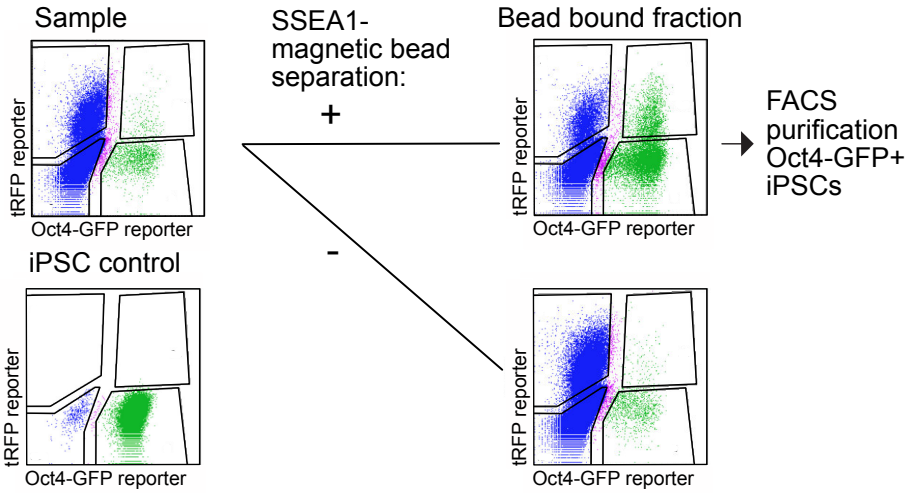


Figure 5. Flow cytometry analysis of samples before and after magnetic-activated cell sorting (MACS) for SSEA1 expression. Prior to the more stringent but less efficient FACS for *Oct4*-GFP+ iPSCs, MACS was performed for stem cell marker SSEA1 to increase the iPSC concentration. These flow cytometry plots show the increase of the *Oct4*-GFP (x-axis) fraction (green population) after MACS. Some iPSC still retain viral tRFP (y-axis) expression, but most silence the viral vector. An iPSC control sample with some contaminating feeder cells is shown for comparison.

interest or gene groups such as kinases or transcription factors might be more appropriate when screening is a subtopic of your studies. It would be interesting to apply our lessons, e.g. on serial screening, to smaller scale shRNA screens.

Take home points:

- Do not underestimate the scale of an experiment needed to properly cover a large screening library. Assume everything loses efficiency when working on a large scale;
- Ensure good library coverage at each step, not just in cell culture;
- Decide on a smart sampling and readout method to save work and materials;
- Consider biases and false-positive or -negative results due to technical reasons or a heterogeneous cell population or process of interest, and think of appropriate controls;
- Consider serial enrichment screening for heterogeneous cells or processes;
- Be prepared to use trial-and-error in your hit selection process, but keep number of hits to be validated realistic;
- Ideally collaborate with someone with experience cloning the library and analyzing the sequencing results.

References

1. Chang, K., Elledge, S. J. & Hannon, G. J. Lessons from Nature: microRNA-based shRNA libraries. Nat. Methods 3, 707–714 (2006).
2. Luo, J. et al. A Genome-wide RNAi Screen Identifies Multiple Synthetic Lethal Interactions with the Ras Oncogene. Cell 137, 835–848 (2009).
3. Hu, G. et al. A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. Genes Dev. 23, 837–848 (2009).
4. Seo, M. et al. RNAi-based functional selection identifies novel cell migration determinants dependent on PI3K and AKT pathways. Nature Communications 5, 5217 (2014).

A Exemplary yield from multiplexed Solexa sequencing

	Lane 1		Lane 2	
Barcoded sample	Reads	%	Reads	%
Total reads	203,794,591		219,854,696	
AT	52,800,457	25.91	50,644,519	23.035
CT	43,178,848	21.19	57,602,239	26.2
TC	47,029,559	23.08	64,420,342	29.3
CA	59,567,901	29.23	46,221,655	21.024
Un-Matched reads	1,217,826	0.60	965,941	0.439

B

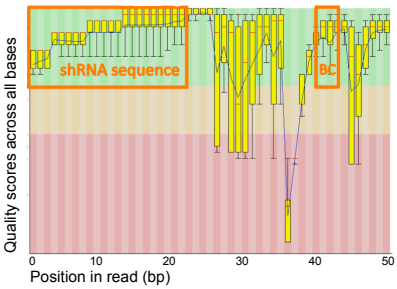


Figure 6. Solexa sequencing output of barcoded libraries. (A) Exemplary yield from Solexa sequencing. Here we used two Solexa flow cells (lane 1 and 2). In each we ran four samples with different barcodes: 'AT', 'CT', 'TC' or 'CA'. Total reads per lane were more than 200 million; of which 43 to 64 million were assigned to each sample. A fraction of reads was unidentifiable (around 1 million per lane). (B) Average quality of the read (y-axis) plotted against the length of the read (50 base pairs, x-axis). This plot shows how well the sequence could be read at each position of the reads, across all reads. The top green zone means good quality; usable for downstream analysis. Highly variable parts such as the individual shRNA sequence (first 20 base pairs) and the barcode (BC) were sequenced at a high quality, whereas the rest of the read, containing sequences common to all the reads, was sequenced with a poor quality.

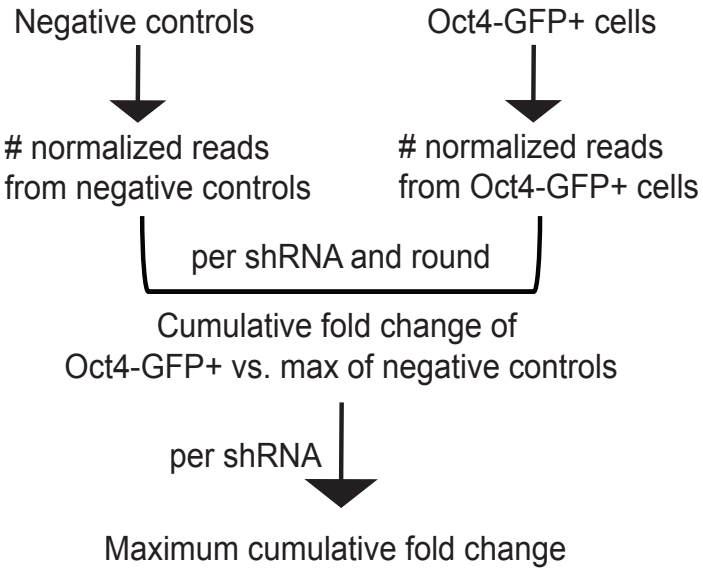


Figure 7. Ranking algorithm shRNAs based on enrichment during reprogramming rounds. Schematic depiction of algorithm used for picking candidates from enrichment data. Overall fold enrichment score was calculated by comparing all three negative controls to the iPSC sample. For this, number of reads were normalized per sample, before calculating fold enrichment per round. Then, overall (cumulative) fold enrichment was calculated for each shRNA, giving it a score to rank them on.

Chapter 6

Reprogramming to pluripotency as a model for oncogenic IDH mutations

Marti Borkent, Kelly Marsh, Eftychia Apostolou, Abhishek Jha, Justin Brumbaugh, Katherine Yen and Konrad Hochedlinger.

Manuscript in preparation

Abstract

IDH mutations are important drivers in a range of malignancies including leukemia, glioblastoma and cholangiocarcinoma. The IDH enzymes, their substrates and their metabolites are highly conserved and present in all cells. Point mutations generate neomorphic IDH enzymes that produce the 'oncometabolite' 2-hydroxy-glutarate (2-HG). Although thought to be early mutations in tumorigenesis, they act in the context of other tumorigenic mutations in genetically unstable cells. To study the effect of mutant IDH (mIDH) expression during transformation, but in an otherwise healthy cell, we used fibroblast to induced pluripotent stem cell (iPSC) reprogramming as a transformation model. We were able to express the most common human mutant enzymes and measure high levels of 2-HG in mouse embryonic fibroblasts (MEFs). Although the efficiency of iPSC formation was unaffected, or even reduced at high 2-HG levels, mIDH did facilitate early reprogramming as characterized by a loss of fibroblast and mesenchymal gene expression. This effect was dependent on reprogramming factor expression. Moreover, mIDH created a population of cells that did not form iPSC but also failed to revert back to THY1 expressing MEF cells. Macroscopically, mIDH leads to the formation of small cells that stain positive for alkaline phosphatase; interestingly, this phenotype was reverted by ascorbic acid. To characterize IDHm expression on a molecular level, we profiled early reprogramming cells both by RNAseq and several metabolic analyses, together revealing the most affected metabolic pathways. These experiments shed light on the effect of IDHm independently of a cancerous genetic background and illustrates the effect of a metabolic perturbance on early reprogramming events.

Introduction

Our understanding of the molecular underpinnings of cancer started with the study of DNA mutations that amplified oncogenes or mutated tumor-suppressor genes. Some cancer associated mutations were found in molecules deemed too essential for cellular survival such as enzymes from the Krebs or TCA cycle, which drives oxidative phosphorylation, the most efficient route of glucose breakdown in terms of ATP (Adenosine Tri-Phoshate, the 'energy currency' of the cell) production. However, mutations of the enzyme isocitrate dehydrogenase are found in glioblastoma multiforme, leukemia, cholangiocarcinoma and others[1-4]. Wildtype IDH1 converts isocitrate to α-ketoglutarate (α-KG) in the cytoplasm, IDH2 has the same function but as part of the Krebs cycle inside the mitochondria. These oncogenic mutations do not require homozygosity; it is not the loss or gain of function of these enzymes that stimulate oncogenesis. Rather, neomorphic mutations in IDH1 or IDH2 lead to a mutant enzyme that catalyzes the 'oncometabolite' 2-hydroxyglutarate (2-HG).

Oncometabolite 2-HG is structurally highly similar to α-KG. It can bind α-KG dependent enzymes including epigenetic regulators such as Jumonji-domain containing histone demethylases and the 5mC hydroxylating Tet enzymes[5].

Mutations in IDH1 and IDH2 have been found to be potent drivers of tumorigenesis[6,1,7]. So far many studies have been conducted in the context of tumor models or patient derived cancer cell lines. These models harbor various other aberrations that arise during tumorigenesis unique to each tumor and patient or mouse. Moreover, there are no marker-defined stages of transformation towards malignancy that can be used to dissect individual oncogenic events.

Our lab specializes in studying the molecular events leading towards pluripotency using controlled, transcription factor induced reprogramming systems. We have defined intermediate stages during the reprogramming of mouse embryonic fibroblasts (MEF) to induced pluripotent stem cells (iPSCs), marked with cell-surface markers and an endogenous pluripotency reporter that can easily be tracked by fluorescence microscopy and flow cytometry. Moreover, we combined an expression cassette carrying all four required transcription factors, *Oct4*, Sox2, *Klf4* and *c-Myc*, under a doxycycline-inducible promoter, along with the endogenous pluripotency reporter *Oct4*-GFP[8,9].

Reprogramming to pluripotency is a process of transformation that shares similarities with tumorigenesis. Importantly, like transforming cancer cells, otherwise healthy reprogramming cells also switch to glycolytic glucose metabolism, downregulate apoptosis pathways, increase cell proliferation while altering DNA methylation patterns and chromatin structure on a large scale. However, the resulting stem cells are capable of generating entire cloned animals. We used the relatively physiological background of reprogramming to dissect the effect of IDH mutations on all of these important molecular processes.

Results

Human mutant IDH expression in MEF leads to high levels of 2-HG which can be reversed by small molecule inhibitors.

In this study we included three different prevalent IDH mutants which are active in the cytosol (IDH1 R132H) or in the mitochondria (IDH2 R140Q and IDH2 R172K) (Figure 1a). The two IDH2 mutants differ effectively in their rate of 2-HG production; IDH2 R172K overexpression leads to higher intracellular 2-HG levels. The IDH enzymes are highly conserved between mouse and human, thus we were able to examine the effect of human mutant IDH expression in our murine model of iPSC formation. To confirm adequate expression of the enzymes in our cells of choice, mouse embryonic fibroblasts (MEFs), we transduced the cDNA in a constitutive lentiviral vector and assessed 2-HG production after

48 hours (Figure 1b). Indeed, 2-HG was produced at high levels in all mutant IDH samples, but not in wildtype IDH samples. The level of 2-HG production was mutant-dependent, IDH2 R172K producing most of the oncometabolite. We reproduced the overexpression of wildtype and mutant IDH in our experimental setup for large scale profiling experiments. There we included mutant-specific small molecule inhibitors or solvent (DMSO) to confirm efficacy of inhibitors in reducing 2-HG production by the mutants and serve as an additional control (Figure 1c). We reached almost normal 2-HG levels by using the inhibitors, except for the IDH2 R172K samples which turned out to be under dosed to fully inhibit the potent mutant. We included these samples in the following experiments as an interesting midway point to very high or normal 2-HG levels.

IDHm expressing MEF form iPSC at the same or lower efficiency

Stable iPSC colonies form after different lengths of time of reprogramming factor exposure, and when looking at marker-based intermediate stages, each step towards pluripotency is increasingly rare to occur. As a consequence, the reprogramming culture contains cells at different stages of reprogramming pretty soon after the induction of OKSM expression. For this reason, bulk-harvested reprogramming cultures are a mixture of reprogramming and refractory cells, unless harvested at an early time point such as day three, when a majority of cells has undergone the first step of losing expression of the MEF marker THY1. Our lab avoids additional heterogeneity due to reprogramming factor delivery by using MEFs from a 'reprogrammable mouse' (repMEFs), a mouse strain that carries all alleles for doxycycline-inducible expression of the OKSM cDNAs (Figure 2a).

Next, we tested the effect of mIDH on cell death and -growth at days 1-3. It is known that cell proliferation and survival positively correlates with reprogramming efficiency, which could create a bias in reprogramming experiments[10]. To this end, we induced reprogramming in repMEFs expressing the different vectors and measured the rate of apoptosis at day 3 (AnnexinV staining, Figure 2b) and the total number of cells at day 1-3 (Figure 2c). There was no discernable effect on the percentage of AnnexinV positive cells by any of the IDH overexpression vectors. However, both wildtype and mutant IDH expression reduced cell growth compared to the empty vector. When comparing mutant expression to corresponding wildtype expression, only IDH2 R172K expression resulted in a clear reduction of cell proliferation.

Next we tested whether wildtype or mutant IDH overexpression would affect reprogramming ability and efficiency of repMEFs. We assessed stable, dox-independent iPSC formation by endogenous *Oct4* expression (by *Oct4*-GFP reporter knock-in allele; Figure 2d) and by colony formation assays (Figure 2e). After 10 days of Dox induction and 4 days dox

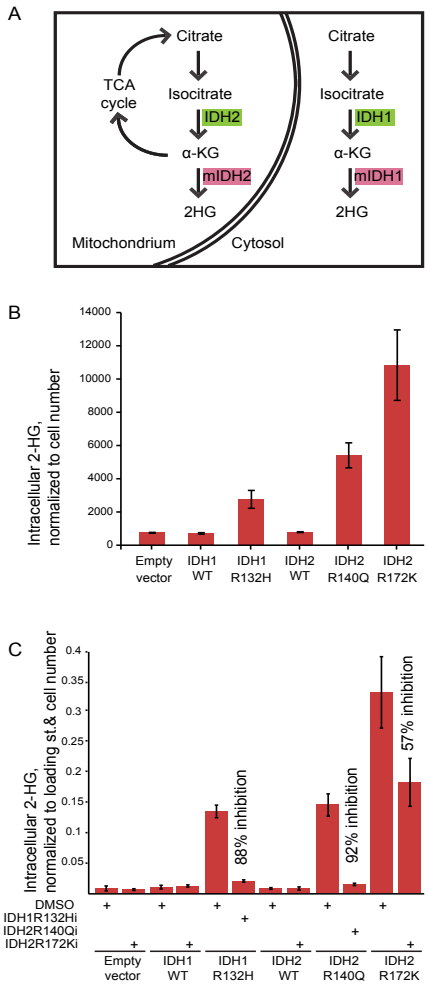


Figure 1. Overexpression of human mutant IDH raises 2-HG levels in mouse embryonic fibroblasts. (A) Mutant IDH (mIDH) produces oncometabolite 2-hydroxyglutarate (2-HG) from physiological metabolite alpha-ketaglutarate (α-KG), either in the mitochondrium (mIDH2) or in the cytosol (mIDH1). (B) Overexpression of mIDH through viral vectors raises intracellular 2-HG levels in mouse fibroblasts; quantification normalized to cell number, n=3, error bars represent standard deviation. (C) Production of 2-HG can be effectively inhibited by small molecule inhibitors specific to the different mIDH. DMSO is the inhibitor solvent used as negative control. Quantification normalized to cell number and to internal loading control; n=5, error bars represent standard deviation.

withdrawal, flow cytometry analysis detected *Oct4*-GFP expressing cells regardless of wildtype or mutant IDH expression. However, IDH2 R172K expression did reduce the abundance of *Oct4* expressing cells. The colony formation assay, using alkaline phosphatase (AP) staining of dox-independent colonies demonstrated that none of the conditions prevented stable iPSC formation. Remarkably, the mutant IDH cells formed more AP staining colonies, although many of these colonies did not show the characteristic strong AP staining and sharp border required to count a colony as an iPSC clone.

Mutant IDH expressing cells create iPSC-like, Oct4-GFP negative colonies that are eliminated by ascorbic acid exposure

Before the AP staining of iPSC colonies from IDH expressing cells, we examined them under light and fluorescent microscopy. We could distinguish colonies with the irregular shaped ESC-like dome shape of primary iPSC clones, from single layer, dense clusters of small cells that lacked ESC or fibroblast morphology. When we quantified the AP staining colony assay, counting only sharp, brightly stained colonies as true iPSC colonies, we saw no significant difference in the colony formation efficiency between mutant IDH or control cells

(Figure 3a, grey bars). When we counted solely the dimmer, fuzzy-bordered colonies, that appeared as dense clusters under the microscope, we noted that overexpression of each of the mutant IDH vectors caused an increase in the formation of these colonies (Figure 3b, grey bars). We previously observed that ascorbic acid in the cell culture media nearly eradicated the presence of 'fuzzy' colonies in unrelated experiments. Also, our lab earlier described ascorbic acid to be a potent enhancer of reprogramming efficiency towards iPSC cells[11]. When we added ascorbic acid to the mutant IDH or control cells, the number of fuzzy colonies was reduced to nearly absent in all samples but IDH2 R172K, where they were significantly reduced. Also, the number of iPSC clones increased (Figure 3a-b, red bars).

To further characterize the cell clusters that formed and were stimulated by mutant IDH expression, we analyzed reprogramming cultures by flow cytometry for MEF marker THY1 and endogenous *Oct4*-GFP reporter expression. After 4 days of dox withdrawal, control cultures contain nearly only THY1 positive or *Oct4*-GFP positive cells, suggesting there are stable iPSC or MEF (Figure 3c). Mutant IDH expressing cells, however, formed many more cells negative for either marker.

IDHm expressing MEF more efficiently lose their MEF identity during early reprogramming

To further study the 'marker-negative population, we analyzed reprogramming cells at intermediate timepoints during reprogramming to see when this population is formed. We noticed that all mutants lose the MEF marker THY1 much more efficiently between day 3 and 6 of reprogramming (Figure 4a; IDH1 in left panel, IDH2 in right panel). Note that THY1 positive cells are decreased in all samples at later timepoints due to the relative expansion of iPSC. The next, intermediate step during reprogramming is the gain of the stem cell marker SSEA1. This was not affected by mIDH expression, except for a reduction in IDH2 R172K expressing repMEFs (Figure 4b).

Finally, when we looked at iPSC formation based on the formation of *Oct4*-GFP positive cells, we did not see any divergence between samples, except a reduced number of cells at all time points for IDH2 R172K expressing cells (Figure 4c). We repeated the day 6 analyses including the mutant specific inhibitors to confirm the significant THY1 loss, and were able to rescue the rate of THY1 loss using the mutant IDH inhibitors (Figure 4d).

To exclude that we were looking at an expansion of a rare THY1 negative population within the heterogeneous MEF prep, we also analyzed non-reprogramming fibroblasts expressing the mutants, but failed to see a clear THY1 loss there (Figure 4e), suggesting that loss of THY1 is a reprogramming factor dependent event.

Transcriptional profiling of reprogramming cultures with mutant IDH expression

Mutant IDH expression leads to tumorigenesis and aids dedifferentiation[12,13]. In our model mIDH expression does not enhance reprogramming to iPSC. However it does enhance formation of cells that lose MEF identity but do not gain any stem cell marker that we normally detect in reprogramming cells. These cells arise between day 3 and 6 of reprogramming and

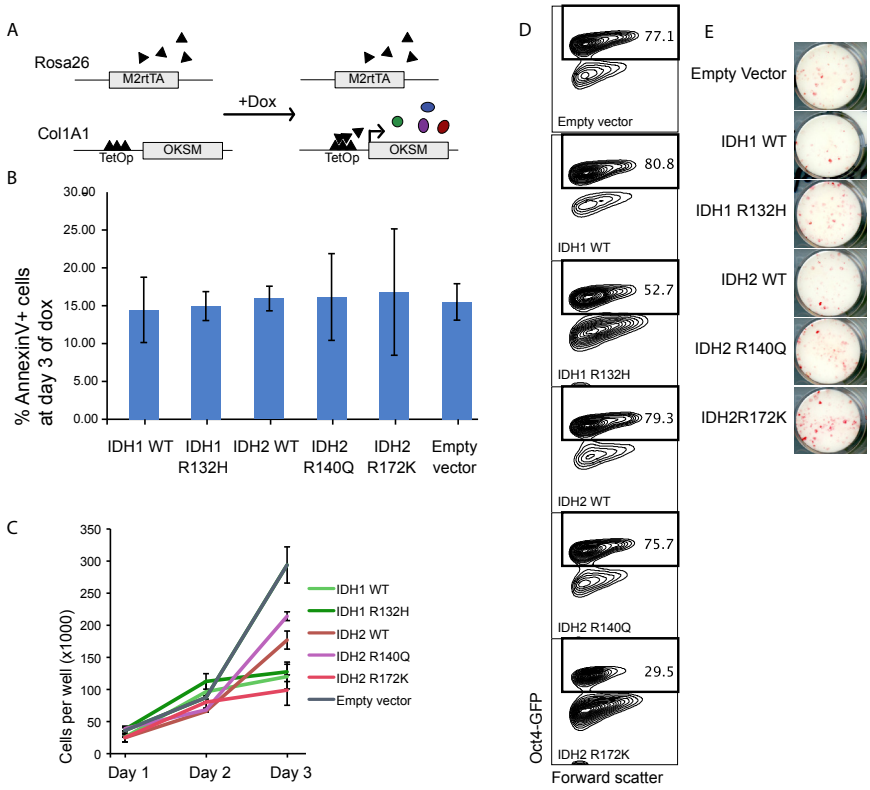


Figure 2. Effect of mIDH on viability and reprogramming potential of reprogrammable mouse fibroblasts. All Figures: error bars represent standard deviation; n=3; WT= wildtype allele. (A) Alleles carried by reprogrammable mouse embryonic fibroblasts (repMEFs). OKSM: Tet-driven polycistronic cassette with *Oct4*, *Klf4*, *Sox2, c-Myc* cDNAs. M2rtTA: reverse tetracycline transactivator. (B) assessment of cell viability by percentage of Annexin V presenting cells at day 3 of reprogramming. (C) Cell proliferation plotted as absolute cell number per well at days 1-3. (D) Flow cytometry analysis of cultures after 10 days reprogramming and dox withdrawal. *Oct4*-GFP: expression of pluripotency reporter allele on y-axis. Numbers in plots: % of viable cells. (E) Wells with reprogramming cultures after 10 days of reprogramming and 4 days dox withdrawal, stained for stem cell marker alkaline phosphatase to detect iPSC colonies.

persist to a larger degree than in controls, even after withdrawal of OKSM and 2 weeks of culture. Thus we wondered what these cells are to understand what mIDH does to somatic cells that are transforming in a context free from other mutations involved in tumorigenesis.

To catch the molecular events just prior to the formation of these cells, and not look at possible indirect effects, we decided to profile reprogramming cultures at day 3 post induction. First we analyzed global transcription by RNA-sequencing. To validate that the loss of MEF identity was already taking place at the transcriptional level, we looked at expression of *Thy1,* collagens and mesenchymal-to-epithelial transition (MET) related molecules (Figure 5a-c). *Thy1* was indeed downregulated in the mutants. Several collagens were downregulated and rescued by the inhibitors of mutant IDH. *Tet1*, which is inhibited by competitive binding to 2-HG instead of its cofactor α-KG, was not transcriptionally affected by mIDH (Figure 5d).

Next we looked at genes of interest based on earlier studies on reprogramming such as mesenchymal genes that are downregulated in the transition to the epithelial like state of pluripotent cells; and genes found to be linked to more pluripotent or more differentiated cells. We found no clear pattern of expression of mesenchymal genes *Snai1, Twist* or epithelial genes such as *Cdh1* (Figure 5e-f). Stem cell related genes were expressed at very low levels in all samples this early in reprogramming (not shown). Another group of genes of interest were those transcriptionally altered during reprogramming in prior reprogramming studies. When we looked for mIDH dependent expression in this group, we found *Prx* and *Ereg*, genes normally downregulated during early reprogramming, and *Myb* which is upregulated during day 6-9 of reprogramming. (Figure 5g-i).

Another group of interesting genes are those described to be differentially methylated upon mIDH2 expression in human blood cell and leukemia cell lines 14. We analyzed expression of the murine homologs in our RNAseq data. Most were not differentially expressed. The only exceptions were *Dok1* which was slightly downregulated, and *Gpc3* which was slightly upregulated in mIDH expressing cells (Figure 5j-k).

Finally, we looked with an unbiased view by selecting those genes that were consistently up- or downregulated between the mIDH DMSO or inhibitor samples. The vast majority of hits were downregulated by mutant expression (Table 1; selected for 2+ fold down- or upregulated in all three mIDH samples compared to inhibitor). This included genes such as Wnt pathway signaling molecules *Wnt2* and *Dkk2*, p53-activated pro-apoptosis signaler *Palmd* (ref24810057), negative stem cell proliferation regulator *Gabre* (ref18185516) and adhesion molecule *Icam1* (Figure 5l-r).

Unexpectedly, some genes were also affected by IDH2 wildtype overexpression. Wildtype IDH2 overexpression does produce some 2-HG as a byproduct, possibly explaining this. Interpretation of some of these expression patterns is complicated by overall low

expression levels. Genes that were upregulated in a mIDH dependent manner were *Pou4f1* (*Brn-3a*, neuronal transcription factor and proposed oncogene, ref 12405260) and *Mob3b* (in family of mitotic regulators; possible human tumor suppressor gene, ref 26130958). Overall, the RNA profiling points to some genes well known to affect oncogenesis and reprogramming, as well as some unexpected genes. The gene expression patterns mostly differ from earlier studies on gene expression upon mIDH expression, which is likely due to the analysis of different cell types.

Metabolic profiling of reprogramming cultures.

We submitted the same day 3 reprogramming cultures to bulk harvest for metabolic profiling. We profiled by several modalities: both targeted (LCMS negative and positive mode) and untargeted analysis (FIA-TOF negative and positive mode). We looked for

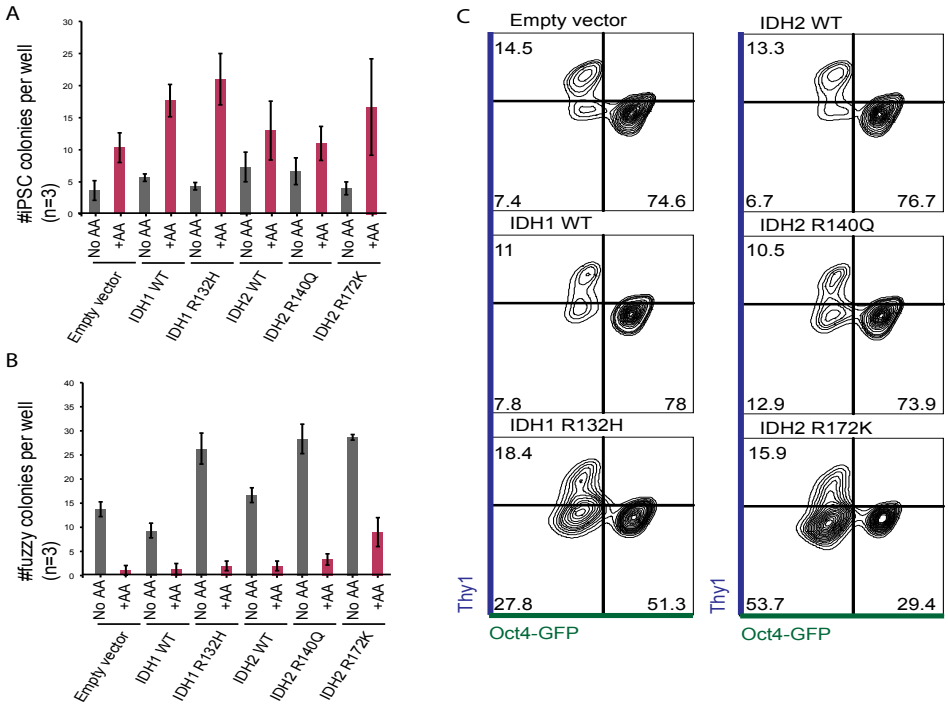


Figure 3. Formation of *Oct4*-GFP, THY1 negative, 'fuzzy' colonies is eliminated by ascorbic acid treatment. Error bars represent standard deviation; n=3; WT= wildtype allele. (A) Quantification of sharp bordered, brightly stained colonies in alkaline phosphatase assay, upon or without ascorbic acid (AA) treatment during reprogramming. (B) Quantification from same wells as in (A), counting fuzzy, dim stained colonies. (C) Flow cytometry analysis of reprogramming cultures after dox withdrawal. MEF marker THY1 (y-axis) plotted against pluripotency reporter *Oct4*-GFP (x-axis). Numbers in plots are % of viable cells.

metabolic pathways affected by either of the IDH mutants but normalized by the mutant inhibitors. Perhaps unsurprisingly, IDH2 mutants, which are localized to the mitochondrium, had a clear effect on TCA cycle metabolites and derivatives downstream of α-KG, such as acetyl-CoA, citric acid, fumaric acid, succinic acid and glutathione (Figure 6A-E). Our reprogramming phenotype was comparable for IDH1 and IDH2 mutants, so these changes may not be causal to that phenotype. Furthermore, phospholipid and glutathione related metabolites were altered such as citicolin and oxidized glutathione.

Hif1α is degraded in α-KG dependent manner, and postulated to be affected by mutant IDH through 2-HG antagonism. A byproduct of the degradation reaction is the metabolite succinate. The transcriptional data showed no difference in the expression of a range of *Hif1α* targets (Table 2) in mIDH expressing reprogramming cells. The metabolic analysis revealed no change in succinate levels either, suggesting that *Hif1α* regulation is unaffected by mIDH in our system (Figure 6D). Individual metabolites affected similarly by IDH1 and IDH2 mutants were glutamate and GABA (both reduced by mutants) (Figure 6H-I).

Finally, we integrated the transcriptional data and metabolic data to detect which metabolic pathways were affected both at the metabolite and metabolic enzyme level. Three pathways were highlighted by this 'COMBI' analysis[15]; metabolism of phytanic acid by PHYH, aminotransferases in serine metabolism and lysine degradation by AASS (Figure 7-9). Of these three, AASS showed the strongest positive correlation with expression of the IDH mutants. AASS is involved in the conversion of lysine to glutamate, which is low in IDH mutant cells, suggesting this is a response to glutamate depletion. Glutamate depletion itself may be caused by failure to use 2-OG to form glutamate instead of 2-HG. Since this is such a central metabolite and affected by all IDH mutants, it could provide metabolic explanation of the reprogramming phenotype.

Discussion

Mutations of IDH1 and IDH2 are potent drivers of tumorigenesis[1,6,7]. In this study we took advantage of our iPSC reprogramming model to study the oncogenic IDH mutations outside a cancer model.

First we noted that mIDH expression has no discernable effect on cell proliferation, cell survival or overall reprogramming efficiency, except for the most potent 2-HG producing mutant IDH2 R172K which reduced cell growth and the efficiency of iPSC formation.

However, a third cell population not resembling iPSC or refractory MEF, was amplified by expression of either mIDH. Like in normal controls, this normally rare population expressed no MEF or stem cell markers, only faintly stained for alkaline phosphatase, and nearly disappeared when ascorbic acid was added to the media. Further analysis during

reprogramming revealed that mIDH causes an increase in loss of MEF marker THY1 between day 3 and 6. This is an early event in reprogramming, normally succeeded by gain of stem cell markers and eventually, a stable pluripotent state or return to a MEF like THY1 positive state. However, mIDH caused larger populations of cells to remain marker-free throughout. We were able to inhibit mIDH with mutant-specific small molecules to rescue these phenotypes.

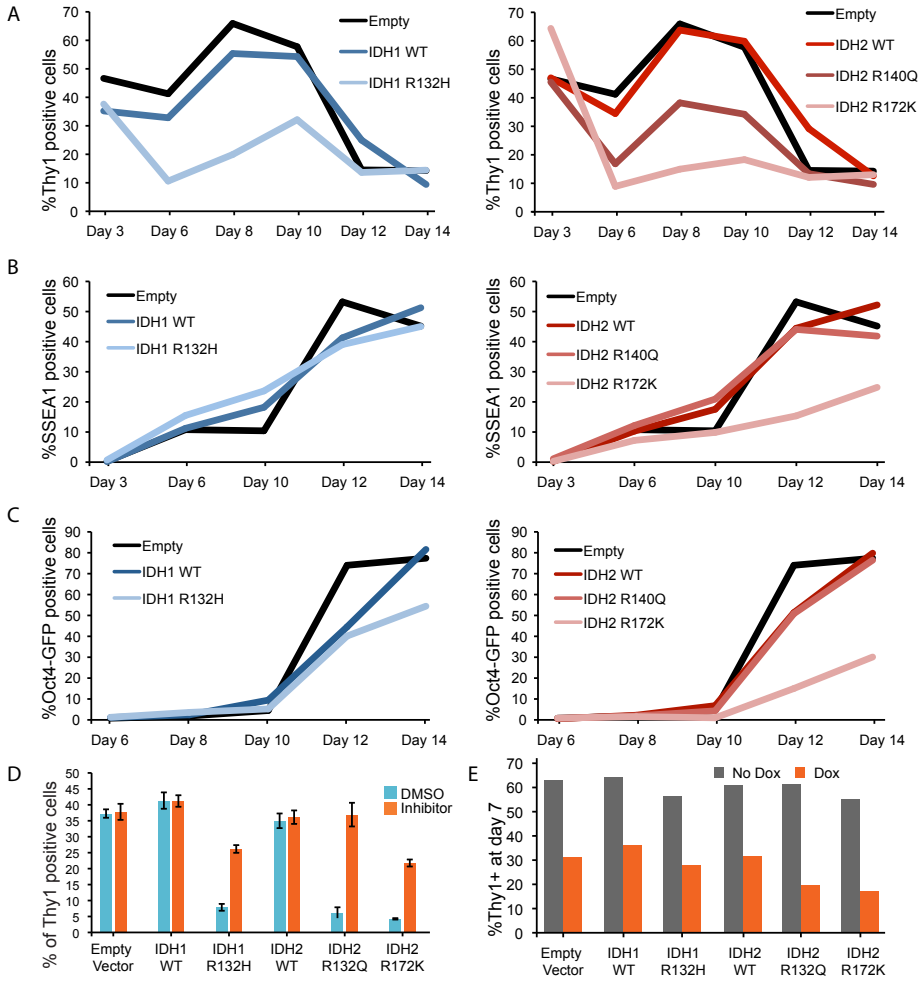


Figure 4. Flow cytometry analysis of reprogramming intermediates from mIDH expressing repMEFs. A-C) Proportion of viable cells expressing THY1, SSEA1 or *Oct4*-GFP respectively, in mIDH1 (left panels) or mIDH2 (right panels), at different timepoints during reprogramming. Dox was withdrawn at day 12. (D) Quantification of the loss of MEF marker THY1 at day 6 of reprogramming of mIDH expressing cells with or without the mIDH specific inhibitors. N=3; error bars represent standard deviation. (E) Quantification of THY1 loss at day 6 of reprogramming or dox-free culturing (grey bars) to exclude dox-independent THY1 loss or expansion of contaminating THY1 positive cells in the MEF culture.

Next we sought to characterize this loss of identity and other effects of mIDH on these transforming cells between day 3 and 6 of reprogramming on a transcriptional and metabolic level. We picked day 3 as a time just prior to the phenotype manifesting, avoiding secondary effects of the phenotype. Transcriptional profiling confirmed that the loss of expression of THY1 and other MEF related genes such as collagens, differentiation related genes and cell growth inhibitors. GO analysis of genes consistently affected by mIDH and rescued by the inhibitors revealed cell signaling (e.g. WNT pathway), morphology, and various metabolic processes.

Earlier studies on the effect of mIDH on tumorigenesis in blood cells suggested that the high 2-HG production is oncogenic by inhibiting differentiation through antagonism of essential epigenetic regulators. In our system, we find that mIDH stimulates a cell population that seems to dedifferentiate from a MEF state but refrain from forming iPSCs unless ascorbic acid is added. Ascorbic acid is a cofactor to some of the same enzymes antagonized by 2-HG. Thus we hypothesize that in our reprogramming model, high 2-HG levels caused by the mIDH steers cells to a state of dedifferentiation, unable to make the epigenetic alterations to stabilize a pluripotent state. This could largely be due to α-KG dependent, 2-HG sensitive enzymes that are activated by ascorbic acid such as the TET enzymes. Ascorbic acid could either overcome 2-HG antagonism or compensate for inactivation of some enzymes by activating others.

Since small changes in levels of metabolites or metabolic enzymes are associated with functional alterations of the system, we combined transcriptional data with extensive metabolic profiling. Looking at the metabolomics alone, most changes were

In DMSO vs inhibitor:			
Upregulated	Downregulated		
Zfp869	Ccdc125	Foxs1	Samd14
Prickle1	Nynrin	Ccdc8	Smpd3
6720427I07Rik	Aspa	Rab27a	D3Bwg0562e
Lum	Pde9a	Icam1	Trh
Pou4f1	Gabre	Nat14	Mustn1
Fbxo40	Klk8	Krtdap	Palmd
Il6	Smoc2	Wnt2	Slc13a5
B230216N24Rik	Serpinb9	Efcab12	Trim30d
Plb1	Slc16a3	Pgam2	Snca
Myb	9630013D21Rik	Phf19	Gm17322
Mob3b	Scarf2	Ccdc116	Efna1
Fndc1	Dkk2	Tubg2	Gm10443
	Acbd4	Fbxo44	Trim45
	Fam46b	Prmt1	Gm10036

Table 1. Transcriptionally affected genes in IDHm expressing reprogramming cells., compared to cells treated with the respective IDHm small molecule inhibitor

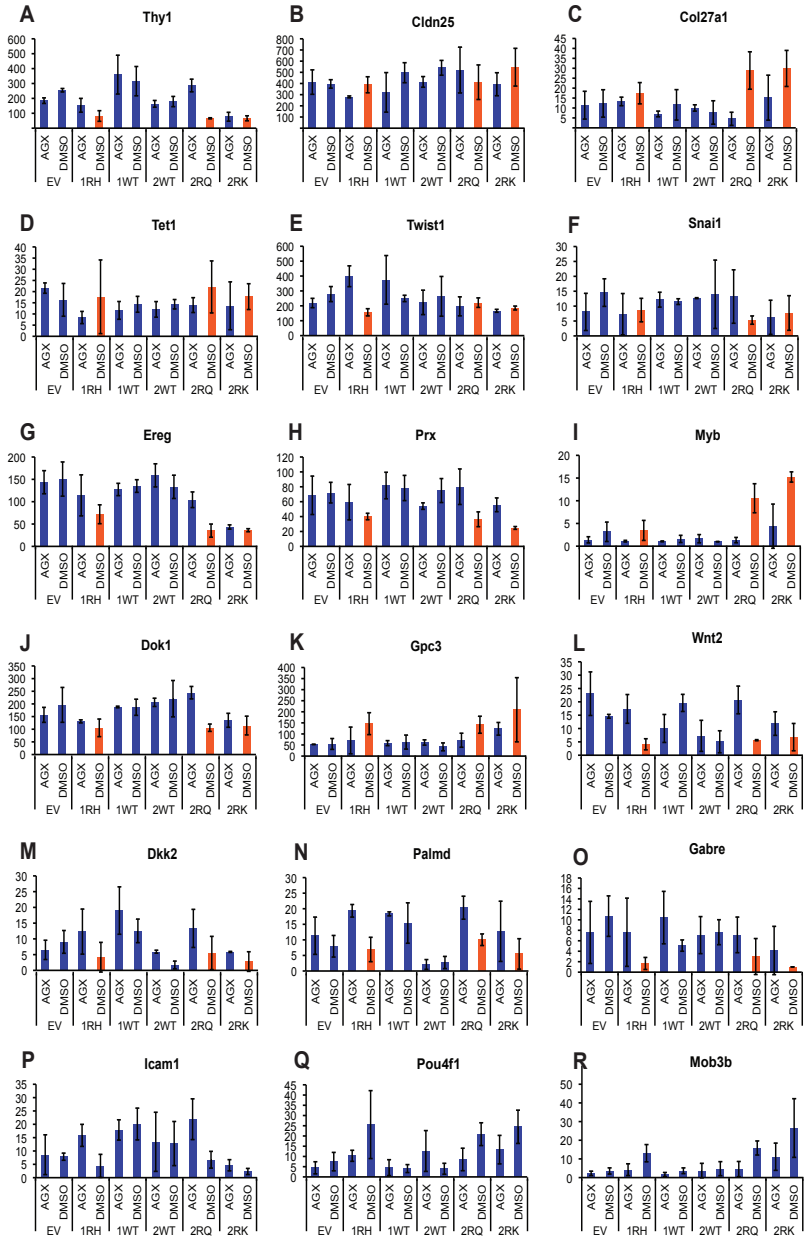


Figure 5. Expression analysis of individual transcripts in day 6 reprogramming cultures expressing mIDH. EV= Empty vector; 1RH= IDH1 R132H; 1WT= IDH1 wildtype; 2WT= IDH2 wildtype; 2RQ= IDH2 R140Q; 2RK= IDH2 R172K; AGX= mIDH specific inhibitor; DMSO=solvent only control. Relative expression levels from RNA-sequencing. Error bars represent standard deviation from 3 replicates. mIDH samples without inhibitor treatment are highlighted in orange.

caused only by the IDH2 mutants: their expression reduced abundance of metabolites in the TCA cycle, glucose derivatives, pentose pathway, and acetylation pathway. However when we combined the metabolite levels and the transcription data of the enzymes that metabolize them, pathways were highlighted in all three mutants. Most notably, the lysine to glutamate degradation pathway involving the Aass enzyme was clearly altered by all mutants. Glutamate is depleted, possibly secondary to α-KG (=2-OG; 2-Oxoglutarate) depletion through the high rate of 2-HG production by mIDH. Most enzymes using glutamate are upregulated without producing higher levels of their end product suggesting that these cells struggle to maintain normal levels of these metabolites.

We entertained the idea that mIDH affects HIF1α targets: 2-OG is turned into succinate by EglN prolyl hydroxylases. In the same reaction the enzyme hydroxylates a proline residue of Hif1α which marks it for degradation. However, we could find no proof suggesting that succinate, Hif1α or downstream factors were affected by the mIDH during reprogramming.

Future directions

The combination of transcriptional and metabolic data can only be performed in two-sample comparisons. We looked at mutants compared to empty vector, but more comparisons could be made. Next, based on these analysis and the ones already described here, more validation experiments can be done, both to confirm technical accuracy (qPCR for transcripts of interest, measuring individual metabolites) and further explore the dynamics of pathways identified as triggered by mutant IDH (metabolite tracing experiments).

The hypothesis of mutant IDH and ascorbic acid having opposing effects, most likely through inhibiting or stimulating the same 2-OG dependent epigenetic enzymes, should be further explored. Possible experiments to this end would be to determine DNA methylation levels at sites known to be targeted by TET enzymes. Also, the hypothetically 'stuck' cells that remain free from MEF or stem cell marker expression after dox withdrawal could be isolated and further characterized for gene expression and oncogenic potential. These and further experiments should help crystallize the molecular consequences of mutant IDH

Hif1a targets examined in RNAseq expression analysis			
Ak3	Hk1	Krt19	Pfkl
Ccng2	Hk2	Ldha	Pkm
Col5a1	Igf2	Lep	Tgfa
Cp	Igfbp1	Lrp1	Tgfb1
Eno1	Igfbp2	Mmp2	Tgm2
Ets1	Igfbp3	Nt5e	Tpi1
Fn1	Krt14	P4ha1	Trf
Gapdh	Krt18	Pfkfb3	Vim

Table 2. Overview of Hif1α targets examined

expression and the oncometabolite 2-HG during reprogramming.

To conclude, our study resulted in functional, transcriptional and metabolic analysis of transforming cells expressing oncogenic IDH mutations. We demonstrate a loss-of-identity phenotype common to both IDH1 and IDH2 mutants. Future directions include metabolic validation of our findings and evaluation of apparent key players that our results point to as effectors of mutant IDH driven oncogenesis.

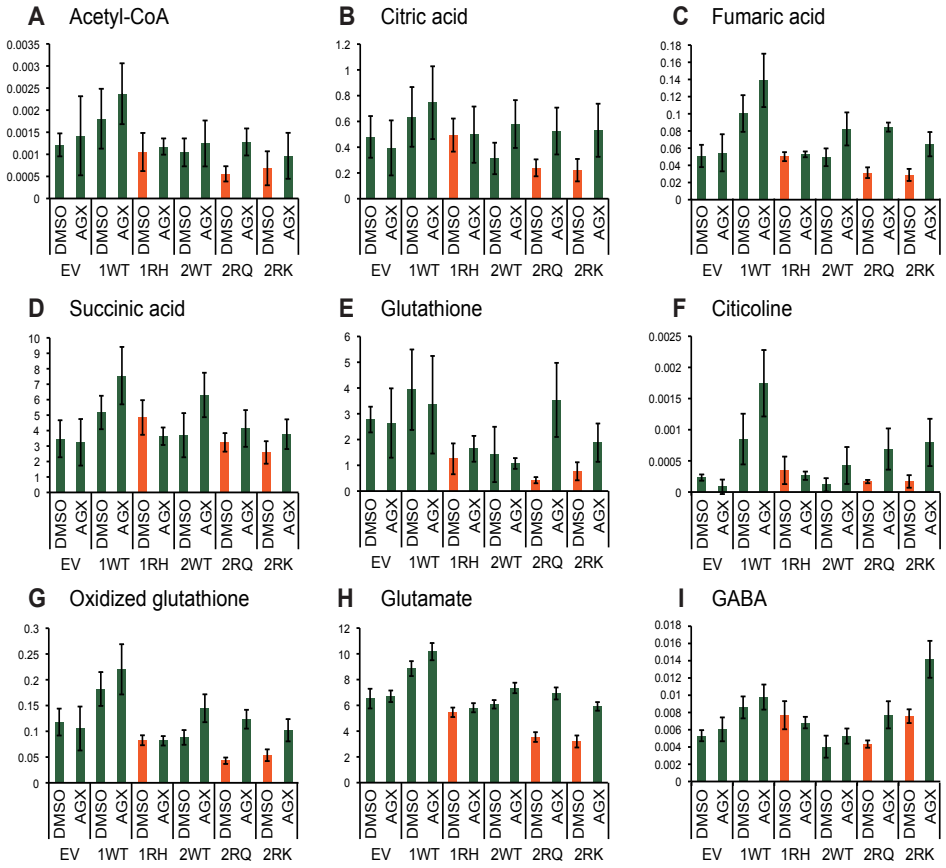


Figure 6. Metabolic profiling of day 6 reprogramming cultures expressing mIDH. EV= Empty vector; 1WT= IDH1 wildtype; 1RH= IDH1 R132H; 2WT= IDH2 wildtype; 2RQ= IDH2 R140Q; 2RK= IDH2 R172K; AGX= mIDH specific inhibitor; DMSO=solvent only control. Relative abundance levels from LCMS metabolomics, normalized to labeled internal control. Error bars represent standard deviation from 5 replicates. mIDH samples without inhibitor treatment are highlighted in orange.

References

1. Mardis, E. R. et al. Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. N. Engl. J. Med. 361, 1058–1066 (2009).
2. Amary, M. F. et al. IDH1 and IDH2 mutations are frequent events in central chondrosarcoma and central and periosteal chondromas but not in other mesenchymal tumours. J. Pathol. 224, 334–343 (2011).
3. Kang, M. R. et al. Mutational analysis of IDH1 codon 132 in glioblastomas and other common cancers. Int. J. Cancer 125, 353–355 (2009).
4. Parsons, D. W. et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. Science 321, 1807–1812 (2008).
5. Xu, W. et al. Oncometabolite 2-Hydroxyglutarate Is a Competitive Inhibitor of α-Ketoglutarate-Dependent Dioxygenases. Cancer Cell 19, 17–30 (2011).
6. Saha, S. K. et al. Mutant IDH inhibits HNF-4α to block hepatocyte differentiation and promote biliary cancer. Nature 513, 110–114 (2014).
7. Kats, L. M. et al. Proto-Oncogenic Role of Mutant IDH2 in Leukemia Initiation and Maintenance. Cell Stem Cell 14, 329–341 (2014).
8. Stadtfeld, M., Maherali, N., Breault, D. T. & Hochedlinger, K. Defining Molecular Cornerstones during Fibroblast to iPS Cell Reprogramming in Mouse. Cell Stem Cell 2, 230–240 (2008).
9. Stadtfeld, M., Maherali, N., Borkent, M. & Hochedlinger, K. A reprogrammable mouse strain from gene-targeted embryonic stem cells. Nat Meth 7, 53–55 (2009).
10. Utikal, J. et al. Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. Nature 460, 1145–1148 (2009).
11. Stadtfeld, M. et al. Ascorbic acid prevents loss of Dlk1-Dio3 imprinting and facilitates generation of all–iPS cell mice from terminally differentiated B cells. Nature Genetics 44, 398–405 (2012).
12. Rohle, D. et al. An Inhibitor of Mutant IDH1 Delays Growth and Promotes Differentiation of Glioma Cells. Science 340, 626–630 (2013).
13. Losman, J. A. et al. (R)-2-Hydroxyglutarate Is Sufficient to Promote Leukemogenesis and Its Effects Are Reversible. Science 339, 1621–1625 (2013).
14. Kernytsky, A. et al. IDH2 mutation-induced histone and DNA hypermethylation is progressively reversed by small-molecule inhibition. Blood 125, 296–303 (2015).
15. Jha, A. K. et al. Network Integration of Parallel Metabolic and Transcriptional Data Reveals Metabolic Modules that Regulate Macrophage Polarization. Immunity 42, 419–430 (2015).

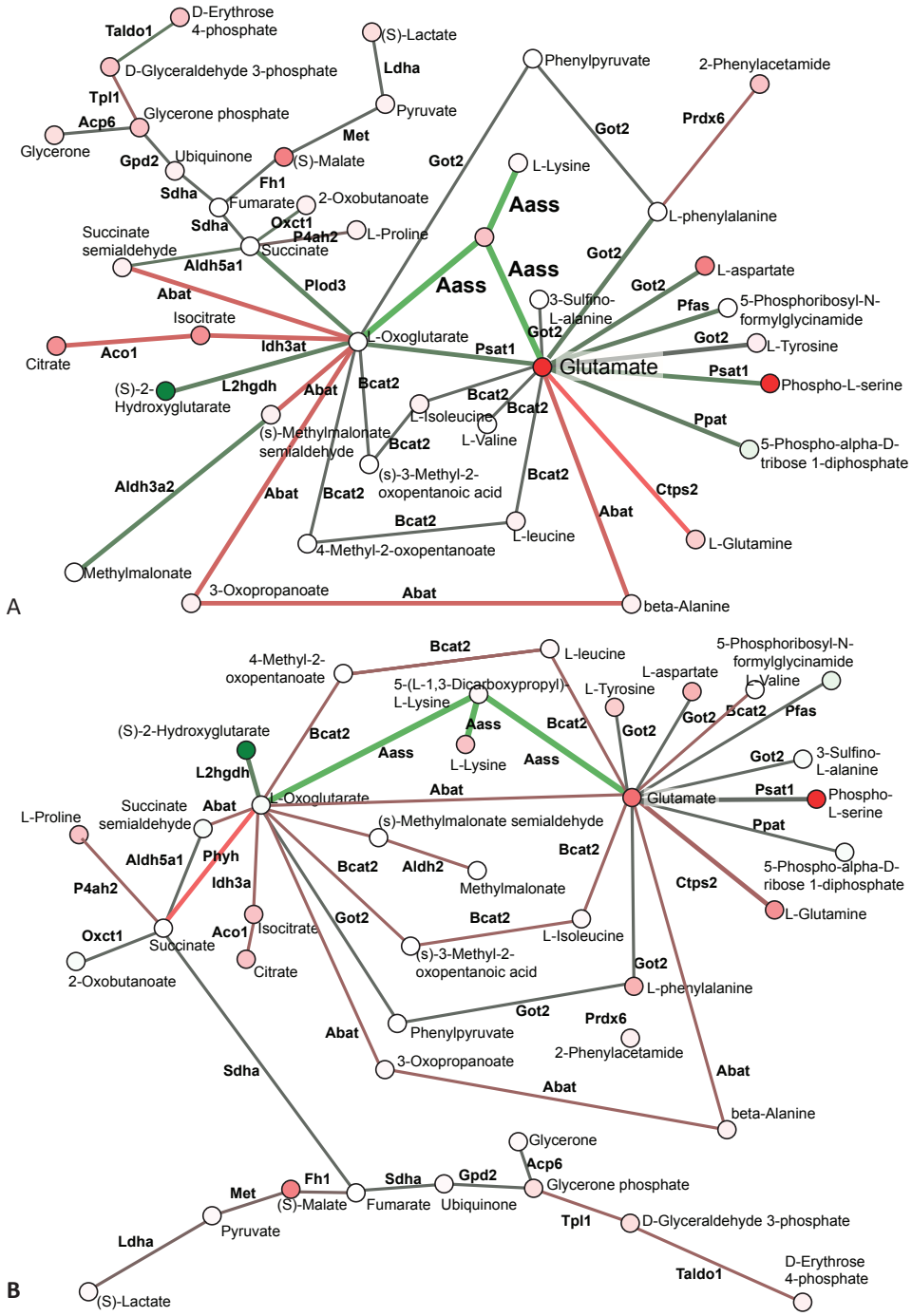


Figure 7 (Continued on next page)

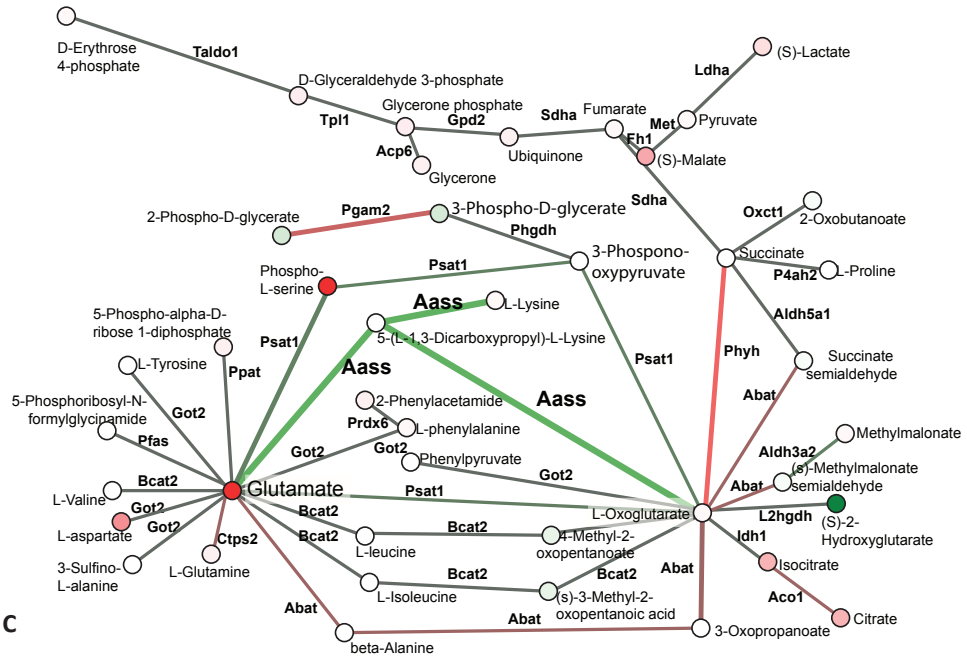


Figure 7. COMBI analysis of transcriptional and metabolic profiling highlights metabolic pathways altered by respectively (A) IDH1 R132H, (B) IDH2 R140Q or (C) IDH2 R172K expression. Lines represent enzymatic conversions by enzymes in bold; dots represent metabolites. Green= higher expressed or more abundant in IDH1 R132H expressing reprogramming cells compared to empty vector control; red= less expressed or abundant. Note that 2-Hydroxyglutarate (2-HG; green dot on left) is overabundant; L2-HGdh converts 2-HG to α-KG and is overexpressed. Aass is overexpressed and its substrate glutamate is low in abundance.

Methods

Tissue culture and virus production

Reprogrammable Mouse Embryonic Fibroblasts (repMEFs) were derived from day E13.5-15.5 mouse embryos carrying the *Col1a1*-tetOP-OKSM and *Rosa26*-M2rtTA alleles[9]. Reprogramming was induced in RepMEFs by adding 20ng/ml doxycycline (dox) and, where indicated, 25ug/ml L-Ascorbic acid (Sigma A4544-25G) to ESC media (Knockout DMEM, 1,000U/ml LIF, 20% FBS). RepMEFs were expanded under hypoxic (4% oxygen) conditions until dox induction. MEFs were infected with overexpression vectors at passage 4, allowed to recover for 16 hours, selected for vector expression by drug selection, harvested, counted and seeded at a density of 20K cells per well of a 6-well plate. Media was changed every 2 days. Dox was withdrawn by removing media, washing with 1xPBS, and continuing

culture in ESC media after 10-12 days. For virus production from overexpression vectors, 293T cells were seeded and transfected at 50% confluency with PEI (Polyethylenimine) and DNA (vector + VSV-G and delta 8.9) at a ratio of 3:1 in Optimem media (Life Technologies). After 24 hours, supernatant was collected through a 0.4um filter and precipitated with PEG (Polyethylglycol) and kept at -80C for later infections. Viral transductions were performed by spin infection for 30 minutes at 2,150rpm at room temperature. Virus was washed away and fresh selection media was added ~16 hours after infections.

Flow cytometry

To determine viral infection efficiency and cell numbers before initiation of reprogramming, infected MEFs were harvested using 0.25% trypsin-EDTA (Life Technologies) and kept at 4C. To prepare growth curves, 3 replicates were harvested at each time point to measure cell counts (DAPI-negative live cells). Intermediates of reprogramming were analyzed by staining with THY1-Viogreen (BD), and SSEA1-APC (Biolegend) (1:200 for 30min. at 4C). To measure the fraction of apoptotic cells, BD's Annexin V kit was used according to the manufacturers' instructions. All cytometry data was analyzed and plotted using FlowJo software.

Quantification of reprogramming efficiencies

For macroscopic detection of iPSC colonies, Alkaline Phosphatase (AP) staining was carried out according to manufacturer's instructions using the Vector Labs AP staining kit (#5100). AP staining was always performed 2 to 4 days after dox withdrawal to eliminate partially reprogrammed colonies and to score for transgene-independent iPSCs.

RNA expression analysis, Metabolic analysis, COMBi analysis

For global gene expression analysis, total RNA was isolated from indicated samples using RNeasy Mini Kit (Qiagen). RNA sequencing, metabolite harvesting and LCMS analysis, as well as COMBI analysis were performed as described before[15].

Chapter 7

Failure to replicate the STAP cell phenomenon

Alejandro De Los Angeles*, Francesco Ferrari*, Yuko Fujiwara, Ronald Mathieu, Soohyun Lee, Semin Lee, Ho-Chou Tu, Samantha Ross, Stephanie Chou, Minh Nguyen, Zhaoting Wu, Thorold W. Theunissen, Benjamin E. Powell, Sumeth Imsoonthornruksa, Jiekai Chen, Marti Borkent, Vladislav Krupalnik, Ernesto Lujan, Marius Wernig, Jacob H. Hanna, Konrad Hochedlinger, Duanqing Pei, Rudolf Jaenisch, Hongkui Deng, Stuart H. Orkin
Peter J. Park & George Q. Daley

*These authors contributed equally to this work.
Received 10 November 2014; accepted 22 July 2015.
ARISING FROM H. Obokata et al. Nature 505, 641–647 (2014) doi:10.1038/nature12968; retraction 511, 112 (2014) doi:10.1038/nature13598;
and H. Obokata et al. Nature 505, 676–680 (2014) doi:10.1038/nature12969; retraction 511, 112 (2014) doi:10.1038/nature13599

Introduction to the publication 'Failure to replicate the STAP cell phenomenon'

Previous chapters described studies into reprogramming mechanisms and factors involved in a well-established reprogramming model, using Oct4, Klf2, Sox2 and c-Myc overexpression to derive pluripotent stem cells. The next chapter describes a concerted effort to validate an alternative reprogramming method, thereby testing the role of cellular stress in reprogramming.

Early 2014, Nature reported in two back-to-back publications the generation of pluripotent cells upon cellular stress induced by acidic conditions or mechanical disturbance. The authors from renowned centers of Brigham and Women's hospital in Boston, USA and the Riken Institute in Kobe, Japan announced that stress induced a pluripotent cell type from murine somatic cells called a 'STAP (stress acquired pluripotent-) cell'. These cells could be converted to STAP-stem cells by specific culture conditions. STAP-stem cells (STAP-SCs) behaved like embryonic stem cells (ESCs) and furthermore, could form intra and extra-embryonic tissues, suggesting they were totipotent. Since the tools and techniques used for these publications were readily available to research groups in the ESC- and reprogramming fields, and the potential implications for reprogramming with such simple triggers were considerable, our groups and many others immediately tried to reproduce the results.

In total, we made 6 attempts to generate STAP or STAP-SCs from fibroblasts or splenocytes. To detect pluripotency, pluripotency reporters Oct4-GFP or Sox2-GFP-containing knock-in lines were used. We used the protocol from the Nature publications, the protocol later published by the Riken Institute as well as the protocol published by the Vacanti laboratory; we also performed the latter protocol in Vacanti's lab under guidance of his labmembers. We analyzed resulting cells by live fluorescence microscopy, with knock-in reporters or cell surface markers, by immunofluorescence staining of fixed cells and by flow cytometry of live cells. Our experiments convinced us that the STAP publications were not reproducible. In July 2017, 6 months after online publication, the papers were retracted by Nature, following investigations of the Riken Institute pointing to several flaws in the data. In the months before, the online community of scientists had discredited the papers already by critical analysis of the data, failed reproduction attempts and even an opinion monitor on a popular stem cell blog (https://ipscell.com/2014/05/does-europe-know-something-about-stap-cells-that-we-dont-latest-poll-results/).

In this chapter, a publication in Nature is presented that resulted from collective data of worldwide stem cell laboratories that tried and failed to make STAP cells; including deeper analysis by the Daley group into alternative explanations for the data presented in the STAP papers.

Table 1. Global efforts to replicate STAP cell generation

Lab	Mouse strain tested	Cell types tested	pluripotency reporter tested	STAP generation protocol tested	Pluripotency assessment essay	≤ of experimental replicates	Conclusion
Jaenisch (MIT, USA)	C57BL6/129 JAX008214	MEFs, Neonatal Spleen, Neonatal adipocytes, Neonatal Fibroblasts	*Oct4*-EGFP Knockin (Jaenisch)	Obokata Article; Obokata, Sasai and Niwa Protocol	Epifluroescence / FACS	17	Negative
Wernig (Stanford, USA)	C57BL6/129 JAX008214	P4 MEFs	*Oct4*-EGFP Knockin (Jaenisch)	Obokata Article	Epifluorescence / FACS	3	Negative
Pei (Guangzhou, China)	C57BL6/CBA JAX004654	Spleen MNCs, CD45+ spleen cells, MEFs, pre-iPSCs, EpiSCs, mammary epithelial cells	GOF18-*Oct4*ΔPE –GFP transgene (Scholer/ Mann)	Obokata Article; Obokata, Sasai and Niwa Protocol	Epifluorescence / FACS	11	Negative
Deng (Peking, China)	C57BL6 x ICR	MEFs, Neonatal fibroblasts, Neonatal bone marrow, Neonatal brain, Neonatal cardiac muscle cells, Neonatal heart, Neonatal lung, Neonatal Spleen, Adult Spleen	GOF18-*Oct4*-GFP transgene (Scholer -Identical to Obokata Article reporter)	Obokata Article; Obokata, Sasai, and Niwa protocol	Epifluorescence/ FACS / qPCR	37	Negative
Hanna (Weizmann, Israel)	C57BL6/CBA JAX 004654	Neonatal CD45+ spleen cells, Neonatal liver, Neonatal brain	GOF18-*Oct4*ΔPE –GFP transgene (Scholer/ Mann)	Obokata Article	Epifluorescence / FACS / STAP-SC	5	Negative
Hochedlinger (Harvard, USA)	C57BL6/129 JAX 008214	Neontal CD45+ splenocytes, E14.5 MEFs	*Oct4*-EGFP Knockin	Obokata Article; BWH protocol; Obokata, Sasai and Niwa protocol	Epifluorescence/ IF / FACS	6	Negative
Daley (Harvard, USA)	C57BL6/CBA JAX 004654	Neonatal CD45+ splenocytes, Neonatal lung cells, MEFs, Neonatal Brain, Neonatal Heart, Neonatal Liver	GOF18-*Oct4*ΔPE –GFP (Scholer/ Mann)	Obokata Article; BWH protocol; Obokata, Sasai and Niwa protocol	Epifluorescence / IF / FACS / qPCR / Teratoma / Chimera	54	Negative

Abstract

Although the reports that stress (such as exposure to acid) can coax somatic cells into a novel state of pluripotency[1,2] have been retracted[3,4], the validity of stimulus-triggered acquisition of pluripotency (STAP) remains unclear. Here we describe the efforts of seven laboratories to replicate STAP, including experiments performed within the laboratory where STAP first originated, as well as re-analysis of the sequencing data from the STAP reports. Neonatal cells treated with two STAP protocols exhibited artefactual autofluoresence rather than bona fide reactivation of an *Oct4* (also known as *Pou5f1*) and green fluorescent protein (GFP) transgene reporter, did not reactivate pluripotency markers towards embryonic stem (ES)-cell-like levels, and failed to generate teratomas or chimaerize blastocysts. Re-analysis of the original RNA sequencing (RNA-seq) and chromatin immunoprecipitation sequencing (ChIP-seq) data identified discrepancies in the sex and genetic composition of parental donor cells and converted stem cells, and revealed a STAP-derived cell line to be a mixture containing trophoblast stem cells, attesting to the importance of validating the properties and provenance of pluripotent stem cells using a wide range of criteria.

Results

Replication efforts

To assess the reprogramming capacity of STAP protocols, we used a transgenic *Oct4*-GFP reporter, which shows GFP reactivation during *Oct4*/*Sox2*/Klf4 reprogramming, in established induced pluripotent stem (iPS) cells and in the gonads of mid-gestation 'all iPS cell' embryos generated by tetraploid complementation[5–7] (Extended Data Figures 1 and 2a). Working within the Vacanti laboratory where the concept of STAP cells originated, and assisted by a co-author of the STAP papers, a Daley laboratory member (A.D.L.A.) attempted to replicate two reported STAP protocols: (1) mechanical trituration and acid treatment of mouse lung cells (Brigham and Women's Hospital (BWH) protocol; see Supplementary Information), and (2) acid treatment of mouse splenocytes (RIKEN protocol; Methods and Extended Data Figure 2b). Seventy-two hours after stress treatment of lung cells, floating spheres appeared amidst cellular debris. Fluorescence microscopy revealed that both *Oct4*-GFP and wild-type spheres emitted lowlevel broad spectrumfluorescence detectable within both green and red filters, indicating autofluorescence (Figure 1a). Untreated *Oct4*-GFP ES cells did not emit the same low-level broad spectrum fluorescence as STAP-treated cells. STAP-treated splenocytes formed spheres with lower efficiency, but also appeared autofluorescent. Flow cytometry indicated STAP-treated *Oct4*-GFP cells did not exhibit *Oct4*-GFP reactivation at levels comparable to control *Oct4*-GFP mouse ES cells, and were indistinguishable from

stressed wildtype controls (Figure 1b). Absence of ES-cell-like levels of *Oct4*, *Sox2* and *Nanog* transcripts and nonspecific immunofluorescence corroborated flow cytometry data (Extended Data Figure 2c, d).

Rare pluripotent cells should generate teratomas in immunocompromised mice[8,9], but STAP cells could not, unlike control ES cells (Extended Data Figure 2e, f). Replication of the poly-L-glycolic acid (PLGA)-based teratoma production method described in the original STAP reports with GFP cells to distinguish host and donor contribution produced distinct masses of connective tissue, muscle and scar, with minimal GFP content, indicating primarily host origin (Figure 1c, d and Extended Data Figure 2g). Rare GFP-positive clusters did not form differentiated tissues characteristic of ES-cell-derived teratomas (Figure 1d). Autofluorescent spheres failed to enter development after morula aggregation or blastocyst injection (Figure 1e and Extended Data Figure 2h–j). vTherefore, pluripotency was undetectable in STAP experiments.

Six other laboratories (Deng, Hanna, Hochedlinger, Jaenisch, Pei and Wernig) also attempted to generate STAP cells (Table 1) andmade the following observations. First, autofluorescent sphere-like aggregates after STAP treatment were universally seen. Second, transgenic reporters used by Obokata and colleagues (GOF18-*Oct4*-GFP, containing the 18-kilobase genomic *Oct4* fragment (GOF18)) and by the Daley, Pei and Hanna laboratories (GOF18-*Oct4*DPE-GFP, lacking the *Oct4* proximal enhancer (PE) element) both exhibit activity in pre-implantation embryos, early post-implantation epiblast cells (embryonic day (E) 5.5), germ cells, and mouse ES/iPS cells; however, differential activity in late post-implantation epiblast (E6.5) and early passage mouse epiblast- derived stem cells has been ascribed to the *Oct4* proximal enhancer[10–12]. Using the same reporter as Obokata and colleagues[1,2], the Deng laboratory observed that the GFP signal in chemical iPS cells was easily distinguishable from the autofluorescence of STAP-treated cells (Extended Data Figure 2k). The Jaenisch, Wernig and Hochedlinger laboratories failed to observe GFP reactivation with *Oct4* or *Nanog* knock-in reporters, excluding a scenario of uncoupling between GFP and endogenous pluripotency expression10. Despite a range of tested reporters, no group documented authentic *Oct4*/*Nanog* reporter activation that resembled bona fide ES cells.

Third, the Deng laboratory failed to observe *Oct4*, *Sox2* and *Nanog* induction 3 and 7 days after STAP treatment, reducing the likelihood that pluripotency was transiently activated and silenced by day 7 (Extended Data Figure 2l). Finally, the Hanna, Wernig and Hochedlinger laboratories failed to generate stem-cell lines by culturing STAP-treated cells in leukaemia inhibitory factor (LIF) and adrenocorticotropic hormone (ACTH)-supplemented medium. In summary, 133 replicate attempts failed to document generation of ES-cell-like cells, corroborating and extending a recent report[13].

(Continued from right)
(A) STAP treatment produces fluorescent signal detected in both FITC (green) and TRITC (red) channels in STAP-treated *Oct4*-GFP and wild-type (WT) cells, consistent with autofluorescence. TRITC signal was not detected in control *Oct4*-GFP mouse ES cells. Note saturation of the green signal in *Oct4*-GFP ES cells at the higher exposure time required to detect FITC from autofluorescent spheres. (B) Absence of ES-cell-like *Oct4*-GFP reactivation. Representative flow cytometry results 7 days after STAP treatment of lung cells or splenocytes (BWHor RIKEN protocol, respectively) without singlet/doublet exclusion and live/dead-cell discrimination. GFP gates were calibrated based on control *Oct4*-GFP ES cells grown on feeders. Whereas control ES cells are bright and situated at approximately 13105 (arbitrary units), no event resembling *Oct4*-GFP ES cells was detected after STAP treatment. One replicate per protocol is shown. iMEFs, irradiated MEFs. (C) STAP-treated cells do not form teratomas using PLGA-based teratoma production methods[1]. Photograph of controlmouse EScell- derived teratoma (top left) and non-teratoma STAP-PLGA mass (bottom left). Representative haematoxylin and eosin (H&E) stainings of a control mouse ES-cell-derived teratoma (top right) and the non-teratomaSTAP-PLGA mass (bottom right). (D) STAP-PLGA mixtures present no indication of ES-celllike in vivo differentiation capacity after injection into immunocompromised mice. Note lack of organization into representative tissue structures typically observed in ES-cell-derived teratomas. DAPI, 49,6-diamidino-2-phenylindole. (E) STAP-treated lung cells fail to incorporate into preimplantation embryos after morula aggregation. (F) Analysis of sequencing data. Samples are classified based on copy number and genotype. STAP cells, STAP stem cells (STAP-SCs) and ES cells share similar characteristics for genotype and copy number of chromosome X. (G) Copy number (CN) profiles, reported as a log2 ratio (observed to expected read counts), derived using ChIP-seq input data. Red/ green correspond to significant amplifications and deletions (log2(CN)≥0.2 or ≤20.2 and P≤0.01), respectively. Grey denotes non-significant variants. Note the amplifications of chromosomes 8 (FI-SCs) and 6/11 (TSCs) and the single copy of chromosome X in STAP cells, STAP-SCs, FI-SCs and ES cells. (H) SNVs inferred from RNA-seq data using the mouse reference genome (derived fromC57BL6 strain). The selected SNVs are classified as homozygous for reference allele (0/0 genotype), homozygous for alternative allele (1/1 genotype) or heterozygous (0/1 genotype). Samples are clustered based on the sum of edit distance between each SNV. Note that each pair of replicates is always grouped together. A subset of samples (CD451, STAP, STAP-SCs and ES cells) shows prevalence of heterozygous alleles (A); FI-SC samples have prevalence of homozygous alleles for the reference variant (B); and, TSC and epiblast stem cell (EpiSC) samples have a larger number of homozygous alternative alleles (C). (I) Contamination in the FI-SC samples with TSCs. The expected frequency of reads covering the alternative allele for heterozygous SNVs is ,50%, which is observed in all samples including TSCs (left). In FISCs, it was ,12% (Extended Data **Figure** 3), suggesting false-positive calls or contamination. The alternative allele frequency distributions of TSC homozygous and heterozygous SNVs sets in FI-SCs (right) show peaks at 9% and 4%, respectively. These results indicate that FI-SC samples are approximately 10% contaminated by TSC samples. Original magnifications, 320 (a, (D) e) and 34 (c).

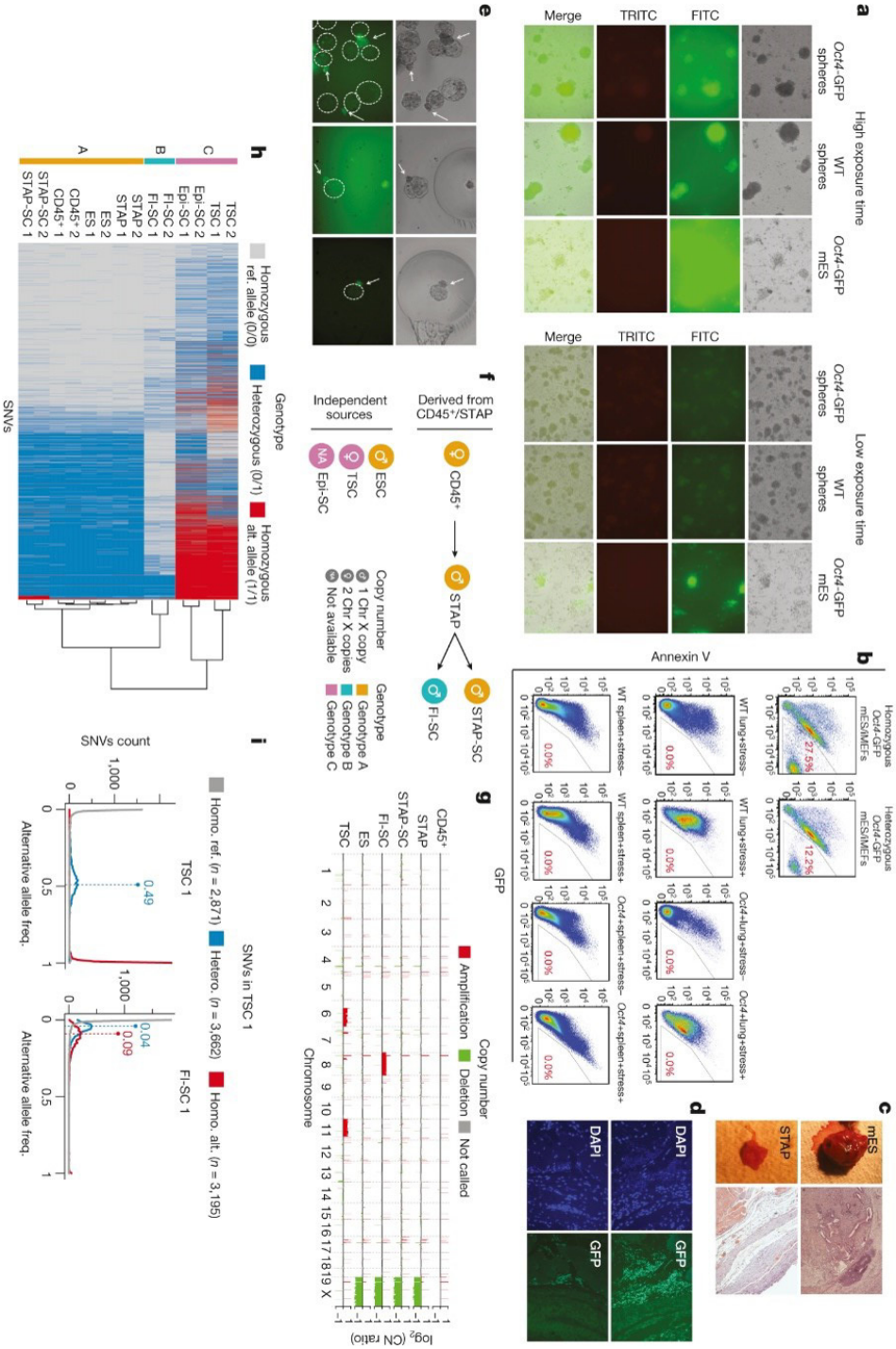


Figure 1. Characterization of the STAP cell phenomenon. See Supplementary Information for further details. (Continued on left)

Computational analysis of sequencing data

We re-examined the high-throughput sequencing data from the STAP reports to investigate the genetic provenance of parental CD451 cells and converted STAP cells, STAP stem cells and FGF4-induced stem cells (FI-SCs) (Figure 1f). Comparative genomic hybridization array data mentioned in the original paper1 were not publicly released. Copy number variation (CNV) analysis conducted using ChIP-seq input samples revealed a discrepancy in sex across samples as well as chromosomal aberrations (Figure 1g). In the original STAP reports, the authors stated that they mixed CD451 cells from male and female mice owing to the small number of CD451 cells retrieved from individual neonatal spleens. However, our analysis indicates that CD451 cells were female, whereas the derived cells (STAP cells, STAP stem cells and FI-SCs) were all male, a clear inconsistency. We note that control ES cells were also male (Figure 1g). FI-SCs possessed trisomy 8, which renders mouse ES cells germline-incompetent14 (Figure 1g). Inferred single nucleotide variants (SNVs) from RNA-seq data allowed classification of samples as genetically similar or dissimilar (Figure 1h). Control ES cells, parental donor female CD451 cells, STAP cells, and STAP stem cells all possessed similar SNV profiles, consistent with their derivation from a first generation hybrid of C57BL6/129 strains, the reported genotype (Figure 1h and Extended Data Figure 3). By contrast, FI-SCs had an SNV profile that matched a single nucleotide polymorphism (SNP) profile of C57BL6 strain origin, indicating distinct genetic provenance from parental CD451 and STAP samples (Figure 1h and Extended Data Figure 3).

Independently sourced control epiblast stem cells and trophoblast stem cells (TSCs) had SNV profiles divergent from the CD451 and STAP sample cohort, as expected (Figure 1h). An anomalous allele frequency distribution observed in FI-SCs, and reciprocal analyses of FI-SC heterozygous SNVs in TSCs and TSC homozygous and heterozygous SNVs in FI-SCs, revealed that FI-SCs were derived from a C57BL6 strain origin, with approximately 10% contamination from TSCs (Figure 1i and Extended Data Figure 3). These are concordant with the findings from a recent RIKEN report (http://www3.riken.jp/stap/e/c13document52.pdf). This contamination with TSCs explains the high-grade placentaforming capacity reported for the Fl-SCs2, an unusual feature that implied totipotency, but which seems to have been due to admixture of cells.

Conclusion

In summary, our replication attempts and genetic analysis indicate that existing STAP protocols are neither robust nor reproducible. To substantiate future claims of reprogramming and alternative states of potency, we urge a rigorous application of several independent means for validating functional pluripotency and genomic profiling to confirm cell line provenance.

ultimately, the essential standard of robustness and reproducibility must be met for new claims to exert a positive and lasting influence on the research community.

Timeline

The STAP papers[1,2] were published in Nature on 29 January 2014 (both retracted[3,4] in July 2014). Below, we provide the time period in which laboratories attempted to replicate STAP cell generation. The Jaenisch laboratory performed STAP replication experiments from February to April 2014. The Hanna laboratory performed STAP replication experiments from February to March 2014. The Wernig laboratory performed STAP replication experiments in February 2014. The Deng laboratory performed STAP replication experiments from February to July 2014. The Pei laboratory performed STAP replication experiments from February to May 2014. The Hochedlinger laboratory started STAP replication experiments on the day of STAP paper publication (29 January 2014) and continued to April 2014. The Daley laboratory performed STAP replication experiments and characterization of the STAP phenomenon from February to November 2014.

Methods

Note on Protocols for STAP Cell Generation

Three different protocols have been reported for generating STAP cells: 1) mechanical trituration and low pH treatment (BWH protocol employed in the Vacanti laboratory as described here); 2) low pH treatment (RIKEN protocol (found at the following URL: http://dx.doi.org/10.1038/protex.2014.008)); and 3) mechanical trituration and ATP treatment (revised BWH protocol released online September 3, 2014 http://research.bwhanesthesia.org/research-groups/cterm/stap-cell-protocol). We note that the data in our study comprises the first two protocols. The BWH protocol refers to an unpublished, in-house protocol created and used by the Vacanti laboratory to generate STAP cells. The RIKEN protocol is the published version described in the original STAP reports (Obokata et al, 2014; Obokata et al, 2014). The Daley laboratory also attempted the revised BWH protocol without success (data not shown).

The three protocols contain differences and similarities. The major differences concern the distinct combinations of physical and chemical stressors. For example, the BWH protocols emphasized a combination of chemical and physical stressors, such as extensive mechanical trituration, whereas the RIKEN protocol primarily consisted of low pH treatment. However, in all three protocols, somatic cells are treated with a transient acid bath, although the composition of the low pH solution varied across protocols. Whereas the

original BWH protocol and RIKEN protocol generate low pH solutions via titration of HBSS with hydrochloric acid, the revised BWH protocol generates a low pH solution via titration with ATP. The experimental procedures followed for the BWH protocol and RIKEN protocol are described below.

Mechanical Trituration and Low pH Treatment of Lung Cells (BWH Protocol)

To trigger STAP cell generation by mechanical trituration and low pH treatment of lung cells, we followed a procedure involving tissue isolation, extensive mechanical trituration, and an acid bath (pH 5.4). The protocol described below was adapted from the website of the Vacanti laboratory, which can be found at the following URL: (http://research.bwhanesthesia.org/research-groups/cterm/stap-cell-protocol) Lungs were dissected from seven- to ten-day old neonatal *Oct4*-GFP mice and wildtype mice. Dissected lungs from multiple animals were pooled and washed twice in 60 mm petri dishes with Hank's Buffered Saline Solution (HBSS without calcium and magnesium, Life Technologies, 14170-112). To enhance enzymatic dissociation, the whole tissue was minced with scissors for approximately 10 minutes. The minced tissue was treated with collagenase P for 30 minutes and placed in an incubator/shaker for 30 minutes at 37 degrees Celsius at 90 revolutions per minute (rpm). Following collagenase P treatment, HBSS was added and the total tissue suspension was centrifuged to create a pellet. After aspirating the supernatant, the pellet was resuspended into a small volume of HBSS (approximately 500 µl to 1000 µl). The resulting suspension was forcefully triturated for five minutes using a Pasteur pipette to further dissociate cell aggregates and debris. The triturated cell suspension was further triturated through a series of flame-pulled Pasteur pipettes. Generally, the trituration was performed for thirty minutes starting with the flame-pulled Pasteur pipette with the largest diameter. After sufficient dissociation was achieved using the large diameter pipette, trituration was performed with pipettes with increasingly smaller diameters.

Following mechanical trituration, erythrocyte depletion was performed. HBSS was added to the triturated suspension to increase the total volume to 5 mL (approximately 3 mL of HBSS). An equal volume of Lympholyte (Cedarlane) (5 mL) was added to the bottom of the tube to create a bilayer. The tube was then centrifuged at 1500 g for 10 minutes. After ten minutes, the tube was rotated 180 degrees and re-centrifuged at 1500 g for an additional 10 minutes. Following the second centrifugation, erythrocytes formed a pellet at the bottom of the tube. The layer between the HBSS and Lympholyte was transferred into a new 50 mL tube and HBSS was added to a total volume of 20 mL. The solution was then centrifuged at 1200 rpm for 5 minutes and the supernatant was aspirated. The cells were re-suspended into acid solution (HBSS titrated to a pH of 5.4 by hydrochloric acid) and placed into an incubator at 37 degrees Celsius for 25 minutes. The pH of the experiment was periodically monitored

using a pH meter to confirm that the pH did not rise during the course of the experiment. If the pH increased significantly, the cellular suspension was centrifuged and the resulting pellet was resuspended into a fresh batch of acid solution. After the 25 minute acid bath, the cell suspension was centrifuged at 1200 rpm for 5 minutes. The supernatant was aspirated and the pellet was re-suspended into sphere medium at a density of approximately 100,000 cells per mL of sphere medium.

Sphere medium consisted of DMEM/F12 medium (Stem Cell Technologies) supplemented with 2% B27 supplement (Life Technologies), penicillin/streptomycin (Life Technologies), bFGF (20 ng/mL, Life Technologies), EGF (20 ng/mL, Peprotech), heparin (0.2%, Stem Cell Technologies), and LIF (1000 U, Sigma). To reduce attachment to the bottom of the non-adhesive plate and facilitate sphere formation, STAP-treated cultures were gently pipetted using a 5 mL pipette, twice per day for 2 minutes. The medium was refreshed every two days, unless otherwise noted.

Low pH Treatment of Splenocytes (RIKEN protocol)

We attempted to follow procedures as closely as possible to the original STAP paper published by Obokata and colleagues and the refined protocol published by Obokata, Sasai, and Niwa. The following protocol is adapted from and is essentially identical to the refined protocol published by Obokata, Sasai and Niwa (found at the following URL: http://dx.doi.org/10.1038/protex.2014.008).

To isolate hematopoietic cells, spleens were dissected and surgically excised from seven- to ten-day old neonatal *Oct4*-GFP mice and wild-type mice. Spleens from multiple animals were pooled and minced with surgical scissors to facilitate tissue dissociation. We mechanically dissociated the spleen tissues by triturating the minced spleens with Pasteur pipettes. Dissociated spleen cells were re-suspended with PBS (Life Technologies) and strained through a cell strainer (BD Biosciences). Pelleted cells were re-suspended in DMEM or HBSS media (Life Technologies).

For erythrocyte depletion, a similar protocol to the one used for processed lung tissues was applied to minced spleen specimens. An equal volume of Lympholyte (Cedarlane) was added to the bottom of the tube, causing two layers to form. The tube containing two layers was then centrifuged at 1000 g for 20 minutes. The lymphocyte layer was taken out and when indicated, stained with a CD45 antibody (PE-conjugated anti-CD45 antibody (Abcam) or PerCP-Cy5.5-conjugated anti-CD45 antibody (eBiosciences)) and sorted for CD45 on a FACSAria II (BD Biosciences). The resulting cells (unsorted splenocytes or CD45-sorted splenocytes) were treated with 500 µl of low-pH HBSS solution (titrated to pH 5.4 by hydrochloric acid (HCl)) for 25 minutes at 37 degrees Celsius. Throughout the experiment, the pH was periodically monitored to confirm that the pH did not increase significantly

during the course of the experiment. If the pH rose significantly, the cellular suspension was centrifuged and the pellet was re-suspended into a fresh batch of acid solution.

After the 25 minute acid bath treatment, treated cells were centrifuged at 1000 rpm for 5 minutes at room temperature. The supernatant was removed by aspiration and pelleted cells were re-suspended in DMEM/F12 medium supplemented with LIF and 2% B27 supplement at a density of approximately 100,000 cells per mL and plated onto nonadhesive culture plates. The medium was refreshed every 2 days until day 7, unless otherwise noted. The spleen data presented in Figure 1B is from unsorted splenocytes because we observed increased sphere-forming efficiencies when the sorting step was omitted.

Imaging of STAP cultures

STAP-treated lung cells were imaged every two days until seven days after initiation with the BWH protocol. Genetically matched, homozygous *Oct4*-GFP mouse ESCs were used as positive controls. In the Vacanti laboratory, for Figure 1A, imaging was performed with a Biorevo Keyence BZ9000 microscope, and BZII Analyzer software was used to acquire and analyze images. For capturing green and red signals, the following filters were used: Green signal: FITC filter; red signal: the TRITC filter.

Flow cytometry

STAP-treated cells were analyzed seven days after STAP treatment by flow cytometry to assess for reprogrammed *Oct4*-GFP+ cells. To distinguish authentic *Oct4*-GFP signal from autofluorescence, STAP-treated samples were compared to treated wild-type cells, untreated controls, and positive control mESCs. Positive control mESCs harbored an identical transgenic reporter, originated from the same mouse strain of STAP-treated somatic cells, and were used to calibrate the GFP gates for assessment of authentic reporter reactivation. STAP treatment was performed in the Vacanti laboratory and flow cytometry analyses were performed seven days after STAP treatment. Flow cytometry was performed on an LSR II at Children's Hospital Boston.

To prepare STAP-treated cells for flow cytometry analyses, cultures were trypsinized to generate a single cell suspension. Single cell suspensions from lungs, splenocytes, STAP-treated cells, and mESCs were pelleted, washed, and stained with PEconjugated Annexin V antibody (BD Pharmingen, Cat no. 51-65875X; and Biovision, Cat No. K128-100) and DAPI (Life Technologies) to detect apoptotic cells and dead cells, respectively. GFP gates were set based on the GFP excitation/emission profile of positive control homozygous and heterozygous transgenic *Oct4*-GFP mESCs (cultured on iMEFs) containing the identical reporter present in STAP-treated somatic cells. The data presented in Figure 1B has no singlet/doublet exclusion or live/dead cell exclusion to enable comprehensive analysis for

any GFP signal that resembles the GFP signal of *Oct4*-GFP mESCs. FACS data were analyzed with FlowJo software (TreeStar).

Oct4-GFP Mouse ESC Derivation and Cultivation

To replicate the original STAP papers as closely as possible, we used a transgenic *Oct4*-GFP reporter (JAX004654). We used this reporter because the identical *Oct4*-GFP reporter used by Obokata and colleagues was not readily available. The use of homozygous transgenic *Oct4*-GFP mice facilitated experimental throughput and was suggested by Niwa and colleagues to promote the detection of reprogramming events. Mouse ESCs were derived from E3.5 blastocysts of mice both homozygous and heterozygous for the transgenic *Oct4*-GFP reporter to generate positive control cells for STAP replication studies. Because the somatic cells used for STAP replication studies were homozygous, the primary positive control cells used throughout this study were the homozygous *Oct4*-GFP ESCs. The transgenic mouse ESCs harboring the GOF-*Oct4*ΔPE -GFP reporter were isolated by explanting the inner cell mass of an E3.5 embryo into KSR-LIF medium on iMEF feeders. Following ICM outgrowth formation, the outgrowth was trypsinized, plated onto irradiated CF1 mouse embryonic fibroblast (iMEF) feeders (GlobalStem) and continuously expanded in classical mESC conditions (serum/LIF medium). Homozygous and heterozygous *Oct4*-GFP mouse ESCs were cultured in standard mouse ES cell medium (serum/LIF conditions) on iMEF feeders in gelatinized tissue culture dishes, unless otherwise noted (such as in Figure S1) when mESCs were cultured in N2B27-2i/LIF to verify that the batch of B27 supplement used in the Vacanti laboratory can sustain the self-renewal and GFP signal of *Oct4*-GFP mouse ESCs). Classical mouse ESC medium (serum/LIF conditions) contains Knockout DMEM (Life Technologies) supplemented with 15% fetal bovine serum, Penicillin / Streptomycin (Life Technologies), nonessential amino acids (Life Technologies), nucleoside mix (Millipore), L-Glutamine (Life Technologies), and 1000 U of mouse leukemia inhibitory factor (LIF) (Gemini Bioproducts). KSR-LIF medium refers to mouse ESC medium containing Knockout DMEM, 10% Knockout Serum Replacer (Life Technologies), Penicillin / Streptomycin, nonessential amino acids, nucleoside mix, L-Glutamine, and 1000 U of mouse LIF. N2B27-2i/LIF medium refers to a 1:1 mixture of DMEM/F12 and Neurobasal media (Life Technologies), supplemented with 1X N2 supplement (Life Technologies), 1X B27 supplement (Life Technologies), Penicillin / Streptomycin, L-Glutamine, 1000 U LIF (Gemini Bioproducts), 1 uM PD0325901 (Stemgent), 3 uM CHIR99021 (Stemgent). When indicated, LIF and B27 supplements used in our laboratory were substituted with Vacanti laboratory LIF and B27 supplement to validate their suitability for STAP replication experiments. The mouse LIF used in Vacanti laboratory was Sigma, Catalog No. L5158-5 ug (Lot ≤ 021M1557V). The Vacanti laboratory B27 supplement was Life Technologies, Catalog No. 12587-010 (Lot≤ 1582961). Notably Vacanti laboratory B27

supplement lacks Vitamin (A) whereas the Daley laboratory B27 supplement has Vitamin A.

Mouse iPSC Generation and Cultivation

Mouse iPS cells were generated to validate the *Oct4*-transgenic reporter for STAP replication studies. Mouse iPS cells were generated using standard methods16. Retroviruses (pMX-*Oct4*, pMX-*Sox2*, pMX-Klf4, Addgene) were packaged in 293T cells using conventional retroviral production methods. MEFs harboring the GOF-*Oct4*ΔPEGFP reporter were infected with retroviruses encoding reprogramming factors. Infected mouse embryonic fibroblasts were cultivated in KSR-LIF media. Two to three weeks after infection, mouse iPSC colonies were selected based on morphological criteria and GFP reporter activity, expanded and continuously cultivated in serum/LIF conditions on iMEF feeders.

Teratoma formation analyses

Positive control mESCs used for teratoma experiments were derived from the same mouse strain used for STAP replication efforts. Teratoma formation assays were performed similarly to previously described methods, unless otherwise noted8,9. Mouse ESCs and STAP-treated cells were re-suspended in Matrigel (BD Biosciences) prior to implantation for teratoma formation. Mouse ESCs grown on irradiated MEFs were collected by trypsin treatment, re-suspended as single cells and injected subcutaneously into the dorsal flanks or into the kidney capsule of NOD-SCID mice. STAP-treated cultures were injected for teratoma generation seven days after being treated with the STAP protocols. STAP-treated cells were not dissociated, and were instead re-suspended as clumps and injected subcutaneously into the dorsal flanks or into the kidney capsule of NOD-SCID mice. For production of teratomas from mESCs, approximately 100,000 mESCs were typically injected. For injection of STAP-treated cultures, approximately 1 – 2 million cells were injected. After 4 – 8 weeks, teratomas were dissected and fixed in 10% formalin. Embedding in paraffin, sectioning of tissue, and hematoxylin and eosin staining were performed by the Rodent Histopathology service of the Dana Farber Cancer Institute.

For subcutaneous transplantation, the following method was used. Briefly, mice were anesthesized with ketamine/xylazine. The dorsal flanks of the mouse were shaved and the skin was swabbed with Betadyne. After the mouse was shaved, mouse ESCs or STAP cells were injected subcutaneously with approximately 100 – 500 µl of the cellular suspension containing either positive control mouse ESCs or experimental STAP treated cells. Mouse ESC-transplanted mice typically formed teratomas within 3 – 4 weeks, whereas mice transplanted with STAP-treated cells were observed for at least 8 – 12 weeks.

For kidney capsule transplantations, the left flank of the anesthesized mouse was shaved and the skin of the mouse swabbed with Betadyne. After the left kidney was

located, a small incision was made in the skin and the peritoneum to expose the kidney. After exposing the kidney, STAP-induced lung cells, STAP-induced splenocytes, or positive control mouse ESCs were injected into the kidney capsule. The peritoneum and skin were then closed using silk sutures. Mouse ESC-transplanted mice typically formed teratomas within 3 – 4 weeks, whereas mice transplanted with STAP-treated cells were observed for at least 8 – 12 weeks. The original Nature Article describing in vivo differentiation of STAP cells entailed the following: STAP cells were seeded onto a sheet composed of a non-woven mesh of polyglycolic acid fibers (PLGA) and cultured for 24 hours in DMEM containing 10% FBS. This mixture was implanted subcutaneously into the dorsal flanks of immunecompromised NOD-SCID mice.

To replicate the in vivo differentiation experiments described by Obokata and colleagues, we obtained PLGA from a member of the Vacanti laboratory. We seeded STAP-treated GFP+ splenocytes onto a sheet of PLGA mesh, cultured the mixtures for 24 hours in DMEM + 10% FBS, and subcutaneously implanted the mixture into the dorsal flanks of NOD-SCID mice. For subcutaneous implantations, the left flank of the anesthesized mouse was shaved and the skin of the mouse swabbed with Betadyne. An incision was made. The mixture containing the STAP-treated cells and PLGA fibers was implanted subcutaneously, and the incision was closed using silk sutures. The implant was typically placed near the incision site to facilitate implant recovery. After approximately 3 weeks or three months, the implant was recovered from the implantation site. Similar data was obtained between three week and three month STAP-PLGA implant experiments. We dissected the opposite side of the mouse to confirm that the same implant structure was not present. The implants were fixed with 10% formalin. Embedding in paraffin, sectioning, and stainings (H & E staining and Masson's staining) were performed by the Rodent Histopathology service of the Dana Farber Cancer Institute. Additional GFP immunohistochemistry was performed in-house or by the Histology Core at the Harvard Stem Cell Institute. Detailed procedures for the GFP staining of paraffin-embedded STAP masses are described below in the immunofluorescence section of the methods. For validating the pluripotency of GOF18-iPSCs by teratoma formation capacity, similar methods to ESC-based teratoma generation were used to generate GOF18-iPSCderived teratomas.

Morula aggregations and blastocyst injections

The developmental potential of STAP-treated cells was further evaluated in vivo by two methods: morula aggregation and blastocyst injection. Morula aggregations were

performed similarly to previously described methods used for mouse ES cells17. Briefly, STAP cells were aggregated as small clumps with E2.5 morula stage embryos in depression wells. The aggregates were cultured in KSOM embryo medium overnight and incorporation/ contribution was assessed 24 hours later (~E3.5) by fluorescence microscopy. Methods for blastocyst injection were similar to those described in the Nature STAP article with minor modifications. Briefly, the STAP-treated spheres were mechanically "crushed" or disaggregated to generate aggregates of reduced size that can fit into a blastocyst injection needle. Following immobilization of an E3.5 blastocyst using a holding pipette, the end of the injection pipette containing the STAP-treated sphere was juxtaposed against the blastocyst surface and swiftly introduced into the blastocyst cavity. The STAP-treated aggregates were released from the injection needle slowly into the blastocyst cavity. Following withdrawal of the needle from the blastocyst, the injected blastocyst was allowed to recover in an incubator. Recovered embryos were transferred into the uterus of 2.5 dpc pseudopregnant females. Seven to eight days later, females were sacrificed by CO2 asphyxiation and embryos were harvested at ~E10.5 for analysis.

Tetraploid complementation (iPS cells)

For validating the pluripotency of GOF18-iPS cells, tetraploid (4N) blastocysts were placed into KSOM embryo medium. iPS cells were injected into the blastocyst cavity using a microinjection pipette. After blastocyst injection, blastocysts were returned to KSOM medium and cultured at 37 degrees Celsius until transferred to recipient females. qPCR analysis Total RNA was isolated as described in the RNeasy Kit (Qiagen) and reverse transcribed using the Superscript III First Strand Synthesis Kit (Invitrogen). Quantitative RT-PCR analysis was performed in triplicate with FAST SYBR Green Mix (Applied Biosystems). Gene expression was normalized to GAPDH. The primer sets used to generate the data in Extended Data Figure 2 are listed below. We note that alternative primer sets also showed lack of ESC-like *Oct4*, *Sox2*, and *Nanog* induction (data not shown). GAPDH-F: GGTTGTCTCCTGCGACTTCAACAGC GAPDH-R: CGAGTTGGGATAGGGCCTCTCTTGC *Oct4*-F: TTGGGCTAGAGAAGGATGTGGTT *Oct4*-R: GGAAAAGGGACTGAGTAGAGTGTGG *Nanog*-F: GGTTGAAGACTAGCAATGGTCTGA *Nanog*-R :TGCAATGGATGCTGGGATACTC *Sox2*-F: GCACATGAACGGCTGGAGCAACG *Sox2*-R: TGCTGCGAGTAGGACATGCTGTAGG Immunofluorescence Standard immunofluorescence procedures were followed. To stain mouse ES cells and STAP-treated cells, cells were fixed with 4% paraformaldehyde, washed and permeabilized with PBS containing 0.1% Triton-X-100. Next, samples were blocked with PBS containing 5% FBS solution. After overnight incubation in primary antibody solution (PBS containing 1% FBS solution) at 4 degrees Celsius, cells were washed three times, and then incubated with antibody solution containing secondary antibodies. Cell nuclei were visualized with DAPI. The following primary antibody dilutions

were used: mouse monoclonal against *Oct4* (1:100) from Santa Cruz Biotechnology (C:10, Cat No.: sc5279), rat monoclonal against *Nanog* (1:500) from eBiosciences (eBioMLC-51, Cat No.:14-5761-80).

GFP immunostaining of paraffin-embedded STAP masses was performed using the following protocol. Slides with paraffin-embedded sections were de-waxed with xylene and rehydrated through a series of washes with decreasing percentages of ethanol. Antigen retrieval was performed in 10 mM sodium citrate buffer (pH 6.0) by placement in a de-cloaking chamber at 95 degrees Celsius for 30 minutes. Immunohistochemistry was performed with Elite ABC kit and DAB substrate (Vector Laboratories) according to the manufacturer's protocol.

Genomic data analysis

We used the RNA-seq and ChIP-seq data (input control) from Obokata and colleagues, available in the Sequence Read Archive (SRA) database (http://www.ncbi.nlm.nih.gov/sra) with accession number SRP038104. Copy number variation analysis Input (control) samples from the ChIP-seq datasets were used to infer copy number profiles. Array CGH experiments were performed in the original publication but the data were not deposited; our attempt to obtain data through the editor was not successful. Reads were aligned to the mm9 genome using BWA18 (v.0.5.9-r16) and filtered to select uniquely aligned reads ("XT:A:U" flag). Copy number analysis was performed on the read counts using BIC-seq19 after normalization by GC-content and sequence context, with parameters lambda=3 and bin-size=10000. We used 10kb bin size since we expect the read coverage in ChIP input to be uneven due variation in global chromatin status and accessibility. Single nucleotide variant analysis As exome or whole-genome sequencing data were not available, single nucleotide sequence variants (SNVs) were inferred from RNA-seq data. RNA-seq data were aligned to the mouse genome mm10 using Tophat 2.0.10[20,21] allowing two mismatches and two indels. The ENSEMBL GRCm38 gene annotation was used. On average 14.6M reads were aligned (12.9M-16.4M), which corresponds to 84.3% of total reads. Uniquely aligned reads were used, and duplicates were removed. Variants were called using Genome Analysis Toolkit22 3.0 UnifiedGenotyper, with local indel realignment and base quality score recalibration23. Only those variants assigned PASS by variant quality score recalibration (VQSR)23 and with quality score ≥100 were retained. Known mouse SNPs for various strains were obtained from the Wellcome Trust Sanger Institute Mouse Genomes Project[24] and used during the recalibration steps. To remove the effect of variable read coverage (due to varying levels of gene expression), only the 9,739 variants with at least 30X coverage across all samples were considered for downstream analyses.

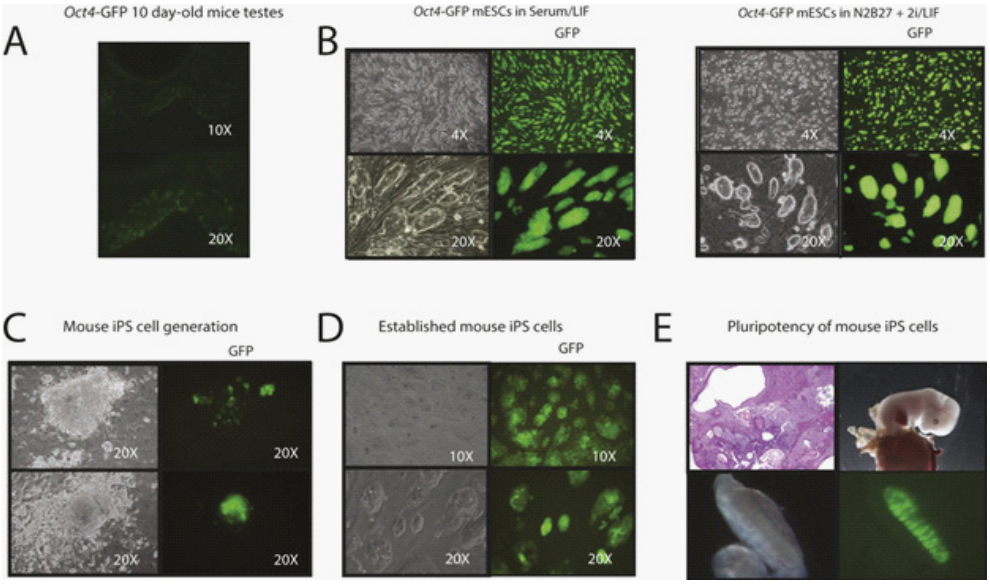
Author Contributions

A.D.L.A. and F.F. contributed equally to this work. A.D.L.A. and
G.Q.D. conceived and designed the project. A.D.L.A. performed experiments within the Vacanti laboratory, where indicated. A.D.L.A., Y.F., R.M., H.-C.T., S.C. and Z.W. analysed STAP experiments. S.R. facilitated teratomainjections. T.W.T., B.E.P., S.I., J.C., M.B., V.K., E.L., M.W., J.H.H., K.H., D.P., R.J. and H.D. contributed STAP replication data. F.F. performed bioinformatics analyses, assisted by So.L. and Se.L., and supervised by P.J.P. A.D.L.A., F.F., P.J.P. and G.Q.D. wrote the manuscript.

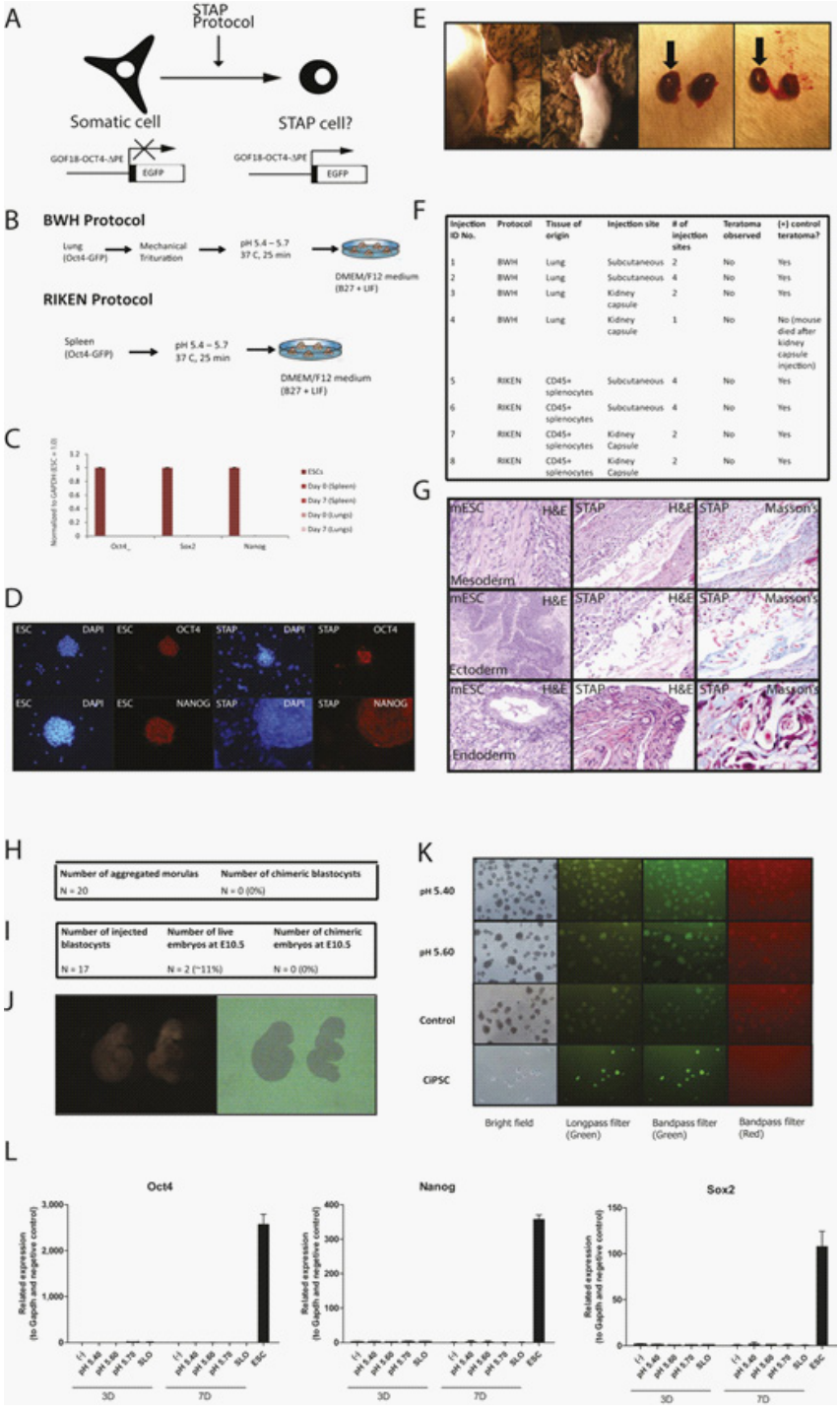
References

1. Obokata, H. et al. Stimulus-triggered fate conversion of somatic cells into pluripotency. Nature 505, 641–647 (2014).
2. Obokata, H. et al. Bidirectional potential in reprogrammed cells with acquired pluripotency. Nature 505, 676–680 (2014).
3. Obokata, H. et al. Retraction: stimulus-triggered fate conversion of somatic cells into pluripotency. Nature 511, 112 (2014).
4. Obokata, H. et al. Retraction: bidirectional potential in reprogrammed cells with acquired pluripotency. Nature 511, 112 (2014).
5. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. Nature 448, 313–317 (2007).
6. Nichols, J. et al. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor *Oct4*. Cell 95, 379–391 (1998).
7. Szabo´ , P. E., Hubner, K., Scholer, H. & Mann, J. R. Allele-specific expression of imprinted genes in mouse migratory primordial germ cells. Mech. Dev. 115, 157–160 (2002).
8. Lawrenz, B. et al. Highly sensitive biosafety model for stem cell-derived grafts. Cytotherapy 6, 212–222 (2004).
9. Cao, F. et al. Spatial and temporal kinetics of teratoma formation from mouse embryonic stem cell transplantation. Stem Cells Dev. 16, 883–892 (2007).
10. Han, D. W. et al. Epiblast stem cell subpopulations represent mouse embryos of distinct pregastrulation stages. Cell 143, 617–627 (2010).
11. Yeom, Y. I. et al. Germline regulatory element of *Oct4* specific for the totipotent cycle of embryonal cells. Development 122, 881–894 (1996).
12. Ohbo, K. et al. Identification and characterization of stem cells in prepubertal spermatogenesis in mice. Dev. Biol. 258, 209–225 (2003).
13. Tang, M. K. et al. Transient acid treatment cannot induce neonatal somatic cells to become pluripotent stem cells. F1000Res. 3, 102 (2014).

14. Kim, Y. M., Lee, J. Y., Xia, L., Mulvihill, J. J. & Li, S. Trisomy 8: a common finding in mouse embryonic stem (ES) cell lines. Mol. Cytogenet. 6, 3 (2013).
15. Hou, P. et al. Pluripotent stem cells induced from mouse somatic cells by smallmolecule compounds. Science 341, 651–654 (2013).



Extended Data Figure 1. Validation of the *Oct4*-GFP transgenic reporter. (A) Context-appropriate expression of the GOF18-*Oct4*DPE-GFP transgene reporter in the testes of 10-day-old neonatal male mice. (B) STAP replication culture reagents sustain *Oct4*-GFP signal in *Oct4*-GFP mouse ES cells. Vacanti laboratory LIF and B27 supplement sustain self-renewal and strong GFP signal of *Oct4*-GFPmouse ES cells in serum/LIF (left) and N2B27 minimal media (see Methods) plus 2i/LIF (MEK inhibitor PD0325901 and GSK3-b inhibitor CHIR99021) (right). (C) Reactivation of the GOF18-*Oct4*DPE-GFP reporter during direct reprogramming of MEFs by *Oct4*, *Sox2* and Klf4. Left, phasecontrast images of founder GOF18-mouse iPS cells. Right, GFP signal in primary GOF18-mouse iPS cells. Note the heterogeneous reactivation of the GOF18-*Oct4*DPE-GFP reporter in primary founder mouse iPS cell colonies (derived in knockout serum replacement/LIF). (D) GOF18-*Oct4*DPE-GFP reporter expression in established mouse iPS cell lines (passage 12). Left, phasecontrast images of established GOF18-mouse iPS cells. Mouse iPS cells were maintained on feeders in serum/LIF media. Right, note GFP signal in GOF18- mouse iPS cells. GFP is observed in essentially all iPS cell colonies and in most cells in each colony. GFP heterogeneity was slightly increased in GOF18-iPS cells compared with GOF18-ES cells. (E) Developmental potential of GOF18-iPS cells. Top left, phase-contrast image of a teratoma generated from GOF18-iPS cells. Original magnification, 20x. Top right, to assess the developmental potential of GOF18-iPS cells, 'all iPS cell embryos' were generated by injection of GOF18-iPS cells into 4N blastocysts ('tetraploid complementation'). A photograph of a live E13.5 embryo generated from GOF18-iPS cells is shown. Bottom row, gonadal contribution in all-iPS-cell embryos indicates GOF18-iPS cells are highly pluripotent. GFP is expressed in E13.5 days post-coitum (dpc) male gonads, and fluorescent cords are visible. The silencing of GFP in surrounding cells re-confirms the context-appropriate expression of the *Oct4*-GFP reporter.

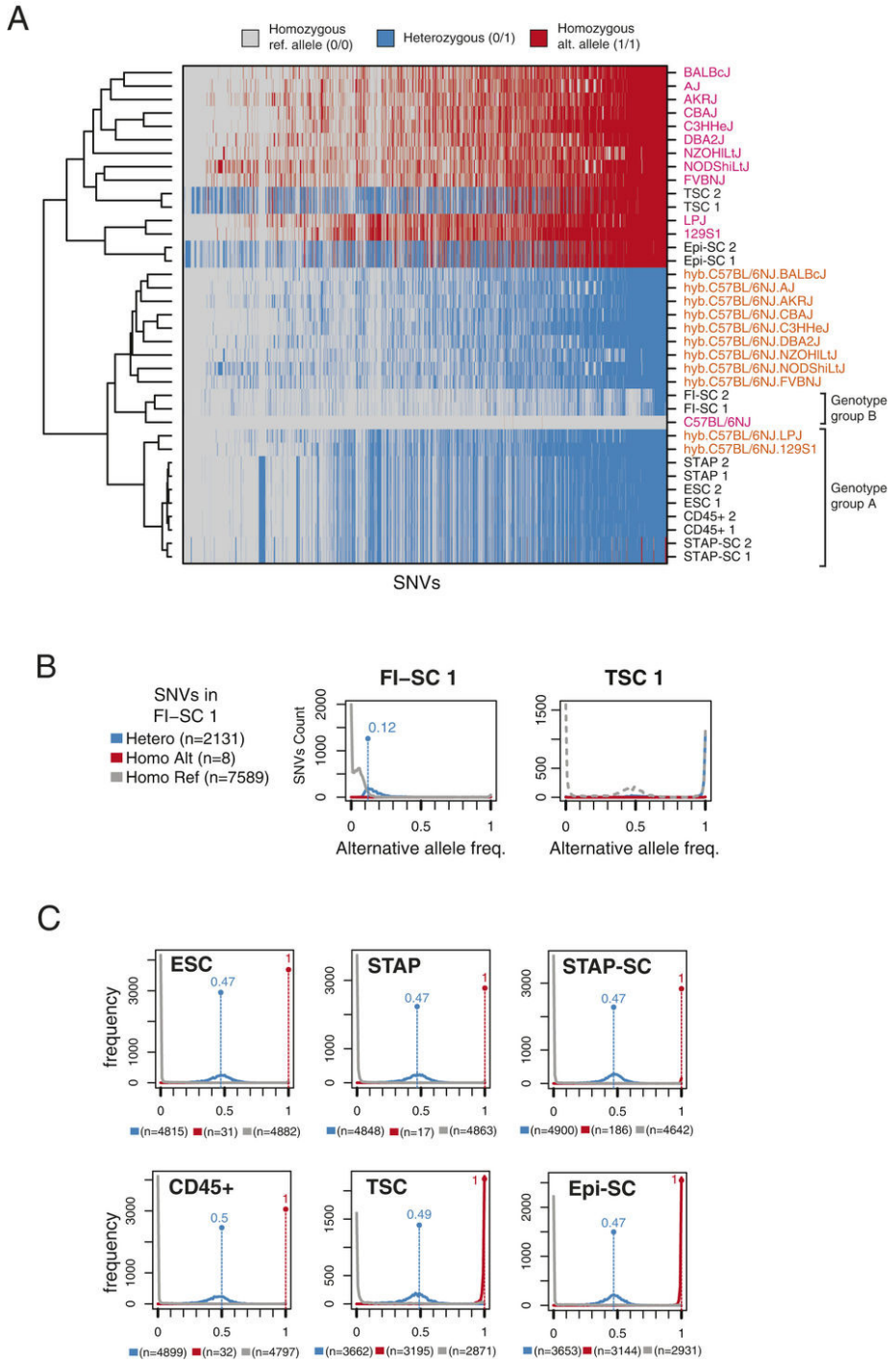


Injection ID No.	Protocol	Tissue of origin	Injection site	# of injection sites	Teratoma observed	(+) control teratoma?
1	BWH	Lung	Subcutaneous	2	No	Yes
2	BWH	Lung	Subcutaneous	4	No	Yes
3	BWH	Lung	Kidney capsule	2	No	Yes
4	BWH	Lung	Kidney capsule	1	No	No (mouse died after kidney capsule injection)
5	RIKEN	CD45+ splenocytes	Subcutaneous	4	No	Yes
6	RIKEN	CD45+ splenocytes	Subcutaneous	4	No	Yes
7	RIKEN	CD45+ splenocytes	Kidney Capsule	2	No	Yes
8	RIKEN	CD45+ splenocytes	Kidney Capsule	2	No	Yes

Extended Data Figure 2. STAP replication data. (legend on next two pages)

Legend extended data Figure 2: (A) Experimental scheme. Reactivation of the transgenic GOF18-*Oct4*DPE-GFP (*Oct4*-GFP) reporter to detect reprogramming after STAP treatment of somatic cells. (B) Two STAP protocol variants. BWH: mechanical trituration and low pH treatment of lung cells. RIKEN: low pH treatment of spleen cells. Stressed cells were plated into non-adhesive dishes and cultured in DMEM/F12 medium plus B27 and LIF. (C) qPCR analysis 7 days after STAP induction. Expression levels of *Oct4*, *Sox2* and *Nanog* transcripts at ES-cell-like levels were not observed in lung cells or splenocytes after treatment with the BWH or RIKEN STAP protocol, respectively. Levels normalized to Gapdh. One replicate per protocol is shown. Although low level *Sox2* and *Nanog* upregulation (1–2 DCt cycles; data not shown) was inconsistently observed, we speculate that minimal induction of *Sox2* and *Nanog* messenger RNA may be due to relaxed transcriptional control in stressed cells. (D) Nonspecific staining observed in STAP-treated cells suggests immunofluorescence artefacts. ES cells and autofluorescent spheres (BWH protocol) were processed in parallel and stained with OCT4 and NANOG antibodies. In contrast to the specific nuclear signal observed in positive control ES cells, nonspecific and non-nuclear staining is observed in spheres generated after STAP treatment. Original magnification, 20x. (E) Assessing the presence of rare ES-cell-like cells in STAP-treated cultures by teratoma formation assays. STAP-treated cells were transplanted subcutaneously or into the kidney capsule to detect rare ES-cell-like pluripotent cells. If ES-cell-like cells are generated after transient low pH treatment with/without mechanical trituration, a teratoma containing elements of all three germ layers should form. STAP-treated cells did not form teratomas using conventional teratoma generation protocols. Left two images, immunocompromised mice injected subcutaneously with STAP-treated cells, which do not exhibit teratoma-like mass formation after approximately 4 months of observation. Right two images, kidneys after STAP-treated cells were transplanted into the kidney capsule indicating lack of teratoma-like formation after 3 months of observation. Black arrows indicate kidney transplanted with STAP-treated cells; second kidney from same mouse not transplanted with STAP-treated cells. (F) Immunocompromised NOD/SCID mice transplanted with STAPtreated cells did not formteratoma-like masses. Summary of teratoma injection experiments. Every assessable injection of mouse ES cells produced teratomas (7 out of 8 positive-control ES-cell-injected mice formed teratomas within 3–4 weeks. The mouse that did not form a teratoma immediately died from surgical complications and therefore was discarded from the analysis). n=58 independent injection sessions; n=521 injection sites. Therefore, STAP-treated cells did not form teratomas using conventional methods. (G) Extended histological analysis of a recovered STAP-PLGA mass (as in Figure 1c). Obokata and colleagues[1,2] reported a distinct teratoma productionmethod that involved seeding STAP-treated cells onto a PLGA scaffold before implantation into immunocompromised mice. Around 10–20 million STAP-treated cells from GFP-positive mice were seeded into PLGA. GFP-positive cells were used to distinguish donor- and host-derived tissues. Left, positive-control ES cells formed teratomas with tissue derivatives of all three germ layers. Left, original magnifications (from top to bottom): 340, 320, 340. Middle, recovered STAP-PLGA mass,H&E staining. Middle, original magnifications (from top to bottom): 320, 340, 340. Right, recovered STAP-PLGA mass, Masson's staining (used to illustrate collagen deposition or presence of an inflammatory reaction, which commonly occur in response to foreign body implants). Right, original magnifications (from top to bottom): 320, 340, 360. All images were obtained from formalin-fixed/paraffin-embedded tissue sections.

Legend of extended data Figure 2 (continued):
(H) STAP-treated autofluorescent spheres failed to re-enter development after morula aggregation. Unlike ES or iPS cells, autofluorescent spheres failed to incorporate into the inner cell mass of the host embryos (n520), suggesting incompatibility with the pre-implantation embryo. (I) STAP-treated autofluorescent spheres failed to re-enter development after blastocyst injection. Mechanically disaggregated autofluorescent spheres were injected into pre-implantation blastocysts and implanted into pseudopregnant mice. From 17 implanted embryos, only two were recovered, which were developmentally abnormal, suggesting that the other 15 embryos died or were resorbed. (J) Contribution of STAP-treated lung cells to chimaeras was not detected after blastocyst injection. Images of two abnormal E10.5 embryos with no obvious GFP signal that would indicate integration of donor test cells into the developing host embryo. Original magnification, 310. (K) Autofluorescence and *Oct4*-GFP fluorescence were distinguishable by fluorescence microscopy in cells containing the same *Oct4*-GFP transgenic reporter used by Obokata and colleagues[1,2] (data fromDeng laboratory). MEFs with the same transgenic *Oct4*-GFP reporter (GOF18-*Oct4*-GFP, intact PE) used in ref. 1 (passage 0) were treated with low pH solutions (pH 5.4 and 5.6, respectively). MEFs without low pH treatment were used as a negative control. After treatment, samples were cultured in suspension. Chemically induced pluripotent stem cells (CiPSC)15 containing the transgenic *Oct4*-GFP reporter were used as a positive control for green fluorescence. GFP fluorescence was detected using a long-pass and band-pass filter. Red fluorescence was also observed in low-pH-treated MEFs, but not in CiPSC, as shown in the right column. Scale bar, 100 mm. (L) ES-cell-like levels of *Oct4*, *Sox2* and *Nanog* mRNA (analysed by qPCR) were not observed 3 days after STAP treatment of MEFs (data from Deng laboratory). MEFs were treated with low pH solutions and cultured in suspension for 3 and 7 days (following the RIKEN STAP protocol) and analysed. R1 ES cells were used as a positive control. MEFs that were not subjected to the RIKEN STAP protocol but cultured in suspension medium were used as the negative control (2).



Extended Data Figure 3 Re-analysis of published STAP RNA-seq data.

Legend extended data Figure 3: (A) SNVs inferred from RNA-seq (from Figure 1h) were further filtere d to select only the known SNPs across mouse strains based on the Sanger database (see Methods). We compared the SNV profile inferred from STAP RNA-seq data to the expected profiles (simulated based on the known SNPs) in different mouse strains (magenta) as well as simulated first-generation hybrids of C57BL/6NJ and each of the other strains considered (orange). Selected SNVs are classified as homozygous for reference allele (0/0 genotype), homozygous for alternative allele (1/1 genotype) or heterozygous (0/1 genotype) at each locus. Samples are clustered based on the sum of edit distance between each SNV using complete linkage hierarchical clustering. Note that replicates of the same experiment are always grouped together. A subset of samples (CD451, STAP, STAP stem cells and ES cells) (genotype A as in Figure 1h) are clustered with simulated first-generation hybrids of C57BL/6NJ and 129S1, in accordance with Obokata et al.[1,2] (LPJ strains have a profile similar to 129S1 for the selected SNVs). Whereas FI-SCs (genotype B as in Figure 1h) are closer to the C57BL/6NJ strain (not the hybrid), EpiSC samples cluster with 129S1 or LPJ simulated SNVs profiles, both with some differences. Again the high similarity between 129S1 and LPJ for these selected SNVs does not allow discriminating which of them is closer to EpiSC samples. Finally, TSC samples are clustered with other strains not mentioned by Obokata et al.[1,2]. Overall, it is clear that TSC (as well as EpiSC) samples are derived from independent sources compared with STAP cells. (B) Allele frequency distribution for SNVs shows number of reads for alternative alleles compared to the total number of reads for each SNV. The frequency of reads covering the alternative allele for heterozygous SNVs is expected to be approximately 50%, but in FI-SCs, it is nearly 12% (left, blue), suggesting false-positive calls or contamination (default thresholds in the variant calling algorithm result in incorrect classification of calls). We found that these FI-SC 'heterozygous' SNVs are predominantly homozygous for the alternative allele in TSCs (right, blue line), suggesting TSC samples as a contamination source in FI-SCs. The additional plots in Figure 1i confirm that FI-SC samples are approximately 10% contaminated by TSC samples. (C) Allele frequency distributions were independently calculated for all samples. As expected, the frequency of reads covering the alternative allele for heterozygous SNVs is approximately 50% (blue line) in all samples except FI-SCs (see b). In these plots, the first replicate (replicate 1) for each RNA-seq sample is reported; an almost identicalprofile is observed in each replicate pair.

Chapter 8

General discussion and future implications

Overview of our findings

Within a few years since its discovery in 2006, iPSC technology graduated from a novelty to a trusted technology to generate pluripotent stem cells. Research groups worldwide are since striving to derive patient-specific cells for transplantation, drug testing, and disease modeling. In this thesis, we describe our efforts to understand how cells can alter their cell fate from somatic- to pluripotent stem cell and what mechanisms are in place to prevent this transformation. We also used reprogramming to pluripotency as a model to study oncogenic mutations. Furthermore, we examined an alternative reprogramming technique of deriving pluripotent stem cells, which did not uphold its promise for the field.

The mouse model

Our lab developed a model system suited particularly well for experiments on murine reprogramming to pluripotency. Our 'reprogrammable mouse', as described in chapter 2, is a transgenic mouse with inducible expression of the four reprogramming factors *Oct4, Sox2, Klf4* and *c-Myc* (OKSM) in all cells. We show that upon doxycycline treatment of 'reprogrammable' cells in culture, iPSCs can be derived at an efficiency of around 1%; 10-100 times more efficient than reprogramming by individual transduction of viral vectors with reprogramming factors. The system helps to achieve more reproducible results as variability due to random viral integrations and different factor ratios in different cells are avoided.

Our mouse model allowed us to derive iPSC in a standardized and scalable way. The model was used to generate large samples of reprogramming intermediates to perform micro-array based expression analysis, as described in chapter 3. Using markers of reprogramming intermediates and fluorescence assisted cell sorting (FACS), our group was able to separate these intermediates at different timepoints and characterize their expression and DNA- and chromatin- methylation status. A similar approach was used in parallel to characterize the proteome of intermediates1. Both analyses revealed which specific factors are upregulated, downregulated, or transiently altered during reprogramming from fibroblasts to iPSCs, shedding some light into the 'black box' of reprogramming.

Thanks to this mouse model, large scale screening experiments were feasible. It removed one source of variability; the variability due to delivery of reprogramming factors. Also, when genes of interest were transduced in reprogrammable cells using viral vectors, this would not interfere with or add to toxicity caused by OKSM-vectors.

Since the discovery of OKSM-mediated reprogramming, other combinations of reprogramming factors were found, as well as factors able to induce transdifferentiation of one somatic cell type to another. A transgenic mouse carrying an inducible polycistronic cassette can be generated for any reprogramming or transdifferentiation protocol, in order

to study the mechanism of the resulting transformation in more detail.

In this thesis, the mouse model was used to identify factors that influence the reprogramming process. Two different approaches were used – a candidate screen using cDNAs (or ORFs), and a large-scale pooled RNAi screen.

A cDNA screen to identify regulators of reprogramming

A screen for enhancers of reprogramming efficiency and speed was performed and described in addendum to chapter 3. A set of 24 cDNAs expressed by viral vectors was prepared and delivered into fibroblasts from our reprogrammable mouse model. The top hit from this screen was the gene Embryonic stem cell specific Ras (*ERas*). Since the expression data had revealed that ERas was strongly upregulated during reprogramming, in a pattern similar to the important pluripotency factor Nanog, these functional overexpression experiments of ERas validated the molecular characterizations performed in our research group (chapter 3).

Another conclusion from this cDNA screen is that, although reprogramming can be enhanced by various factors, none of the candidates were able to remove all roadblocks to pluripotency. That is, no single factor enhanced reprogramming efficiency to 100%. Likely, many factors need to be manipulated in parallel to achieve this, or perhaps an overarching epigenetic regulator that can manipulate many factors indirectly. Manipulation of the chromatin binding factor MBD3 has been suggested to be capable of such dramatic enhancement of reprogramming but is limited by specific timing requirements[2]. Indeed, timing of depletion or overexpression of factors may be key to navigate the road to pluripotency: much like during development, this transition may depend on different conditions depending on the stage of reprogramming.

Building on our understanding of pluripotency regulators, our candidates were selected as potential reprogramming factors. This 'biased' approach has proven successful before; most notably by Takahashi and Yamanaka when searching for the original reprogramming factors and Yu, Thomson and colleagues[3] when identifying an alternative reprogramming factor combination. To find factors that modulate 4-factor induced reprogramming, biased screens have been tested primarily on chromatin modifying factors[2,4,5]. However, the 4-factor reprogramming model allows to search beyond the 'usual suspects' for new genes not previously implied in pluripotency- or reprogramming regulation. These genes could nuance our idea of pluripotency regulation beyond known pathways and transcriptional programs. To this end we decided to conduct a genome-wide RNA-inhibition screen.

RNAi screen to identify regulators of reprogramming

A search for inhibitors of reprogramming to pluripotency inspired the main project for this thesis. Not all cells exposed to the four reprogramming factors become iPSCs. Naturally, cells must be equipped with a response to exogenous signals or DNA mutations that promote loss of cell identity. The identification of roadblocks to iPSC formation provides insight into the mechanism of reprogramming to pluripotency. Possibly this information informs us about the maintenance of cell identity and the way cells combat signals that drive malignant transformation.

For this screen we employed a pooled, genome-wide library containing 63,000 shRNAs. This screen was unbiased, allowing us to find novel regulators that were previously not associated with reprogramming/pluripotency. The challenges faced when performing such a screen are detailed in the addendum to chapter 4 and 5.

We identified two hits, which are described in chapters 4 and 5. The top hit, the small peptide SUMO2, is one of several SUMO peptide tags that affect protein function and are regulated in a way very similar to the better known ubiquitination system. Sumoylation of a target protein can inhibit its function by blocking sites of interaction with other proteins; it can prohibit co-localization of other proteins, it can mark the protein for degradation or sequestration to a particular cellular compartment[6,7,8-11]. Hence, it is still unclear how *Sumo2* downregulation promotes reprogramming. The effect is likely to be dose- and/or SUMO paralog -dependent; when all sumoylation is inhibited through knockout of the essential sumoylation enzyme UBC9, reprogramming is inhibited[12]. Proteomic analysis of sumoylation patterns during reprogramming, with or without *Sumo2* knockdown, should be revealing. Also, the sumo-removal enzymes (SENP1 through SENP7) are targets for further study; current knowledge suggests that not the sumoylating enzymes, but the de-sumoylating enzymes determine which target is sumoylated with which SUMO paralog and whether one or multiple SUMO peptides are attached[13]. Manipulation of SENP enzymes could help to further dissect the role of the different paralogs and mono-, di- or trisumoylation during reprogramming.

The second hit was *Nudt21* (chapter 5), encoding for CPSF5, a factor involved in alternative polyadenylation. Alternative polyadenylation (APA) occurs when a messenger RNA (mRNA) can be cleaved and polyadenylated at different sites in its distal untranslated region (3'UTR). The 3'UTR influences protein levels by being inhibited by microRNAs (miRNAs) or inhibited or stimulated by RNA binding proteins (RBPs). Some mRNAs even have a potential cleavage- and polyadenylation site (PAS) in an intron upstream of the last coding exon, resulting in a mRNA encoding a truncated protein. Recent papers have highlighted a possibly strong effect of alternative polyadenylation on regulation of cell proliferation and

fate, as well as oncogenesis[14,15]. Around 50% of all mRNAs are suspected to be regulated by APA, making specifying a mechanism challenging. Initial analysis of RNAsequencing aimed at identifying PAS site usage (PAS-Seq) was performed and pointed to genes affected by *Nudt21* knockdown on the level of PAS choice. Further analysis of PAS-Seq and resulting downstream effectors of the *Nudt21* knockdown phenotype is indicated.

Many mRNAs are regulated by APA. Our preliminary PAS-Seq analysis pointed to a large group of targets affected by *Nudt21*. The effect of alternative PAS choice depends on how miRNA and RBP binding is affected and whether the resulting protein is truncated as a result. Thus there is no homogeneous effect of *Nudt21* knockdown on protein abundance of a large group of proteins. This complicates any search for a mechanistic explanation of the *Nudt21* knockdown phenotype. Although single genes can likely be identified that impact reprogramming when they undergo APA, none might recapitulate the phenotype of *Nudt21* knockdown. More detailed PAS-Seq analysis combined with proteomic analysis should provide a better view of which proteins are downstream effectors.

Knockdown of *Nudt21* or *Sumo2* significantly increased reprogramming efficiency and -speed. When combining these treatments with known stimulants of reprogramming, such as chemical inhibition of GSK3b and DOT1l[16,17], as well as ascorbic acid treatment[18], we could reduce duration of OKSM expression to 38 hours. This may become relevant for reprogramming approaches that rely on transfection of reprogramming factors. Approaches which aim to avoid genetic manipulation by viral vectors or transgenics, have used transfection of modified messenger RNA or protein for the reprogramming factors. Since reprogramming cells divide at a high rate, transfected molecules are swiftly diluted, making multiple transfections necessary before reaching a factor-independent state of pluripotency. If the duration of dependence on exogenous reprogramming factors can be reduced to less than 2 days, only a single transfection could reprogram cells. Thus, protocols that currently require with several transfections could be made much more attractive when combined with *Sumo2* or *Nudt21* inhibiting siRNAs or small molecules.

This would benefit human reprogramming experiments most, where genomic alterations are a problem for downstream applications. We have shown that *Sumo2* knockdown aides enhanced human somatic cell reprogramming; a next step would be to perform reprogramming of human cells with *Nudt21* suppression.

It is interesting to note that both genes are normally expressed at high levels throughout reprogramming. Gene expression analysis during reprogramming has been applied widely to identify regulators of reprogramming[19,20], but these results proves that genes that are expressed in a manner similar to 'housekeeping genes', can still be important factors. Consequently, expanding studies beyond transcriptional regulators and factors whose abundance are strongly regulated is justified.

Considerations when using a genome-wide RNAi screen

With the popularity of RNAi screens, it is important to note the challenges in the design of such screens. Performing a genome-wide RNAi screen can take years to optimize; it requires careful design, analysis, and validation before one can start looking at the biology behind the results. In our case, our system required more thorough selection of true-positive hits, resulting in multiple extra rounds of screening. Two groups have published RNAi screens during reprogramming, although without clear overlap of resulting candidate genes. This points to the limitation of each of these screens in fully covering the genome during the screen. In our case, we did not find candidates all we would have expected to find. For example, inhibition of p53 facilitates reprogramming, but we found that coincidentally, the shRNAs targeting p53 had not made it into the screening library during the preparation.

Also, differences in the screening approach can explain different outcomes. In 2010, Samavarchi-Tehrani et al[21] performed a screen during the initiation phases of reprogramming. They transfected individual wells with small inhibitory RNAs (siRNAs) targeting 4000 genes. Transfected siRNAs are expressed for several days thus mostly early modulators of reprogramming were found. Although comprehensive and unbiased, the library was not genome-wide. This is actually a strong point of the study: combining a limited library with state-of-the-art automated single well culture and analysis made individual siRNA testing possible. This ensures that each siRNA is appropriately tested. However, the resources required will not be available to most research groups.

A more recent shRNA screen by Qin et al[22] looked at human iPSC formation, co-transducing the four reprogramming factors with an shRNA vector. In a method similar to ours, shRNAs were retrieved through amplification from genomic DNA of iPSCs and sequenced to reveal enriched shRNAs. Unlike our approach, only a single round was performed with a much more complex library of 600,000 shRNAs. The multiplicity of shRNAs targeting the same gene allowed them to identify true hits. This could compensate for the lowered specificity from screening one round, where shRNAs may get enriched by hitchhiking onto an already reprogramming cell. However, the coverage of 50 cells per shRNA in a pooled screen, in an experiment with low overall reprogramming efficiency likely resulted in low coverage of the shRNAs.

For future projects, achieving good coverage of the library should be a primary focus. Therefore, the necessity and added value of genome-wide screening versus using select libraries or handpicked candidates should be considered. Next, an efficient screening model is required. For this, one should consider what percentage of cultured cells realistically participates in the screen. In our case, we were interested in screening in reprogramming cells. Of our cultures cells, in optimal conditions, near 1% would reprogram into iPSCs. Even

if reprogramming efficiency would be boosted several fold by our candidate shRNAs, the vast majority of cells would still not contribute to the screen because they are not reprogramming regardless. In screens testing cell death or proliferation, this is not as problematic; every cell should be capable of dividing or dying. The efficiency of whichever biological process one is interested in, should be optimized prior to a large screen.

To conclude, a future screen should use a highly complex library covering a limited number of genes, with high degree of coverage (depending on the base efficiency of the process of interest), in one or two rounds of screening.

Reprogramming to pluripotency as a model for oncogenesis

As discussed in the introduction, reprogramming to pluripotency has clear similarities to malignant transformation; by definition, pluripotent cells are tumorigenic themselves. However, cancer cells often harbor genetic aberrations that make each tumor unique. Reprogramming to iPSC does not require DNA mutations, eliminating a source of bias for experiments. In chapter 6, we use reprogramming as a model to study cancer by introducing a cancer-related mutated gene into reprogramming cells. The gene of interest is Isocitrate Dehydrogenase or IDH1 and IDH2, in which a neomorphic mutation provides the ability to produce the 'oncometabolite' 2-HG instead of the physiological metabolite α-KG. One problem we did encounter when modeling cancer with reprogrammable fibroblasts, is that key cancer related genes in some cases are very specific to the tissue from which the cancer arises. Only gene expression of genes that are not tissue-specific, but rather cellular process-specific, such as DNA methylation, can be compared between our reprogramming model and cancer cells.

Our observations during reprogramming suggest that expression of human mutant IDH1 or IDH2, in reprogrammable fibroblasts during reprogramming, facilitates dedifferentiation but mildly inhibits iPSC formation; iPSC formation takes longer and only in case of the strongest mutant is reprogramming efficiency affected. This could be caused by inhibition of DNA demethylation through the Tet enzymes; these enzymes are alpha-ketoglutarate (α-KG) dependent and inhibited by competitive binding of 2-HG[23]. DNA methylation patterns of mutant IDH-derived iPSC or intermediates should reveal how mutant IDH affects cell fate in this setting.

Failure to reproduce STAP phenomenon

Finally, we tried to reproduce the STAP phenomenon of inducing pluripotency or even omnipotency by applying chemical or mechanical stress to somatic cells. We have performed several protocols, with several independent colleagues in the lab, and even with on-site help

from the Vacanti laboratory; all without success. As we know now, alternative explanations to the derivation of true pluripotent stem cells were found for the key data in the STAP papers, and they were subsequently retracted. In the future, such impactful papers might require the journal to consult independent technical experts, such as flow cytometrists, bioinformaticians and pathology technicians who encounter highly technical data daily, unlike many reviewers. This should ensure an objective standard of data quality and quickly identify intended or unintended errors in key data.

References

1. Hansson, J. et al. Highly Coordinated Proteome Dynamics during Reprogramming of Somatic Cells to Pluripotency. Cell Reports 2, 1579–1592 (2012).
2. Rais, Y. et al. Deterministic direct reprogramming of somatic cells to pluripotency. Nature 502, 65–70 (2013).
3. Yu, J. et al. Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. Science 318, 1917–1920 (2007).
4. Onder, T. T. et al. Chromatin-modifying enzymes as modulators of reprogramming. Nature 483, 598–602 (2012).
5. Cheloufi, S. et al. The histone chaperone CAF-1 safeguards somatic cell identity. Nature 528, 218–224 (2015).
6. Bertolotto, C. et al. A SUMOylation-defective MITF germline mutation predisposes to melanoma and renal carcinoma. Nature 480, 94–98 (2011).
7. Morris, J. R. et al. The SUMO modification pathway is involved in the BRCA1 response to genotoxic stress. Nature 462, 886–890 (2009).
8. Subramanian, L. A Synergy Control Motif within the Attenuator Domain of CCAAT/Enhancer-binding Protein alpha Inhibits Transcriptional Synergy through Its PIASy-enhanced Modification by SUMO-1 or SUMO-3. Journal of Biological Chemistry 278, 9134–9141 (2003).
9. Ouyang, J., Shi, Y., Valin, A., Xuan, Y. & Gill, G. Direct Binding of CoREST1 to SUMO-2/3 Contributes to Gene-Specific Repression by the LSD1/CoREST1/HDAC Complex. Molecular Cell 34, 145–154 (2009).
10. Sachdev, S. PIASy, a nuclear matrix-associated SUMO E3 ligase, represses LEF1 activity by sequestration into nuclear bodies. Genes Dev. 15, 3088–3103 (2001).
11. Galanty, Y. et al. Mammalian SUMO E3-ligases PIAS1 and PIAS4 promote responses to DNA double-strand breaks. Nature 462, 935–939 (2009).
12. Tahmasebi, S., Ghorbani, M., Savage, P., Gocevski, G. & Yang, X.-J. The SUMO conjugating enzyme Ubc9 is required for inducing and maintaining stem cell pluripotency. Stem Cells

32, 1012–1020 (2014).
13. Kumar, A. & Zhang, K. Y. J. Advances in the development of SUMO specific protease (SENP) inhibitors. Comput Struct Biotechnol J 13, 204–211 (2015).
14. Ji, Z., Lee, J. Y., Pan, Z., Jiang, B. & Tian, B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proceedings of the National Academy of Sciences 106, 7028–7033 (2009).
15. Mayr, C. & Bartel, D. P. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. Cell 138, 673–684 (2009).
16. Onder, T. T. et al. Chromatin-modifying enzymes as modulators of reprogramming. Nature 483, 598–602 (2012).
17. Silva, J. et al. Promotion of Reprogramming to Ground State Pluripotency by Signal Inhibition. PLoS Biol 6, e253–11 (2008).
18. Stadtfeld, M. et al. Ascorbic acid prevents loss of Dlk1-Dio3 imprinting and facilitates generation of all–iPS cell mice from terminally differentiated B cells. Nature Genetics 44, 398–405 (2012).
19. Polo, J. M. et al. A Molecular Roadmap of Reprogramming Somatic Cells into iPS Cells. Cell 151, 1617–1632 (2012).
20. Mikkelsen, T. S. et al. Dissecting direct reprogramming through integrative genomic analysis. Nature 454, 49–55 (2008).
21. Samavarchi-Tehrani, P. et al. Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming. Cell Stem Cell 7, 64–77 (2010).
22. Qin, H. et al. Systematic identification of barriers to human iPSC generation. Cell 158, 449–461 (2014).
23. Xu, W. et al. Oncometabolite 2-Hydroxyglutarate Is a Competitive Inhibitor of α-Ketoglutarate-Dependent Dioxygenases. Cancer Cell 19, 17–30 (2011).

Chapter 9

Summary / Samenvatting

English

A pluripotent stem cell represents one of the earliest stages of development of an embryo, when choices for intra-embryonic lineages are yet to be made. During development, cells become restricted to their lineage and eventually cell type. However, genetic information for the pluripotent state remains in place and is occasionally reactivated unintentionally, with factors such as *c-Myc* contributing to cancer formation. The pluripotent state can be induced by reactivating pluripotent stem cell factors experimentally. A couple of years before work on this thesis started, a relatively simple and highly reproducible technique was discovered to derive induced pluripotent stem cells (iPSCs) by overexpression of transcription factors *Oct4, Sox2, Klf4* and *c-Myc* (OKSM). In this thesis, we used somatic cell reprogramming as a model to identify enhancers and inhibitors of reprogramming.

The 'reprogrammable mouse' model aides reprogramming experiments by allowing exogenous factor expression without transfection of viral vectors or other methods that may add to interexperimental variation. Cell lines from this mouse contain a doxycycline-inducible polycistronic cassette with the OKSM cDNAs, as well as an endogenous *Oct4*-driven fluorescent (GFP) reporter.

Reprogramming to iPSC takes 8-12 days in most protocols, during which cells pass through distinct molecular states. Transcriptional analysis as well as analysis of epigenetic marks identifies which genes and loci are activated, inactivated, or transiently regulated during reprogramming. Embryonic Ras (*ERas*) was identified as a strong enhancer in a gain-of-function screen for enhancers of reprogramming and is upregulated during reprogramming based on transcriptional data.

Although a molecular roadmap of reprogramming can be drawn, the response of individual cells to reprogramming cues is heterogeneous and in general low efficiencies are achieved. This complicates large scale screens. However, a genome-wide loss-of-function screen was successful by using a serial-enrichment approach. *Sumo2* and *Nudt21* were identified as inhibitors of reprogramming whose downregulation markedly increases reprogramming efficiency. Inhibition of *Sumo2* or *Nudt21* enhanced reprogramming speed even in high-efficiency protocols using small molecule inhibitors such as DOT1l inhibitor and the ascorbic acid/ GSK3b inhibitor combination, leading to iPSCs formed after just 38-48 hours of OKSM expression. Also, reprogramming without oncogene *c-Myc* was greatly enhanced. Derived iPSCs proved to be pluripotent in teratoma formation assays and chimera contribution assays. Also human reprogramming was enhanced by *Sumo2* downregulation. Transcriptional analysis showed an upregulation of pluripotency associated factors in reprogramming cells with downregulated *Sumo2* or *Nudt21* but did not point to specific responsible downstream regulators. SUMO2 is a small peptide which

can be conjugated to target proteins in a process called 'sumoylation', which is molecularly similar to ubiquitination. Based on literature, many cellular processes are affected by sumoylation, including epigenetic regulators of cell fate and proliferation. *Nudt21* encodes for CPSF5, involved in regulation of the length of the 3'UTR of messenger RNAs. Analysis of 'alternative polyadenylation' of 3'UTRs transcriptome-wide shows that multiple sites can be used to cleave the pre-mRNA and start the poly-A tail; modulating translational regulation by 3'UTR binding miRNAs and RNA binding proteins. In this thesis, we show that *Nudt21* downregulation causes alternative polyadenylation, including stem cell factor *Wdr5*, pointing to the likely mechanism of the reprogramming phenotype.

Questions of cell fate changes beyond the stem cell field can be found in oncology. Oncogenic mutations in the isocitrate dehydrogenase enzymes (mIDH), important in e.g. the citric acid cycle, occur in glioblastoma multiforme, leukemia and cholangiocarcinoma, amongst others cancer types. Mutant IDH was previously shown to induce an undifferentiated state in leukemic cells. This thesis provides evidence that expression of mIDH in reprogramming cells causes a loss of differentiation markers but not a gain of pluripotency. Analysis of the metabolome of these cells suggests possible metabolic mechanisms. Moreover, a mechanism of inhibiting α-KG dependent epigenetic regulators was suggested in experiments with the enzyme co-factor ascorbic acid (vitamin C).

Finally, a new reprogramming method was presented during work on this thesis. Efforts to reproduce the method were unsuccesfull, regardless of the protocol used and assistance from the laboratory behind the findings. In a collaboration between seven stem cell laboratories, the irreproducibility and faults in the supporting data were identified.

Nederlands

Een pluripotente stamcel vertegenwoordigt een vroeg stadium in de ontwikkeling van een embryo, als de differentiatie naar verschillende cellijnen nog volgt. Uiteindelijk leggen cellen zich toe op een cellijn en celtype. Echter, de genetische informatie voor een pluripotente cel blijft aanwezig in elke cel en kan per abuis gereactiveerd worden. Omdat pluripotentie-gerelateerde genen zoals *c-Myc* de proliferatie en overleving van een cel stimuleren, dragen zij bij aan oncogenese. Tegelijkertijd kan ook een gezonde pluripotente cel gevormd worden uit lichaamscellen. Enkele jaren voordat het hier beschreven onderzoek van start ging, werd ontdekt dat pluripotente stamcellen experimenteel geïnduceerd kunnen worden door geforceerde expressie van vier pluripotente stamcel-gerelateerde transcriptiefactoren: *Oct4, Sox2, Klf4* en *c-Myc* (OKSM). De verworven cellen worden geïnduceerde pluripotente stamcellen genoemd (induced pluripotent stem cells, iPSCs). In dit proefschrift gebruiken we deze 'herprogrammerings' techniek als model om het herprogrammeren naar een

pluripotente stamcel te bestuderen en factoren te identificeren die daarin een rol spelen.

Voor de meeste experimenten gebruiken wij de 'herprogrammeerbare muis' als model. Cellen van deze transgene muis hebben geen virale vectoren met OKSM nodig om OKSM productie in de cel te veroorzaken en hierdoor de cel te herprogrammeren. Ze hebben een doxycycline-induceerbare polycistronische cassette met de cDNAs van de 4 factoren. Dit betekent dat cellen van deze muis de vier herprogrammeringsfactoren gaan produceren zodra ze in contact komen met doxycyline. Met behulp van dit model is vastgesteld dat embryonale fibroblasten via specifieke tussenstadia gaan gedurende herprogrammering. Deze tussenstadia zijn bestudeerd door transcriptie en DNA methylatie genoom-wijd te analyseren en vergelijken met andere tussenstadia, fibroblasten en de uiteindelijke iPSCs. Een groep genen is getest op hun vermogen herprogrammering te bevorderen, waarop onder andere embryonale Ras (*ERas*) werd gevonden; *ERas* wordt normaliter geleidelijk aan meer tot expressie gebracht in herprogrammerende cellen zoals bleek uit analyse van de tussenstadia.

Met behulp van RNA-inhibitie konden ook inhiberende factoren worden gezocht op grote schaal; genoom-wijd. Omdat herprogrammeren met OKSM een langzaam, inefficiënt en heterogeen verlopend proces is, is het lastig om op grote schaal een screen te doen zonder verlies van gevoeligheid en specificiteit van de screen. Hierop werd een zichzelf herhalende screen ontworpen om vals-positieve resultaten te vermijden. Twee top kandidaten bleken *Sumo2* en *Nudt21*. RNAi remming van *Sumo2* of *Nudt21* zorgde voor sterke toename van snelheid en efficiëntie van herprogrammeren, ook in de afwezigheid van *c-Myc*. Efficiënte herprogrammeringsprotocollen konden hiermee worden verbeterd en de benodigde duur van OKSM expressie kon worden teruggebracht tot onder de 48 uur (normaal 6-8 dagen). De iPSCs die daarmee werden gevormd doorstonden alle kwaliteitstesten.

SUMO2 is een kleine peptide die aan andere eiwitten geconjugeerd kan worden in een proces genaamd 'sumoylatie', wat erg lijkt op ubiquitinatie. Sumoylatie lijdt tot het stabiliseren of juist vernietigen, aantrekken van andere eiwitten of juist afstoten ervan, volledig afhankelijk van het doeleiwit. Het fenotype van *Sumo2* remming in herprogrammering kan dan ook op veel mechanismes berusten. Echter is een rol in epigenetische regulatie, wat belangrijk is voor herprogrammeren, wel reeds beschreven.

Nudt21 codeert voor CPSF5, een factor die helpt bepalen waar het distale eind, de onvertaalde regio (3'UTR) van een pre-mRNA wordt geknipt en verlengd met adenosines (polyadenylatie). Alternatieve polyadenylatie zorgt dat delen van een mRNA die gevoelig zijn voor regulatie door micro-RNAs of RNA bindende eiwitten, wel of niet aanwezig blijven; dit bepaalt dan de mate van translatie van het mRNA naar eiwit. Analyse van de 3'UTRs van alle mRNAs liet zien dat *Nudt21* remming tijdens herprogrammering zorgt voor alternatieve polyadenylatie, onder andere van stamcelfactor *Wdr5*.

Tijdens het ontstaan van kanker (oncogenese) ondergaan cellen ook grote veranderingen van karakter, inclusief het weer meer kunnen delen en minder gedifferentiëerd raken. Derhalve werden herprogrammerende cellen gebruikt als model voor oncogene mutaties in de Isocitraat dehydrogenase (IDH) enzymen die onder andere in de citroenzuurcyclus een rol spelen. Van mutante IDH (mIDH) was al bekend dat het helpt om leukemiecellen in een ongedifferentieerde staat te houden. Tijdens herprogrammeren viel op dat de fibroblasten met expressie van mIDH wel sneller fibroblastmarkers verloren, maar niet makkelijker iPSCs werden. De staat van het metabolisme en van transcriptionele activiteit in deze cellen werd geanalyseerd. Ook werd de interactie van mIDH met vitamine C getest; hieruit bleek dat vitamine C een deel van het fenotype kon terugdraaien. Dit wijst erop dat enzymen die vitamine C als co-factor gebruiken en gevoelig zijn voor de metaboliet van mIDH mogelijk een belangrijke rol spelen. Onder deze enzymen zijn enkele belangrijke epigenetische factoren.

Tot slot werd een nieuwe methode om pluripotente stamcellen te maken getest. Toen bleek dat deze methode niet gereproduceerd kon worden, restte de vraag hoe dat kon en hoe de ontdekkers hun data hadden verworven. Een groep van ervaren stamcellaboratoria groepeerde hun pogingen om verschillende varianten van de methode te testen, met de conclusie dat geen ervan werkte.


A

Appendices

- List of abbreviations
- Curriculum Vitae
- PhD Portfolio
- List of publications

Acknowledgements

List of abbreviations

2-HG	β-hydroxyglutarate
3'UTR	3 prime Untranslated Region (downstream of protein coding sequence)
α-KG	α-ketoglutarate
AA	Ascorbic Acid (Vitamin C)
AGi	Combination of AA and GSK3b inhibitor
Alk5i	Alk5 inhibitor (TGF-beta signalling inhibitor)
AP	Alkaline Phosphatase
APA	Alternative Polyadenylation
ATP	Adenosine Tri-Phosphate
CPSF5	Cleavage and Polyadenylation Specificity Factor 5
D1li	Dot1l inhibitor
dox	Doxycycline
dsRNA	double stranded RNA
ECC	Embryonal Carcinoma Cell
EGC	Embryonic Germ Cell
EMT	Epithelial-to-Mesenchymal Transition
ERas	Embryonic Ras
ESC	Embryonic Stem Cell
FACS	Fluorescence Activated Cell Sorting
FF	Firefly (negative control for RNAi experiments)
gDNA	genomic DNA
GFP	Green Fluorescent Protein
ICM	Inner Cell Mass; clump of cells within a blastocyst-stage embryo
IDH	Isocitrate Dehydrogenase
iPSC	induced Pluripotent Stem Cell
LCMS	Liquid chromatography–mass spectrometry
MACS	Magnetic Activated Cell Sorting
MEF	Mouse Embryonic Fibroblast
MET	Mesenchymal-to-Epithelial Transition
mIDH	mutant Isocitrate dehydrogenase
miRNA	microRNA
MSC	Mesenchymal Stem Cell
Nudt21	Nudix hydrolase 21, gene which encodes CPSF5
OKS-mCherry	polycistronic cassette with Oct4, Klf4, Sox2 and mCherry reporter cDNAs
OKSM	(polycistronic cassette with cDNAs of) Oct4, Klf4, Sox2 and c-Myc
ORFs	Open Reading Frames
PAS	Polyadenylation site
PASSeq	PAS sequencing

PCR	Polymerase Chain Reaction
PPP	Pentose Phosphate Pathway
PSC	Pluripotent Stem Cell
repMEF	Reprogrammable MEF (containing dox-inducible OKSM)
RFP	Red Fluorescent Protein
RNAi	RNA inhibition
RNAseq	RNA sequencing
SCNT	Somatic Cell Nuclear Transfer
Senp	Sentrin-specific protease; Sumo protease
shRNA	short hairpin RNA
siRNA	small interfering RNA
STAP	Stimulus Triggered Acquisition of Pluripotency
STAP-SC	STAP-Stem Cell
STEMCCA	Stem Cell Cassette; contains OKSM
SUMO2	Small Ubiquitin-like modifyer 2
TCA cycle	Tricarboxylic Acid cycle
Tcl1	T Cell Lymphoma 1
TF	Transcription Factor
WB	Western Blot
WT	Wild Type

Curriculum Vitae

Marti Anne Borkent was born on January 24th in Oegstgeest. She graduated high school in 2004 at the Stedelijk Gymnasium Leiden (cum laude). She chose to study medicine at Erasmus University Rotterdam. In the second year she started to pursue the Master of Science 'Molecular Medicine' in parallel to the medical curriculum. She concluded the doctoral degree in Medicine with a 6 month internship in the laboratory of prof. Riccardo Fodde. Here she studied tumor forming capability of colon cancer cell lines upon active or inactive canonical Wnt signaling.

Next, she took an extracurricular internship in the laboratory of prof. Konrad Hochedlinger at the Massachussetts General Hospital / Harvard Medical School in Boston, supported by a VSB fonds grant. The next year, 2009, she returned to complete the 12 month research internship for the Master of Science, and to start her graduate research project in 2010. She received a Boehringer Ingelheim Fonds PhD fellowship to cover 2,5 years of her work. She returned to the Netherlands in 2014 to return to the medical curriculum and start the medical internships (2015-2017). She graduated medical school in September 2017.

In her free time, she took part in a variety of committees and activities within the student society 'RSG', student volleyball club 'Snoopy' and medical student organizations 'IFMSA' and 'MFVR'. She served as treasurer at the volleyball club. She served as an editor at a local student radio, and at science communication channels 'Science In The News' and 'Dokter Media'. She participated in the 'Academische Jaarprijs' competition with her Rotterdam department of Reproduction and Development. She is a member of the board and webmaster of the national society of nominees of the Chris Gips Geneeskundeprijs (LVGGG). She is married and has a 1-year-old daughter.

Phd Portfolio

Summary of PhD training and teaching activities

Name PhD candidate	Marti Anne Borkent
Research location	Massachusetts General Hospital (MGH) Boston, MA, USA
Principal Investigator	Prof.dr. Konrad Hochedlinger
Affiliations	Cancer Center, Center for Regenerative Medicine, MGH Harvard Stem Cell Institute (HSCI), Harvard dept. of Stem Cell and Regenerative Biology (SCRB)
PhD Supervision	Dept. of Development and Reproduction, Erasmus Medical Center; Erasmus University Rotterdam, the Netherlands
Promotor	Prof. dr. Joost Gribnau
Co-Promotor	Prof. dr. Konrad Hochedlinger

PhD Training	*Year*
<u>Courses</u>	
Molecular Diagnostics IV, MolMed Erasmus MC	2011
Ensembl Workshops, MolMed Erasmus MC	2011
Communicating Science Course for BIF fellows; Adobe Photoshop, Illustrator; Presentation and Writing skills	2012
PhD level courses Harvard Medical School (audited); 'Genetics 216', 'Stem Cell Biology' and 'Chromatin Dynamics'	2011-2013
Operation of the BD LSR2 Cell Analyzer, Cancer Center Flow Core	2009
Operation of Miltenyi MACSQuant Cell Analyser	2012
Radiation Safety Course, MGH	2009
Lentiviral Safety Course, MGH	2010
Animal Handling and Regulations, MGH	2009
Training, Nikon fluorescence microscope	2011

Phd training	*year*
<u>Seminars</u>	
Floor meetings and faculty candidate talks:	2008-2014
HSCI, Cancer Center, Center for Regenerative Medicine (CRM), Department of Molecular Biology	
Clinical Grand Rounds MGH	2008-2014
Scientific retreats	2009-2014
HSCI, Cancer Center, CRM, Harvard Dept. for Regenerative Biology (DRB)	
HSCI Malkin Yearly Retreats	2009-2014
Summer Internship Seminar Series	2012
<u>Meetings</u>	
International Society for Stem Cell Research (ISSCR)	2009-2013
Yearly meetings (Poster in 2009)	
Wood's Hole Meeting for BIF Fellows and Alumni (Oral presentation)	2012
IRIC Stem Cell meeting, Montreal (poster, won posterprize)	2013
Dutch Society for Stem Cell Research (DSSCR)	2015-2016
Yearly meetings (Oral presentation 2015; 2nd prize)	
<u>Supervising/ tutoring</u>	
Supervision of rotation (PhD) student Harvard Medical School	2013
<u>Science related facultative activities</u>	
Editor, 'Science in the News';	2014
Student-led organization to provide context to news articles about science.	
Editor, 'Dokter Media' ;	2016-2017
Dutch review website to add nuance and clarification to popular media reports on biomedical publication.	

List of publications

'A serial shRNA screen for roadblocks to reprogramming identifies the sumoylation effector protein Sumo2'; Marti Borkent, Brian D. Bennett, Brad Lackford, Justin Brumbaugh, Ori Bar-Nur, Li Wang, Ying Du, David C. Fargo, Effie Apostolou, Sihem Cheloufi, Nimet Maherali, Stephen Elledge, Guang Hu, Konrad Hochedlinger, Stem Cell Reports May 2016

'Failure to replicate the STAP cell phenomenon'; De Los Angeles A, Ferrari F, Fujiwara Y, Mathieu R, Lee S, Lee S, Tu HC, Ross S, Chou S, Nguyen M, Wu Z, Theunissen TW, Powell BE, Imsoonthornruksa S, Chen J, Borkent M, Krupalnik V, Lujan E, Wernig M, Hanna JH, Hochedlinger K, Pei D, Jaenisch R, Deng H, Orkin SH, Park PJ, Daley GQ, Nature 2015 Sep 24

'A molecular roadmap of cellular reprogramming into iPS cells'; Jose M. Polo, Endre Anderssen, Ryan M. Walsh, Benjamin A. Schwarz, Christian M. Nefzger, Sue Mei Lim, Marti Borkent, Effie Apostolou, Sara Alaei, Jennifer Cloutier, Ori Bar-Nur, Sihem Cheloufi, Matthias Stadtfeld, Maria Eugenia Figueroa, Daisy Robinton, Sridaran Natesan, Ari Melnick, Jinfang Zhu, Sridhar Ramaswamy, Konrad Hochedlinger, Cell December 2012

'A reprogrammable mouse strain from gene-targeted embryonic stem cells'; Matthias Stadtfeld, Nimet Maherali, Marti Borkent, Konrad Hochedlinger, Nature Methods January 2010

Acknowledgements

Bedankt!

Thank you! Let me thank a wide variety of people- not just the people I directly worked with for this thesis, but everyone who was part of my life during my Phd work, it takes a village!

Allereerst, dank aan iedereen van de **Erasmus Universiteit en het Erasmus Medisch Centrum**, die het mogelijk maakten om een onderzoeksmaster en promotieonderzoek te doen, naast het ambitieuze medische curriculum.

Bedankt **Riccardo**, ik zie je nog voor de collegezaal staan om ons te vertellen over de onderzoeksmaster **Molecular Medicine**. Jij inspireerde me niet alleen om deze master te volgen, maar ook om onderzoek te doen naar stamcellen en kanker en gaf me mijn eerste werkervaring in het lab.

Bedankt **Elaine Dzierzak, Dik van Gent, Niels Galjart en alle andere betrokkenen**, voor het organiseren van een geweldige onderzoeksmaster en dat jullie mij ervoor aannamen natuurlijk..

Bedankt **Mehrnaz**, voor jouw passie en geduld als supervisor, van jou leerde ik de basis! Bedankt **Sabrina, Patrick, Yaser, en alle Fodde collega's**, van jullie leerde ik dat een lab gezellig en georganiseerd kan zijn.

Benno, jouw mailtje over een beurs maakte dat een vage dagdroom ging leven en ik uiteindelijk voor het eerst naar Boston ging. Heel erg bedankt dat je naast de goede administratieve zorg voor de master, ook steeds essentiële extra hulp bood om het woud van papierwerk in orde te krijgen. Dank aan het **VSB fonds** dat mijn eerste stage in Boston mogelijk maakte.

Marike, dank voor je geduld en praktische kennis om deze promotie te kunnen afronden!

Dhr. Bollen, bedankt voor het bieden van een luisterend oor, praktische informatie en het geven van richting als ik door de verschillende studies en persoonlijke omstandigheden even niet meer wist hoe ik het verder moest organiseren.

Bedankt collega's van de interne geneeskunde en MDL in het **Elisabeth**, bij jullie wist ik zeker dat ik klinisch dokter wilde worden, maar tegelijkertijd mocht ik mijn symposiumpraatje bij jullie oefenen.

Bedankt collega's van de **MDL in het Erasmus MC** voor een gaaf oudste coschap maar ook de ruimte om waar nodig de laatste zaken voor mijn promoveren rond te krijgen.

Bedankt **professor Chris Gips, de stichting en de LVGGG** voor de steun voor mijn

werk middels de nominatie en de inspirerende ontmoetingen sindsdien, ik hoop dat we onze groep van internationale arts-onderzoekers kunnen blijven versterken.

A big thank you to everyone at the **Boehringer Ingelheim Fonds**, for the fellowship, but even more for the meetings, courses and amazing network that you maintain. I am forever honoured and grateful to have been a PhD fellow. Thank you to my fellow PhD students at BIF for being such an inspiring and social group!

Hartelijk dank aan professor **Sjaak Phillipsen, dr. Raymond Poot, Joost, Niels en Konrad** voor het lezen en bekritiseren van mijn proefschrift en deelnemen in de kleine commissie. Bedankt professor **Frank Grosveld en Willy Baarends** voor het deelnemen aan de grote commissie. Ik kijk uit naar jullie vragen en meningen!

Prof. Joost Gribnau, Joost, bedankt dat je mij schijnbaar onvoorwaardelijk hebt gesteund sinds ik voor het eerst begon over de VS. Jouw aanstekelijke enthousiasme, hartelijke begeleiding en enorme schat van kennis en contacten waren onmisbaar voor dit boekje en wat eraan voorafging. Bedankt dat je mijn promotor wilde zijn!

Prof. Konrad Hochedlinger, Konrad, thank you for calling me, that random weekday early 2008, I was having a pasta meal in my windowless student kitchen in Rotterdam and got a call from a Harvard PI in Boston saying I was welcome. Over the next 6 years your hospitality launched me into a different life in Boston, you took a leap of faith initially and consequently funded my PhD. You helped me acquire a PhD fellowship that is a lifelong treasure. You let me work amongst amazing colleagues, with constant access to courses and seminars by leaders of the field. You let me pursue ambitious projects and connected me to essential collaborators. You gave me space when I needed it to spend time with my dad and later process the loss. Thank you for being my PhD supervisor!

Thank you to all Hochedlinger **labmates! Matthias**, thank you for being this super friendly, social, hyper intelligent workbeast who took the time to teach me how to do a mini prep (and simply grinned at me when I accidently flushed a whole set of preps down the drain). **Jiho**, you were a great baymate, taught me elementary Korean script and helped me to attend some amazing PhD courses. **Katrin,** you took me along with your rich social life, shaping mine then and now. **Abby**, thanks for being a great labfriend, sharing our joys and struggles of the PhD life and always being supportive. **Jose**, you set the bar high both for scientific pursuits and social coherence, and along with Matthias one of the first and most appreciative guests of the pancake parties. **Mary Anna**, thank you for your hard work for the lab, your exciting and friendly personality outside of work and of course, the cheese parties. **Sihem, Eftychia and Inna**, thank you for your friendship and collegiality from the start, you inspired me, collaborated with me and cheered me on during the hardest parts of this thesis.

Ori, no need to spell it out, you know you are an integral part of what I enjoyed about

Boston and the lab. Thank you for your never ending generosity and contagious energy towards all parts of life. As I write this, I'm hoping Europe wins in the end ;).

Ryan, Justin, and Aaron, Ben, Daniel, Sara and floormates from the neighbouring labs including **Adlen, Rushika, Jean Pierre, Meredith, Charlie**, **Ondre**, you are a brilliant bunch of people, thank you for being my colleague and contributing to such an exciting research environment. I'm excited to see where your careers take you and will be following your work! The same counts for the many colleagues who presented their work, had beer and pizza or simply played volleyball with me at one of the department events of the **HSCI, CRM, HSCRB, MolBio or Cancer Center**. I am incredibly grateful to have been working and living amongst such diverse, international, but universally brilliant and accessible colleagues for all this time. **Raul**, thank you for the excellent course, but also telling us about good chocolate and decorating the hallways, therefor reminding us that it is possible to be succesfull, social and have a life outside of the lab.

Thank you to my collaborators! **Guang**, thank you for your knowledge, ability to make everything work and patience to teach me how. It has been great to work with you and you guided me through the toughest parts of my biggest project. **Yongsheng**, thank you for collaborating and helping me explore the amazing world of alternative polyadenylation. Thanks for your ambition to make the project even bigger. I look forward to meet you in person one day..

Kelly, Kate, Abhishek and all the other collaborators at **Agios**, thank you for an amazingly interesting collaboration, Kelly I enjoyed working on the bench with you a lot, thank you for teaching me about all the possibilities and impossibilities of metabolomics!

Mona, Chris, Dave and Konrad's administrative assistants, thank you for your help, patience and tolerance of these sometimes eccentric European scientists- it is no small feat. I'm sorry if I forgot to clear out all the hagelslag and espresso coffee from the kitchen cabinets after I left.

Prof. Niels Geijsen, Niels, wat een geluk heb ik gehad jou nog in Boston te leren kennen! Jij doet geweldig onderzoek, bent van de nieuwste markten thuis, hebt in het buitenland gewoond, hebt met gezin een intercontinentale move gemaakt om hier in Nederland weer toonaangevend te zijn in het veld maar ook in communicatie naar de buitenwereld toe, en over dat alles kan je dan vertellen met een luchtigheid en toegankelijkheid waar een lokale winnaar van een taartenbakwedstrijd nog wat van kan leren. Heel erg bedankt voor jouw steun, inspiratie en dat je onderdeel van dit project wilde zijn!

Annie, Nade and Kat, thank you for being nice and helpful colleagues but moreover really good friends outside the lab and in Europe, I hope we keep sharing this semi-long distance-, but ever enjoyable friendship for many years to come. **Sofia**, I forgive you for moving back to the US. Thank you for your mouse drawing and friendship in and out the lab!

Nimet, I'll mention you separately, thank you for being a great colleague, setting up and inspiring my favourite project, meanwhile combining your serious talent and collegiality with unexpected dirty humor over lunch. Outside the lab you have always been a support and great friend, regardless what else happened, so thank you for everything!

Thank you to my amazing roommates, **Winston and Nadia**, we were an interesting bunch (attempting to-) discover Cambridge's nightlife; the New Orleans Harvard master student, the fancy London lawyer with an Italian temperament and the young Dutch intern. And we had a great time. Thank you to the Winston family for an amazing and heart-warming American Christmas experience, including that German Christmas cake you found to enjoy with the gumbo. Thank you Nadia for your wise comments on life, Boston Legal and that beer right off the train in Londen. Thank you **Hootan** for letting me into your house before meeting me in person, for the good conversations, your tolerance of pancake-parties and the home-cinema setup with **Netflix** I enjoyed very much.

Thank you to all the wonderful friends I've met in Boston, you are what I miss most of that time and place. Thank you **Rick and Sheila Berry, Luke, Kate and Norah**, for becoming our friends, for letting us into your lives and deepening ours with your amazing spirits. Thank you for your wonderful art in this publication and on the journal cover.

Thank you **Emily and the Sparer family** for your friendship and your incredible hospitality at Thanksgiving, we hope to see you there often as possible in the future. Alies loves the Woodstock onesie!

Thank you **Sarah, Erich, the SuperAwesome squad, including band, Isaac, Catia, Dusko, Gwen, Jolijn, Bart, Jan Willem, Danielle and Mike, Julia, Carlijn, Doro, Derek, Tim, Cenk, Bob and Lisa, Gustavo, Bent and Margie, Boaz, Laila, Vivienne, the Appleton Dutchies, the Forro dancers, the LosstUnnown dancers,** for the friendship, fun, support and distraction when I needed it.

Thank you to all the athletic people at **MIT Women's volleyball club**, thank you **Tony** for being an amazing coach and inspiration as fanatic cook and top-notch superpostdoc, I envy those that still play with you or get to enjoy your guidance in the lab. Thank you assistant coaches! Thank you teammates at all stages, from C- minus champions (cool sweater!) to Nationals B champions (cool medal and cool canyon!).

Thank you **dragonboaters of Harvard Dudley/ Boston1 teams**, we worked hard in a mind-relaxing way, suffered the pain of salt water in blistering hands, and continuously raised inner arms, together, using it as an excuse to eat a lot (dim sum party!), as a way to keep up a fit body and a lot of social gatherings for us labrats. Thank you **Vince and Jo and cats** for your hospitality before our departure!

Bedankt **Karin, Bjorn, Uli en Marthe** voor alle dans- en eetfeestjes en gezelligheid, wat bof ik dat jullie nu ook in de buurt wonen!

Bedankt mastergenootjes! Bedankt **Martin** voor je vriendschap en advies, jouw werkethiek inspireert mij! **Lalini**, bedankt voor je vriendschap door de jaren heen, ik hoop in jouw voetstappen te treden en nog veel over werk-leven-gezin te kunnen hebben..

Lieve **Eric**, bedankt voor je steun en vriendschap door de jaren heen, en jouw gezellige bezoek aan Boston waar je me duidelijk maakte dat pittige tijden onderdeel zijn van een interessant mens worden..

Lieve meiden van **Idunn**, **Mirjam, Irene, Myrrhe, Lucienne, Claire, Sanne, Mieke, Maayke, Lieke, Talitha, Tjitske, Simone, Geerdina, Emma, Felice,** dames, moeders, collega's, dispuutgenootjes maar vooral ook vriendinnen ondanks- en na een lange tijd weg te zijn geweest. Ik heb veel van jullie trouwerijen, geboortes, afstuderen, promoveren, verjaardagen, verhuizen en meer gemist en het is me ontzettend dierbaar dit weer te hebben kunnen oppakken in Nederland. Dit is de waarde van ons dispuutje en onze vriendschap!

Anna en Cassandra, lieve paranimphen, bedankt dat jullie deze rol wilden aannemen, met jullie aan mijn zijde durf ik overal naar binnen!

Anna, als we elkaar zien is het altijd of het een dag terug was dat we elkaar het laatst zagen. Dank voor behalve je vriendschap, ook je steun, praktisch in de vorm van een logeeradres, of superattent met een kaartje naar Boston.

Cassandra, we leerden elkaar kennen als collega's in het lab, maar na de eerste kennismaking over Skype verbaasde het me niet dat we zulke goede vrienden werden. Jij combineert op een ondoorgrondelijke manier het luisteren van Disney liedjes in de celkweek, ronddansen op hoge hakken en altijd übervrolijk zijn, met een oprechte vriendelijkheid en behulpzaamheid maar ook ontzettend betrouwbare, hardwerkende, ambitieuze werkhouding. Ik weet niet waarom, maar jij hebt me een enorme loyaliteit en gastvrijheid laten zien, en als jij loyaal aan iemand bent dan is dat ook onbetaalbaar. Bedankt voor jouw vriendschap en daarmee ook van jouw 'clan' **Andre, Ben en Corette**!

Eve, wie start er nou een vriendschap met de afspraak niet over werk te praten? We waren immers altijd aan het werk, gelukkig was een kop thee met gezelligheid maar een straat verderop te krijgen. Jij bent een deel van mijn gezin geworden omdat je er voor me was als een zus. Jou mocht ik bellen als ik midden in de nacht een telefoontje uit Nederland kreeg. Bedankt voor jouw steun, gezelligheid, ondernemingslust en eetfeestjes met bitterballen; bedankt dat je zo'n goede vriendin was en bent ten alle tijden.

Emma, bedankt voor jouw vriendschap door de jaren heen, wat een lange weg sinds de brugklas.. Bedankt dat voor jou de afstand ook niet uitmaakte, dat je langskwam en dat het altijd vertrouwd voelt.

Lieve schoonfamilie, **Kier, Edith, Lies, Wil, Ineke, Corrie en Bob, Daphne, Felix, Esther, Marcel, Stijn, Tycho, Matthijs, Stefan en Linda,** bedankt voor jullie steun, bezoek en dat ik jullie liefste broertje/zoontje/neefje mee mocht nemen naar Boston!

Lieve familie, **Erik, Daan, Juul, Noor, Syb, Frans en Liesbeth, Gaby, Jane, Maja, Natasa en Snezana, Dick, Alexandra, Rein, Wieke, Geerke en Andy**, bedankt voor de steun en contact al konden we lang niet altijd komen, bedankt voor alle lieve kaartjes, gastvrijheid als we er wel waren, Juul, opa en Liesbeth bedankt dat je naar Boston kwam!

Mam, jij was op persoonlijk en organisatorisch vlak onmisbaar al die tijd. Dank je, dat je me met zelfvertrouwen de wereld instuurde, nooit klaagde al was altijd duidelijk dat je me graag meer zag en hoorde. Dank je dat ik naar Boston kon in een van de moeilijkste periodes van jouw leven. Dank je voor je bezoekjes, hulp met ingestorte kledingkasten en vele, vele rijden om spullen en mensen in Nederland op de juiste plek te krijgen. Dank je voor de inspiratie, dat ik naar je op kan kijken als professional en van je kan leren hoe je jezelf herpakt als het leven je overkomt. Dank je dat ik altijd onvoorwaardelijk bij je terecht kan. Dank je **Ron**, dat jij er voor deze lieve topperd bent, haar veel groente en fruit laat eten maar niet hebt geprobeerd de koffie en andere hardnekkigheden te verwijderen.

Pap, dank je voor de eigenzinnigheid, nieuwsgierigheid en interesse in de wereld die je me hebt geleerd. Dank je dat je me naar Boston stuurde ook al was je ziek. Dank je dat je al wist dat het goed ging komen met mij, dank je dat je me liet zien wat leven waard is. Ik mis je vandaag meer dan ooit.

Alies, lieve kleine boef van me, ik kan niet wachten tot jouw wereld weer groter wordt, dat je (dit) kan lezen en er ook op uit trekt, niet beperkt door iets anders dan je wil en fantasie. Weet dat ik nu al ontzettend gek op je ben.

Lieve **Aerjen**, wisten wij veel hoe onze eerste 10 jaar samen eruit zouden zien- de lange afstanden, oneindig veel Skypegesprekken en eenzame vliegreizen om elkaar te zien, jarenlang managen met weinig en improviseren om alles mogelijk te maken. Dat krijg je als je verliefd wordt met een trekkersrugzak aan.. Bedankt lieverd, voor jouw geduld, vertrouwen, jouw liefde en doorzettingsvermogen. Bedankt, dat je jouw comfortzone verliet om onze avonturen mogelijk te maken. Bedankt dat je me de ruimte gaf om mijn plannen uit te voeren, met de werkuren en geografische implicaties van dien. Jij bent mijn rots!

Liefs, Marti