THE UNIVERSITY *of* EDINBURGH

Novel cell surface markers identify routes to iPS cells

James O'Malley

Thesis presented for the degree of Doctor of Philosophy

MRC Centre for Regenerative Medicine

University of Edinburgh

2013

Declaration

I declare that the work presented in this thesis is my own, unless otherwise stated, and has not been submitted for any other degree or professional qualification.

James O'Malley

Acknowledgements

I would like to thank my supervisor Kei for devising and supporting this project, and supporting me financially during the long process of paper re-submission. He encouraged me to think critically and demand high standards of my research. Also, he enabled my attendance at numerous international conferences and was often there to put me in a taxi at the end of the night!

I would also like to thank Kat for helping me get through the initial few years of my PhD by providing valuable information and encouraging me to keep going, and finding the silver lining in every grey cloud. Thanks also must go to all the members of the old ISCR in the Roger Land building; I spent many afternoons walking around in that square building asking for help, reagents, sweets and comic relief and it was always to be found. Thanks to my first rotation hosts, the Brickman lab, especially Nahuel who put up with my many, many, mistakes but was always patient. Also thanks to the Chambers lab across the hall who provided reagents, advice and who had the little punching-bag that dispensed words of wisdom. Thanks also to Vaila and Pat for sorting out my money and life when I first arrived – especially when Pat shouted at Edinburgh council until they agreed they were wrong! Finally, a massive thank you to all the people who helped me to understand what was happening and what to do next at PhD club, DB club and Stem Cell club, and of course the legendary Happy Hour where problems disappeared proportional to my gin intake (usually the Blackburn lab's fault).

At our new location in Little France, I wish to say thank-you for all the input I received from the new labs I interacted with, especially the ffrench-Constant group. A massive thank you also to all the administration and science communications team, especially Ingrid for providing massive support and encouragement to get involved in public engagement and publicise my work.

Thanks also to all the staff at the transgenic unit, Jan, Sal and Lynsey, and all the team in the "mouse-house"; this project would have been impossible without

their help. Thanks also to Marilyn, Helen, Jonathan, Audrey, Ewan and Robbie for all their hard work to keep everything working for us researchers.

I also wish to thank my amazing friends that I made in the ISCR/SCRM. Laura, thank you so much for all your hard work on this project, and I'm so happy to see you move on to bigger and better things, and for all the nights out and providing the Greek perspective! Alessia, not only did we share an office, but also a fear of wasps – I knew we'd be great friends! You have been such a massive support and I hope I was the same for you. You are an amazing scientist (although you still haven't explained to me what an IP is), but more importantly you are a generous, kind and loving person, and I was delighted to celebrate your numerous milestones with you this year including your publication and finishing your thesis. You also got married. That was good too I suppose. Nicola you have been a huge inspiration; your knowledge and skill encouraged me to try and keep up with you, and I look forward to following the many accolades that are certainly in your future. You are a great friend, and I can't believe the three of us finally made it to Italy together!

Thanks especially to my friends outside the institute; Ale, Cristina, Pedro and Rafael. You have been fantastic friends and we have had great times together and I'm sure we will again soon. I wish you all the best in the future in San Francisco, Oxford and London. Thanks also to Gary for many, many nights out and being my first friend here in Edinburgh - summer 2008: "The best summer ever!"

To my amazing lab who kept me sane, made fun of me often and kept me sane! Sara/Beyoncè: You were the first one to join me in the lab and I cannot express how much fun it was to spend the last 4 years with you. So many highlights, including bonding at Happy Hour over beers, going to Glasgow for gigs (Lissie!) and general singing/dancing in the lab – Kei was so lucky to have such multi-talented students! Eleni, you proved my theory about Greek hair to be correct, but more importantly you've been an amazing Postdoc and friend, always ready with advice, insight, and support and when none of that worked: sarcasm and shots!

Tyson, your enthusiasm for science is inspiring, and you made the lab a happier place. Also thank you for taking up the difficult mantle of lab party-guy! Sergio, thank you for all your encouragement and providing a fresh perspective on my project - your insights helped and improved my work immensely. Thank you to Kumiko for your hard work which was extremely important in strengthening our work, I wish you all the best in Boston. Finally, thank you to Luca for being hilarious. You're a star, and don't worry, you'll always be the boss' favourite!

Thanks you to my parents for all their emotional and financial support and encouragement. Thank you for everything you have done for me over the years, and have always told me that I can do anything.

Finally to Mike, thank you for being there to support me throughout the last 5 years. You're generous and kind-hearted, and I feel very lucky to have had you and your family's encouragement throughout this process. Although I spent many late nights in the lab or in front of my computer working, I hope the trips to Cuba, Japan, California, Germany, France, Poland, Italy and of course Ireland balanced out all that!

This work was supported by a MRC PhD studentship, ERC grants, and the Anne Rowling Regenerative Neurology Clinic.

Abstract

The generation of induced pluripotent stem cells (iPSCs) presents a challenge to normal developmental processes. The low efficiency and heterogeneity of most methods have hindered understanding of the precise molecular mechanisms promoting, and roadblocks preventing, efficient reprogramming. While several intermediate populations have been described, it has proved difficult to characterize the rare, asynchronous transition from these intermediate stages to iPSCs. The rapid expansion of a minor population of reprogrammed cells can also obscure investigation of relevant processes. Understanding of the biological mechanisms essential for successful iPSC generation requires both accurate capture of cells undergoing the reprogramming process and identification of the associated global gene expression changes. Here we demonstrate that reprogramming follows an orderly sequence of stage transitions marked by changes in cell surface markers CD44 and ICAM1, and a Nanog-GFP reporter. RNA-sequencing (RNA-seq) analysis of these populations demonstrates two waves of pluripotency gene up-regulation, and unexpectedly, transient up-regulation of multiple epidermis-related genes, demonstrating that reprogramming is not simply the reversal of normal developmental processes. This novel high-resolution analysis enables the construction of a detailed reprogramming route map, and this improved understanding of the reprogramming process will lead to novel reprogramming strategies.

Table of Contents

Index of Figures

Index of Tables

CHAPTER 1 - Introduction

1.1 Changing cellular potential: The road to iPS cells

1.1.1 Early developments in re-establishing pluripotency

The idea that a differentiated cell, with defined characteristics and limited, if any, plasticity could in some way change its cellular identity and become pluripotent was first demonstrated in the seminal work of Prof. John Gurdon (Gurdon, 1962). It was shown that introduction of nuclei from tadpole intestinal epithelial cells to unfertilized, enucleated *Xenopus* eggs could lead to the generation of normal tadpoles, from which in turn further nuclei were isolated and the experiment successfully repeated. These results suggested that cells retain their genetic information throughout development, and that the effects of differentiation on somatic nuclei could be reversed. Subsequently, the birth of Dolly the sheep, cloned from an adult cell, provided clear evidence that there is no permanent, irreversible genetic modification of genetic information in cells, and these cloned cells can generate viable offspring (Wilmut et al., 1997). However, the low frequency of successful cloning using nuclear-transfer techniques led to speculation that contaminating, or rare, tissue stem cells may be responsible for this phenomenon. Wabl et al. (1975) generated cloned tadpoles from lymphocytes and used karyotyping to ensure the cloned tadpoles were derived from the differentiated nuclei. The Jaenisch group (Hochedlinger and Jaenisch, 2002) demonstrated that differentiated cells could be used to generate adult cloned mice. In this study, terminally differentiated B-cell nuclei were transferred into enucleated oocytes from which cloned blastocysts and embryonic stem (ES) cells could be isolated. These ES cells carried the characteristic immunoglobulin locus rearrangement of the original B-cells, and could be used to generate cloned mice.

1.1.2 Embryonic stem cells

The derivation of mouse embryonic stem (ES) cells was a major breakthrough in developmental biology. The discovery of these cells stemmed from the identification of the pluripotent nature of cells isolated from teratocarcinomas, termed embryonic carcinoma (EC) cells (reviewed in Evans, 2011). ES cell lines are derived from the inner cell mass (ICM) of the pre-implantation embryo which continues to develop into the epiblast and primitive endoderm *in vivo*, with the latter cells in turn giving rise to the visceral and parietal endoderm (Gardner, 1982). Given their similarity to EC cells, ES cells were likewise initially maintained and expanded on fibroblast feeder cells (Evans and Kaufman, 1981; Martin, 1981). It was subsequently discovered that the spontaneous differentiation induced upon feeder-free culture of ES cells could be prevented by addition of leukaemia inhibitory factor (LIF) and foetal calf serum (FCS) (Smith et al., 1988; Williams et al., 1988; Ying et al., 2003). These advances enable long term maintenance and expansion of ES cells in a culture in the absence of feeder cells, which, unlike EC cells, retain normal karyotypes upon extended passaging. Additionally, their ability to contribute to chimeric mice and for germline transmission heralded ES cells as a powerful tool with which to investigate the mechanisms of development and differentiation (Bradley et al., 1984).

It was a number of years later before similar humans embryonic stem (hES) cells could be isolated (Thomson et al., 1998). While it was shown that these cells could give rise to cells from all three germ-layers, hES cells differed from their murine counterparts in that they could not be maintained by LIF and BMP, and instead required basic fibroblast growth factor (FGF2) for maintenance of the undifferentiated state (Thomson et al., 1998; Amit et al., 2000; Reubinoff et al., 2000; Daheron et al., 2004) Subsequent experiments determined the transforming growth factor beta (TGF-β)/Activin signalling pathway is required to maintain hES in feeder- and serum-free conditions (Amit et al., 2004).

Interestingly, it was subsequently demonstrated that culturing post-implantation mouse embryos in FGF2/Activin A enabled the derivation of cell lines which shared similar characteristics with hES cells, including morphological appearance, inability to be passaged at single-cell density, and differentiation to primitive endoderm and trophectoderm upon exposure to BMP4 (Brons et al., 2007; Tesar et al., 2007). Notably these cells, although capable of differentiation into cell types from all three germ-layers, did not contribute at all to developing blastocysts in one of these studies, and only at a very low frequency in the other. These cells were also shown to display some similarity in gene expression to mES cells, but also up-regulated expression of epiblast associated factors, and were termed epiblast stem cells (EpiSCs). It was also shown that murine genetic backgrounds, as well as other species, which were previously thought to be non-permissive for ES cell derivation, could be induced to produce these EpiSCs (Wang et al., 2008a; Alberio et al., 2010). Most recently, it has been shown that EpiSCs can in fact be derived from pre-implantation embryos, even those which fail to give rise to ES cell lines (Najm et al., 2011).

The isolation of pluripotent cells from both mouse and human sources has lead to the development of myriad techniques and culture systems which have enable derivation of a variety of differentiated cell types (reviewed in Keller, 2005). While this has no doubt greatly increased our understanding of development and enabled screening of compounds which can be used in patients, certain limitations remain. For example, the generation of human ES cell lines, which requires generation and subsequent destruction of a fertilized embryo, is ethically contentious, and also somewhat limits the availability of diseased cell lines for study. One approach to overcome the latter obstacle has been the use of ES cell fusion. It was initially shown that fusion of a male mouse ES cell with a female thymocyte carrying an Oct4-GFP reporter lead to GFP activation, destabilisation of the X-chromosome silencing factor *Xist* and the fused cell could contribute to chimeric mice (Tada et al., 2001). This reactivation of pluripotency genes was

subsequently demonstrated in human cells by fusing hES cells and human fibroblasts carrying a Rex1-GFP reporter (Cowan et al., 2005). More certain analysis of the modification of the somatic genome is made possible through the use of heterokaryons which involve the fusion of cells from different species. Fusion of human B-lymphocytes or fibroblasts with mouse ES cells enabled identification of nascent pluripotency gene expression from the somatic genome, and identification of factors required for reacquisition of a pluripotent state (Pereira et al., 2008; Bhutani et al., 2010). However, these techniques are limited in their appeal in that the resulting cells carry genetic content from both fused cells, and the success of the technique appears to be correlated with the relative differentiation of the non-ES cell, with embryonic germ cell (GC) and neural stem cell (NSC) fusions more successful than fully differentiated cells (Silva et al., 2006). However, as outlined by the work of Gurdon, these results confirmed that differentiated cells are capable of acquiring an earlier, ES cell-like state both in terms of pluripotency gene re-expression and also therefore at an epigenetic level. However, identification of the precise mechanism by which this occurred was hindered by the sheer number of factors, both known and unknown, present in oocytes and ES cells that may play a role in this process.

1.1.3 iPS cells

In 2006, many of the above issues were addressed by the discovery that just four transcription factors could be used to induce an ES cell-like state in a somatic cell (Takahashi and Yamanaka, 2006). In this groundbreaking study, a set of 24 candidate genes, identified based on their expression and documented importance in ES cells, were over-expressed through retroviral infection of both mouse embryonic fibroblasts (MEF) and adult fibroblasts. Subsequent culture of the infected cells in the ES cell conditions lead to major changes in cell morphology, and eventually clusters of ES cell-like colonies were identified, from which cell lines could be derived and maintained similarly to ES cells. These were termed induced

pluripotent stem (iPS) cells. Through the process of elimination the initial list of candidates was reduced to just four factors: *c-Myc*, *Oct4*, *Sox2* and *Klf4*. Some of the iPS cell lines showed reactivation of ES cell associated genes, demethylation of certain pluripotency promoters and the ability to give rise to cells from all three germ layers in teratomas and *in vivo* from iPS cell-injected blastocysts, although no full-term chimeric mice were obtained. This result provided, for the first time a defined set of factors which could be used to induce pluripotency in the process of somatic cell reprogramming. Further work from this group and by Jaenisch and colleagues reported the generation of iPS cells which displayed more fully demethylated Oct4 and Nanog promoters and were both chimera and germline competent (Okita et al., 2007; Wernig et al., 2007).

It was subsequently shown that iPS cells could be derived from cultures of human dermal fibroblasts, neonatal fibroblasts, human mesenchymal stem cells (MSCs) and foetal lung fibroblasts (Takahashi et al., 2007; Park et al., 2008). These human iPS (hiPS) cells could be maintained in human ES cell culture conditions, and similarly displayed demethylation of pluripotency gene promoters, ES cell-like morphology, and ability to differentiate into derivatives of all germ layers in both tumour formation and embryoid body differentiation.

It had been noted that in mice generated using iPS cells, failure to appropriately silence expression of the *c-Myc* transgene resulted in the development of tumours (Okita et al., 2007). Thereafter it was reported that both mouse and human iPS cells could be generated by *Klf4*, *Oct4* and *Sox2* alone (Nakagawa et al., 2008). Notably, in all cases the removal of *c-Myc* resulted in a lower total number of iPS cell colonies, but those that did develop were of a high quality. It was also demonstrated that with the exception of *Oct4*, each of the reprogramming factors could be individually replaced by a related family member such as the substitution of *Klf4* with *Klf1*, *Klf2* or *Klf5* (Nakagawa et al., 2008). Further substitution of reprogramming factors in human iPS cell generation was demonstrated by the use of NANOG and LIN28 with OCT4 and SOX2 (Yu et al., 2007). The orphan nuclear

receptor Esrrb was reported to substitute Klf4 during reprogramming and iPS cells were also generated in combination with Oct4 and Sox2 (Feng et al., 2009a). Another screen for factors that replace Klf4 identified the maternal factor Glis1 as improving reprogramming in both MEF and human fibroblasts (Maekawa et al., 2011). Most significantly, it was shown that Oct4 could be replaced by the nuclear receptor Nr5a2, which enabled reprogramming of MEF when introduced with Klf4 and Sox2 (Heng et al., 2010). This was an important milestone as Oct4 activity during reprogramming cannot be substituted by its closely related family members Oct1 and Oct6 (Nakagawa et al., 2008). Nr5a2 was found to activate both Oct4 and Nanog during reprogramming from MEF, with Nanog over-expression capable of rescuing the reduced iPS cell colony formation observed upon endogenous Nr5a2 knockdown (Heng et al., 2010).

The advent of defined transcription factor based reprogramming of cell fate also lead to attempts to interconvert different multipotent stem cells to an earlier developmental state. It was shown that neural stem (NS) cells which express high levels of endogenous Sox2 could be reprogrammed to an ES cell-like state using only *Oct4* and either *Klf4* or *c-Myc* (Kim et al., 2008b). Subsequently reprogramming was demonstrated using both mouse and human NS cells using *Oct4* alone (Kim et al., 2009b; Kim et al., 2009c). A number of factors were also found to be capable of inducing an ES-like state from EpiSCs, including Nanog, Esrrb, Klf4 and Nr5a2 (Guo et al., 2009; Silva et al., 2009; Guo and Smith, 2010; Festuccia et al., 2012). Interestingly, the latter two factors, Klf4 and Nr5a2, were also capable of inducing both mouse ES cell-like clonogenicity and LIF/Stat3 dependence when utilised during human iPS cell derivation (Hanna et al., 2010; Wang et al., 2011b).

These results revealed that reprogramming to iPS cells is an important tool for investigation of the interactions between and requirements for various factors in the maintenance and acquisition of pluripotency.

1.2 Strategies to generate iPS cells

1.2.1 Virus based delivery of reprogramming factors

The first set of reprogramming studies utilised gammaretroviruses to deliver the reprogramming factors to the target cells, a system which was shown to also be successful for the derivation of human iPS cells (Takahashi and Yamanaka, 2006; Maherali et al., 2007; Okita et al., 2007; Takahashi et al., 2007; Wernig et al., 2007; Park et al., 2008).

Subsequently it was shown that lentiviruses could also be used to generate human iPS cells, which unlike gammaretroviruses do not require cells to be dividing for successful transduction (Yu et al., 2007; Mali et al., 2008; Zhao et al., 2008). In order to reduce abrogation of the host cell genome, many laboratories developed polycistronic vectors which encoded all the reprogramming factors in one transcript (Carey et al., 2009; Shao et al., 2009; Sommer et al., 2009). In these vectors the reprogramming factors are separated by 2A peptide linkers which enable "ribosome skipping" and the translation of multiple proteins from a single mRNA molecule (Donnelly et al., 2001). iPS cells were also generated by lentiviruses using vectors carrying the individual reprogramming factors encoding a loxP site in the 3′ LTR (Soldner et al., 2009). Upon proviral replication this site was duplicated into the 5′ LTR, thus resulting in integrated, loxP flanked transgenes which could subsequently be removed upon transient Cre recombinase expression. Interestingly, this study revealed that post-excision iPS cells were more similar to hES cells than before the removal of the transgenes. This method was similarly applied to the polycistronic lentivirus vectors, enabling the generation of iPS cells with a low number of transgene integrations, thus reducing the chance of introducing detrimental genomic rearrangements upon Cre-mediated excision of the reprogramming factors (Chang et al., 2009; Somers et al., 2010; Sommer and Mostoslavsky, 2010; Papapetrou and Sadelain, 2011).

Two of these studies delivered the Cre recombinase vector through the use of the non-integrating adenovirus (Chang et al., 2009; Sommer and Mostoslavsky,

2010). This strategy has also been used for the delivery of the reprogramming factors, and iPS cells have been derived from mouse hepatocytes, fibroblasts and human fibroblasts using adenoviruses (Stadtfeld et al., 2008b; Gonzalez et al., 2009; Zhou and Freed, 2009). However, the rapid clearing of viruses from dividing cells means that reprogramming via this method requires multiple subsequent rounds of infection, and iPS cells are generated at a low frequency. While the removal of transgenes from the host genome through the use of Cre-recombinase removes the majority of foreign genetic information, the inactive LTR fragment (~250bp) remains.

The sendai virus, which replicates exclusively in the cytoplasm has also been used to generate iPS cells. A number of human cell-types have been shown to undergo reprogramming using sendai virus vectors (Fusaki et al., 2009; Seki et al., 2010). However, as outlined in the former study, high levels of the virus persist in the reprogrammed cells. Two recent studies have addressed this issue with the use of temperature sensitive viral vectors or by using viral replication- and transcription-interfering iPS cells (Ban et al., 2011; Nishimura et al., 2011). These techniques allow relatively efficient generation of virus-negative iPS cell colonies.

1.2.2 Non viral delivery of reprogramming factors

In order to generate truly fully transgene-free iPS cells, the *piggyBac* transposon has been used for delivery of reprogramming factors (Kaji et al., 2009; Woltjen et al., 2009; Yusa et al., 2009). Transposons are genetic elements which occur naturally in the genome and can introduce mutation and genomic rearrangement as they are mobile, moving from one position in the genome to another (McClintock, 1950). The *piggyBac* transposable element is flanked by 13-bp terminal inverted repeats, and insertion into the genome duplicates its TTAA target site (Cary et al., 1989). Its activity is reliant on co-expression of the transposase enzyme, with early experiments showing seamless insertion and excision of a lacZ reporter gene into the viral genome, and subsequently it was used for efficient germ-line mutagenesis

in the mouse (Fraser et al., 1996; Ding et al., 2005; Wu et al., 2007). In mouse ES cell colonies bearing an integrated *piggyBac* transposon, it was shown that upon expression of the transposase enzyme roughly 40% of colonies displayed re-insertion of the transposon, with complete excision from the host genome in the remainder (Wang et al., 2008b). Kaji et al. (2009) demonstrated virus-free reprogramming of MEF via nucleofection of a floxed polycistronic vector which was subsequently flanked by *piggyBac* transposons and delivered to human embryonic fibroblasts to generate iPS cells. Woltjen et al. (2009) also utilised *piggyBac* technology to deliver separate vectors encoding each of the Yamanaka reprogramming factors to both MEF and human embryonic fibroblasts. Upon re-expression of transposase enzyme, complete transposon removal occurred in 66% of the mouse iPS cell colonies tested, and 90% showed no post-excision-aberrations at the site of transposon insertion. The use of negative selection to enrich for successfully excised colonies has also been demonstrated (Yusa et al., 2009).

It has been demonstrated that the direct delivery of proteins fused to a poly-argenine tag which enables crossing of the cellular membrane can be used to generate mouse iPS cells, albeit at a highly reduced efficiency compared to other methods (Zhou et al., 2009). Reprogramming of human fibroblasts was also demonstrated using this technique, however the argenine-tagged reprogramming factors were expressed in a human embryonic kidney (HEK) cell-line from which whole cell extracts were isolated (Kim et al., 2009a). Whole mouse ES cell extracts, delivered by streptolysin O-mediated reversible permeabilisation, were also demonstrated to generate mouse iPS cells from fibroblasts which were competent to generate chimera (Cho et al., 2010). A major caveat with these techniques is that all operate at a low frequency of success and also require large amounts of starting protein or cell extract.

More recently, using *in vitro* transcription (IVT) to generate modified mRNA molecules, Warren et al. (2010) demonstrated highly efficient derivation of iPS cells from human fibroblasts. In this study it was found that transfection of the synthetic mRNA lead to activation of single strand RNA (ssRNA) sensor RIG-I, resulting in a

high level of cytotoxicity. To reduce this immune response a number of modifications were made to the synthetic mRNAs including treatment with phosphatase and substitution of cytidine and puridine with 5'methylcytidine and pseudouridine respectively. Culture media was also supplemented with an inhibitor of interferon signalling. In these conditions transfection efficiencies of up to 90% were achievable and ES cell-derived, foetal, postnatal and adult human fibroblasts were all successfully reprogrammed, with the ES cell-derived fibroblasts shown to generate iPS cell colonies at a faster rate (two-fold greater) and higher efficiency (36-fold greater) than fibroblasts infected with the reprogramming factors in parallel. However successful reprogramming using this technique requires daily transfection of the reprogramming factor mRNA due to the low stability and high turnover of these molecules, and involves a high level of technicality thus making this system less widely accessible.

Another approach has been the use of oriP/EBNA1 vectors which have been derived from the Epstein-Barr virus which can be introduced to cells in the absence of viral packaging. In the presence of drug selection these vectors replicate extra-chromosomally once per cell cycle and are established stably in ~1% of transfected cells, and lost at a rate of 5% per cell cycle upon removal of selection (Leight and Sugden, 2001; Nanbo et al., 2007). Yu at al. (2009) demonstrated that using a combination of *OCT4*, *SOX2*, *NANOG*, *LIN28*, c-MYC, *KLF4*, and *SV40LT*, iPS cell colonies could be derived from human foreskin fibroblasts, and upon removal of selection about 30% of established iPS sub-clones lost the episomal vectors. While successful, this technique was limited in that the efficiency of reprogramming was more than three orders of magnitude lower than reprogramming using lentiviruses in the same study, and utilized the oncoprotein SV40LT. It has since been demonstrated that higher efficiency reprogramming using episomal vectors can be achieved without the use of SV40LT by either the inclusion of shRNA against p53 or the replacement of *NANOG* and *c-MYC* by *L-MYC* (Okita et al., 2011). Another recent study used minicircle episomal vectors and demonstrated reprogramming of human adipose stem cells (Jia et al., 2010). These are stripped-back versions of the

larger standard episomal vectors and contain only a eukaryotic expression cassette, lacking the bacterial backbone elements enabling both an increased transfection efficiency and duration of transgene expression.

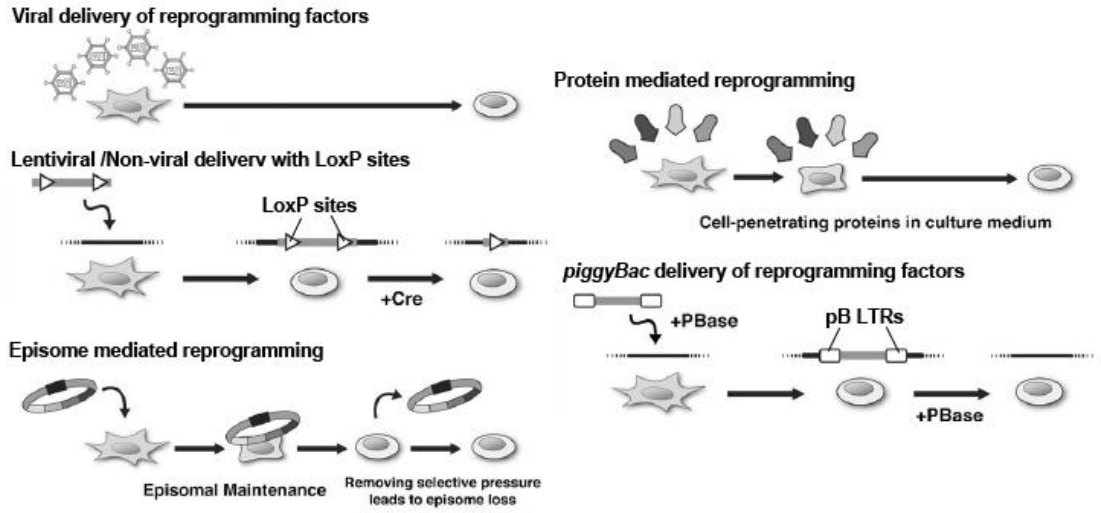


Figure 1.1 Overview of methods used for delivery of reprogramming factors. Adapted from O'Malley et al. (2009).

1.2.3 Secondary reprogramming systems

The previously described lentiviral reprogramming strategy was demonstrated to successfully generate mouse iPS cells with expression of the reprogramming factors under the control of the tetO doxycycline (dox) inducible promoter (Brambrink et al., 2008; Stadtfeld et al., 2008a). This approach was subsequently used to develop a system termed secondary (2°) reprogramming (Wernig et al., 2008). This system is based on the requirement for doxycycline binding of the reverse tetracycline transactivator protein for interaction and expression from the tetO operator sequence (Tet-On system) (Gossen and Bujard, 1992). In this 2° reprogramming system, iPS cells generated using the dox-inducible lentiviruses from transgenic MEF expressing rtTA from the Rosa26 locus, were used to generate chimeric mice. Cells isolated from these chimeras could subsequently be cultured in dox-containing media, thus inducing re-expression of the

reprogramming factors and generating 2° iPS cells. A variety of cell types were successfully reprogrammed using this system including MEF, MSCs, neural progenitors, epithelial cells and keratinocytes. Interestingly, despite the genetic homogeneity of these cells, the maximum efficiency of MEF reprogramming was roughly 1% (although it should be noted that this was 30 times greater than parallel retrovirus reprogramming), and the reprogramming of various cell types revealed greatly different temporal requirements of transgene expression before iPS cell colonies could be isolated. The results of this study indicated 2° reprogramming was a useful system for investigating the mechanism of reprogramming, and was subsequently used to reprogram fully mature B-cells (Hanna et al., 2008). Secondary reprogramming systems were also developed for human cells, with reprogramming of fibroblasts and keratinocytes derived from iPS cells generated using dox-inducible lentiviruses (Hockemeyer et al., 2008; Maherali et al., 2008).

Polycistronic vectors carrying dox-inducible versions of the reprogramming factors were also demonstrated, both in viral and *piggyBac* transposon systems which generated 2° iPS cells from MEF (Sommer et al., 2009; Woltjen et al., 2009). Further to this two studies used targeting of the Col1a1 locus in mouse ES cells with dox-inducible polycistronic vectors containing the reprogramming factors (Carey et al., 2010; Stadtfeld et al., 2010b). Carey at al. demonstrated that while a single copy of the transgene-containing Col1a1 locus was sufficient to generate iPS cells from MEF, iPS cells could not be derived from any adult tissue. Crossing of transgenic mice produced mice carrying two copies of the transgenic locus and adult liver and keratinocytes readily underwent reprogramming. However, reprogramming of other adult cell types including fibroblasts, macrophages and pro-B cells did not require an additional copy of the reprogramming factor containing locus, but rather two copies of the rtTA expressing Rosa26 locus. Furthermore, two copies of both transgenic loci were required before MSC and gut epithelial cells gave rise to iPS cells. Similarly, while Stadtfeld et al. also generated iPS cells from MEF bearing single copies of the reprogramming factor- and rtTA-loci, MEF homozygous for rtTA generated three-times more iPS cells. In addition, echoing a study in which it

was discovered that more differentiated hematopoetic lineages undergo reprogramming less efficiently, it was found that while heterozygous rtTA 2° hematopoietic stem cells (HSCs) gave rise to iPS cells, this was only possible in B-cells, T-cell and granulocytes carrying two rtTA loci (Eminli et al., 2009). Taken together, these results that suggest that high expression of the reprogramming factors is important for successful generation of iPS cells, and certain cell types undergo reprogramming more readily than others. However it is important to note that the "Yamanaka factors" used in these studies were identified for their ability to reprogram MEF, and as has been demonstrated, other combinations of factors could increase the efficiency of iPS cell generation from other cell types (Takahashi and Yamanaka, 2006; Hanna et al., 2008). Nevertheless, 2° reprogramming systems have been widely implemented and have aided identification of important aspects of reprogramming including the role of BMP, the stochiometry of the reprogramming factors, splicing mechanisms and epigenetic regulation (Samavarchi-Tehrani et al., 2010; Carey et al., 2011; Gabut et al., 2011; Koche et al., 2011). More recently, 2° systems have been used to combine large-scale analysis and modern high-throughput technology to generate vast amounts of information about the process of reprogramming (Buganim et al., 2012; Golipour et al., 2012; Polo et al., 2012; O'Malley et al., 2013).

1.2.4 Micro-RNA mediated reprogramming

Micro-RNAs (mi-RNAs) are short (20-25 nucleotide) non-coding RNAs which repress mRNAs by interfering with the expression and stability of their targets. It was shown that over-expression of mouse ES cell specific mi-RNAs during retroviral reprogramming of MEF enhanced iPS cell generation, and when used in the absence of *c-Myc*, produced colonies which homogeneously expressed an Oct4-GFP reporter (Judson et al., 2009). Human orthologs of these mouse ES cell mi-RNAs were also shown to enhance reprogramming factor-mediated iPS cell generation from human fibroblasts, and were reported to affect multiple targets and

pathways associated with reprogramming (Subramanyam et al., 2011). However, the most striking results were reported by Anokye-Danso et al. (2011) in a study which demonstrated reprogramming of both mouse and human somatic cells via lentiviral over-expression of the miR302/367 cluster in the absence of exogenous reprogramming factors. This method increased the number of iPS cell colonies generated in the order of two to four magnitudes greater compared to lentiviral reprogramming using the Yamanaka factors. Interestingly the reprogramming of MEF, but not human fibroblasts, via this method was dependent on the presence of the histone deacetylase inhibitor (HDAC) Valproic acid (VPA), and it was suggested that the lower levels of the VPA target HDAC2 in human cells compared to MEF was responsible for this difference. A second study which transfected mouse and human cells with mature double-stranded mi-RNAs also demonstrated reprogramming in the absence of exogenous transcription factors, but this non-integrative approach was much less efficient (Miyoshi et al., 2011). mi-RNA-mediated reprogramming, despite these reports, has yet to become commonplace, possibly reflecting a difficulty in replicating these results in other labs and under different experimental conditions.

1.3 Cell culture conditions and reprogramming

1.3.1 Mouse embryonic stem cell culture conditions

Initial cultures of mouse ES cells could be maintained only by culturing in the presence of mitotically inactivated fibroblast feeder layers. Addition of leukaemia inhibitory factor (LIF) was found to prevent the characteristic spontaneous differentiation which occurred in feeder-free culture of ES cells (Smith et al., 1988; Williams et al., 1988). The exact mechanism by which LIF was able to maintain ES cells was subsequently elucidated through investigation of the LIF receptor (LIF-R) dimerisation partner gp130 (Niwa et al., 1998). Phosphorylation of specific cytoplasmic domains of gp130 as a result of LIF activity results in interaction with and activation of STAT3 causing its dimerisation, relocation to the nucleus and transcription of its associated targets. Subsequently, serum-free culture of mouse ES cells was demonstrated using the chemically defined N2B27 media supplemented with LIF and BMP4 or BMP2 (Ying et al., 2003). Derivation of ES cell lines from blastocysts was also demonstrated in this serum-free culture, and it was observed that undifferentiated ES cells expressed high levels of BMP receptor (Bmpr)Ia and BmprII as well as Bmp4, highlighting the important role of BMP signalling in stem cell maintenance. In addition, it was found that BMP signalling was responsible for induction of the helix-loop-helix inhibitor of DNA binding (Id) proteins Id1/2/3, with Id1 over-expression and LIF sufficient for ES cell self-renewal in serum-free conditions (Ying et al., 2003). Interestingly, in N2B27 alone, ES cells differentiate into neural lineages, however in the absence of LIF Id1-overexpressing cells differentiated into non-neural lineages, indicating the requirement for both BMP and LIF signalling in ES cell maintenance.

1.3.2 2i and reprogramming

A number of small molecules have increasingly been utilised in order to enhance or assist the reprogramming process from somatic cell to iPS cell. Ying et al.

defined “ground state pluripotency” in ES cells as the innate programme for self replication independent of extrinsic instruction (Ying et al., 2008). By culturing ES cells in a defined media in the presence of the selective Glycogen Synthase Kinase-3 (GSK3) inhibitor CHIR99021 (CHIR), an inhibitor of FGF receptor tyrosine kinase activity SU5402 and an inhibitor of the ERK cascade PD184352 (3i conditions), it was shown that small molecules could maintain pluripotency and self-renewal similarly to less defined culture factors e.g. serum (Ying et al., 2008). Subsequently, culture in the presence of the more potent ERK inhibitor PD0325901 (PD03) and CHIR (2i conditions) enabled neural stem cells (NSCs) and MEF-derived partially reprogrammed cells which had failed to undergo complete reprogramming to gain expression of endogenous Oct4, reactivate inactive X chromosomes in XX cells, and contribute to chimeras, thus becoming fully reprogrammed (Silva et al., 2008). It was also highlighted in this report that genetic background had a significant effect on reprogramming frequency, with inbred strain 129 yielding more iPS cells compared to hybrid background cells (129/outbred MF1). The Non-obese Diabetic (NOD) mouse is an animal model of the human disease type I diabetes which has enabled analysis of the onset of this disease (Wicker et al., 1995). This strain of mouse was widely considered as non-permissive for the derivation of iPS cells, but permissive for Epiblast Stem Cell (EpiSC) isolation (Brons et al., 2007). It was found that culture of inner cell masses (ICMs) from NOD mice in 2i conditions plus LIF enabled the derivation of ES cell lines from this strain (Nichols et al., 2009a). Hanna et al. also showed derivation of NOD ES cell lines, however, this was achieved via infection of ICM outgrowths with constitutive lentiviruses encoding Klf4 and c-Myc (2009a). This group also generated iPS cells from NOD MEF, which were similarly shown to require constitutive expression of Klf4 and cMyc. These iPS cells could be stabilised in mouse ES medium (mESM) in the presence of the GSK3 and CDK1/cyclin B inhibitor Kenpaullone (KP) which had previously been shown to replace Klf4 during reprogramming (Lyssiotis et al., 2009).

1.3.3 Small molecules reprogram rat and human ES/iPS cells to a mES cell state

Human ES/iPS cells are more similar to mEpiS cells than mES cells; in terms of cellular signalling, LIF is insufficient to support self-renewal, rather bFGF and Activin A are required, and BMP induces their differentiation (Yu and Thomson, 2008). Tesar et al. (2007) demonstrated that SB431542, an inhibitor of activin receptor-like kinase (ALK)5/4/7, induced differentiation in EpiS cells but not ES cells. Another ALK receptor inhibitor, A-83-01, was subsequently used in order to develop culture conditions suitable for the propagation of human iPS (hiPS) cells and rat iPS (riPS) cells in a more mES cell-like context (Li et al., 2009b). In this study, 2i conditions were shown to be insufficient to maintain riPS cells in mES cell media, even in the presence of the FGF receptor inhibitor PD17304. However a combination of CHIR, PD03 (2i), ALKi and LIF enabled long-term propagation and survival of these "mES cell conditioned" h/riPS cells, as well as expression of the mES cell associated transcription factor Rex-1 (which is not expressed in regular h/riPS cells), and resistance to differentiation despite inhibition of the FGF and Activin signalling pathways. More recently, constitutive activity of the reprogramming factors cMyc/Klf4/Oct4/Sox2 (MKOS) in addition to PD03, CHIR and LIF were also shown to support propagation of hiPS/ES cells in mES cell-like conditions, with the adenylate cyclase activating compound Forskolin (FK) shown to replace MKOS (Hanna et al., 2010). Transcriptional profiling showed these "naïve" hiPS/ES cells were more similar to mES/iPS cells than more traditional hiPS/ES cells. These reports show that the culture conditions used to propagate or isolate pluripotent cells whether they are iPS or ES cells can be suitably manipulated using small molecules in order to increase or enable full reprogramming to a "naïve" mES cell-like state, and thus lead new insights into the acquisition of pluripotency.

1.3.4 The mesenchymal-to-epithelial transition and reprogramming

The above studies highlight that in naïve rat and human iPS cells the reprogrammed state displayed continued reliance on small molecules, therefore it

may not be the case that these cells have altered intrinsically and rather what occurs may simply be selection of cells which can appropriately respond to such signalling cues. This selective process could therefore not only be detrimental to those cells which have not yet reached an appropriate stage of the reprogramming process, thus reducing iPS cell yield, but may also generate unstable iPS cells with incorrect or insufficient (re-)establishment of ES cell-like signalling pathways. In reprogramming from mouse cells, conversely, it has been shown that relatively short-term (24hr-4d) inhibition of the TGFβ signalling pathway during reprogramming appears to enhance the generation of iPS cells (Ichida et al., 2009; Maherali and Hochedlinger, 2009). Both studies used "RepSox" a molecule reported by both to have the ability to substitute for Sox2 during reprogramming with c-Myc, Klf4 and Oct4. Although these studies presented conflicting data with regards the timing of inhibitor treatment, in both cases the iPS cells generated were stable, with Maherali and Hochedlinger (2009) even increasing the yield of reprogrammed cells obtained in the absence of the reprogramming factor c-Myc. Ichida and colleagues (2009) also demonstrated that similar results could be acquired through the use of another TGFβ receptor inhibitor or receptor-blocking antibodies. This study also isolated reprogramming "intermediates"; iPS cell-like colonies which could be expanded, but failed to express an Oct4-GFP reporter, similar to those reprogrammed by culture in 2i conditions (Silva et al., 2008). Expression of the Oct4-GFP reporter was demonstrated upon culture of these cell lines in the presence of RepSox, and although this effect appeared to be highly line-specific, interestingly, kinase activity assays indicated 2i targets were not inhibited.

The importance of TGFβ/BMP signalling during reprogramming was further highlighted with the back-to-back publication of two recent studies identifying the mesenchymal-to-epithelial (MET) transition as a conserved stage of reprogramming from MEF (Li et al., 2010; Samavarchi-Tehrani et al., 2010). Samavarchi-Tehrani and colleagues identified up-regulation of a large number of epithelial-associated genes and down-regulation of mesenchymal-associated transcription factors within the

first five days of reprogramming. Mimics of miRNAs miR200b and miR200c, which are induced upon reprogramming factor expression in MEF, were shown to down-regulate mesenchymal-associated factors Zeb1 and Zeb2, as well as up-regulate epithelial factors Cdh1, Epcam and Ocln. This switch in cell character was also linked with a gain in BMP signalling-associated factors, with a threefold increase in reprogramming efficiency upon addition of BMP7 to culture media, while inhibition of BMP signalling decreased reprogramming efficiency. Li at al. (2010) demonstrated a similar gain of epithelial and loss of mesenchymal gene expression during reprogramming from MEF. This study also demonstrated inhibition of colony formation through the addition of TGFβ ligands TGFβ1 and TGFβ2 to the culture media, and reported high levels of TGFβ1 in bovine serum commonly used in media for iPS cell generation and maintenance. It was also reported that Klf4 activity during reprogramming is important in inducing Ecad expression and this reprogramming factor could be substituted by an Alk receptor inhibitor, A-83-01, during reprogramming of mammary gland epithelial cells (MECs). While these studies highlighted the importance of this transition, it was reported to occur within the first 5 days of reprogramming factor expression and before activation of most key pluripotency-associated factors.

1.4 Epigenetics of Reprogramming

1.4.1 Modulation of epigenetic factors to influence reprogramming

During reprogramming cells not only acquire a similar transcriptional state as ES cells, but also a similar epigenetic landscape. Chromatin modifications have been intensively studied in order to determine the extent and effect of epigenetic regulation during reprogramming to iPS cells. The initial iPS cells generated by Takahashi and Yamanaka (2006) would not be considered of a high-quality today, displaying a number of differences compared to ES cells including high levels of DNA methylation at the promoters of pluripotency genes. Subsequent studies generated iPS cells which displayed full ES cell-like demethylation of promoters while maintaining imprinting patterns, ES cell-like tolerance to global DNA demethylation and loss of X inactivation in XX cells (Maherali et al., 2007; Wernig et al., 2007).

A recent study attempted to identify epigenetic changes occurring within the first cell-cycles after reprogramming factor expression in MEF (Koche et al., 2011). This study reported that histone 3 lysine 4 dimethylation (H3K4me2), the precursor to H3K4me3 which is associated with locus accessibility and active gene expression, is gained and increased at loci crucial for reprogramming, before changes to other histone modifications. However some of these loci do not display a similar pattern of H3K4me2 as in ES cells, and after just three cell divisions the level of K4me2 detected is identical to that of ES cells, thus given the low efficiency of reprogramming, it is not certain that this event represents a significant barrier to reprogramming. In agreement, a recent study based on isolation of intermediate reprogramming populations using the ES cell-associated marker stage-specific embryonic antigen 1 (SSEA-1) identified major changes in H3K4me3 at early stages of reprogramming, but also identified a second, later wave of H3K4me3 acquisition when cells were closer to a more reprogrammed state, and this may bear more correlation to reprogramming success (Polo et al., 2012). This study also revealed

both gain and loss of the gene expression silencing-associated marker H3K27me3 early in reprogramming, and an early establishment of bivalency - the co-localization of both active (H3K4me3) and repressive (H3K27me3) histone marks at developmentally important genes which is a feature of ES cells (Bernstein et al., 2006). Bivalency is established and maintained by the activity of two distinct complexes; the trithorax group (TrxG) which is involved in establishing H3K4me3 and the polycomb group proteins (PcGs) which form polycomb repressive complexes (PRCs) which mediate H3K27me3 generation. A number of studies have shown the important role these factors play in reprogramming. Knockdown of the TrxG member Wdr5 in ES cells resulted in a loss of self-renewal, global loss of H3K4me3, and inhibition of iPS cell generation (Ang et al., 2011). Interestingly, depletion of another sub-unit of H3K4 methylation complexes, Dpy-30, in mouse ESCs leads to a defect in lineage specification but does not significantly affect ESC self-renewal (Jiang et al., 2011). In addition, ES cells which lack the CpG-binding factor Cfp1 and show loss of H3K4 methylation are viable and can self-renew, but again demonstrate differentiation defects (Clouaire et al., 2012). These findings suggest that while the establishment of H3K4 methylation may be important during reprogramming, formation of bivalent domains may not be crucial for the maintenance of pluripotency. Also a number of members of PRC1 and PRC2 have also been shown to be vital for reprogramming from human fibroblasts (Onder et al., 2012). The nucleosome remodelling and deacetylation (NuRD) complex deacetylates H3K27, thus enabling binding of PRC2. Mbd3, which is a component of the NuRD complex and has been shown to be important in controlling differentiation of ES cells, has been shown to inhibit the efficiency of reprogramming from MEF, however in the above study Onder et al. (2012) found that knockdown of this factor actually slightly reduced human iPS cell colony formation (Kaji et al., 2006; Kaji et al., 2007; Luo et al., 2013; see Section 6.3 for dicussion of novel data regarding Mbd3 and reprogramming). Despite this accumulated evidence, Sridharan et al. (2009) reported pre-iPS cells which fail to express key pluripotency genes had already acquired ES cell-like bivalent

methylation patterns, and while it has been shown that the efficiency of reprogramming from MEF is unaffected in the absence of the *de-novo* DNA methyltransferases Dnmt3a and Dnmt3b, preventing maintenance of DNA methylation via inhibition of Dnmt1 transferase activity can increase reprogramming efficiency from pre-iPS cells (Mikkelsen et al., 2008; Pawlak and Jaenisch, 2011). This suggests failure to demethylated DNA represents a greater barrier to reprogramming than establishment of bivalent domains. Interestingly, in human fibroblasts the initial binding of the reprogramming factors c-Myc, Klf4, Oct4 and Sox2 occurs across both methylated and unmethylated DNA (Soufi et al., 2012). However, megabase regions which are bound by the reprogramming factors in ES and iPS cells were found to be refractory to binding in the early stages of reprogramming. These regions were enriched for H3K9 methylation, and knockdown of H3K9 methyltransferases enhanced Oct4 and Sox2 binding at these sites and increased reprogramming efficiency. Other factors which influence chromatin have also been implicated in reprogramming including the SWI/SNF complex that destabilises histone-DNA interactions, members of which have been shown to increase the efficiency of reprogramming upon over expression in MEF (Singhal et al., 2010). One other group of factors recently gaining much interest are the ten-eleven translocation (TET) family methycytosine hydroxylases TET1 and TET2. These proteins modify DNA via changing 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), with the latter modification being found at pluripotency gene regulatory elements in ES cells (Ito et al., 2010). It was reported that knockdown of TET2 prevented acquisition of 5hmC at the loci of pluripotency associated genes Nanog and Esrrb, and resulted in a reduction in the efficiency of reprogramming from MEF (Doege et al., 2012). More recently Costa et al. (2013) reported that TET1 and TET2 mediate reprogramming from intermediate cell types as well as MEF but only in the presence of Nanog which is required to direct their enzymatic activity to pluripotency-related loci. The significance of this modification compared to other epigenetic changes occurring during reprogramming has yet to be fully explored, however there is evidence to suggest that in conventional

reprogramming the modification of 5mC to 5hmC occurs early in reprogramming as it appears to precede H3K4me3 recruitment to loci (Costa et al., 2013). Interestingly, the TET enzyme's catalytic domain binds iron and therefore their activity may be enhanced by the presence of Vitamin C in reprogramming cultures (see section 1.4.3).

1.4.2 Small molecule modification of epigenetic state during reprogramming

One of the first studies of the epigenetic state of cells undergoing reprogramming identified that fully reprogrammed iPS cells re-established ES cell-like H3K27 and H3K4 bivalent methylation at the promoters of developmentally important genes (Mikkelsen et al., 2008). This study also discovered that in partially reprogrammed iPS cell-lines genes which carried K4 methylation or bivalent marks in the parental MEF cells were more likely to be re-activated, and this intermediate cell type had a lower percentage of bivalent promoters compared to fully reprogrammed iPS cells. Addition of the DNA methyltransferase (DNMT) inhibitor 5'-azacytidine (Aza) for 48 hours to the culture media increased the percentage of colonies reactivating expression of an Oct4-GFP reporter, with the resulting iPS cells displaying CpG demethylation at pluripotency associated genes. In addition, siRNA against Dnmt1 increased the efficiency of reprogramming, indicating loss of DNA methylation is an important epigenetic barrier to reprogramming. Interestingly, partially reprogrammed iPS cells derived from mature B-cells did not reprogram in Aza alone, but did do so when the inhibitor was used in combination with knockdown of aberrantly expressed lineage-specific transcription factors. Reprogramming from MEF was also shown to be enhanced by short-term Aza treatment, and it was separately reported to be capable of replacing c-Myc in three factor (Oct4, Sox2 and Klf4) reprogramming of MEF (Huangfu et al., 2008a).

Huangfu et al. (2008a) also demonstrated that the histone deacetylase (HDAC) inhibitors suberoylanilide hydroxamic acid (SAHA), trichostatin A (TSA) and valproic acid (VPA) also increased the efficiency of reprogramming from MEF,

with more than a 100-fold increase in colony formation with VPA. VPA was also able to facilitate reprogramming in the absence of c-Myc, and a further study from this group demonstrated reprogramming of human fibroblasts with only Oct4 and Sox2 in the presence of VPA (Huangfu et al., 2008b). Another HDAC inhibitor, sodium butyrate (NaB) was also shown to aid reprogramming of both human and mouse cells (Liang et al., 2010; Mali et al., 2010). Mali et al. (2010) demonstrated the highly dynamic nature of epigenetic remodelling during reprogramming, as the use of the histone acetyltransferase (HAT) inhibitor C646 abolished the positive effect of NaB. This study also reported NaB capable of replacing c-Myc or Klf4 during three factor reprogramming of human fibroblasts. In contrast, Liang et al. (2010) observed a loss of reprogramming efficiency when NaB was used during reprogramming from MEF in the absence of c-Myc, with microarray analysis comparing NaB-treated three (without c-Myc) and four factor reprogramming cultures identifying a number of ES cell-associated genes which fail to up-regulate in the former.

Mali et al. (2010) also reported that the effects of NaB could be enhanced by the inclusion of the G9a histone methyltransferase inhibitor BIX-01294 (BIX). This molecule was also identified to increase the efficiency of reprogramming from mouse neural progenitor cells (NPCs) in a drug screen (Shi et al., 2008b). This molecule enabled Oct4 and Klf4 or Klf4, Sox2 and c-Myc mediated reprogramming of NPCs. Subsequently it was also demonstrated that Oct4 and Klf4 in combination with BIX could generate colonies from MEF (Shi et al., 2008a). Further screening with this two factor +BIX system also identified the DNMT inhibitor RG108 as increasing the efficiency of reprogramming.

1.4.3 The role of Vitamin C in reprogramming

Another small molecule which was found to improve the efficiency of generating iPS cells from MEF, Vitamin C (VitC) was identified from a mixture of antioxidant compounds (Esteban et al., 2010). Surprisingly, other antioxidants present in the mixture were unable to similarly enhance reprogramming; indicating

VitC functioned via another mechanism. Conversion of pre-iPS cells to fully reprogrammed iPS cells was also demonstrated by the use of VitC in serum-free conditions, however unlike the similar conversion observed using 2i culture conditions, Erk signalling was found to remain fully active (Silva et al., 2008). Proliferation of MEF and pre-iPS cells was also enhanced by the use of VitC, and decreased levels of the senescence related proteins p53 and p21 were detected compared to untreated cultures.

It was subsequently discovered that VitC aided reprogramming via its activity as a co-factor of iron-containing enzymes. The activity of these enzymes results in their iron atom attaining an oxidative state which is higher than the state compatible with continued activity of the enzyme. VitC functions as an electron donor, adjusting the redox state of the iron atom and thus enabling further catalytic activity of the enzyme. The Jumonji C (JmjC)-domain containing enzymes are one such family of enzymes, a sub-class of which display lysine demethylation activity (Wang et al., 2011a). Knockdown of Jhdm1b and Jhdm1a, which catalyse H3K36me2/3 demethylation, resulted in a 50% decrease in reprogramming from MEF. In contrast, over-expression of either factor, but especially Jhdm1b, increased the efficiency of reprogramming, and a further increase in reprogramming was achieved if VitC was also present in the cell medium and the combination of Jhdm1b and VitC was sufficient to enable reprogramming of MEF using Oct4 alone. Furthermore, as described by Esteban et al. (2010), VitC enhanced the proliferation of MEF in culture, and this was abolished by knockdown of Jhdm1b. This modulation of the proliferative potential was found to be as a result of the binding of Jhdm1b to the *Ink*/*Arf4* locus resulting in a loss of H3K36 methylation and a gain in H3K27 methylation, echoing the results of an earlier study which described up-regulation of PRC2 component Ezh2 and promotion of PRC1 binding to this locus mediated by Jhdm1b activity (Tzatsos et al., 2009). This is significant as products of this locus are involved in stabilisation of p53 which has been reported to act as a barrier to successful reprogramming (see Section 1.4.3.1). This study also demonstrated physical interaction of Oct4 and Jhdm1b and binding in close

proximity to each other at the promoter of the micro-RNA cluster mi-R302/367 which has also been implicated in enhancing reprogramming efficiency from both mouse and human cells (see Section 1.2.4).

Other members of the JmjC family have also been found to positively influence reprogramming, dependent on the presence of VitC (Chen et al., 2013). This study identified that the ability of VitC to direct pre-iPS cells to a more reprogrammed state was inhibited by the action of BMP signalling, originating either from the reprogramming culture (FBS) or upon addition of BMP4 in serum-free conditions. The repressive activity of BMP4 was found to be linked to maintenance of H3K9 methylation, and knockdown of the K9 histone methyltransferase Setdb1 increased the efficiency of reprogramming from both pre-iPS cells and from MEF. Down-regulation of Setdb1 was found to synergise with VitC-mediated activation of the lysine 9 demethylases Jhdm2a, Jhdm2b, Jhdm4b and Jhdm4c to enable reprogramming, highlighting the dynamic nature of epigenetic modification with factors responsible for gain and loss of methylation both contributing to the outcome of reprogramming. Demethylation of H3K27me2/3 mediated by JmjC family member Utx was also found to be crucial for reprogramming (Mansour et al., 2012). MEF and pre-B cells which did not express Utx could only undergo reprogramming in the presence of shRNA against Eed, a member of the PRC2 complex responsible for establishing H3K27me3 at developmentally important loci. In contrast to wt MEF, Utx null cells displayed a gain of K27 methylation at a number of ES cell-associated genes during reprogramming and chimeras generated from null ES cells failed to contribute to the germline. Notably, during reprogramming, Utx expression is not required once the pluripotency network has been established, reflecting the precise nature of epigenetic regulation during iPS cell generation.

Imprinting is the epigenetic mechanism by which expression of certain genes occurs only from either the maternally or paternally inherited allele, for example differentially methylated regions (DMRs) in the *Dlk-Dio3* locus are heavily methylated on the paternal allele and expression of *Gtl2* and *Rian* transcripts occurs

only from the maternal allele. Recently it has been shown that iPS cell lines display a maternal allele with a methylation pattern similar to the paternal allele. These cell lines lack expression of Gtl2 and Rian and contribute poorly to chimeric mice and are incapable of contributing to 4n, "all-iPS cell", mice via tetraploid complementation (Stadtfeld et al., 2010a). Treatment of these paternalised cells with the histone deacetylase inhibitor VPA could rescue this phenotype in a small percentage of sub-clones which became 4n competent; however a failure to correctly re-establish the imprinted state and maternally expressed imprinted genes in the Dlk1-Dio3 cluster resulted in non-viable mice. Subsequently this group also showed that reprogramming in serum-free conditions generated more iPS colonies with correct imprinting at this locus (Stadtfeld et al., 2012). It was found that VitC in the serum free culture condition was the component responsible for this effect, and in its absence Dnmt3a mediated hypermethylation of the differentially methylated regions (DMRs), which regulate imprinting, occurred at the latter stages of reprogramming after loss of histone 3 K4me2 and a failure to acquire H3K4me3. Importantly, it was also demonstrated that while iPS cells generated in the serum-free condition could be used to generate live born 4n mice, aberrantly imprinted iPS cells could not be rescued by VitC treatment. Interestingly, another JmjC family-member Kdm1b/Aof1 was found to be required for establishment of maternal imprinting in oocytes via H3K4me2/1 demethylase activity, suggesting that other family members may be required for maintenance of imprinting during reprogramming, and are reliant on VitC for their activity (Ciccone et al., 2009). It was also discovered that altering the stochiometry of the reprogramming factors could also influence the maintenance of proper imprinting at the *Dlk1-Dio3* locus, and in the presence of elevated levels of Oct4 and Klf4 the number of iPS cell lines displaying CpG hypermethylation was decreased (Carey et al., 2011).

1.4.3.1 The Ink4/Arf locus during reprogramming to iPS cells

A number of studies have recently identified p53 and associated factors as representing a major barrier to successful reprogramming. Hong et al. (2009) reported p53 null ($p53^{-/-}$) MEF underwent reprogramming 7-12 times more efficiently than their wt counterparts, and could even be used to generate reprogramming factor-integration-free iPS cells. Human dermal fibroblasts in which p53 was knocked down by shRNA also displayed increased reprogramming efficiency, and it was discovered that knockdown of p53 target p21 could also increase the efficiency of reprogramming. This study also demonstrated that over-expression of the E3 ligase Mdm2 which degrades p53 also led to greater colony formation. Similar results were obtained by Kawamura et al. (2009) who introduced a non-degradable version of Mdm4. This study also highlighted the importance of the products of the *Ink4/Arf* locus in reprogramming; Mdm2 and Mdm4 are destabilised by $p19^{Arf}$, and $p16^{Ink4a}$ activity is responsible for stabilisation of the cell cycle checkpoint protein Retinoblastoma. Knockdown of both of these factors resulted in a 5-fold increase in iPS cell colony formation from MEF. Interestingly, Li et al. (2009a) compared the activity of both factors and reported that while $p19^{Arf}$ increased reprogramming efficiency in MEF, human fibroblasts reprogrammed more efficiently in response to $p16^{Ink4a}$ knockdown. It was also demonstrated that the promoters of these genes carry both H3K4 and H3K27 methylation in ES and iPS cells, and this is regained during reprogramming. These results highlighted the important role that activation of cell-cycle checkpoint regulators plays in reprogramming.

Another aspect of these studies concerned DNA damage and how it affects the reprogramming process. Marion et al. (2009) reported that the efficiency of generating iPS cells from $p53^{-/-}$ MEF could be further increased if the cells also lacked the telomerase enzyme which is responsible for reducing DNA damage and aberrations via the regulation of telomere length. It was also shown that the introduction of DNA mutations via UV or ionising radiation resulted in a decrease in reprogramming efficiency of wt MEF, but this could be reversed by the

knockdown of p53 or over-expression of Bcl2 which inhibits the mechanism of apoptosis in cells. Finally, MEF homozygous null for the DNA damage-repair factors Atm and 53BP1 displayed decreased reprogramming efficiency compared to wt MEF indicating that DNA damage can limit the efficiency of iPS cell generation and the damage response may be mediated by p53 activation. While p53 knockdown can increase the efficiency of reprogramming as described above, the resulting iPS cells display damage to their genomic integrity. Utikal et al. (2009) demonstrated short term inhibition of p53 activity, even at relatively late time-points can still increase the efficiency of reprogramming, and so transient inhibition may increase reprogramming efficiency without accumulation of drastic DNA damage.

1.5 Markers of the reprogramming process

1.5.1 Cell surface markers

A number of studies of the reprogramming process have utilised the ES cell-associated marker SSEA-1/Fut-9 which was identified from the antisera of mice immunized with irradiated EC cells, and was found to be expressed on ES cells (Solter and Knowles, 1978). Similarly, the expression of the MEF-associated marker Thy1 (CD90) has also been used to determine reprogramming stage. Stadfeld et al. (2008a) identified the expression kinetics of these markers during reprogramming of MEF. Thy1 was rapidly down-regulated followed by up-regulation of SSEA-1 expression, and reprogramming associated retrovirus silencing increased as cells gained a Thy-1^{-}/SSEA-1^{+} phenotype. The SSEA-1 positive population appeared to give rise to cells which had re-activated the inactive X chromosome of XX MEF, and an increase in SSEA-1 expression was correlated with an increase in colony forming potential before the activation of endogenous Oct4 activity. However, it was subsequently reported that pre-iPS cells which are reliant on continued expression of exogenous reprogramming factors express SSEA-1 and that introduction of c-Myc alone can result in down-regulation of Thy1 expression (Sridharan et al., 2009). In addition, sorting of pre-iPS cells based on SSEA-1 expression revealed that both fractions carried reprogramming inhibitory DNA methylation at pluripotency gene loci, and responded equally to treatment with a DNA methyltransferase inhibitor, suggesting that SSEA-1 expression has a poor correlation to successful iPS cell generation in the latter stages of reprogramming (Mikkelsen et al., 2008). Finally, during reprogramming of p53 null cells, Thy1^{-}, Thy1^{+} and SSEA1^{+} cells were found to have very similar colony forming potential, suggesting in highly efficient systems these markers may not be useful (Utikal et al., 2009). Recently a study utilised both of these markers to carry out thorough analysis of the reprogramming process (Polo et al., 2012). This study identified SSEA-1 positive populations which eventually give rise to cells expressing endogenous pluripotency genes, and these were used for comparison across different reprogramming time-points. Principal component

analysis carried out in this study not only included these subpopulations but also pre-iPS cells derived from a variety of sources. Interestingly, despite also expressing SSEA-1, the pre-iPS cells were found to cluster completely separately from the SSEA-1^{+} populations identified in this study, casting some doubt on the use of this marker to define populations undergoing reprogramming. Another study of populations undergoing reprogramming from MEF identified up-regulation of epithelial cell adhesion molecule (EpCAM) to be correlated with a higher level of Nanog expression compared to SSEA-1 sorted cells, and EpCAM over-expression has also been reported to increase the efficiency of reprogramming (Chen et al., 2011; Huang et al., 2011). Another recent study aimed to identify ES cell-specific glycoprotein-binding surface proteins and generated a cluster of differentiation (CD) "barcode" for pluripotent cells (Gundry et al., 2012). A number of these markers, including CD326 and CD31 were described to enrich for cells expressing higher levels of endogenous pluripotency genes during reprogramming, however how this correlated to iPS cell generation was not demonstrated.

1.5.2 Pluripotency genes as reporters of reprogramming

The first iPS cells generated by Takahashi and Yamanaka (2006) were isolated based on selection for reactivation of Fbx15, an ES cell-associated gene, and downstream target of Oct4. However, these iPS cells were found to display a number of differences compared to ES cells, including an inability to give rise to adult chimeric mice. However, selection for expression of Nanog or endogenous Oct4 was shown to enable the isolation of iPS cells that could contribute to chimeras which also displayed germline transmission (Maherali et al., 2007; Okita et al., 2007; Wernig et al., 2007). It was also discovered that delayed selection resulted in a higher yield of iPS cells, highlighting the heterogeneous nature of reprogramming from MEF (Wernig et al., 2007). Brambrink et al. (2008) integrated expression of these pluripotency gene reporters into the timeline of SSEA-1 expression during reprogramming and discovered poor enrichment for Nanog and Oct4 expressing

SSEA-1^{+} cells. Isolation of colonies based on morphology alone without using selective agents was also carried out, and upon expansion these cells expressed an Oct4-GFP reporter and chimera competent iPS cell lines generated (Blelloch et al., 2007). Finally, this non-selective approach in combination with an Oct4-GFP reporter was also used to isolate lines which could contribute to 4n, "all-iPS cell", embryos (Meissner et al., 2007).

1.5.3 Nanog in ES cells and reprogramming

Reprogramming to the pluripotent state requires re-establishment of the network of factors responsible for maintenance of the pluripotent self-renewing state. Central among these factors are Oct4, Sox2 and Nanog. Nanog was identified in both a screen of factors capable of bestowing LIF-independent self-renewal on LIF receptor null ES cells upon over-expression, and a screen of factors that are over-represented in ES cells compared to somatic cells (Chambers et al., 2003; Mitsui et al., 2003). Nanog null ES cells could be generated via gene targeting and maintained on feeders but null embryos generated by crossing heterozygous null mice were not viable due to loss of the ICM, highlighting the importance of this factor in development. Subsequently it was demonstrated that Nanog null ES cells aggregated with wt cells could contribute to chimeras, but primordial germ cells (PGCs) failed to mature, in line with the expression of Nanog in these cells (Chambers et al., 2007). Niwa et al. (2000) used an inducible system to show that Oct4 expression must be maintained at precise levels in ES cells with differentiation to primitive endoderm and mesoderm upon Oct4 over-expression, and under-expression inducing cells to differentiate to trophectoderm. Similarly, Sox2 silencing causes differentiation of ES cells to multiple lineages including trophectoderm (Masui et al., 2007). Interestingly, Sox2 and Oct4 have been shown to co-bind the promoters of many genes expressed in pluripotent cells, but these factors were not down-regulated in Sox2 null cells due to substitution by Sox family members Sox4, Sox5 and Sox11. However, loss of Sox2 decreases expression of positive- and

increases expression of negative-regulators of Oct4 and genes associated with differentiation, resulting in the observed differentiation phenotype, which can be rescued by Oct4 over-expression. The interaction between these factors is further strengthened by identification of SOX2 binding to the *Pou5f1* promoter which regulates expression of Oct4, and binding of OCT4 to the *Sox2* promoter in mouse and human ES cells (Chew et al., 2005; Okumura-Nakanishi et al., 2005). Subsequently the Oct4/Sox2 binding motif was identified in the promoter region of Nanog in mouse and human ES cells, mutation of which lead to a decrease in Nanog expression, as did knock-down of Oct4 or Sox2 (Rodda et al., 2005). This finding linked these three crucial factors and indicated the importance of their interaction in order to maintain ES cell identity. Interestingly, Nanog expression was found to be heterogeneous in ES cells, with cells transitioning between Nanog negative and positive states (Chambers et al., 2007). Navarro at al. (2012) demonstrated this phenomenon occurred independently of Oct4 and Sox2, and NANOG protein binding to the *Nanog* promoter prevents Nanog expression in an auto-repression loop. It was also shown that Zfp281 knockdown increased Nanog expression and was required for recruitment of the silencing-associated NuRD complex to the *Nanog* locus (Fidalgo et al., 2012). Similarly, knockdown of Zfp281 enhanced 2i mediated reprogramming of pre-iPS cells to iPS cells by enhancing Nanog expression without the NANOG associated self-inhibition.

Chromatin immunoprecipitation (ChIP) analysis revealed that about half of identified Oct4 targets in ES cells were also bound by Nanog (Loh et al., 2006). Another analysis of Sox2 and Oct4 binding sites in human ES cells identified Nanog co-occupation at a high percentage (>90%) of targets including factors involved in chromatin remodelling and ES cell specific transcription factors as well as factors important for lineage specification, indicating Nanog, Oct4 and Sox2 may negatively, as well as positively regulate their common targets (Boyer et al., 2005). ChIP analysis carried out via pull-down of biotinylated Nanog, Oct4 and Sox2 identified enrichment for bivalent marks at common target sites, with a correlation found between the number of pluripotency associated genes binding to a site and

presence of H3K4me3 (Kim et al., 2008a). Interestingly, genes rapidly and highly up-regulated upon ES cell differentiation were found to usually be bound by only a single ES cell-associated factor, and enriched for K27me3 in ES cells, indicating while ES cell-factor interaction may be required to maintain the pluripotent state, individual factors may repress key lineage specifiers. Multiple ES cell factor binding was also identified via ChIP-sequencing (ChIP-seq) analysis carried out in ES cells, and it was also discovered that short fragment sequences derived from sites of Nanog, Oct4 and Sox2 co-binding contained enhancer activity (Chen et al., 2008). The large number of individual and common targets for these genes in ES cells highlights the important role they each play in maintenance of the pluripotent state. The central role of Nanog has been investigated with a large number of factors reported to comprise the NANOG interactome, including proteins related to processes as diverse as RNA processing, cell cycle control and DNA replication and repair (Gagliardi et al., 2013).

The role of Nanog in reprogramming is also a source of much interest. Silva et al. (2006) demonstrated that reprogramming of neural stem (NS) cells, MEF and thymocytes via ES cell fusion was enhanced upon over-expression of Nanog. Reprogramming from EpiSCs to iPS cells was also demonstrated in the presence of Nanog over-expression which was possible in both 2i and LIF/serum ES cell conditions (Silva et al., 2009). Subsequently it was also found that Nanog could reprogram pre-iPS cells and EpiSCs in 2i media in the absence of LIF, echoing its activity in fully reprogrammed ES cells (Theunissen et al., 2011).

Another marker of completed reprogramming to an ES cell-like state is the reactivation of the inactive X (Xi) chromosome in XX cells. X inactivation is a compensatory mechanism in female embryos which is reversed in cells of the ICM. One of the active X (Xa) chromosomes is inactivated randomly during differentiation of the epiblast, resulting in all somatic differentiated cells bearing a Xa and Xi. It has been shown that culturing blastocysts in 2i conditions results in an increase in the number of NANOG positive cells, enabling expansion of the epiblast (Nichols et al., 2009b). In XX embryos Nanog expression after E4.5 was identified to

correlate with reactivation of the inactive X (Xi) chromosome, and mark cells with both active X (Xa) chromosomes (XaXa cells) (Silva et al., 2009). In the absence of Nanog, the ICM was found to rapidly degenerate and Xi failed to reactivate. The link between these events was confirmed by Navarro et al. (2008) who identified not only NANOG, but also OCT4 and SOX2 binding and repression of the promoter region of the non-coding RNA *Xist* which is responsible for coating and inactivating the paternal X chromosome from which it is expressed. While Nanog null ES cells up-regulate *Xist* expression, this is only to a level many fold lower than that observed under normal conditions. It was found that effective levels could be achieved by the disruption of the hierarchy of repressive-factor removal from the promoter via rapid loss of OCT4, indicating that loss of Nanog, Oct4 and Sox2 expression and subsequent differentiation regulates *Xist* expression. In addition, it has also been demonstrated that expression of the repressor of *Xist* activity, *Tsix*, is reliant on other ES cell factors Rex1, Klf4 and c-Myc (Navarro et al., 2010). This demonstrates the numerous and inter-connected roles that the core factors Nanog, Oct4 and Sox2 carry out in co-operation with other ES cell-associated proteins, and suggests in order to achieve complete reprogramming a highly complex network must be re-established.

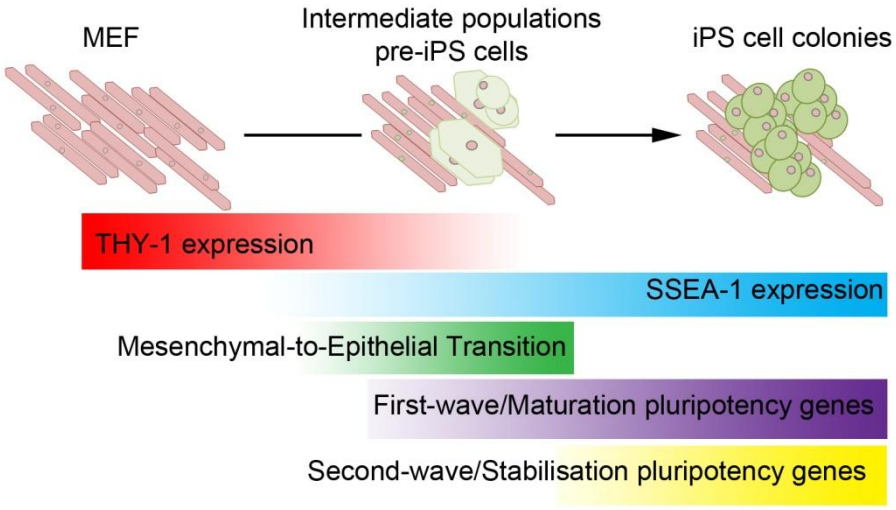


Figure 1.2 Overview and timing of appearance of previously described markers and indicators of the reprogramming process.

1.6 Transdifferentiation and reprogramming

1.6.1 Inducing changes in cell behaviour

The ability to use transcription factors to change the fate of a cell to that of another somatic cell, transdifferentiation, was demonstrated by Davis et al (1987) with the identification of MyoD. This was the result of comparing proliferating myoblasts and fibroblasts in order to identify factors capable of inducing myogenic lineage specification from mouse fibroblasts at a high efficiency. This was based on the earlier observation that myofibroblast DNA treated with the methyltransferase inhibitor Aza was capable of converting fibroblasts to myofibroblasts at a low efficiency (Lassar et al., 1986). Together this data indicated that transdifferentiation requires cells to overcome the epigenetic barriers required for the normal maintenance of cellular lineage. The over-expression of Gata-1 was found to similarly convert myeloblasts to hematopoetic lineages in the chicken (Kulessa et al., 1995). Other transcription factors identified to induce transdifferentation include C/EBPα and C/EBPβ which induced B cell transformation to macrophages, and mature B cell conversion was demonstrated *in vitro* and *in vivo* (Kan et al., 2004). Subsequently it was shown that these factors along with their interacting partner PU1 were capable of transdifferentiating primary mouse fibroblasts to macrophages, indicating how these studies can enhance understanding of the processes required for cell specification and lineage determination (Feng et al., 2008). However, continued expression of the exogenous factors was required as endogenous levels failed to become sufficiently up-regulated, highlighting a potential issue for the use of transdifferentiated cells in a therapeutic context. Interestingly, another study which demonstrated transdifferentiation of exocrine pancreatic cells into β cells found that cells did not appear to de-differentiate into a common progenitor, and cell behaviour was altered in a direct manner (Zhou et al., 2008).

1.6.2 Transdifferentiation towards specific lineages

Transdifferentiation to cardiomyocytes has been intensively studied due to the potential use of these cells for the treatment of infarct damaged cardiac tissue. Screening of transcription factors using a cardiomyocyte-specific reporter identified three factors capable of generating beating cardiomyocytes from mouse dermal fibroblasts (Ieda et al., 2010). A further study also identified micoRNAs involved in cardiac muscle development and differentiation capable of inducing direct reprogramming to cardiomyocyte-like cells (Jayawardena et al., 2012). *In vivo* injection of these miRNAs into injured cardiac tissue showed these cells were capable of incorporation to the heart. Similarly, cardiac fibroblasts could be transformed into cardiomyocytes *in vivo* using the factors identified by Ieda et al. (2010), and this was shown to improve heart function after infarction, with this beneficial affect enhanced by the addition of an additional factor (Qian et al., 2012; Song et al., 2012).

Induction of neural lineages from more readily accessible tissues has also been heavily investigated. Wernig and colleagues identified three transcription factors (3Fs) capable of transdifferentiating MEF to mature neural cells (Vierbuchen et al., 2010). This conversion occurred relatively rapidly with morphology changes obvious after three days, indicating this method may be relatively epigenetically favourable compared to reprogramming to iPS cells from MEF. Human fibroblasts could be induced to become neuronal cells by altering one of the previously identified 3Fs (Qiang et al., 2011). These cells were capable of migrating from ventricles throughout injected mouse brains indicating their complete gain of functional properties of neuronal cells. In addition, this study reported generation of neurons from Parkinson's Disease (PD) patients which were subsequently used to identify processes pertinent to treatment of the diseased cells, indicating the usefulness of this technique. Both transcription factors NeuroD1 and NeuroD2 and microRNAs, factors required for neuronal specification, were subsequently identified to induce neurons from both murine and human fibroblasts (Vierbuchen

et al., 2010; Ambasudhan et al., 2011). Transdifferentiation towards more specific neural cell types has also been demonstrated, providing a more accurate method of production of desired lines, including dopaminergic and spinal motor neurons (Caiazzo et al., 2011; Pfisterer et al., 2011; Son et al., 2011).

While these reports stressed that cells did not pass through a de-differentiated, neural stem cell-like state, the generation of multipotent neural progenitor cells (NPCs) has also been demonstrated. Two factors were shown to generate bi-potent NPCs from mouse fibroblasts, and the addition of an extra factor was shown to be capable of generating tri-potent NPCs (Lujan et al., 2012). One of these factors included the commonly used reprogramming factor Sox2. Subsequently it was shown that Sox2 alone could induce NPCs from human and mouse fibroblasts (Ring et al., 2012). In combination with another reprogramming factor, c-Myc, Sox2 could induce neurons from cord blood, and both of these factors with yet another reprogramming factor, Klf4 could induce MEF to become NSCs capable of *in vivo* differentiation to all three neural lineages (Giorgetti et al., 2012; Han et al., 2012). Interestingly, Oct4, another reprogramming factor, could be used to transdifferentiate cells from human fibroblasts to hematopoetic fates (Szabo et al., 2010). Hematopoetic growth factors were used to induce fibroblasts to attain a hematopoetic stem cell (HSC) marker, CD45. By modulating the culture environment progenitors could be expanded and many lineages established including monocytes which matured to macrophages. These progenitors could engraft *in vivo* and were functionally competent. A combination of all the above reprogramming factors (c-Myc, Klf4, Oct4 and Sox2) has also been used in transdifferentiation of MEF to both NPCs and cardiomyocytes (Efe et al., 2011; Kim et al., 2011). In both cases lineage specific reporters and transient expression of the factors was required to generate the desired cell types at a high efficiency. These results demonstrated that the reprogramming factors may be used to induce de-differentiation to a required point from which progenitor cells can be derived, without complete reprogramming to pluripotency. Transdifferentiation provides

another method by which differentiation and derivation of desired cell types can be improved, and with some methods reporting efficiencies of over 20% for certain lineages, this technique should be investigated in a complimentary manner with iPS cell studies in order to understand the plasticity and potential of cell fate.

1.7 Aims of this thesis

The aim of my thesis is to identify and utilise novel cell surface markers in order to accurately investigate the molecular mechanism of reprogramming from mouse embryonic fibroblasts (MEF). Despite intensive study of reprogramming, generation of induced pluripotent stem (iPS) cells remains a highly inefficient process.

Microarray analysis of transgenic MEF undergoing reprogramming previously carried out in our lab identified a number of cell surface markers which displayed dynamic expression patterns during the reprogramming process. In order to investigate the potential usefulness of these markers I initially employed an *in vitro* reprogramming system using Nanog-GFP reporter MEF and monitored changes in marker expression in relation to the widely-used ES cell marker, SSEA-1. This analysis was capable of correlating the expression of these novel markers with iPS cell colony formation and re-acquisition of Nanog-GFP expression, and identified a number of sub-populations which arose during reprogramming.

In order to validate these observations during MEF reprogramming, a highly efficient secondary system was utilised. Reprogramming sub-populations were isolated from a single time-point based on expression of the novel markers and Nanog-GFP. Each of the sub-populations displayed distinct potentials to generate iPS cells. The transition from one sub-population to the next was monitored and revealed the major routes available to cells undergoing reprogramming from MEF.

RNA sequencing analysis of the isolated sub-populations revealed the differentially expressed genes between each stage of reprogramming. A number of gene expression patterns could be identified from this set of genes. Further investigation revealed that epidermis-associated genes are transiently up-regulated exclusively in the intermediate stages of reprogramming. This revealed that the process of reprogramming is more complex than the loss of MEF genes and gain of pluripotency genes.

In addition, this reprogramming marker system was also shown to be capable of identifying differences between different reprogramming systems and culture conditions. Comparison of a less efficient system also identified two novel genes which could be used to increase the efficiency of reprogramming.

Further investigation of the features of reprogramming identified in this study can potentially contribute to important insights into the molecular mechanism of the process of iPS cell generation.

CHAPTER 2 - Materials and Methods

2.1 Culture and Maintenance of mammalian cell lines

2.1.1 Cell culture materials

GMEM Complete medium:

Glasgow Minimal Essential Medium (GMEM, Sigma G5154)

Foetal Calf Serum (10%)

Non-essential amino acids (1x, Gibco 11140-035)

L-Glutamine (2mM, Invitrogen)

Sodium pyruvate (Invitrogen)

β-mercaptoethanol (100μM, BDH 441413)

1000U LIF (human recombinant)

Penicillin/Streptomycin (100 units/100ug, Sigma P4333)

MEF expansion medium:

GMEM complete medium

bFGF/FGF2 (5ng/mL, Peprotech 100-18-B)

Heparin (1ng/mL, Sigma)

Reprogramming medium:

GMEM complete medium

A-83-01/ALKi (500nM, TOCRIS Bioscience #2934)

Ascorbic Acid/VitaminC (10ug/mL, Sigma 1000731348)

Doxycycline (dox) (various concentrations as outlined in results, Sigma)

Freezing solution:

DMSO (10%, VWR International)

FCS (90%)

Other cell culture solutions:

Gelatine (0.1% in PBS, Sigma G5154)

Trypsin (0.25%,, Gibco 15090-046)

EDTA (0.1%, Sigma 03620)

PBS (Sigma D8537)

2.1.2 Cell lines used in this study

E14 ES cell line

Mouse embryonic stem cell line which demonstrates high contribution to chimeric embryos derived in by Dr. Martin Hooper (Hooper et al., 1987).

6c iPS cell line

iPS cell line generated through the use of *piggyBac* mediated delivery of individual reprogramming factors from Rosa26$^{M2rtTA/+}$ MEF (Woltjen et al., 2009).

IRI1 iPS cell line

iPS cell line derived from primary reprogramming of Rosa26$^{M2rtTA/+}$ MEF using polycistronic vector pB TAP IRI MKOSimO.

D6s4B5 iPS cell line

iPS cell line derived from Rosa26$^{M2rtTA/+}$ Nanog$^{+/GFP}$ MEF using polycistronic vector pB TAP IRI MKOSimO. Vector was integrated to *Sp3* locus as identified via splinkerette PCR.

TANGO ES cell line

iPS cell line derived from targeting Rosa26$^{M2rtTA/+}$ Nanog$^{+/GFP}$ ES cells with pB TAP IRI MKOSimO vector to the *Sp3* locus.

129 MEF cell line

MEF derived from the widely used 129/SvEv strain of mice.

D6s4B5 and TANGO chimera-derived MEF lines

2.1.3 ES cell and iPS cell culture technique

ES cells and established iPS cell line D6s4B5 were cultured in GMEM complete medium. Cells were cultured at 37°C at 7.5% CO_2 in humidified incubators. Cells were passaged upon reaching ~80% confluency via harvesting in trypsin, re-suspension in GMEM complete medium, collection by centrifugation at 300g and resuspension in GMEM complete medium. Cells were passaged every 2-3 days.

2.1.4 Reprogramming from retinoic acid differentiated iPS cells

iPS cell clones were plated in at a density of $1X10^4$ cells per well of a gelatinized 6 well dish in low LIF (10 units) GMEM complete media supplemented with retinoic acid (10^{-6}M, Sigma R2625) and cultured for a period of five days (Smith, 1991). Thereafter the medium was changed to GMEM complete medium supplemented with dox (1.5ug mL^{-1}) and cultures observed. Cells were harvested for flow cytometric analysis and fluorescence-activated cell sorting (FACS) at desired time-points via harvesting as for ES/iPS cells but re-suspended in PBS at a concentration of $1X10^6$ cells per mL for antibody staining.

2.1.5 Generation of chimeric embryos

ES and iPS cell lines used to generate chimeras were cultured in GMEM complete medium. 48 hours before use, $2X10^6$ cells in 2mL of GMEM complete medium were plated in a gelatinized 6-well plate in a doubling dilution series. Colonies of the appropriate size (5-8 cells for aggregation, 10-15 cells for injection) were identified from each well, harvested and introduced to morulas or blastocysts

of C57B1/6 mice. All manipulation of embryos was carried out by the staff of the Transgenics Unit including Jan Ure, Lynsey Robertson and Sally Inverarity.

2.1.6 Mouse embryonic fibroblast isolation and cell culture technique

Mouse embryonic fibroblasts (MEFs) were isolated from embryonic day 12.5 (E12.5) embryos. Embryos were decapitated, eviscerated, dissociated with 0.25% trypsin, 0.1% EDTA and plated in MEF expansion medium in $10cm^2$ dishes (Iwaki) MEF were cultured at 37°C at 7.5% CO_2 in humidified incubators. Upon reaching confluence MEF were stocked in the following manner: cells were washed in PBS, harvested in trypsin, collected via centrifugation at 300g and re-suspended in freezing solution at a concentration of 2 - $5X10^6$ cells per mL. Cells were stored at -80°C for 24 hours and transferred to liquid nitrogen (LN) thereafter. Cells were defrosted at 37°C, re-suspended in MEF expansion medium, collected via centrifugation and re-suspended and re-plated as required.

2.1.6.1 Quantification of transgene containing MEF

For MEF isolated from chimeric embryos: one twentieth of the isolated cells were plated in reprogramming medium in one well of a 12-well culture dish (Iwaki). The total percentage of transgenic cells as indicated by mOrange reporter expression was quantified after 48 hours via flow cytometry (BD Fortessa).

2.1.7 Reprogramming from MEF

The following techniques were used depending on the required outcome of the reprogramming experiment. As a general rule, transgenic MEF and 129 stocks were plated two (D-2) to four (D-4) days before the start of reprogramming in MEF expansion medium and passaged as required. MEF beyond passage 3 (p3) were not used for sorting experiments.

2.1.7.1 Reprogramming for time-course analysis

This technique was generally used in order to carry out flow cytometric analysis at multiple time-points of the reprogramming process. $1X10^5$ MEF containing 5% transgenic cells were plated in each well of a gelatinized 6-well tissue culture plate (Iwaki). 129 (wt) MEF were used in order to dilute transgenic stocks to the required percentage. Cells were plated directly into reprogramming medium (2mL/well) = Day 0 of reprogramming (D0). Culture medium was replaced every two days. Cells were harvested using trypsin at each desired time-point, collected via centrifugation at 300g, re-suspended in an appropriate volume of PBS, total cell number quantified, centrifuged again, and re-suspended in PBS for antibody staining at a concentration of $1X10^6$ cells per mL.

2.1.7.2 Reprogramming for endpoint analysis

This technique was used in order to identify the effect of changes to reprogramming conditions e.g. additional small molecules, genes of interest, or to examine colony formation during reprogramming. $3X10^5$ MEF containing 1% transgenic cells (adjusted as before) were plated in gelatinized $10cm^2$ tissue culture plates, directly into reprogramming medium (8mL/well) = D0. Culture medium was replaced every three days.

2.1.7.3 Reprogramming for fluorescence-activated cell sorting (FACS)

This technique was used in order to generate sufficient numbers of cells for isolation for RNA extraction and functional analysis of desired populations. $2X10^5$ MEF containing from 10% to 50% transgenic cells were plated on gelatinized $10cm^2$ tissue culture plates directly into reprogramming medium. The number of dishes required was based on the day of sorting, enrichment of the population

required at that time-point (as calculated via time-course experiments) and the starting percentage of transgenic MEF. Medium was changed every three days to D5 and every two days thereafter. Cells were harvested for antibody staining as per time-course analysis (see above).

2.1.8 Transfection of MEF with genes of interest

TNG MEF were harvested and re-plated in MEF expansion medium in the absence of Pen/Strep at a density of $1.5X10^5$ cells per well of a 6-well dish. The following day medium was removed and replenished with 500uL fresh medium. For each well per over-expression construct the following DNA master mix was used: PB-TAP IRI 2LMKOSimO (0.5µg), CAG-rtTA (0.5µg), HyPBase transposase helper plasmid (0.5µg). 1ug per well of the over-expression construct was added to this mix. 110uL OptiMEM (Sigma 31985-062) was added to this mix and mixed well. 10uL FugeneHD (Promega E2311) per well was added to this mix and mixed well. 115uL of this solution was immediately transferred to each well of MEF. The following day medium was changed to 2mL of reprogramming medium (1ug/mL dox). Wells were observed for differences in colony emergence and mOrange and Nanog-GFP expression compared to empty vector control wells. Each experiment was carried out in duplicate wells.

2.2 Flow Cytometry and Immunohistochemistry

2.2.1 Flow cytometry materials

FACS Buffer (FB): 2% FCS in PBS

Cell strainers (40um, Stemcell Technologies 27305)

FACS tubes (BD Falcon 3520)

2.2.2 Antibodies used for flow cytometry and FACS analysis

Antibody	Clone	Host + Isotype	Working Dilution	eBioscience Cat. Number
Anti-Mouse CD54 (ICAM-1) Biotin	YN1/1.7.4	Rat IgG2b, kappa	1/100	13-0541
Anti-Human/Mouse CD44 APC	IM7	Rat IgG2b, kappa	1/300	17-0441
Anti-Human/Mouse CD44 Biotin	IM7	Rat IgG2b, kappa	1/100	13-0441
Anti-Mouse Ly-6A/E (Sca-1) Biotin	D7	Rat IgG2a, kappa	1/100	13-5981
Streptavidin PE-Cy7	-	-	1/1500	25-4317
Anti-Human/Mouse SSEA-1 647	eBioMc	Mouse IgM	1/50	51-8813
Anti-Mouse/Rat CD90.1 (Thy-1.1) APC	HIS51	Mouse IgG2a, kappa	1/100	17-0900
Anti-CD324 (E-cadherin) Biotin	DECMA-1	Rat IgG1	1/100	13-3249

2.2.3 Antibody staining technique for flow cytometry and FACS analysis

Harvested cells in PBS were spun down and re-suspended in FB at a concentration of $1X10^6$ cells per mL. Primary antibodies were added and samples were stored on ice for 15-30 minutes. Cells were collected by centrifugation and re-suspended in FB and this was repeated. Secondary antibodies were then added to the FB-re-suspended samples and samples were incubated for 5-10 minutes on ice.

Wash steps were repeated as before. Finally, cells were transferred to FACS tubes at a concentration of $2X10^6$ cells per mL for flow cytometry time-course analysis (BD LSRFortessa). For FACS analysis (BD FACSAriaII) cells were re-suspended at a concentration of $5X10^6$ cells per mL and passed twice through a cell strainer to ensure single-cell suspensions of cells for analysis.

2.2.4 Instrument settings for flow cytometric and FACS analysis

Flow Cytometry BD LSR Fortessa		**Excitation Line**		
		488 nm	561 nm	640 nm
Band pass (BP) filter	530±30	GFP		
	582±15		mOrange	
	780±60		PECy7	
	670±30			APC

FACS Analysis BD FACSAria		**Excitation Line**			
		488 nm	405 nm	561 nm	640 nm
Band pass (BP) filter	450±50		DAPI		
	525±50	GFP			
	582±15			mOrange	
	780±60			PECy7	
	670±14				APC

Figure 2.1 Settings for flow cytometry and FACS analysis. Excitation line refers to the laser used to excite the fluorophore of choice. Band pass (BP) filter indicates the wavelength of light detected by the instrument, with the range either side of this specific wavelength represented (i.e. 530±30 = range of 515-545 nm).

2.2.5 Data processing and gating strategy of flow cytometric and FACS analysis

Gating of cells using forward- and side-scatter (FSC, SSC) parameters was broad in line with the diverse morphologies present in cultures of cells undergoing reprogramming i.e. larger fibroblastic/differentiated cells and smaller reprogramming intermediates/iPS cells. Transgenic cells were identified based on mOrange expression. Transgenic cells were then gated based on Nanog-GFP reporter expression, with Nanog-GFP^{+} and –GFP^{-} cells pseudo-coloured green and red respectively. An example of the typical gating strategy used, and control MEF population is shown in Figure 2.2. Compensation, the correction of overlap or interference between different fluorophores during acquisition, was carried out electronically using DIVA software (BD). Sample data was analysed using FlowJo software (Tree Star). Operation of FACSAria was carried out by Simon Monard and Olivia Rodriges.

2.2.6 Collection, processing and analysis of FACS sorted populations

2.2.6.1 Cells for colony formation assay

This procedure was carried out in order to assess differences in the colony forming potential between assorted populations. $1X10^{4}$ cells were sorted into FACS tubes containing 2mL GMEM complete medium and stored on ice until all populations were isolated. Cells were centrifuged and re-suspended in an appropriate amount of reprogramming medium. $3.5X10^{3}$ cells were plated in a total volume of 1mL reprogramming medium on irradiated feeder MEF (4000 rads, prepared in-house by tissue culture staff) in a single well of a 12 well plate, in duplicate. Medium was changed every two days. mOrange and Nanog-GFP expression in colonies was quantified on D10 post-sort via fluorescent microscopy (Olympus IX51).

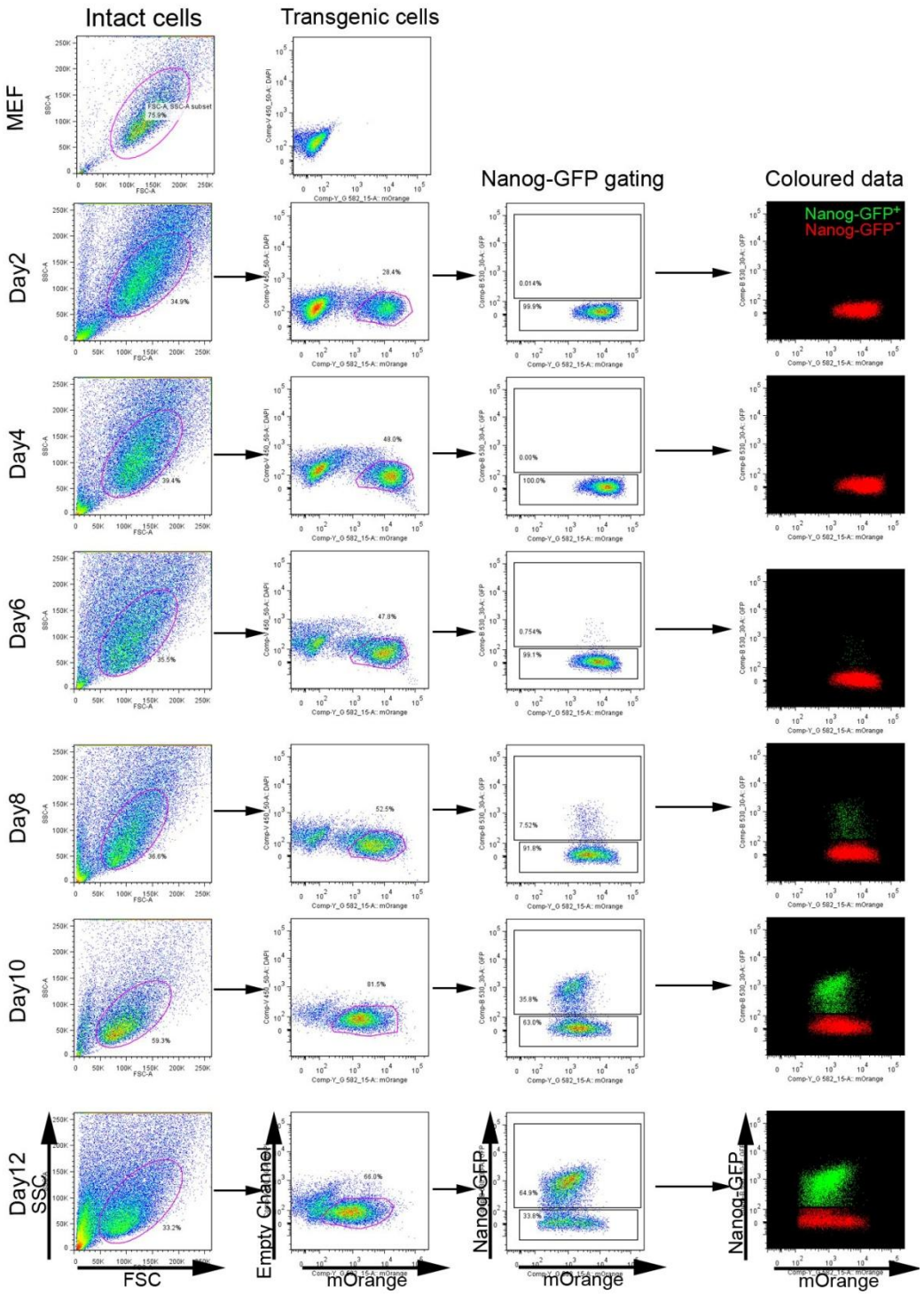


Figure 2.2 Gating strategy for flow cytometry and FACS analysis. Cells were gated based on size (Intact Cells), mOrange expression (Transgenic Cells), Nanog-GFP expression (Nanog-GFP gating) and pseudo-coloured based on Nanog-GFP expression (Coloured data). SSC = Side scatter. FSC = Forward scatter.

2.2.6.2 Cells for population movement analysis

This technique was used in order to investigate changes in cell surface marker and reporter expression from isolated cell populations. $1X10^4$ cells per population, per post-sort time-point required were sorted into 2mL GMEM complete medium and stored on ice until all populations were isolated. Cells were centrifuged and re-suspended in reprogramming medium. $1X10^4$ cells per post-sort time-point were plated in a single well of a gelatinized 48-well plate in a total volume of 1mL reprogramming medium. Medium was changed every two days. Cells were harvested for flow cytometric analysis as per time-course analysis.

2.2.6.3 Cells for RNA isolation

This technique was used to ensure sufficient quantities of cells from which RNA could be isolated for further processing. A minimum of $2X10^5$ cells and a maximum of $7X10^5$ cells were sorted from the desired populations into 1mL GMEM complete medium. Cells were stored on ice until all populations were isolated. Cells were washed twice via centrifuge collection and re-suspension in PBS. After the second wash the cell pellet was re-suspended in 500uL Trizol (Invitrogen) and immediately transferred to -80°C for storage.

2.2.7 Immunohistochemistry solutions

PBS

PFA/Paraformaldehyde (4% in PBS, Sigma)

Triton-X (0.1% in PBS, Sigma)

Blocking solution: Skimmed milk powder (1%) + Tween 20 (0.01%) in PBS

2.2.8 Immunohistochemistry antibodies

Antibody	Clonality	Host + Isotype	Working Dilution	Supplier/ Cat. Number
Anti-Mouse/Rat SFN	Polyclonal	Rabbit IgG2b,k	1/400	Sigma/ HPA011105
Anti-Human/Mouse KRT17	Monoclonal	Rabbit IgG2b,k	1/250	LSBio/ LS-C24979
Anti-Rabbit IgG cf-633	Polyclonal	Goat IgG	1/1000	Sigma/ SAB4600404

2.2.9 Immunohistochemistry methods

For staining of cells with keratin markers, transgenic MEF were plated as for time-course reprogramming. For staining, cells were washed in PBS twice and fixed in PFA at room temperature (RT) for 10 minutes. PFA was removed, Triton-X added for 1 hour at RT. Triton-X was removed and cells were incubated at RT for 2-3 hours in blocking solution. Primary antibodies were diluted in blocking solution and added to the fixed cells for incubation overnight at 4°C. The following day the cells were washed in PBS three times, with 5 minutes incubation in PBS at RT between washes. Secondary antibody was diluted in the blocking solution and incubated with the washed cells at 4°C for 1 hour. Cells were washed in PBS three times as before and imaged immediately. Cells were stored at 4°C in PBS.

2.3 RNA sequencing and analysis

2.3.1 RNA isolation

Trizol suspended samples were defrosted on ice. 100uL of chloroform was added and samples were vortexed. Samples were left at RT for 15min and centrifuged at 12,000g for 15min at 4°C. The upper aqueous layer containing RNA was carefully separated from the protein/lipid fraction and the DNA interphase, and transferred to a 1.5mL Eppendorf. 250uL of isopropanol was added and samples vortexed. Samples were incubated at RT for 10min and centrifuged at 12,000g for 20min at 4°C. The supernatant was aspirated and the remaining RNA pellet was washed in 500uL 75% ethanol, vortexed and centrifuged at 12,000g for 5min. The supernatant was removed and the pellet allowed to dry at RT for 5-10min. The pellet was re-suspended in 15uL sterile, RNAse-free water and stored at -80°C.

2.3.2 Preparation of samples for RNA sequencing

2.3.2.1 Measuring sample concentration

Due to the highly accurate requirements of the RNA sequencing (RNAseq) procedure used in this study, the concentration and quality of the RNA samples were measured for using the Bioanalyzer 2100 (Agilent). 1uL of RNA in water was diluted in 5uL sterile, RNAse-free water for analysis. In brief, samples are loaded onto a chip and run through a polymer gel and mixed with a fluorescent dye. Current applied to the chip enables separation of the charged, dye-intercalated, RNA based on size. Good quality, intact, RNA will display two distinct peaks corresponding to 18S and 28S ribosomal RNA (rRNA) subunits, with a 28s/18S rRNA ratio greater than 2. In addition, a software algorithm devises the RNA Integrity Number (RIN) which determines the quality of the sample with an RIN of

10 representing the most intact, highest quality RNA. All samples used for RNAseq analysis in this study had RINs of ≥8.9.

2.3.2.2 Sample dilution for RNAseq analysis

Samples were diluted to a concentration of 10ng RNA per μL in the running buffer for RNAseq: 10mM Tris pH 7.6 with 0.05% Tween-20. 10uL of each diluted sample was aliquoted into RNAse-free 200uL PCR tubes and stored at -80°C.

2.3.3 Multiplexed RNAseq protocol

All RNAseq analysis was carried out by the lab. of Dr. Sten Linnarsson as previously described (Islam et al., 2011; Islam et al., 2012). Briefly, for each sample, polyA-carrying mRNA undergoes reverse transcription to cDNA via a universal oligo-dT primer. The reverse transcriptase enzyme reads through to the 5′ end of the mRNA, and is induced to undergo template switching to read from a second primer carrying a unique 6-base barcode sequence and sequence for cDNA amplification, thus integrating these components into the nascent cDNA strand. The bar-coded samples are then pooled therefore all subsequent steps occur simultaneously, reducing sample-to-sample variation. The cDNA is then bead purified and full length cDNA amplification occurs, and the quality of the cDNA is checked at this point. The full-length cDNA is then fragmented, blunted and dA-tailed in order to enable ligation of an amplification oligo to the 3′ end of the molecule. cDNA library amplification then occurs, and a final oligo introduces the 5′ sequencing motif to the molecules. Samples then underwent sequencing with Illumina HiSeq 2000 (Illumina).

2.3.4 RNAseq data analysis

All RNAseq data analysis was carried out by members of the Bioinformatics core under the supervision of Dr. Simon Tomlinson at the SCRM. Quality control of the obtained reads and alignment to the mouse reference genome (NCBI37/mm9) was performed using the GeneProf web-based analysis suite with default parameters (Halbritter et al., 2012).

Differentially expressed genes (DEGs) were identified, using the edgeR and DESeq Bioconductor libraries (Gentleman et al., 2004; Anders and Huber, 2010; Robinson et al., 2010). For both methods, low expression transcripts (less than 13 reads in all samples) were filtered out and P-values were adjusted using a threshold for false discovery rate (FDR) ≤ 0.05. Genes listed as DEGs by both methods in any two subpopulation comparison (Figure 5.3a, total 3,171) were used for further analysis. Hierarchical clustering and K-means clustering (K=5) of DEGs was performed using Cluster 3.0 and Java Treeview was used for visualisation (de Hoon et al., 2004; Saldanha, 2004).

Principal Components Analysis (PCA) was performed in R and plotted with the scatterplot3d library (Ligges, 2003). Gene ontology (GO) enrichment was calculated using the DAVID functional annotation bioinformatics tool (Huang da et al., 2009). GO term enrichment analysis was carried out with a modified Fisher Exact p-value.

The three additional published studies (Sridharan et al., 2009; Samavarchi-Tehrani et al., 2010; Polo et al., 2012) (GEO accession number GSE21757, GSE14012, GSE42379) were analysed in a similar way. For the time course data the analysis was performed as following: data were RMA (Irizarry et al., 2003) normalised using the Expression Console from Affymetrix and, since no replicates were provided, fold changes (FC) between each two samples were calculated in Excel. For the Plath and Polo dataset, data were RMA normalised using the 'affy' package (Gautier, 2004) in R. Selected gene expression data shown as relative expression against the highest

signal among the samples using an averaged signal value (reads per million) of duplicates/triplicates.

2.4 Molecular Biology Techniques

2.4.1 Plasmids

ZeroBluntII TOPO	Plasmid containing ccdB expression cassettes for negative expression, ColE1and F1 origins of replication. Kanamycin resistant.
pENTR hCD2	Plasmid containing IRES-linked human CD2 cDNA followed by bovine growth hormone polyadenylation site (bpA), and flanked by the Gateway recombination sites attL1 and attL2. Kanamycin resistant.
pB TAP IRI	Plasmid containing the doxycycline-inducible TetO element up-stream of the mCMV promoter. Contains Gateway recombination sites attR1 and attR2 which flank the bacterial-toxic ccdB gene. A human lamin B2 replicator sequence is located downstream of the attR2 site. Upstream of the TetO element and downstream of the replicator sequence are chicken β2 insulator sequences. All of these elements are flanked by *piggyBac* transposon long terminal repeats. Ampicillin resistant.
pB TAP GOI hCD2	PB TAP IRI plasmid after Gateway recombination with pENTR hCD2 plasmid containing gene-of-interest (GOI) for over-expression. Ampicillin resistant.

2.4.2 cDNA synthesis

cDNA was synthesised from E14 ES cells RNA isolated using Trizol (see section 2.3.1). 1ug RNA was used. Oligo dT (100uM, T24) was added along with separate deoxyribonucleotides (100mM each, Invitrogen55082/3/4/5). Samples were

made up to 10uL and incubated at 65°C for 5min, followed by 37°C. Buffer (1X, Invitrogen Y02321), DTT (Invitrogen Y00147), RNaseIN (NEB, M03141) and the reverse transcriptase MMLV (Invitrogen 28025-013) were added. Samples were incubated at 37°C for 1hour, 90°C for 10min and cooled to 4°C. Samples were stored at -80°C.

2.4.3 Gene of interest cDNA cloning

2.4.3.1 Primers used for gene amplification

Gene	Primer Name	Sequence
Eras	Eras-Sal-BstBI F	GTCGACTTCGAAATGGCTTTGCCTACAAAGTC
	Eras-BstZ17I R	GTATACACCGGTTCAGGCTACAGAGCAGCCAC
Prdm14	Prdm14-Sal-BstBI F	GTCGACTTCGAAATGGCCTTACCGCCCTCTGG
	Prdm14-BstZ17I R	GTATACACCGGTCTAGCAGGTTTTATGAAGCC
Mcm3	Mcm3-Sal-BstBI F	GTCGACTTCGAAATGGCGGGCACAGTAGTGCT
	Mcm3-BstZ17I R	GTATACACCGGTTCAGATAAGGAAGACGATGC
Gpr19	Gpr19-Sal-BstBI F	GTCGACTTCGAAATGGTTTTTGCTCACAGAAT
	Gpr19-BstZ17I R	GTATACACCGGTTCAGACAAAAGTGTTTGGAG
Phc1	Phc1-Sal-BstBI F	GTCGACTTCGAAATGGAAACGGAGAGTGAGCA
	Phc1-BstZ17I R	GTATACACCGGTTTAGGTCTCCTTGAGGACAT

2.4.3.2 cDNA amplification reaction

PCR reactions were prepared in 50uL volumes with 200uM dNTP (NEB), and 0.3uM of each primer. For the DNA polymerase used, PrimeSTAR HS (Takara), concentration was specified by the manufacturer and reactions were supplemented with the reaction buffer supplied with the polymerase, 5X PrimeSTAR Buffer (Mg^{2+} plus). 1-5uL of undiluted cDNA was used for each reaction. All reactions were performed on the DYAD DNA Engine thermal cycler. Three step PCR was used with a denaturation temperature of 98°C for 10sec, an annealing temperature 5 degrees lower than the lowest melting temperature for each set of primers (as supplied by the oligo provider, IDT) for 5 sec, and an extension temperature of 72°C for 1min per kb of template to be amplified. This cycle was repeated 30 times, followed by a final incubation at 72°C for 10 minutes, followed by incubation at 4°C. Samples were stored at -20°C.

2.4.3.3 Blunt cloning of PCR products

PCR products were directly cloned using the Invitrogen Zero Blunt TOPO PCR Cloning Kit (Invitrogen 45-0245) according to manufacturer's instructions.

2.4.3.4 Transformation of bacteria

DH5α E. coli bacteria were routinely transformed with the following amounts of reaction mixture: <1ug plasmid DNA; 5uL ligation mix; 5uL TOPO cloning reactions. DNA and freshly thawed bacteria were incubated on ice for ~5 minutes, bacteria were heat shocked at 42°C for 30sec and immediately incubated on ice for 10-15min. 400uL LB broth was added to the tube and incubated at 37°C for 1hr. 10-300uL (depending on the source of the DNA) was plated on LB plates containing the appropriate antibiotic. Plates were incubated at 37°C overnight.

2.4.3.5 Bacterial plasmid isolation

Individual bacterial colonies were picked from plates and used to inoculate 4mL of LB broth containing the appropriate antibiotic. Colonies were also stabbed onto an LB antibiotic plate in triplicate ('master plate'). The master plate was incubated at 37°C overnight, while the inoculated LB broth cultures were incubated at 37°C with agitation at 200rpm for 16-18hours. The following day bacteria were centrifuged at 4000g for 10min. The bacterial plasmid DNA was isolated using the QIAprep Spin Miniprep kit (Qiagen 27106). For larger preparations of DNA, 50mL of broth was inoculated and plasmid DNA isolated using a QIAprep Spin Midiprep kit (Qiagen 12243).

2.4.3.6 Restriction enzyme digestion of DNA

All restriction enzymes used were supplied by NEB. The amount of DNA digested varied depending on the quantity of digest product required for the next step of cloning, except for confirmation digests for which usually 100-200ng of DNA was used. The reaction buffer used was as recommended by the manufacturer. The amount of restriction enzyme used was based on the quantity of DNA present and was adjusted accordingly in digests using a non-optimal buffer (e.g. two different enzymes being used). Reactions were performed at 37°C for 1hr.

2.4.3.7 Isolation of digested DNA fragments

DNA fragments were separated via agarose gel electrophoresis. Depending on the size of the fragment to be isolated, 0.8-2% (w/v) gels were used. Gels contained GelRed (1X, Cambridge BioScience) for visualisation of DNA under UV illumination. Samples were run in 1x sucrose loading buffer (OrangeG, NEB, + 40% sucrose) alongside a 1Kb Plus (Life Technologies) DNA marker. Samples were run at 120V for 30-45mins. Samples were extracted from gels using a scalpel and gel fragments stored at -20°C if required. DNA was extracted from gels using

Zymoclean Gel DNA Recovery Kit (Zymo Research) according to manufacturer's instructions. DNA was eluted in 8uL sterile water.

2.4.3.8 Ligation of DNA fragments

Ligations were carried out using T4 DNA Ligase (NEB) as per manufacturer's instructions. For ligations, a molar ratio of insert: plasmid of 3:1 or 5:1 was used. Reactions were carried out at RT for 1hr, or at 16°C overnight.

2.4.3.9 Gateway Cloning Reaction

Gateway (Invitrogen) cloning was carried out according to manufacturer's instruction.

CHAPTER 3 - Identification of novel cell-surface markers to isolate intermediate stages of reprogramming

3.1 Introduction

Cell surface markers have long been used in many diverse areas of biology to identify distinct populations of cells at different developmental time-points or which display specific characteristics, such as identification of hematopoietic lineage components (Lin and Goodell, 2011), tracing the maturation of T-cells in the thymus (Gordon and Manley, 2011) and isolation of putative tumorigenic cells in a variety of cancers (Al-Hajj et al., 2003; Visvader and Lindeman, 2008; Botchkina et al., 2009). However, while there have been attempts to apply this approach to reprogramming studies, with an indication that known somatic surface-markers are rapidly down regulated and the mES cell-associated marker stage-specific embryonic antigen 1 (SSEA-1) up-regulated early in the reprogramming process, this occurs well in advance of pluripotent gene expression, and may not accurately capture intermediates stages of reprogramming (Stadtfeld et al., 2008a). In addition, if the theory that reprogramming occurs in a stochastic manner is to be believed, it suggests that there may be no one suitable surface-marker to identify those cells which have the potential to become iPS cells, as this process is highly asynchronous (Hanna et al., 2009b). Despite these indications, a number of molecular events have been described during reprogramming to occur in an ordered manner, for example a mesenchymal-to-epithelial transition (MET) always takes place before cells can become fully reprogrammed (Li et al., 2010; Samavarchi-Tehrani et al., 2010). We theorized that if such ordered changes can occur on a molecular level, there may also be cell surface-markers which also show sequential changes during

reprogramming whose expression correlates with changes in iPS cell-forming potential.

3.1.1 Background work leading to this project

3.1.1.1 Microarray analysis of secondary reprogramming.

A *piggyBac* transposon-generated iPS cell line, **6c**, was used previously in our lab for microarray analysis (Woltjen et al., 2009)(Microarray data unpublished, sample preparation carried out by Keisuke Kaji). This line carries numerous integrations of the individual reprogramming factors *c-Myc*, *Klf4*, *Oct4* and *Sox2* which are under the control of the Tet-O, doxycycline (dox)-inducible promoter, with rtTA expressed from the Rosa26 locus. This clone was used for morula aggregation, and MEF were isolated from chimeric mice at embryonic day 12.5 (E12.5). These MEF were used for "secondary" (2°) reprogramming, with cells cultured in the presence of dox (1.5ug mL^{-1}) to induce re-expression of reprogramming factors from the transgene containing cells (Figure 3.1a). Samples for microarray analysis were isolated at various time-points throughout the time-course (Summarized in Table 3.1).

Sample Name	Description
Secondary (2°) MEF	6c MEF isolated from chimeric embryos
Day 5	6c MEF cultured for 5 days in dox, sorted for SSEA-1^{+} cells
Day 8	Day 5 sorted cells, plus 3 days in dox
Day 11	Day 5 sorted cells, plus 6 days in dox
Day 14	Day 5 sorted cells, plus 9 days in dox
Day 17	Day 5 sorted cells, plus 12 days in dox, harvested and floating after 45min on gelatine
2° iPSC	Dox independent iPSC derived from day 23 by passaging for 20 days in the absence of dox

Table: 3.1: 6c reprogramming samples for microarray analysis. Samples were isolated at the time-points indicated

3.1.1.2 Identification of cell surface-markers.

Bioinformatic processing of the **6c** 2° reprogramming microarray data was carried out by Dr. Simon Tomlinson. From this data, we could observe a number of interesting results (Figure 3.1b). Firstly, despite including only cells positive for the ES cell-associated marker SSEA-1 from day 5 onwards, it was observed that pluripotency gene up-regulation did not show a significant increase until around day 11. This suggested that *SSEA-1* expression is an early marker of cells undergoing reprogramming.

In addition, the expression of the pluripotency genes themselves, despite modestly increasing up to day 17, did not reach the high levels of expression found in ES cells until after numerous passages in the absence of dox, possibly corresponding to the expansion of the already-reprogrammed cells present at day 17. Of note, the previously described MEF marker *Thy1* was found to be rapidly down-regulated, with expression dropping within the first five days of the

reprogramming process. *Thy1* expression was also down-regulated in the SSEA-1 negative cells at day 5, indicating the loss of this marker may not represent a significant barrier to reprogramming (data not shown). Similarly, the epithelial cell marker *Ecad* was found to be rapidly up-regulated, with expression reaching fully reprogrammed iPS cell levels by day 8 in the SSEA-1^{+} population. These results indicated that neither of these markers would be suitable for examining cells at intermediate or later-stages of the reprogramming process as they would not be capable of discriminating between the large numbers of cells which would ultimately fail to generate iPS cells.

It was possible to identify a number of cell-surface markers whose expression, in contrast to known or associated markers of reprogramming, appeared to be far more dynamic in nature. *Sca1* expression was found to gradually increase until day 14 of reprogramming, after which expression rapidly decreased as cells reached an iPS cell-like state. *CD44* expression appeared to be relatively highly expressed in the starting MEF population, and this expression was rapidly down-regulated throughout the reprogramming process, showing a much slower loss in expression when compared to fellow MEF marker *Thy1*. Finally, *Icam-1* expression appeared to be relatively low for the majority of the reprogramming time-course, but was greatly up-regulated at later stage, echoing the expression pattern of key pluripotency genes *Nanog*, *Utf1* and *Dppa4*. A number of other cell surface-markers were also identified to display similar patterns of expression, however the majority of these were not suitable for further investigation as either their fold expression change was low, commercial antibodies were not available for immediate use or the detected transcript level did not correlate with the expressed protein (data not shown).

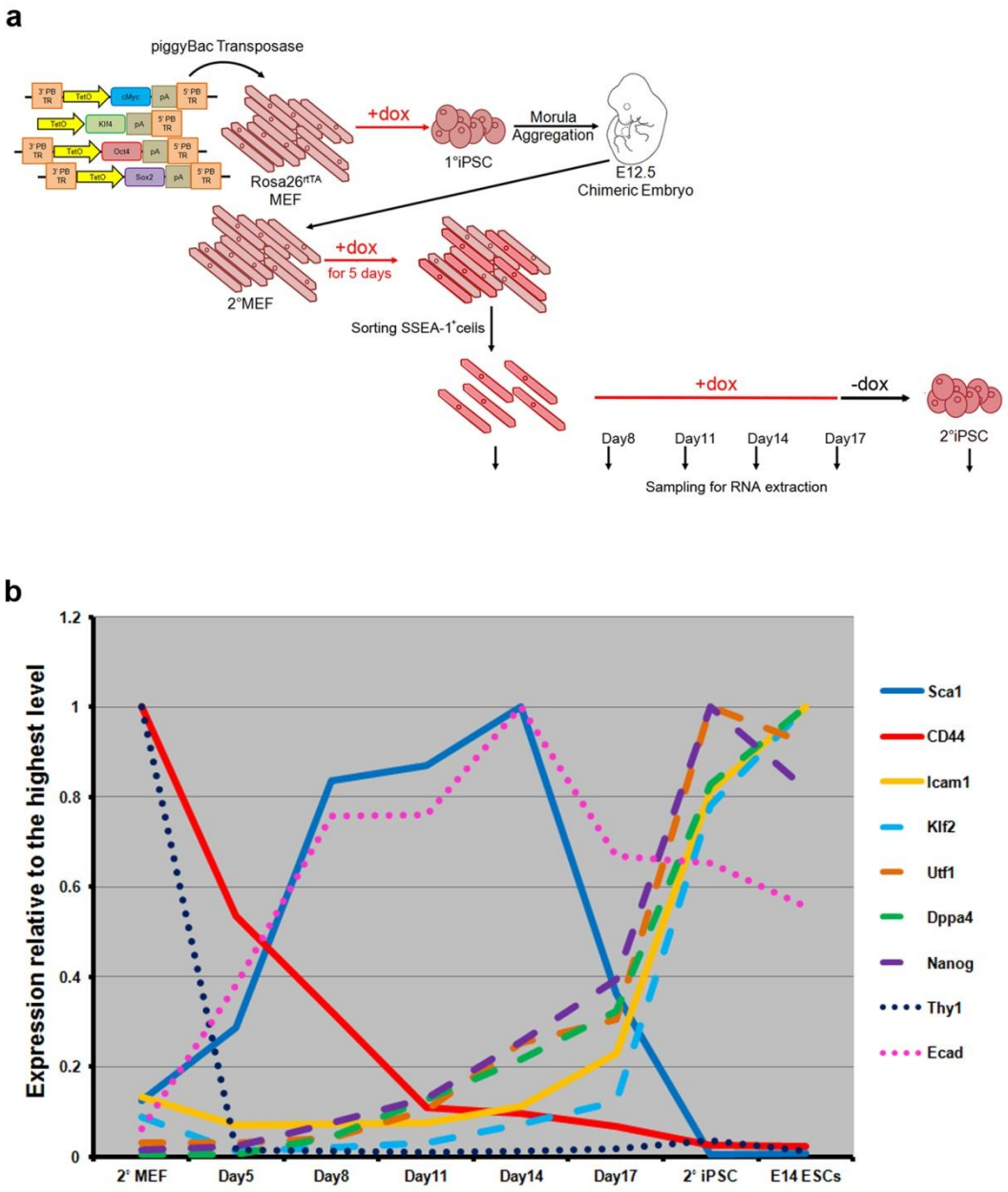


Figure 3.1. Identification of cell surface markers from secondary reprogramming time-course analysis. a. Schematic overview of *piggyBac* transposons mediated generation of 6c cell line, secondary reprogramming samples for microarray analysis and E14 ES cells. **b.** Expression pattern of pluripotency genes *Klf2*, *Nanog*, *Utf1*, *Dppa4*, previously described surface markers *Thy1*, *Ecad* and novel markers *Sca1*, *CD44* and *Icam-1*. Expression level normalized to highest value detected.

3.1.1.3 Retinoic Acid differentiation as an in vitro model of reprogramming

Using *Sca1*, *CD44* and *Icam-1* as candidate genes for further analysis, it was next key to establish a 2° reprogramming system which could ensure consistent, high expression of the reprogramming factors after differentiation. It had previously been shown that 6c-derived MEF cells could successfully reactivate expression of the reprogramming factors (Woltjen et al., 2009). However retinoic acid (RA)-mediated differentiation of iPS cell lines generated using the dox-inducible MKOS-ires-geo polycistronic vector indicated poor re-expression and low efficiency reprogramming upon culture in dox (Kaji et al., 2009) (data not shown). This may have been due to the presence of the *lacZ* reporter component of the β-geo gene which is –C-phosphate-G- (CpG) rich and may be a target of silencing methylation during *in vivo* or *in vitro* differentiation (Chevalier-Mariette et al., 2003).

In order to maintain high levels of transgene expression throughout reprogramming, new iPS cell lines were generated from Rosa26$^{M2rtTA/+}$ MEF nucleofected with a modified version of the previously described vector pCAG2LMKOSimO (Kaji et al., 2009). This modified vector, PB-TAP IRI 2LMKOSimO retained the 2A-peptide-linked reprogramming cassette *c-Myc-Klf4-Oct4-Sox2* (MKOS), and the mOrange transgene reporter which followed, but was placed under the control of the dox-inducible TetO promoter and also contained a number of genetic elements which have been reported to maintain transgene expression (Figure 3.2a). Insulator sequences are commonly found in the genome and depending on the type, generally have two major functions; protecting genes from external signals which may aberrantly effect their normal expression and acting as barriers which prevent the spread of heterochromatin and epigenetic silencing (Gaszner and Felsenfeld, 2006). Transcriptional silencing has been implicated in a shift to replication at late S-phase of the cell cycle (Gilbert, 2002). Replicators are sequences which have been shown to influence the initiation of replication, and can prevent the replication delay and instability associated with long-term transgene expression (Fu et al., 2006). By utilising both of these elements

in our reprogramming vectors, we aimed to achieve efficient induction of reprogramming factors after generating differentiated cells as well as throughout the process of reprogramming. iPS cell lines generated from Rosa26$^{M2rtTA/+}$ MEFs were screened by Southern blot to determine the number of integrations of the reprogramming cassette (data not shown), and a number of low copy (1-3) clones were identified and expanded for further use.

In order to carry out preliminary investigation of the usefulness of the identified novel cell surface markers, it was determined that generation of chimeras from these iPS clones would be both costly and time inefficient. Thus a system by which clones were differentiated in RA and low concentration-leukaemia inhibitory factor (LIF) media was used (Figure 3.2b). Clones were plated in at a density of $1X10^4$ cells/well of a 6 well dish in the presence of RA (10^{-6}M) and LIF (10U) and cultured for a period of five days (Smith, 1991). Changes in morphology became most obvious after three days of culture, and by day 5 the cells acquired a fibroblastic-like morphology, with very few undifferentiated cells (Figure 3.2c, Day 5+RA). Adding media containing higher levels of LIF (1000U mL^{-1}) and dox (1.5ug mL^{-1}) to these cultures resulted in gradual mOrange expression, morphology changes and later the appearance of colonies indistinguishable from the original iPS cells from which they were derived (Fig 3.2c). One clone, **IRI1** showed efficient reprogramming and was used for further study under these conditions. Some cells were found to maintain the ability to generate colonies in the presence of high (1000U mL^{-1}) LIF-containing media in the absence of dox even after five days of differentiation, however ensuring the absence of clumps of iPS cells at the initial seeding prevented this from occurring at a high frequency (Figure 3.2d). It was also found that extending differentiation beyond five days resulted in a dramatic decrease in the number of alkaline phosphatase (AP) positive colonies that could be generated in the presence of dox (Figure 3.2e), thus five days of differentiation was considered optimal for initial experiments.

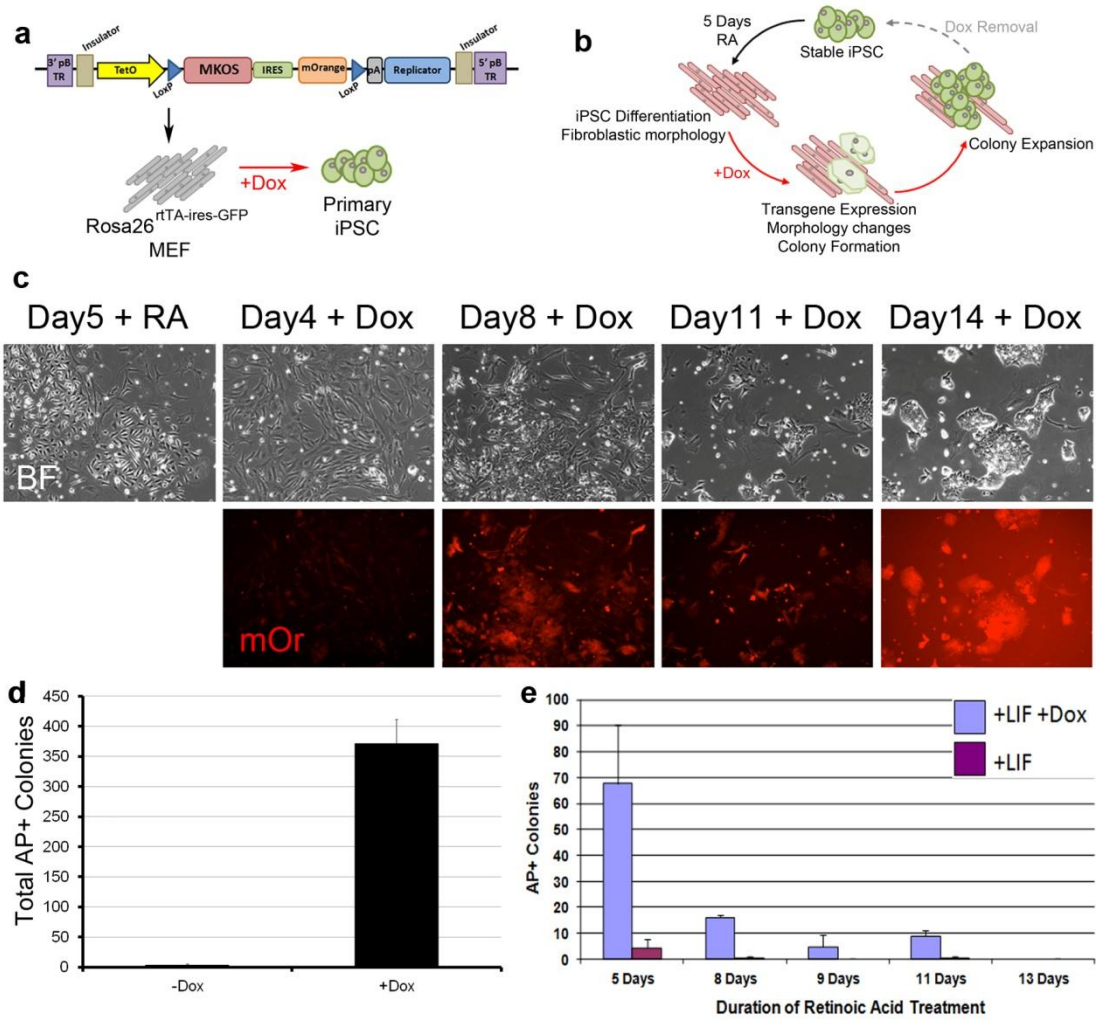


Figure 3.2 A retinoic acid based efficient secondary reprogramming system. a. Schematic of pBTAP IRI 2LMKOSimO and generation of iPSC lines. **b.** Schematic of retinoic acid based differentiation and reprogramming system. **c.** Images of typical morphology of RA differentiated cells undergoing reprogramming. BF = Bright Field. mOr = mOrange. **d.** Number of alkaline positive colonies obtained in the presence and absence of dox after 5 days of RA differentiation. **e.** Number of AP^{+} colonies obtained at day 10 of reprogramming after differentiation of ES cells in RA for various periods. Error bars represent standard deviation of biological replicates. n=2.

3.1.1.4 Marker expression during secondary reprogramming

During 2° reprogramming of RA-differentiated IRI1, in order to establish how the novel cell surface markers identified from the 6c microarray behaved during reprogramming, a time-course of flow cytometric analysis was carried out (Figure 3.3).

CD44 expression was found to be at a moderate level in the RA differentiated iPS cells, in contrast to the high transcript levels detected in the microarray (Figure 3.3, top panels). However, upon initiation of reprogramming it was observed that CD44 levels increased somewhat, with more than 50% of total cells highly positive for this marker by day 2. Expression was gradually lost however, and by day 12 the reprogramming profile appeared dramatically different with a loss of nearly all $CD44^+$ cells.

SCA1 expression was initially slightly lower than CD44 in the majority of differentiated cells (Figure 3.3, middle panels). However, unlike CD44, at day 5 SCA1 expression was almost homogeneously high. A gradual loss of expression was subsequently observed, however unlike CD44, by day 12 had not quite reached low, ES cell-like, levels of expression.

ICAM-1 expression appeared to be quite broad in the RA differentiated iPS cells, with both high and low expressing populations (Figure 3.3, bottom panels). Expression of this marker appeared to be quite static with no clear emergence of separate populations until day 9 cells appeared to gain ICAM-1 expression. By day 12 the majority of cells expressed ES cell-like levels of ICAM-1.

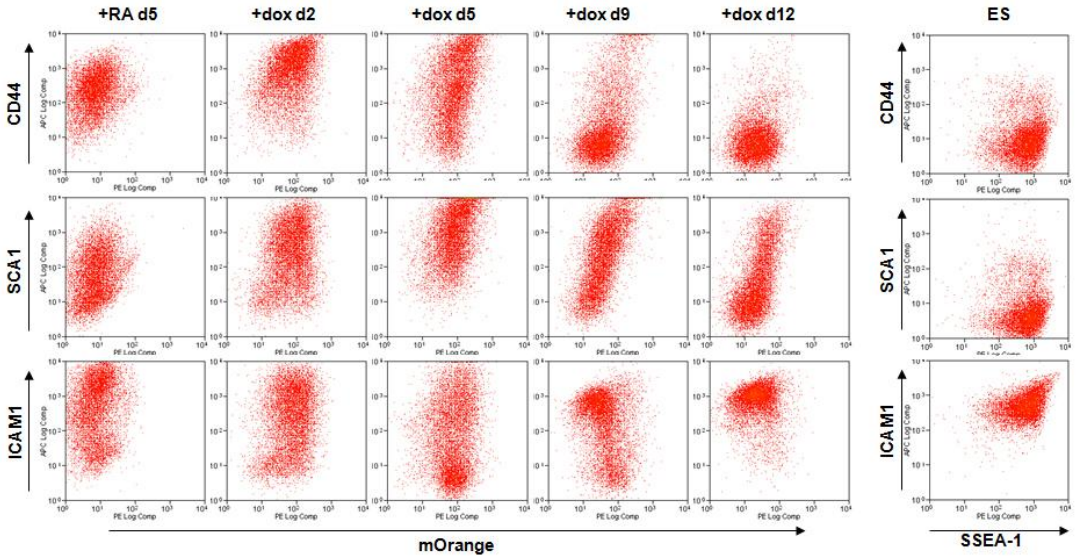


Figure 3.3 Expression of cell surface markers during secondary reprogramming of RA differentiated IRI1 iPS cells. Cells were differentiated in the presence of retinoic acid (RA) for five days and cultured in the presence of dox. Cells were harvested at the indicated days (d) and used for FACS analysis of the expression of cell surface markers CD44 (top), SCA-1 (middle) and ICAM1 (bottom). The expression of these markers was compared to reprogramming factor expression (mOrange). ES cells were stained for the novel cell surface markers and SSEA-1.

3.1.2 Aims of this chapter

The aim of this chapter is to outline the identification of novel cell surface-markers which displayed dynamic expression patterns during iPS cell generation. Based on this work, a series of validation experiments were carried out to identify their robustness and usefulness in marking distinct intermediates of reprogramming, and how these populations behave upon isolation. Furthermore, the use of a transgene expression reporter led to preliminary observations regarding the requirement of reprogramming factor expression during iPS cell generation. Finally, expression of the novel surface-markers was compared to expression of a pluripotency gene reporter, giving further insight into the dynamics of reprogramming.

3.2 Results

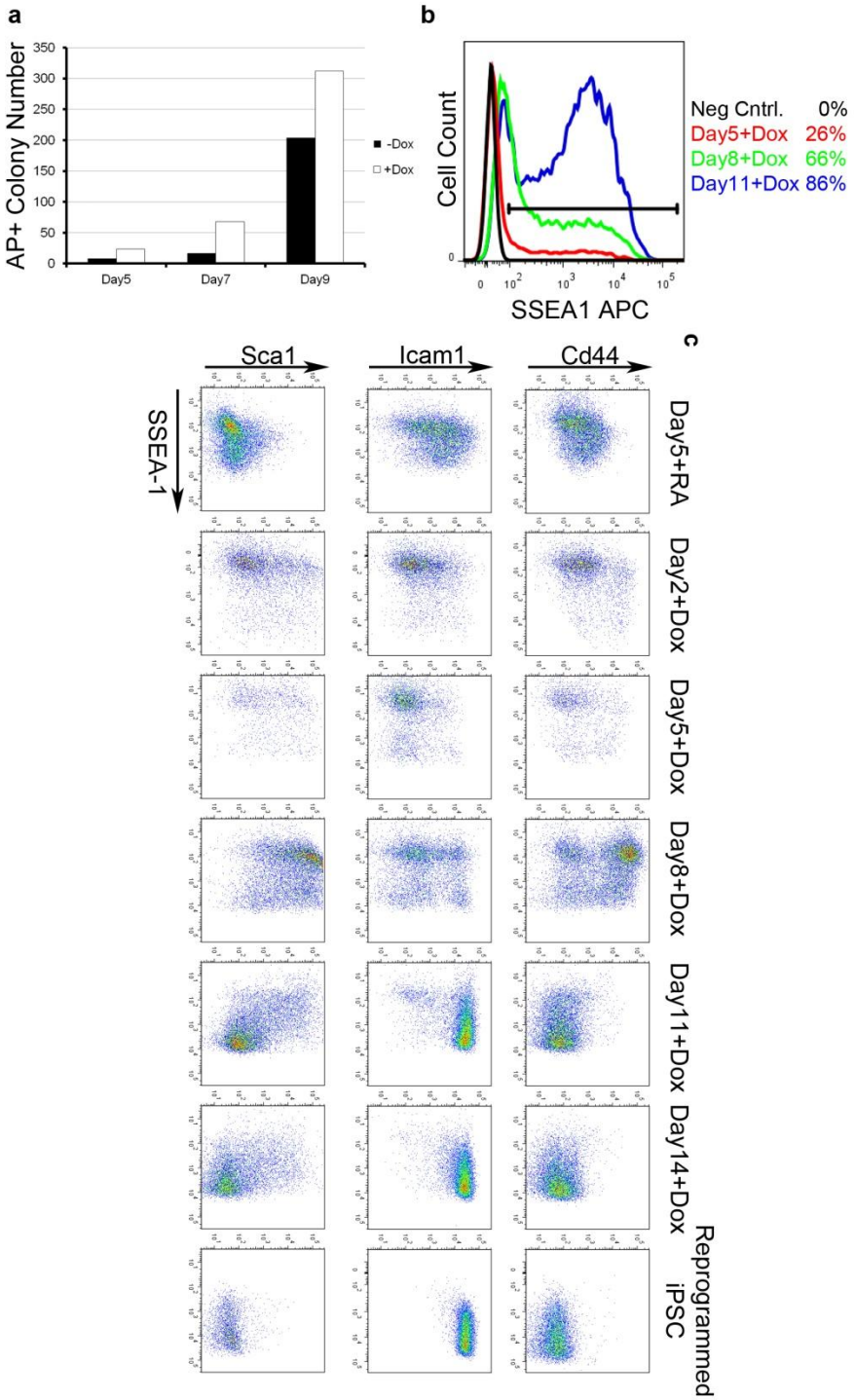
3.2.1 Comparison of novel cell surface markers to SSEA-1 during reprogramming

During 2° reprogramming of IRI1, cells were harvested and stained for expression of the ES cell marker SSEA1 (Figure 3.4a). As expected, SSEA1 expression increased with prolonged reprogramming factor expression, with more than 80% of cells expressing this marker by day 11 of dox administration. To investigate the correlation of SSEA1 expression and cell behaviour during reprogramming, cells were re-plated at clonal density in the absence of dox at different time-points and stained for AP^+ colony formation after five days. From this data it was found that even at day 9, when >60% of cells were $SSEA1^+$ only 40% of cells were capable of successfully generating AP^+ colonies (Figure 3.4b). This supported the idea that SSEA1 was not an appropriate marker of fully reprogrammed cells, and indeed there appeared to be no clear correlation between SSEA1 expression and successful completion of reprogramming. A time-course of flow cytometric analysis was carried out with expression compared to that of SSEA1 which acted as a reference for the usefulness of the novel markers (Figure 3.4c).

CD44 expression was found to be at a moderate level in the RA differentiated iPS cells, in contrast to the high transcript levels detected in the microarray (Figure 3.4c, top panel). However, upon initiation of reprogramming it was observed that CD44 levels rapidly increased with more than 50% of total cells highly positive for this marker by day 8. This increase in CD44 expression was mirrored by an increase in SSEA1, and four populations of cells could be observed by day 8, with both CD44 expressing and non-expressing cells containing both SSEA1 positive and negative fractions. However by day 11 the reprogramming profile appeared dramatically different with a loss of all $CD44^+$ cells. This population remained absent for the remainder of the reprogramming process, while SSEA1 expression gradually increased.

ICAM-1 expression appeared to be quite broad in the RA differentiated iPS cells, with both high and low expressing populations (Figure 3.4c, middle panel). Upon initial transgene expression this marker was relatively static with no clear emergence of separate populations until day 11 when the majority of cells became ICAM-1^{+}. Similarly to CD44 staining at day 8, there appeared to be both positive and negative populations expressing SSEA1. By day 14 a very minor ICAM-1 negative population was observed with most cells expressing an iPS cell-like phenotype.

SCA1 expression was initially very low if not absent in the majority of differentiated cells (Figure 3.4c, bottom panel). However, unlike the other novel markers, it appeared to show a rapid response to transgene induction with a large up-regulation in expression even after 2 days. Expression appeared to increase further in the SSEA1^{-} cells by day 8, however it was apparent that SSEA1^{+} cells had reduced expression of this marker. This pattern became much more apparent by day 11 when there was a clear profile of SCA1^{+}SSEA1^{-} versus SCA1^{-}SSEA1^{+} cells. By day 14 the majority of cells had down-regulated Sca1 expression, and acquired a fully reprogrammed profile.



Legend Overleaf

Figure 3.4. SSEA1 expression compared to novel reprogramming markers. a. Total number of colonies generated after dox treatment for the indicated days and AP stained after a five further days of culture in the presence of dox (+dox) or after dox washout (-dox). **b.** SSEA1 expression measure by flow cytometry during secondary reprogramming. **c.** Flow cytometry profiles of novel cell surface markers compared to SSEA1 during time-course analysis of 2˚ reprogramming.

3.2.2 Reprogramming factor expression during iPS cell generation.

During reprogramming using the RA based 2° reprogramming system, not only could the expression pattern of cell surface markers be identified, but also the expression level of the reprogramming factors themselves due to the presence of the ires-mOrange cassette contained within the vector. One startling observation however, was that transgene expression levels did not appear to remain constant in the cells undergoing reprogramming, despite maintenance of dox at an identical concentration throughout the entire culture period (Figure 3.5a). In fact, while the majority of cells displayed high levels of mOrange at the mid-point of reprogramming (day 8) the later stages (day 11) reduced expression to an identical level to that of cells at earlier time-points (day 5).

This finding was further explored in cells undergoing reprogramming stained for SSEA1 (figure 3.5b). In these cells it was found that there appeared to be a negative correlation between expression of this marker and the mOrange reporter, with the $SSEA1^{high}$ cells (green) displaying lower reporter expression than their $SSEA1^{neg/low}$ counterparts (red). However, as described previously, SSEA1 expression may not correlate well with the reprogramming process, thus CD44 expression was compared to that of mOrange in order to further understand this observation (Figure 3.5c). At the early stages of reprogramming (day 5) there did not appear to be a difference in mOrange expression between $CD44^{high}$ and $CD44^{low}$ cells. A more significant difference became apparent at day 8 when the expression of CD44 appeared to correlate well with mOrange expression. By day 11 the majority of cells had decreased CD44 expression as previously described, and similarly, expressed low levels of mOrange. However it was observed that the minor population which maintained high levels of CD44 continued to express higher levels of the reprogramming factors. This result provided evidence that there may be some form of reprogramming factor down-regulation required for the completion of reprogramming, irrespective of attempts to drive further expression e.g. maintaining high concentration of dox.

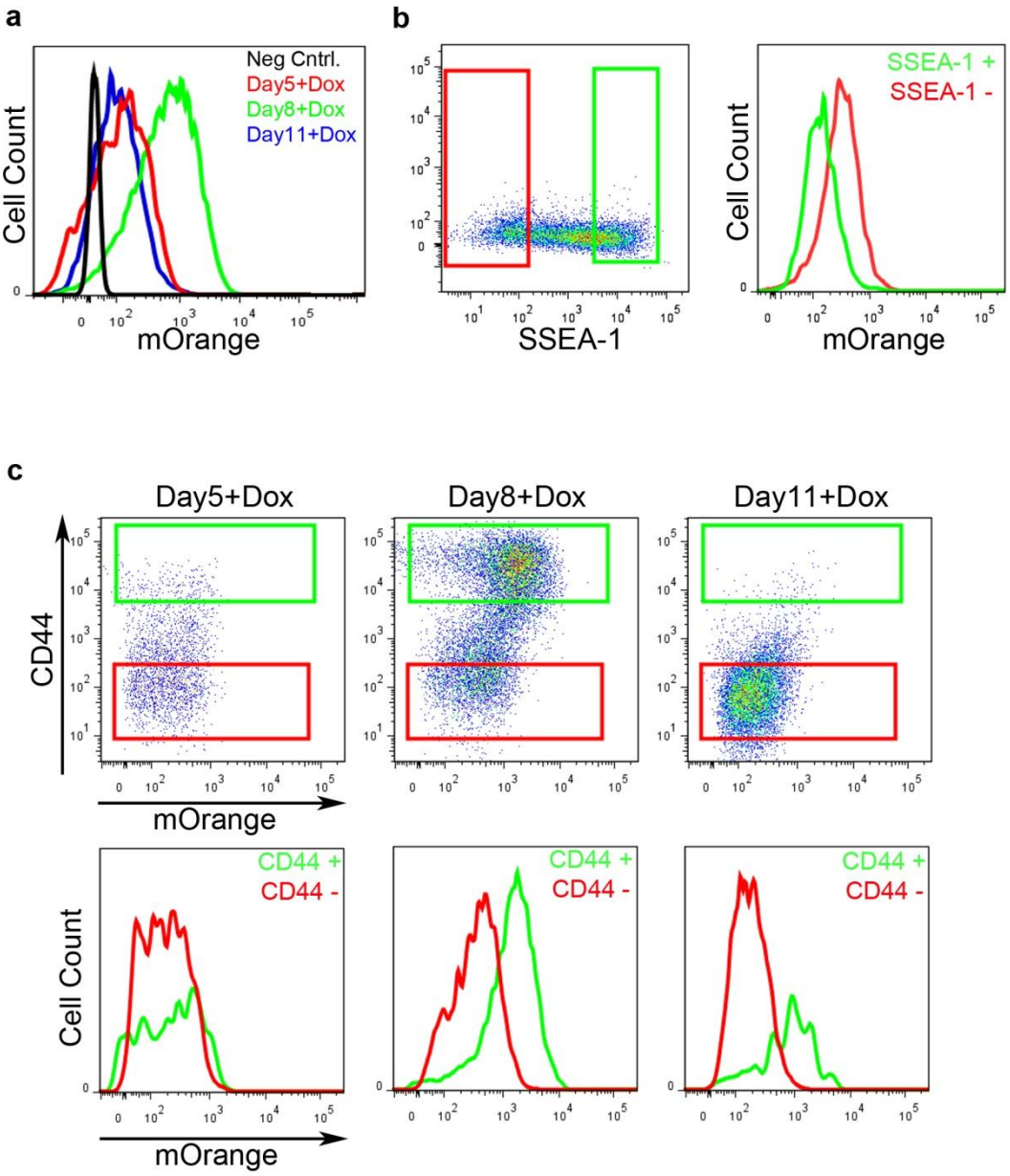


Figure 3.5. Down-regulation of reprogramming factor expression during secondary reprogramming. a. mOrange expression level of total population undergoing reprogramming. **b.** SSEA1 expression identifies differences in mOrange expression. Cells were analysed at day 8 post reprogramming factor induction. **c.** CD44 expression is negatively correlated with mOrange expression during reprogramming.

3.2.3 CD44 expression predicts differences in reprogramming ability

In order to assess the functional significance of the observed differences in CD44, SSEA1 and mOrange expression, cells were isolated from each of the observed populations previously described at day 8 of reprogramming (Figure 3.6a). Sorting purity (typical for all experiments) was more than 80% (Figure 3.6b). Sorted cells were plated on gelatine in the absence of dox in order to assess their ability to form iPS cell-like colonies (Figure 3.6c). Strikingly, differences among the populations could be observed even after only three further days of culture. The CD44 positive populations were unable to give rise to any colonies; rather all cells appeared to differentiate and give rise to a fibroblastic-like culture. In contrast, the CD44 negative sorted populations were capable of producing a number of iPS cell-like colonies capable of maintaining an undifferentiated state, however some differentiated cells could also be observed in these cultures.

The significance of SSEA1 expression could also be examined, and it was observed that in the case of the differentiated cultures produced by the CD44 positive sorted cells, cell survival appeared to be slightly poorer in cells also positive for SSEA1 compared to their negative counterparts. In the CD44 negative populations, there appeared to be fewer differentiated background cells in the SSEA1 positive sorted cells. These results suggested that CD44 expression is a more useful marker in terms of predicting which cells are closer to an iPS cell-like state compared to the commonly used marker SSEA1. In addition, the presence of a population expressing both a marker negatively correlated with reprogramming success (CD44) at the same time as an ES cell marker (SSEA1) suggested there may be more than one route to the fully reprogrammed state, and further reinforced the requirement for a system which can accurately untangle the heterogeneity of cell cultures undergoing reprogramming.

It was observed, as previously described (Figure 3.5), that at the time of sorting both of the CD44 negative populations were found to be expressing lower levels of the reprogramming factors as measured by mOrange expression, compared

to the CD44 positive populations (Figure 3.6d). Coupled with the result that the CD44 positive populations were unable to generate iPS colonies, and therefore were delayed in the reprogramming process compared to the CD44 negative cells, this indicated that some form of transgene down-regulation is indeed required for successful reprogramming. However, this appeared to be an intrinsically determined process requiring some key change/s to occur within the cells themselves before transgene expression can be down-regulated, as these cells did not complete reprogramming when dox (and therefore transgene expression) was removed.

3.2.4 Comparison of novel cell surface marker and Nanog-GFP expression

In order to enhance analysis of the reprogramming process of the novel markers, and investigate correlation with pluripotency gene expression, iPS cell lines were generated by Dr. Kaji using **pBTAP IRI 2LMKOSimO** from $Rosa^{rtTA/rtTA}$ MEF carrying a Nanog-GFP reporter (Chambers et al., 2007). A single integration clone, D6s4B5 was identified, expanded and used for further experiments.

After RA differentiation, the D6s4B5 line efficiently underwent reprogramming in the presence of dox (Figure 3.7a). In addition to the previously described morphology changes and mOrange up-regulation associated with this reprogramming system, a small number of Nanog-GFP^+ cells could be identified by day 8 of reprogramming. These cells rapidly expanded and large, Nanog-GFP^+ iPS cell-like colonies could be obtained. In addition it was observed that mOrange expression appeared to be down-regulated in the cells which had acquired GFP expression. This down-regulation could also be observed by flow cytometry (Figure 3.7b). Over a time-course of cells undergoing reprogramming it was found that GFP^+ cells appeared to arise not from the cells expressing the highest levels of mOrange, but rather from a population of cells with a mid-range of expression

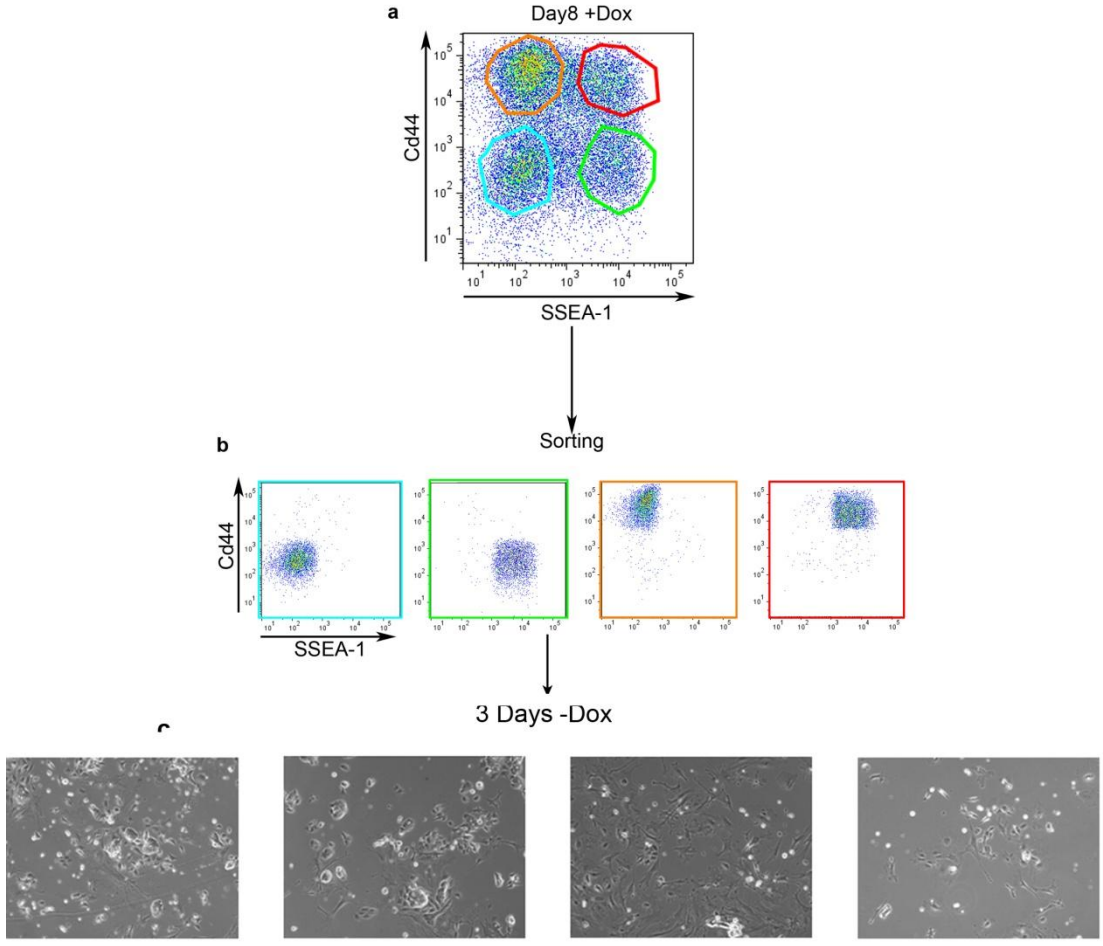


Figure 3.6: CD44 expression as a marker of reprogrammed cells. a. Populations sorted at day 8 of reprogramming. **b.** Purity of sorted populations. **c.** Typical appearance of sorted cells cultured in the absence of dox. **d.** mOrange expression of sorted populations at the time of sorting. Colours indicated sorted population.

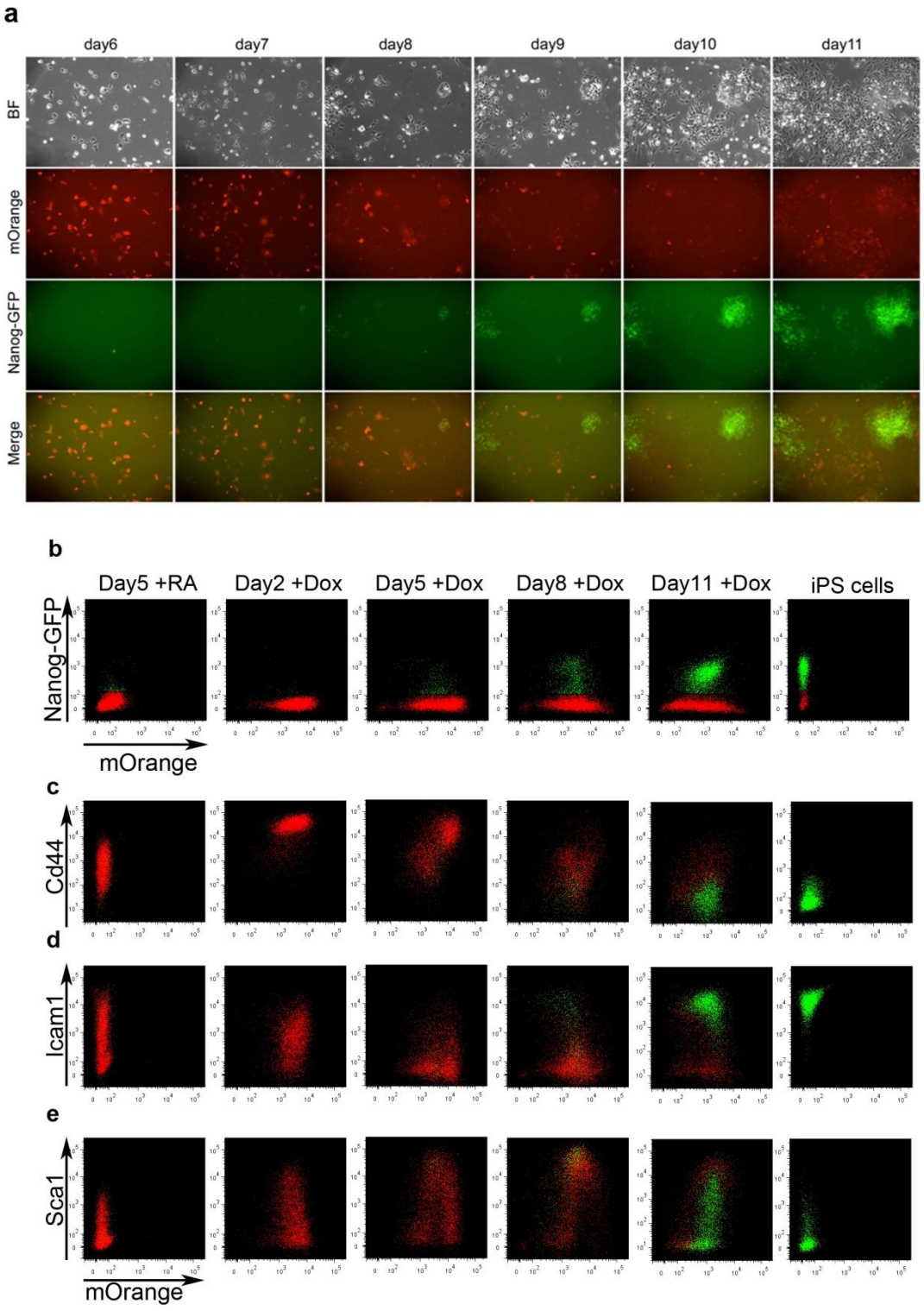
(Figure 3.7b, Day 5, Day 8+Dox). Also of note was that the cells expressing the highest levels of GFP appeared to eventually down-regulate expression of mOrange to an even lower level (Figure 3.7b, Day 11+Dox). This again supported the idea that reprogramming factor down-regulation was required for successful iPS cell

generation. Flow cytometry was also used to investigate the correlation of pluripotency gene expression and the novel cell surface markers identified from the microarray data.

CD44 expression was found to have a negative correlation with Nanog-GFP expression (Figure 3.7c). There was almost a complete absence of $CD44^{+}GFP^{+}$ cells during the time-course, with GFP^{+} cells only appearing once CD44 expression had been almost totally down-regulated. This finding went some way towards explaining the previous observation that cells which had lost CD44 expression, but not yet gained SSEA1 expression, could give rise to reprogrammed colonies (Figure 3.6d).

The more heterogeneous pattern of **ICAM-1** expression compared to CD44 at the later stages of reprogramming was also reflected in the appearance of Nanog-GFP^{+} cells at Day 8, a significant minority of which were observed in an $ICAM^{-}$ population (Figure 3.7d). However, at the later stages of reprogramming the majority of the GFP^{+} cells were also ICAM-1^{+}. Interestingly, this appeared to suggest that cells which do not have an iPS cell-like expression profile can express Nanog-GFP.

Nanog-GFP^{+} cells were initially observed in **$SCA1^{+}$** cells at day 8, however as time passed they appeared to lose SCA1 expression (Figure 3.7e). This again indicated that even cells which have not yet acquired an iPS cell-like phenotype can express pluripotency genes. The wide distribution of SCA1 expression among the Nanog-GFP^{+} cells was unique compared to ICAM-1 and CD44 in which it appeared only minor populations of Nanog-GFP^{+} cells did not express an ES cell-like phenotype (e.g. ICAM-1^{+} or $CD44^{-}$).



Legend Overleaf

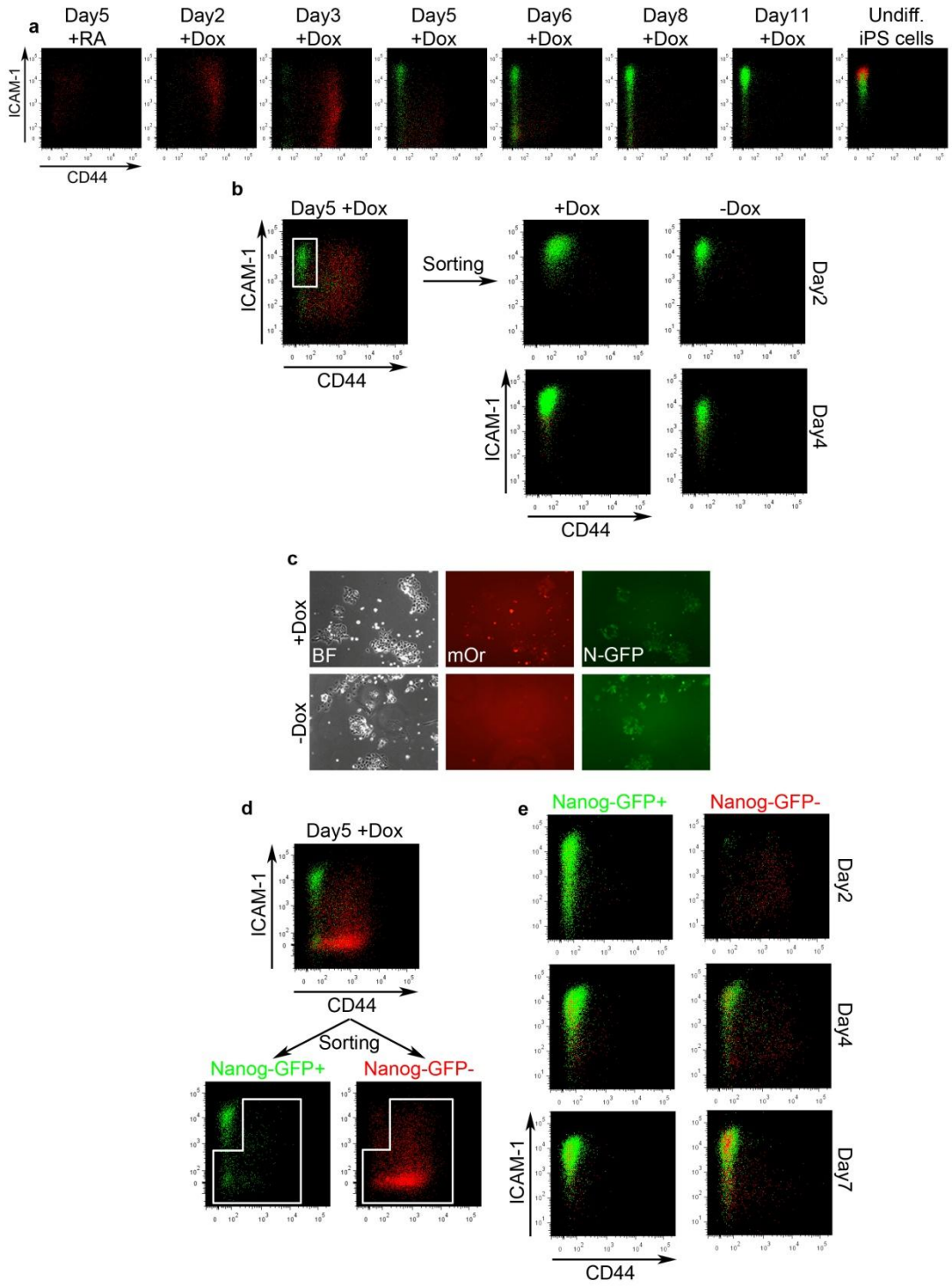
Figure 3.7: Nanog-GFP expression correlates with an ICAM-1^{+}CD44^{-} phenotype. a. Morphology, mOrange and Nanog-GFP expression of RA differentiated iPS cell clone D6s4B5 during 2° reprogramming. Time-course analysis by flow cytometry of mOrange expression compared to **b.** Nanog-GFP, **c.** CD44, **d.** ICAM-1 and **e.** SCA1 expression during 2° reprogramming.

3.2.5 ICAM-1 and CD44 expression defines a stable reprogrammed population

The ability to identify a population of cells in a heterogeneous culture which behaved similarly to fully reprogrammed iPS cells would be an extremely useful tool. However such systems would be required to be reliable and reproducible, and as it had previously shown a good correlation with iPS cell colony formation (Figure 3.6c), the loss of CD44 expression was further investigated. Similarly, the close correlation between ICAM-1 expression and a gain of Nanog-GFP expression was also of interest.

Time-course analysis during reprogramming of RA differentiated cells was carried out and highlighted a pathway to an iPS cell-like profile of gradual CD44 loss followed by a gain of ICAM-1 expression (Figure 3.8a). Loss of CD44 expression was strongly correlated with gain of Nanog-GFP expression, although a significant number of Nanog-GFP^{-}ICAM-1^{-}CD44^{-} cells could still be observed.

Cells expressing the iPS cell phenotype of ICAM-1^{+}CD44^{-} were isolated from a culture undergoing reprogramming (Figure 3.8b). iPS cell-like colonies which expressed Nanog-GFP arose from this sorted population both in the presence and absence of dox, with low levels of background differentiation (Figure 3.8c). This indicated that this profile may correctly identify cells that are either fully reprogrammed or at least in the latter stages of reprogramming displaying reprogramming factor-independent maintenance and self-renewal.



Legend Overleaf

Figure 3.8: ICAM-1 and CD44 expression identifies an end point of reprogramming. a. Reprogramming time-course of RA differentiated cells stained for ICAM-1 and CD44 expression. **b.** Sorting strategy 5 days post dox addition of ICAM-1^{+}CD44^{-} cells (white box) and profile of sorted populations cultured +/-dox as indicated. **c.** Colony formation observed in sorted population +/-dox after 7 days. **d.** Sorting strategy for cells not expressing an iPS cell profile. **e**. Profile of sorted populations 2, 4 and 7 days post sort.

To further dissect the relevance of ICAM-1 and CD44 expression before the acquisition of an iPS cell-like profile, the GFP+ and GFP- populations outside of this gate were isolated and re-plated (Figure 3.8d). Flow cytometric analysis of the isolated populations revealed a conserved pattern of changes in cell surface marker expression (Figure 3.8e).

After 2 days of further culture in the presence of dox, the majority of the **Nanog-GFP^{+}** sorted cells were found to be CD44^{-}, and the majority of cells had acquired ICAM-1 expression. By day 4 post sort most cells had up-regulated ICAM-1 expression and cells maintained this profile after a further 3 days. Some Nanog-GFP^{-} cells could be observed, however the vast majority of cells maintained GFP expression.

The **Nanog-GFP^{-}** cells appeared to maintain a similar profile post sort, with a minority of GFP^{+} cells expressing an iPS cell profile after 2 days of culture. This population appeared to increase over the following four days, with a gradual loss of CD44^{+}GFP^{-} cells and a gain in ICAM-1^{-}CD44^{-} cells, of which 1/3 expressed Nanog-GFP. By day 7 there was a loss of the ICAM-1^{+}CD44^{+} population with most GFP^{-} cells ICAM-1^{-}. Interestingly, a significant number of Nanog-GFP^{-} cells could be observed to display an ICAM-1^{+}CD44^{-} iPS cell profile by day 7 post sort.

These results indicated that cells reproducibly acquire an ICAM-1^{+}CD44^{-} phenotype during reprogramming, irrespective of the initial starting expression

profile. Nanog-GFP expressing cells appeared to behave in a more homogeneous manner with down-regulation of CD44 rapidly followed by up-regulation of ICAM-1. The GFP$^-$ cells behaved in a more heterogeneous manner, and it was not clear from which population ICAM-1$^+$CD44$^-$ cells arose: either being derived in a similar manner to the GFP$^+$ cells from cells which lost CD44 expression; from the double positive cells; or indeed from both. In addition, the gain of Nanog-GFP expression did not occur homogeneously, which indicated there may be more than one pathway towards the reprogrammed state in cells which had not yet acquired Nanog-GFP expression.

3.2.6 Defined populations to identify reprogramming pathway

The heterogeneity of the Nanog-GFP$^-$ cells indicated that further subdivision of the non-reprogrammed cells was required to further dissect the reprogramming process. In both the GFP+ and GFP- sorted cells, it was clear that CD44 down-regulation was conserved thus this marker was used to further define sub-populations. Cells which did not display an iPS cell profile of ICAM-1$^+$CD44$^-$ were isolated from a culture undergoing reprogramming, and were separated based on differences in CD44 expression alone (Nanog-GFP$^+$ cells), or on CD44 and ICAM-1 (Nanog-GFP$^-$ cells) (Figure 3.9a). Cells in each population were isolated and re-plated in the presence of dox, and all populations gave rise to Nanog-GFP$^+$ iPS cell colonies after 4 days of culture in the presence of dox (Figure 3.9b).

The non-reprogrammed Nanog-GFP$^+$ cells which did not express CD44 (Figure 3.9c 2NG+) were found to rapidly up-regulate ICAM-1 expression as expected. These cells maintained an iPS cell phenotype with low levels of Nanog-GFP$^-$ cells arising. The CD44 expressing cells (1NG+) appeared to have poorer growth (data not shown) and after two days a small number of cells were observed to express an iPS cell profile. The remaining cells appeared to lose GFP expression, and surprisingly, after a further 2 days of culture an ICAM-1$^-$CD44$^-$ population was

apparent, the majority of which expressed Nanog-GFP. These cells appeared to behave similarly to the 2NG+ population and rapidly acquired ICAM-1 expression after a further two days.

The Nanog-GFP$^-$ populations which had been divided into ICAM-1$^-$CD44$^-$ (3NG-), ICAM-1$^-$CD44$^+$ (2NG-) and ICAM-1$^+$CD44$^+$ (1NG-) were similarly analysed in order to determine if the heterogeneity of this population had been reduced (Figure 3.9c). Similarly to 2NG+, the 3NG- population dramatically up-regulated ICAM-1 expression after 2 days of culture, with GFP$^+$ cells identified in both ICAM-1$^{+/-}$ cells. Longer culture continued this trend, with increasing numbers of GFP$^+$ cells and an increase in ICAM-1 expression. Population 2NG+ had been identified as being of interest, as unlike the Nanog-GFP$^+$ cells, of which the CD44$^+$ population was very minor, in GFP$^-$ cells it was unclear if loss of ICAM-1 expression was required for the CD44$^+$ cells to undergo reprogramming (Figure 3.8e, Nanog-GFP$^-$). It was observed that after two days of culture roughly 1/3 of cells had down-regulated CD44 expression and about half of these had acquired an ICAM-1$^+$CD44$^-$ profile. Few cells appeared to express Nanog-GFP outside of this iPS cell-like phenotype. This loss of CD44 expression continued and by day 6 post sort the majority of cells were CD44$^-$/Nanog-GFP$^+$. Surprisingly, the 1NG- population appeared to behave in a similar manner to 2NG-, and there was a clear loss of the ICAM-1$^+$CD44$^+$ cells after 4 days while the ICAM-1$^-$ population appeared to be maintained, while this result did not preclude the possibility that double positive cells contribute to reprogramming, they may represent a more minor, non-sustained route in iPS cell formation.

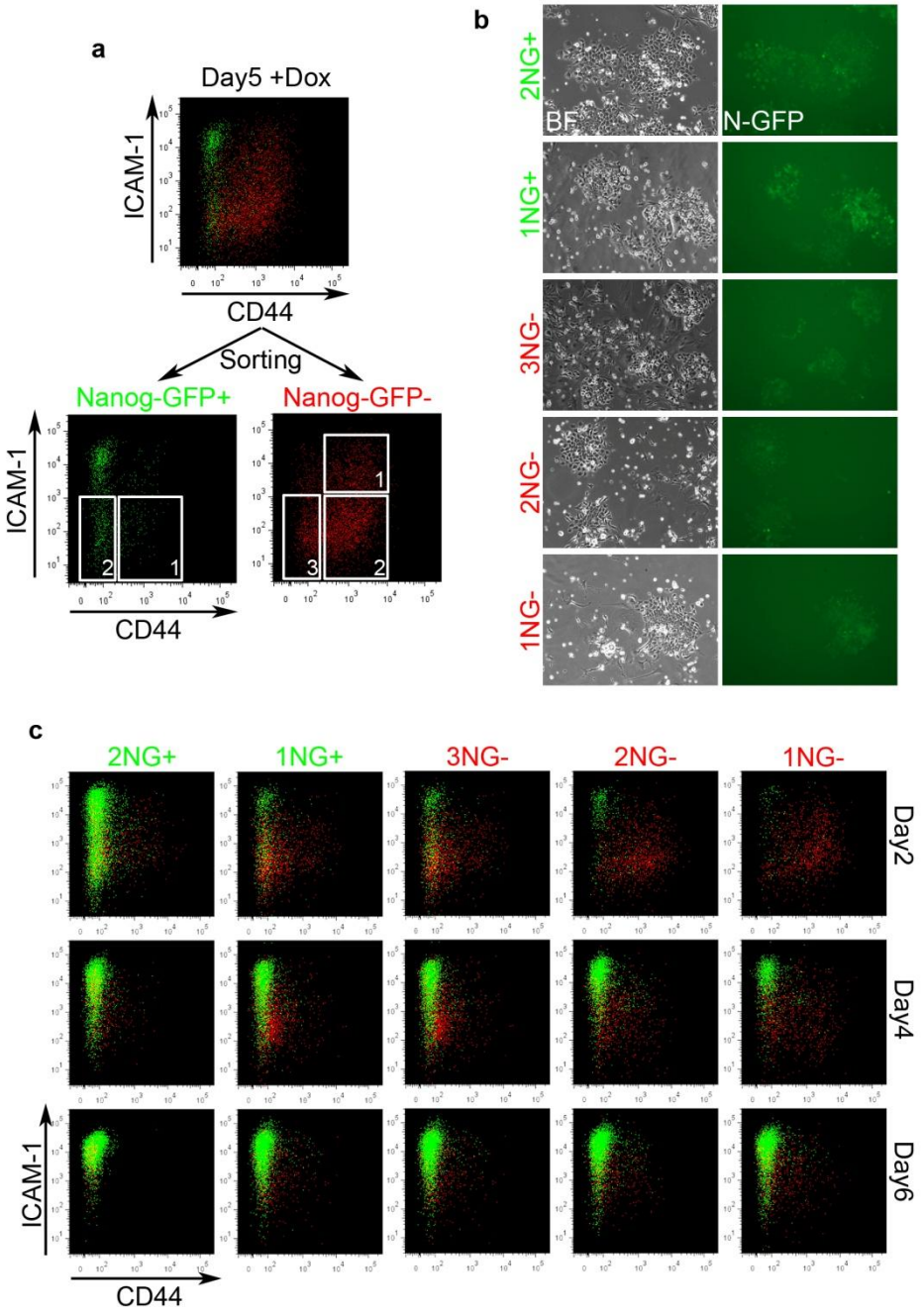


Figure 3.9: Identifying a route to the reprogrammed state using ICAM-1 and CD44. a. Sorting strategy 5 days post dox addition of ICAM-1^{+}CD44^{-} cells subdivided into Nanog-GFP$^{+/-}$ populations (white boxes). Nanog-GFP+: 2 = ICAM-1^{-} CD44^{-}; 1 = ICAM-1^{-}CD44^{+}. Nanog-GFP-: 3 = ICAM-1^{-}CD44^{-}; 2 = ICAM-1^{-}CD44^{+}; 1 = ICAM-1^{+}CD44^{+}. **b.** Colonies formed from each sorted sub-population after 4days culture in the presence of dox. **c.** Analysis of ICAM-1 and CD44 expression profile 2,4 and 6 days post sorting of sub-populations. NG+, NG- refers to Nanog-GFP positive and negative sorted sub-populations, respectively.

3.3 Discussion

3.3.1 Novel cell surface markers for reprogramming intermediates.

The microarray analysis based on reprogramming of the 6c cell line was found to accurately provide an insight into the dynamics of cell surface marker expression. SSEA1 expression had previously been identified as an early marker of reprogramming intermediates (Stadtfeld et al., 2008a). However, my data indicated that further discrimination of cells within the SSEA1^{+} fraction is required, with one population of cells found to be expressing ES cell-like levels of this marker completely failing to generate reprogrammed colonies. This finding suggests that during reprogramming alongside successful intermediates, populations arise which may not be capable of generating iPS cells. This highlights the importance of understanding the mechanism of reprogramming and using the correct markers to do so.

Unlike SSEA-1 expression, the novel markers ICAM-1, CD44 and SCA1 were found to show differences in their expression pattern later in reprogramming time-course experiments. SCA1 expression displayed a pattern of early up-regulation followed by rapid down-regulation. A recent study attempted to use SCA1 expression to discriminate between populations in a starting culture of MEF (Nemajerova et al., 2012). It was found that cells negative for the MEF cell-surface marker THY-1and either positive or negative for SCA1 were capable of undergoing reprogramming more efficiently; suggesting THY-1 is a better marker of the reprogramming state of MEF then SCA1. In addition, also among their findings it was indicated that expression of this marker can be driven by the reprogramming factor KLF4. Therefore the initially high level and subsequent down-regulation of Klf4/reprogramming factor expression, may be responsible for inducing the similar pattern of SCA1 detected during reprogramming in the ICAM-1/CD44 system.

ICAM-1 and CD44 expression, in contrast to SCA1, displayed a more simple gain and loss pattern respectively which proved useful for identification of

reprogramming intermediates. ICAM-1 expression appeared to be up-regulated almost exclusively at the latter stages of reprogramming. Some slight down-regulation of ICAM-1 was observed when sorted cells were plated in the absence of dox, however these cells still generated iPS cell-like colonies and maintained Nanog-GFP expression. The down-regulation of CD44 appeared to be a key event during reprogramming, as evidenced by the fact that cells positive for this marker could not efficiently generate iPS cell colonies in short term culture. Although a minor double positive population was observed, it was unclear if these cells lost ICAM-1 expression, and in general CD44 down-regulation appeared to be a prerequisite for ICAM-1 up-regulation. Combined staining of cells with ICAM-1 and CD44 highlighted an apparent route to a fully reprogrammed ICAM-1^{+}CD44^{-} Nanog-GFP^{+} state as evidenced by the stable, transgene independent nature of this sorted population. This retinoic acid based differentiation system may not accurate represent changes that occur during reprogramming from more differentiated cell types however, and further work was required to confirm these observations in somatic cell reprogramming (see Chapter 4).

3.3.2 Down-regulation transgene expression during reprogramming

While the commonly used marker SSEA1 failed to accurately identify cells at different stages of the reprogramming process, down-regulation of the mOrange transgene reporter appeared to correlate with reprogramming success. Similarly, cells expressing low levels of ICAM-1 or high levels of CD44 were found to have high levels of mOrange. This further supports the use of these markers to untangle the process of reprogramming, and indicates that continued high-level reprogramming factor expression is incompatible with successful reprogramming. This is similar to the observations of many viral reprogramming studies which reported incomplete reprogramming associated with continued viral transgene expression, and more recently a secondary reprogramming system which demonstrated that continued transgene expression may suppress full acquisition of

the reprogrammed state (Hotta and Ellis, 2008; Stadtfeld et al., 2008a; Golipour et al., 2012). This suggests another advantage of using a polycistronic vector for delivery of reprogramming factors is the ability to couple their expression to a reporter which itself can be used as an indicator of the progression of reprogramming. It should be noted that complete down-regulation of the reprogramming factors was not required to generate iPS cells using pBTAP IRI 2LMKOSimO. This has also been observed in viral mediated reprogramming of human cells with a polycistronic vector (Papapetrou et al., 2009). However, high transgene expression does appear to be required for successful reprogramming, as at least some portion of sorted $CD44^+$ cells were capable of generating Nanog-GFP^+ colonies.

3.3.3 Nanog-GFP expression during RA differentiated cell reprogramming

During reprogramming, populations of Nanog-GFP^+ cells could be detected outside of the iPS cell-like profile of ICAM-1^+CD44^-. Isolation of these cells highlighted that they achieved a reprogrammed profile more rapidly than their Nanog-GFP^- counterparts; however, these cells also appeared to follow the process of CD44 down-regulation followed by ICAM-1 up-regulation. Further experiments were required to establish the relevance, if any, of these changes in cell surface marker expression (see Chapter 4). Also of note was the observation that while GFP^- sorted cells appeared capable of giving rise to a minor population of ICAM-1^-CD44^- GFP^+ cells, $CD44^+GFP^+$ cells were almost undetectable. This indicated that these cell surface markers may be capable of dissecting the latter stages of reprogramming in detail, even after the expression of key pluripotency genes, and suggested the potential for heterogeneity in the state or behaviour of Nanog expressing cells during reprogramming, as observed in live cell imaging of the reprogramming of human cells (Chan et al., 2009). In addition, how cells achieve a fully reprogrammed state could be inferred, with some, rarer cells activating Nanog-GFP expression early ($CD44^+$) and rapidly completing the reprogramming process compared to the apparent majority of cells that only up-regulate GFP after CD44 expression is lost in

a process that is slightly delayed. One caveat of these results however, is the possibility that the *in vitro* differentiation model used to obtain the above results may not accurately reflect the kinetics of Nanog-GFP expression and up-regulation during reprogramming, and as with all of the optimisation experiments described in this chapter, required re-examination in a more practical, biologically relevant context (see Chapter 4).

.

CHAPTER 4 - Identification of a route to iPS cells from MEF

4.1 Introduction

Mouse embryonic fibroblast (MEF) cultures are among the most commonly used cell types to generate iPS cells, due to their use in the original study of reprogramming and ease of preparation (Takahashi and Yamanaka, 2006). The vast majority of subsequent work using this cell type has identified a number of key features of reprogramming, some of which have been successfully observed in the reprogramming of other cell types or cells from other species (Takahashi et al., 2007; Hanna et al., 2008; Li et al., 2009b; Ye et al., 2009). MEF have also been widely used in the screening and identification of small molecules which have been shown to increase the reprogramming efficiency, in terms of either increasing the total number of iPS cell colonies produced or accelerating the appearance of iPS cell colonies (reviewed in Feng et al., 2009b). For the majority of compounds identified, for example the use of "2i" (GSK3 and FGFR inhibitor) conditions to generate fully reprogrammed cells from pre-iPS cells, it is unclear as to whether their use enhances the process (i.e. more cells successfully complete reprogramming) or rather their use simply selects for those cells which are most capable of adapting in drug-supplemented culture (Silva et al., 2008). In addition, it may be possible that cells achieve a fully reprogrammed state by a variety of pathways, with small molecules enhancing one route over another. This may be especially relevant in the case of reprogramming MEF cultures which are typically heterogeneous in nature, and therefore an accurate system to monitor changes in cells undergoing reprogramming is required (Nemajerova et al., 2012).

4.1.1 Aims of this chapter

The aim of this chapter is to assess the use of the identified novel cell surface markers ICAM-1 and CD44 in reprogramming from MEF. In addition, other reprogramming systems were utilised to examine the reproducibility of this marker system. Changes in reprogramming efficiency and how this related to expression of these markers was also carried out. Finally, based on observations from these experiments a number of intermediate reprogramming populations were identified, isolated and scrutinised in order to determine how reprogramming progresses and the pathway/s MEF take in order to become iPS cells.

4.2 Results

4.2.1 Efficient generation of iPS cells from transgenic MEFs

The iPS cell line D6s4B5 was used to generate chimeric embryos from which secondary (2°) MEF were isolated (Figure 4.1a). Varying concentrations of dox (1.5ug mL^{-1} to 60ng mL^{-1}) were tested to determine what concentration was optimal for 2° MEF reprogramming, with 300ng mL^{-1} found to generate the greatest number of colonies (data not shown). During reprogramming cells were observed to up-regulate reprogramming factor expression, undergo morphological changes, generate large iPSC-like colonies and eventually express Nanog-GFP (Figure 4.1b).

In order to investigate if the reprogramming efficiency of this secondary system could be further increased, 2° MEF were cultured in the presence of vitamin C (VitC, 10ug mL^{-1}) and/or A83-01, an inhibitor of the activin receptor-like kinase receptor 4,5,7 (ALKi, 500nM) which is involved in the TGF-β signalling pathway, both of which have been shown to have a positive effect on reprogramming efficiency (Esteban et al., 2010; Li et al., 2010). Both VitC and ALKi dramatically increased the total number of AP^+ colonies generated compared to dox alone (~8 fold), and surprisingly, an even greater increase (~25 fold) was observed when these small molecules were used in combination (Figure 4.1c). This culture condition (dox/VitC/ALKi) was used for all further experiments with the D6s4B5 iPS cell line, unless otherwise stated.

Colony formation was closely monitored in order to understand the nature of reprogramming in this secondary system (Figure 4.1d). After initial up-regulation of reprogramming factor expression it was observed that cells rapidly clustered in small groups. Shortly thereafter cell death was observed among certain cells in the colonies. Cells which survived this period then rapidly expanded and eventually up-regulated Nanog-GFP expression. A direct correlation could be made between the time at which colonies reached this minimum size and the subsequent up-regulation of the Nanog-GFP reporter, suggesting that not all cells are capable of

successfully completing reprogramming even after successful transgene up-regulation and clustering, and that a type of "selection" occurs during the reprogramming process with only successful cells surviving (Figure 4.1e).

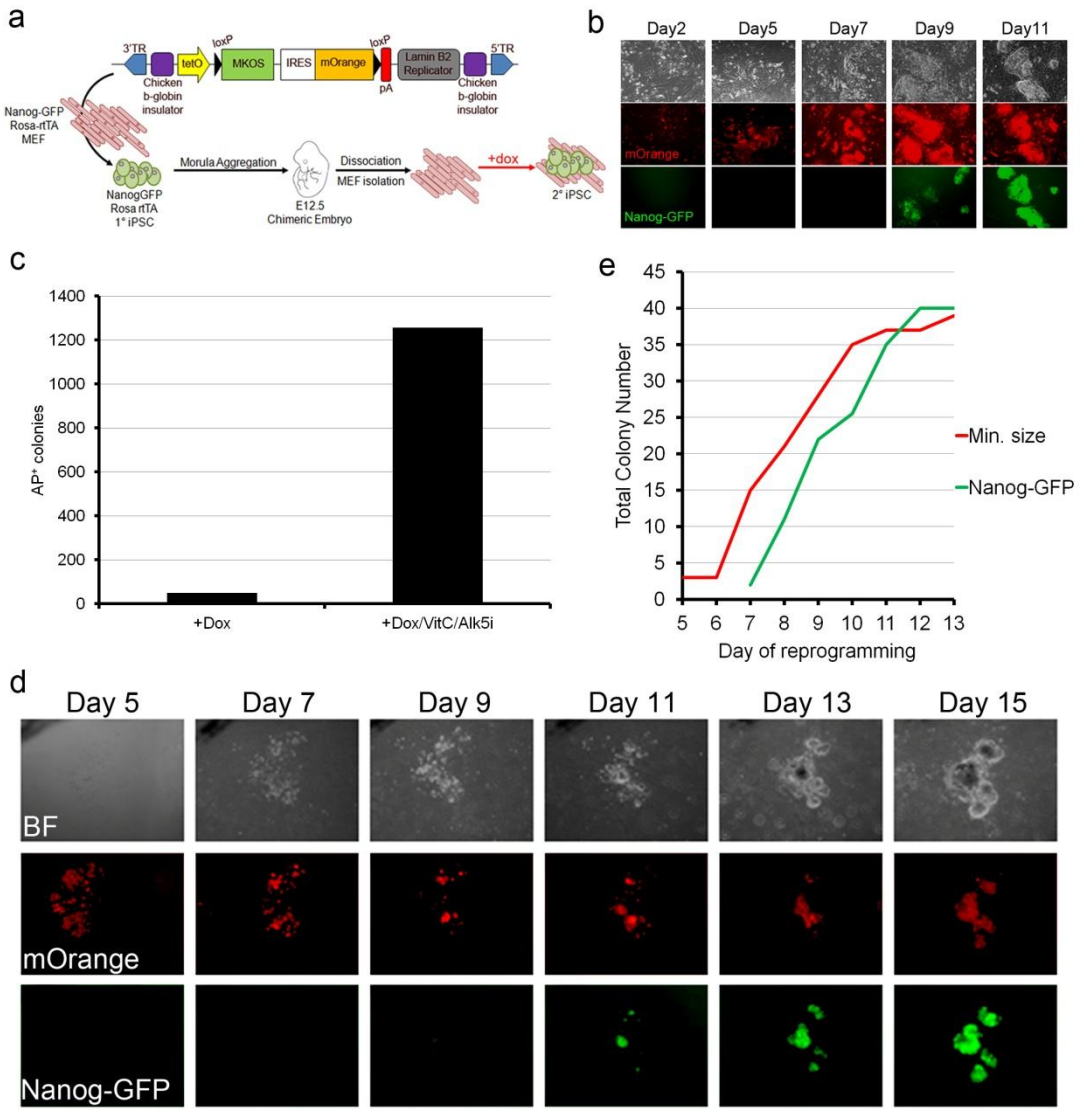


Figure 4.1: Efficient reprogramming system to generate iPS cells. a. Overview of secondary reprogramming system from MEF. **b.** Colony formation in the presence of dox. **c.** Colony formation was greatly enhanced in the presence of VitC and ALKi. **d.** Monitoring of colony formation during reprogramming revealed clustering of transgene expressing cells with only some cells capable of completing successful reprogramming (Nanog-GFP up-regulation). **e.** The timing of minimum colony appearance was correlated with subsequent Nanog-GFP expression.

4.2.2 Previously reported cell surface markers are not closely correlated with Nanog-GFP expression

MEFs undergoing reprogramming were initially examined via FACS analysis for the dynamics of mOrange and Nanog-GFP expression (Figure 4.2a). It was observed that mOrange expression was rapidly up-regulated and maintained for about 7 days before Nanog-GFP expressing cells were observed. Interestingly Nanog-GFP$^+$ cells did not appear to arise from cells expressing the highest level of the reprogramming factors; rather they appeared from a population of cells expressing slightly lower levels of mOrange (Day 8). This underlines the importance of establishing the correct level of reprogramming factor expression for successful reprogramming. As the Nanog-GFP$^+$ population matured and expanded, these cells displayed a further gradual down-regulation of mOrange (reprogramming factor) expression. This is similar to the loss of transgene expression widely associated with completed reprogramming in retroviral reprogramming systems (Maherali et al., 2007; Okita et al., 2007; Wernig et al., 2007). Next, reprogramming cells were examined for their expression of the widely used, previously reported MEF associated marker THY-1, as well as the ES cell-associated cell surface markers E-CADHERIN and SSEA-1.

THY-1 expression in the transgenic MEF was found to be quite broad, with both THY-1^{high} and THY-1^{low} populations (Figure 4.2b). After 4 days of dox addition and reprogramming factor expression, a loss of the THY-1^{high} population could be observed. By day 6 of reprogramming almost all cells had lost THY-1 expression, and no further THY-1 expression could be detected during the reprogramming process. As expected, all Nanog-GFP$^+$ cells arising during reprogramming were THY-1$^-$.

It has been widely reported that during reprogramming cells undergo a mesenchymal-to-epithelial (MET) transition (Li et al., 2010; Samavarchi-Tehrani et al., 2010). **E-CADHERIN** is a marker of cells which have successfully undergone this transition (Figure 4.2c). As expected, MEF, as a mesenchymal cell type, were

negative for this marker. However, 2 days of transgene expression was sufficient to generate E-CADHERIN expressing cells, resulting in almost 100% of cells positive for this marker by day 6 of reprogramming. Some residual cells failed to up-regulate E-CADHERIN expression, however Nanog-GFP$^+$ cells were observed to only arise from the E-CADHERIN$^+$ population.

SSEA-1 is a classical marker of undifferentiated ES cells (Solter and Knowles, 1978). After four days of transgene expression a minor population of SSEA-1$^+$ cells were observed. By day 6 about 50% of the cells expressed SSEA-1, and this proportion was maintained throughout the remainder of the reprogramming process and in fully reprogrammed iPS cells. Interestingly, unlike the markers THY-1 and E-CADHERIN, it appeared that Nanog-GFP$^+$ cells arose from both SSEA-1 negative and positive populations (Day 8, Day 10), consistent with broad SSEA-1 expression levels in undifferentiated ES/iPS cells (Figure 4.2d). In order to determine the source of SSEA-1$^-$/Nanog-GFP$^+$ cells, we isolated cells at day 5 of reprogramming, before the appearance of Nanog-GFP$^+$ cells (Figure 4.3a). These cells were sorted based SSEA-1 expression and re-plated in reprogramming conditions (Figure 4.3b). SSEA-1 expression was observed to fluctuate in both sorted populations, indicating that SSEA-1 expression is not stable in the latter stages of reprogramming. In addition, Nanog-GFP$^+$ cells arose from both sorted populations by day 2 after sorting, indicating SSEA-1 expression does not predict the appearance of Nanog-GFP expression during reprogramming.

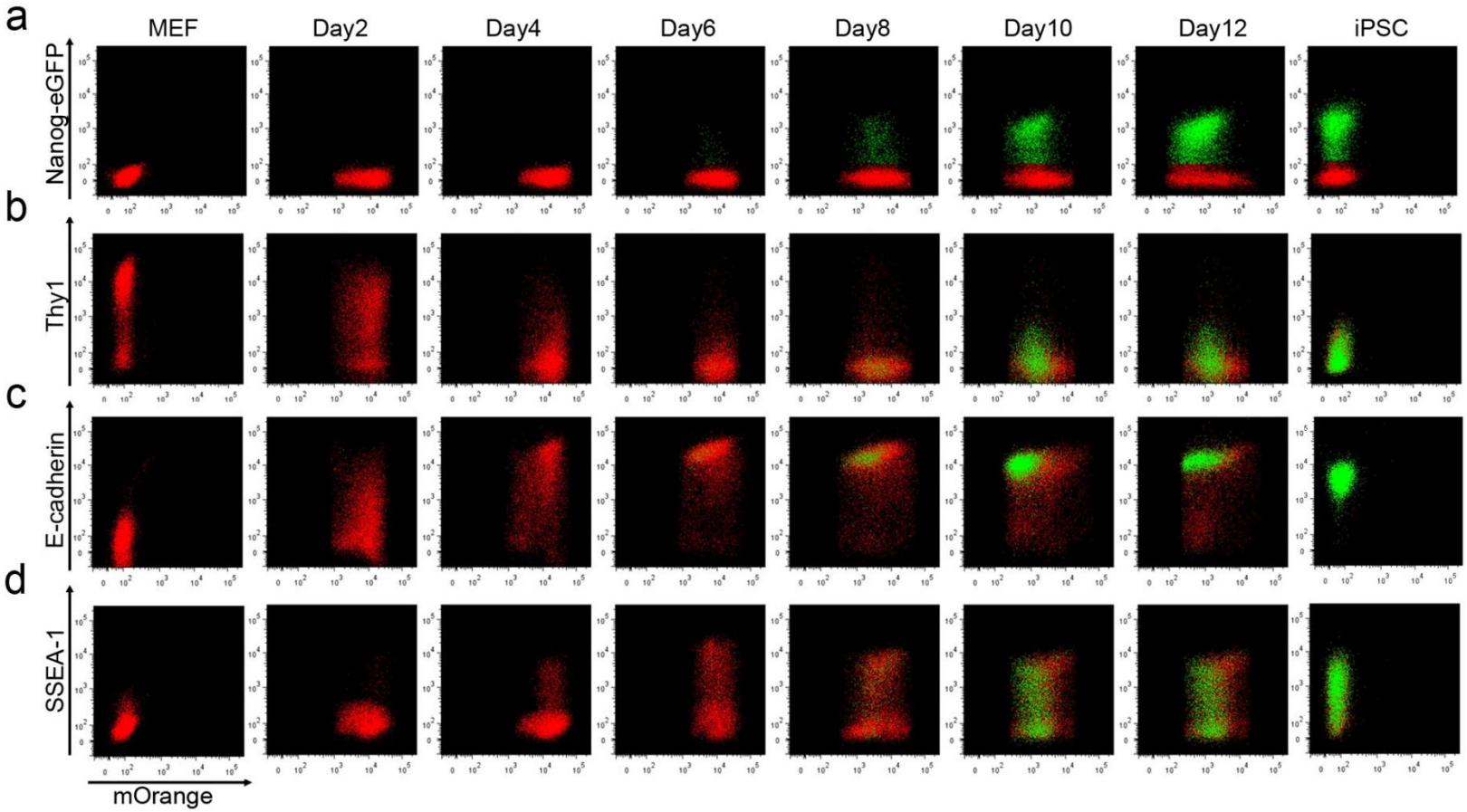


Figure 4.2: FACS analysis of Nanog-GFP and conventional reprogramming marker expression during secondary reprogramming. a. Expression of Nanog-GFP during secondary reprogramming was monitored upon transgene expression indicated by the reporter mOrange. Expression of THY-1 (**b.**), E-CADHERIN (**c.**) and SSEA-1 (**d.**) expression was similarly monitored. Red; Nanog-GFP^{-} cells, Green; Nanog-GFP^{+} cells.

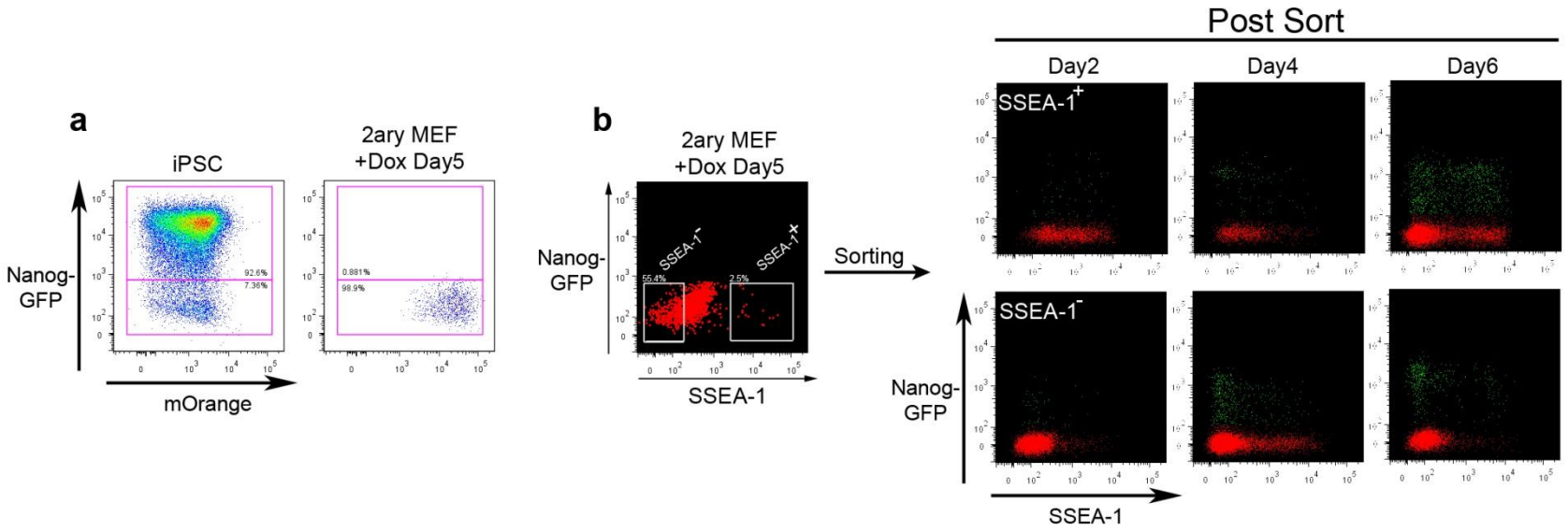


Figure 4.3: SSEA-1 expression does not predict the appearance of Nanog-GFP$^+$ cells. a. Gating strategy for Nanog-GFP$^-$ cells at day 5 of reprogramming. **b.** Sorting strategy for Nanog-GFP$^-$, SSEA-1$^{+/-}$ cells at day 5 of reprogramming. Cells were isolated, re-plated in reprogramming conditions and reanalyzed every 48hours. Red; Nanog-GFP$^-$ cells, Green; Nanog-GFP$^+$ cells.

4.2.3 CD44, ICAM-1 and Nanog-GFP expression dynamics during reprogramming reveal distinct population changes

MEF undergoing reprogramming were examined for CD44 and ICAM-1 expression in order to confirm the previous observations made using the *in vitro* retinoic acid differentiation system (Chapter 3).

CD44 expression was found to be relatively high and homogeneous in the undifferentiated MEF (Figure 4.4a). Expression of this marker was found to be maintained at this high level until at least day 6 of reprogramming. At day 8, a population of $CD44^-$ cells could be observed. Nanog-GFP^+ cells were found almost exclusively in this $CD44^-$ population. This trend was confirmed upon continued reprogramming, with the vast majority of Nanog-GFP^+ cells found in the $CD44^-$ population.

The pattern of **ICAM-1** expression in MEF was observed to be relatively broad (Figure 4.4b). Upon initiation of reprogramming factor expression this broad expression is maintained. However, around day 6 there is a marked decrease in the ICAM-1^{high} population. This was followed, at day 8, by an increase of the ICAM-1^{high} population, concurrent with a gain in Nanog-GFP expression. By day 12 of reprogramming the majority of Nanog-GFP^+ cells were ICAM-1^+, similar to fully reprogrammed iPS cells.

Unlike the previously described markers (Section 4.2.2), the changes in both the loss of CD44 and gain of ICAM-1 expression appeared to be closely correlated with a gain in Nanog-GFP expression. This suggested that these markers may be more suitable to identify intermediate stages of the reprogramming process by providing an insight to events occurring between the gain of E-CADHERIN expression and up-regulation of pluripotency genes. Also of note is that the expression pattern of both CD44 and ICAM-1 closely matched that observed in the retinoic acid differentiation system, suggesting that this *in vitro* method is suitable for screening other potential markers of the reprogramming process.

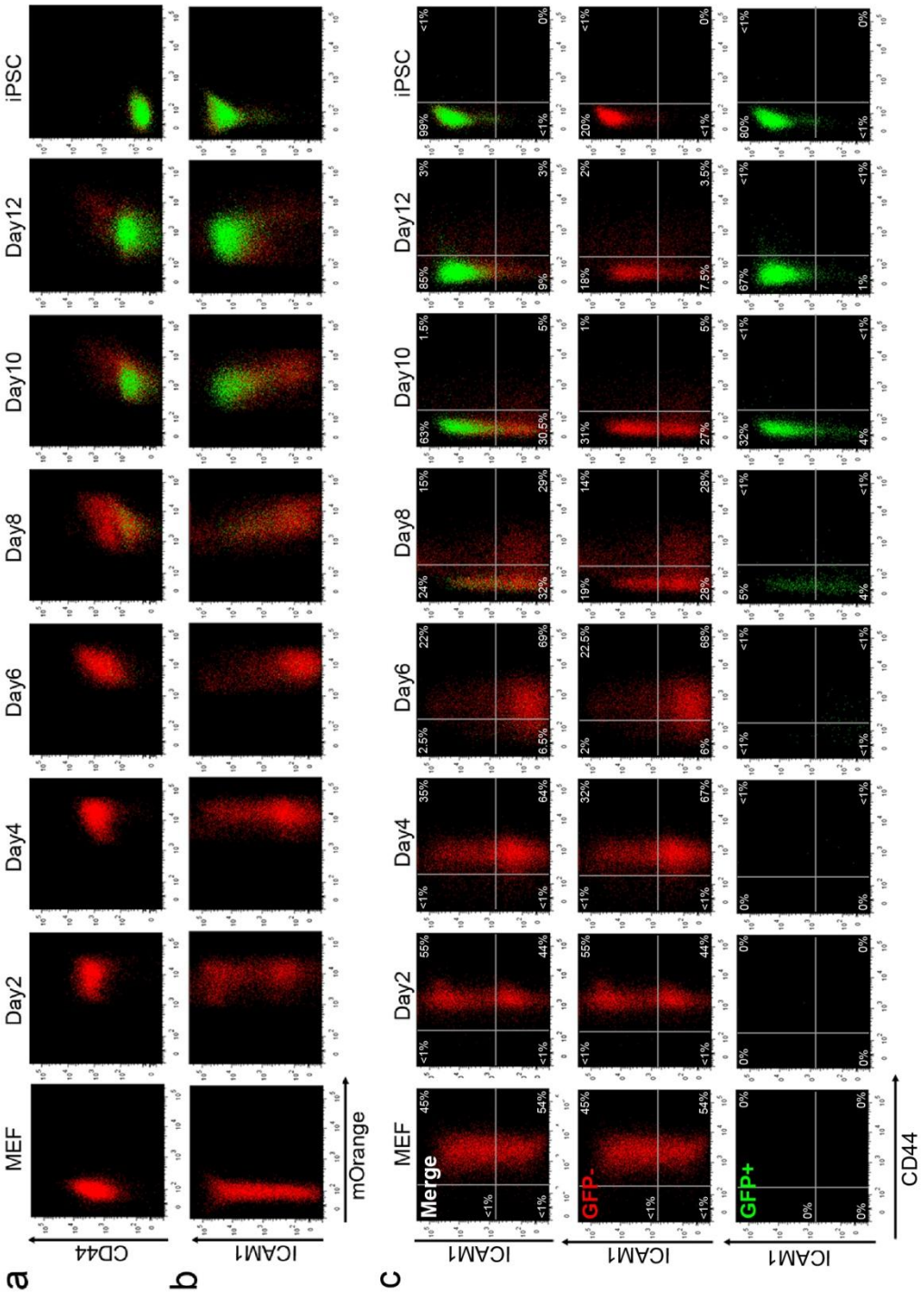


Figure 4.4: FACS analysis of CD44 and ICAM-1 expression during secondary reprogramming from MEF. a. Expression of CD44 was monitored upon transgene induction. Note correlation of CD44 down-regulation and Nanog-GFP appearance at Day 8. **b.** ICAM-1 expression during reprogramming. **c.** CD44/ICAM-1 double staining was carried out during secondary reprogramming. Percentages indicate proportion of cells in each gate. Red; Nanog-GFP$^-$ cells, Green; Nanog-GFP$^+$ cells.

In order to further investigate the significance of CD44 and ICAM-1 expression changes, cells undergoing reprogramming were double-stained for these markers (Figure 4.4c). As expected MEF displayed high CD44 and broad ICAM-1 expression. By day 4 more than 50% of cells had down-regulated ICAM-1 expression. At day 6 a small proportion (~6%) of cells had also down-regulated CD44 expression, giving rise to a double negative population. At this stage, a small proportion of ICAM-1 single positive cells were also observed (~2%). This population, which displayed an iPS cell-like $CD44^-/ICAM\text{-}1^+$ phenotype, appeared more prominent at day 8, with almost a quarter of cells found within this gate. FACS analysis at subsequent time-points (Day 10, 12) revealed an increase in this ICAM-1 single positive population concurrent with a decrease in the $CD44^-/ICAM\text{-}1^-$ population.

Interestingly, we identified cells expressing the Nanog-GFP reporter before the acquisition of a $CD44^-/ICAM\text{-}1^+$ surface marker profile. A small proportion (<1%) of $CD44^{+/-}/ICAM\text{-}1^-$ cells at day 6 were found to $Nanog\text{-}GFP^+$. While the $Nanog\text{-}GFP^+/ CD44^+$ population remained minute, at days 8, 10 and 12 a significant population of $Nanog\text{-}GFP^+/CD44^-/ICAM\text{-}1^-$ cells could be observed. At day 8 a similar proportion of $Nanog\text{-}GFP^+/CD44^-/ICAM\text{-}1^+$ cells were detected however, unlike their $ICAM\text{-}1^-$ counterparts, cells in this gate appeared to readily expand, and at later stages represented the largest population present in the reprogramming cultures (Day 10, 12). Significantly however, despite the appearance of GFP^+ cells, GFP^- cells could always be detected in the same gates, usually in greater proportion, and even at the later stages of reprogramming (e.g. Day 8, 10, 12).

$CD44^+/ICAM\text{-}1^+$ cells appeared to rapidly disappear from cultures undergoing reprogramming. In order to investigate if the heterogeneity of ICAM-1 expression in the starting population of MEF influenced the dynamics of the reprogramming process, cells were sorted based on ICAM-1 expression and re-plated in reprogramming conditions (Figure 4.5). FACS analysis indicated that both $ICAM\text{-}1^+$ and $ICAM\text{-}1^-$ MEF followed a similar pattern of cell surface marker

changes. The kinetics of the reprogramming process were also largely identical with a very slight delay in ICAM-1^{+} sorted MEF in CD44 down-regulation. This data indicated that ICAM-1 expression does not distinguish the reprogramming potential of heterogeneous MEF.

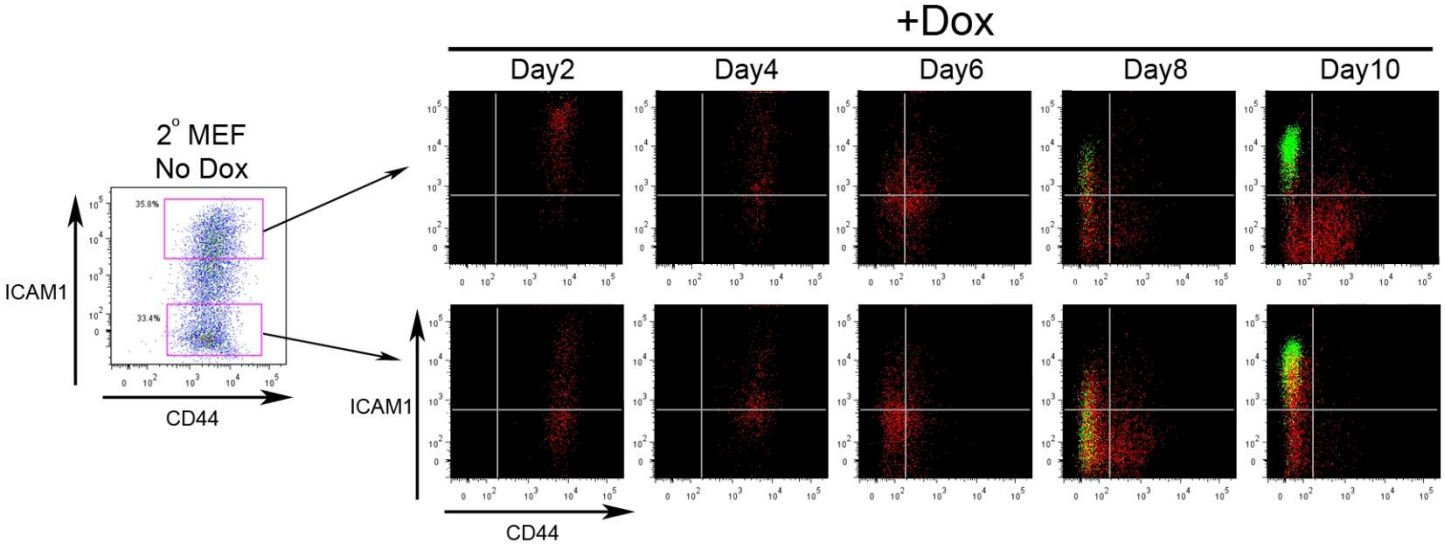


Figure 4.5. Similar reprogramming kinetics with ICAM-1$^{+/-}$ MEFs. ICAM-1^{+} and ICAM-1^{-} secondary MEF were sorted before initiating reprogramming. CD44, ICAM-1, Nanog-GFP expression was monitored every two days during reprogramming. Red; Nanog-GFP^{-} cells, Green; Nanog-GFP^{+} cells.

These results suggested that during reprogramming from MEF, cells follow a distinct series of changes in CD44 and ICAM-1 marker expression, with ICAM-1 down-regulation followed by gradual CD44 down-regulation which in turn is followed by a rapid up-regulation in ICAM-1 expression. The dynamics of Nanog-GFP reporter activation in relation to these markers did not appear to be as straightforward however, with the appearance of both Nanog-GFP$^+$ and –GFP$^-$ cells in populations with identical ICAM-1/CD44 expression profiles. Therefore more accurate analysis in order to determine how these markers correlated with the reprogrammed state was required.

4.2.4 Identification of intermediate populations of the reprogramming process

The previous results indicated that during reprogramming from MEF, ICAM-1 and CD44 expression could potentially be used to identify different stages in the reprogramming process. However, FACS analysis reveals only the total number of cells within each gate at the time of analysis. This means that it is not possible to accurately determine if the observed changes in population dynamics are the result of one population arising from another, or is simply the rapid expansion of one population over another. This may in fact be a common issue with many reprogramming studies, and the idea that "more" reprogrammed cells may grow more favourably in stem cell conditions is plausible. To this end, we attempted to isolate the individual populations identified using CD44 and ICAM-1 and closely monitor their behaviour. In addition, in order to identify the expression dynamics of the Nanog-GFP reporter, cells in Nanog-GFP$^+$ and GFP$^-$ sub-populations were isolated.

Day 10 of reprogramming was selected as this time-point consistantly displayed enrichment for all identified populations of interest (Figure 4.6a). Nanog-GFP$^+$ (NG+) or Nanog-GFP$^-$ (NG-) cells were further subdivided based on their expression of CD44 and ICAM-1: CD44$^+$ICAM-1$^+$ (Gate 1 with 1NG-, 1NG+

populations); CD44^{-}ICAM-1^{-} (Gate 2 with 2NG-, 2NG+ populations); and CD44^{-} ICAM-1^{+} (Gate 3 with 3NG-, 3NG+ populations.) The CD44^{+}/ICAM-1^{+} populations were not isolated due to their very low numbers and later experiments indicated these cells do not represent a significant reprogramming pathway (see Figure 4.6b).

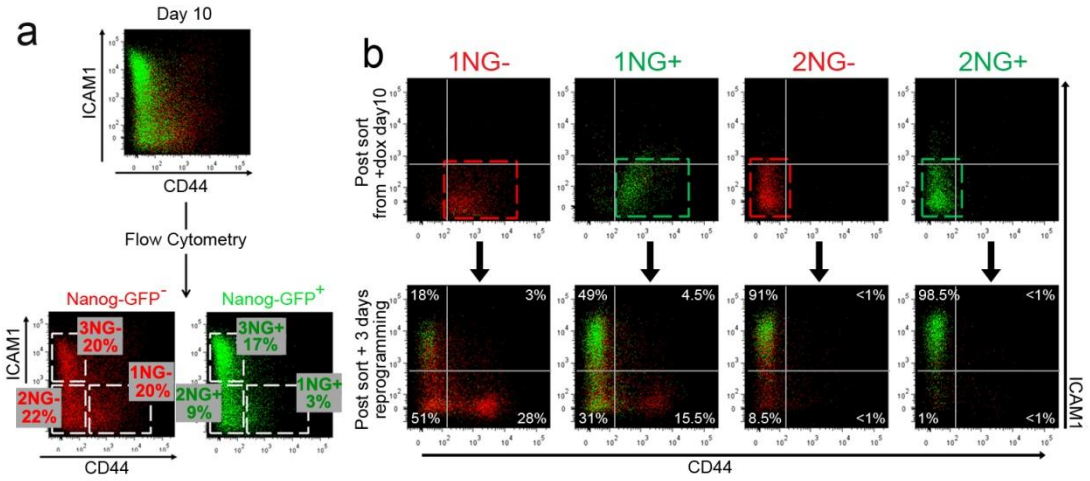


Figure 4.6. CD44/ICAM-1 sub-populations represent distinct stages of reprogramming. a. Nanog-GFP^{+} (NG+) and Nanog-GFP^{-} (NG-) cells were subdivided into CD44^{+} ICAM-1^{-} (Gate1), CD44^{-} ICAM-1^{-} (Gate 2) and CD44^{-} ICAM-1^{+} (Gate 3) populations at day 10 of reprogramming. **b.** FACS analysis of sorted sub-populations after 3 day culture in the presence of dox. Red; Nanog-GFP^{-} cells, Green; Nanog-GFP^{+} cells.

Cells were sorted from 1NG+/- and 2NG+/- in order to monitor their progression during the reprogramming process. 3NG+/- populations were not sorted as they already expressed the iPS cell CD44^{-}ICAM-1^{+} phenotype and represented a stable end point with regard to the novel cell surface markers' profile. The sorted populations were re-plated in reprogramming conditions and re-analysed for CD44, ICAM-1 and Nanog-GFP expression after a further 3 days of culture (Figure 4.6b). 2NG+ cells were found to maintain Nanog-GFP expression and rapidly gained ICAM-1 expression, entering Gate 3. Their Nanog-GFP^{-}

counterparts (2NG-) also gained ICAM-1 with apparent ease, however, in contrast to the 2NG+ population where around 1% of cells retained a double negative profile (ICAM-1$^-$/CD44$^-$), 8.5% of 2NG- cells maintained this phenotype. Surprisingly, 1NG+ cells, which represented the smallest population at day 10, showed down-regulation of CD44 expression and produced a large 3NG+ population. Interestingly, unlike the 2NG+ populations not all cells appeared to maintain Nanog-GFP expression, most notably those cells which did not downregulate CD44 expression, although GFP$^-$ cells could be found in all three gates. Finally, analysis of the 1NG- population revealed that although the majority (69%) of cells were CD44$^-$, few of these cells had also gained ICAM-1 expression (18%), and fewer still Nanog-GFP expression (~5%). In addition, the reprogramming kinetics of the 1NG- population are delayed, with greater numbers of 1NG- and 2NG- cells observed compared to the 1NG+ sub-population.

Taken together, these results suggested that reprogramming proceeded in the order of Gate 1→Gate 2→Gate 3. However, it was also clear that expression of the Nanog-GFP reporter had an effect on the kinetics of the movement from each gate/stage to the next. In order to determine if these kinetic differences reflect the probability of each sub-population to successfully undergo reprogramming, cells were again isolated at day 10, including 3NG+/- populations, and plated at clonal density (Figure 4.7). Cells were cultured for an additional ten days after reprogramming and Nanog-GFP positive colonies were quantified and compared to a sorted, fully reprogrammed iPS cell line. From this analysis we identified the 3NG+ population as having a similar colony forming potential (cfp) as fully reprogrammed iPS cells, suggesting this population was close to a completely reprogrammed state. The other isolated populations displayed a range in their cfp which appeared to be correlated to how distant each population was from attaining a 3NG+ profile. Interestingly, 3NG- cells, although bearing an iPS cell-like CD44 and ICAM-1 profile, had a much reduced cfp compared to 3NG+. Similarly, among populations with the same CD44/ICAM-1 cell surface marker profile, in all cases the

Nanog-GFP+ cells displayed a greater ability to generate colonies, consistant with the observed differences in reprogramming kinetics between NG+ and NG- sub-populations.

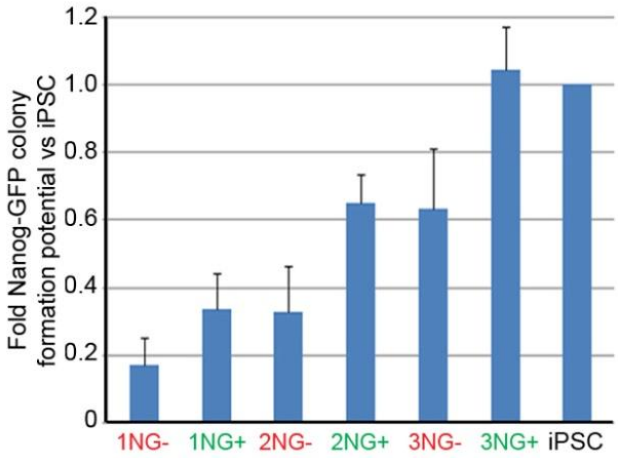


Figure 4.7: Relative probability to generate Nanog-GFP$^+$ iPS cell colonies. Day 10 sorted sub-populations were plated and GFP$^+$ colonies counted after 10 days of culture in reprogramming conditions and compared to fully reprogrammed iPS cells. Error bars represent standard deviation, n=3.

4.2.5 Behaviour of sorted sub-populations reveals distinct characteristics of reprogramming

While analysis at 72 hours revealed the macro-behaviour of the sorted populations, this time period was sufficient to allow more rapid proliferation of one population over another. This obscured the actual transition rate of each individual sub-population. Therefore, a greater level of accuracy was required to understand the mechanism of reprogramming. This was achieved by sorting all populations at day 10 and carrying out re-analysis after just 24 hours of continued culture in reprogramming conditions (Figure 4.8). Surprisingly, the overall result of this experiment did not vary greatly from that carried out at 72 hours, thus confirming

the original analysis on the characteristics of each sub-population. Of note however, is the sorting of the 3N populations. While 3NG+ barely changed after 24 hours, supporting the idea that this is a stable and reliable marker set with which to define the endpoint of reprogramming, 3NG- cells showed a low rate of conversion to 3NG+.

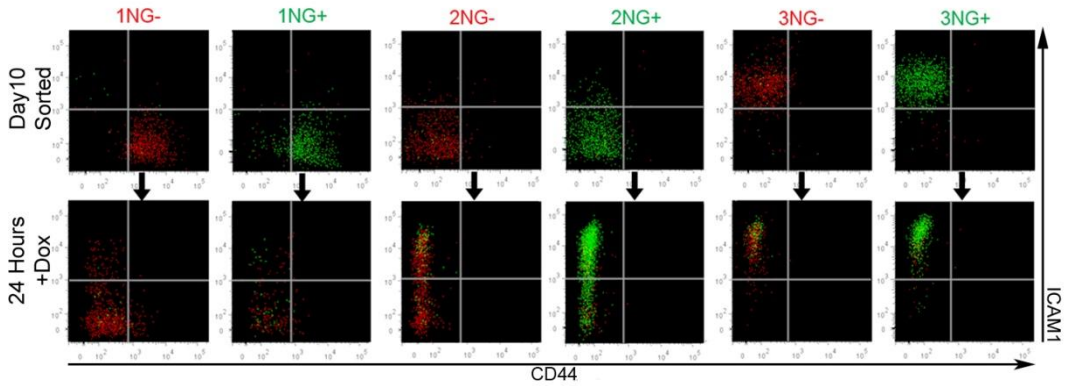


Figure 4.8: CD44/ICAM-1/Nanog-GFP expression re-analysed 24 hours after sorting. Sub-populations were sorted at day 10 and underwent FACS analysis after 24 hours of culture in reprogramming conditions. Red; Nanog-GFP$^-$ cells, Green; Nanog-GFP$^+$ cells.

This analysis also enabled the identification and quantification of the total number of cells arising from each sub-population and their tendency to move to the next stages of reprogramming as defined by the CD44/ICAM-1/Nanog-GFP marker system (Figure 4.9). Further supporting the identification of the 3NG+ population as being very similar to fully reprogrammed iPS cells, we discovered that this population maintains its phenotype and rapidly expands in culture. In contrast, 3NG- cells have a lower capacity to proliferate, and a small number of cells can successfully transition to a 3NG+ state. 2NG+ cells rapidly gain 3NG+ status and a minor number of cells remain 2NG+ or lose GFP expression to become 2NG-. 2NG- cells tend to transition rapidly to 3NG-, with low Nanog-GFP activation. This

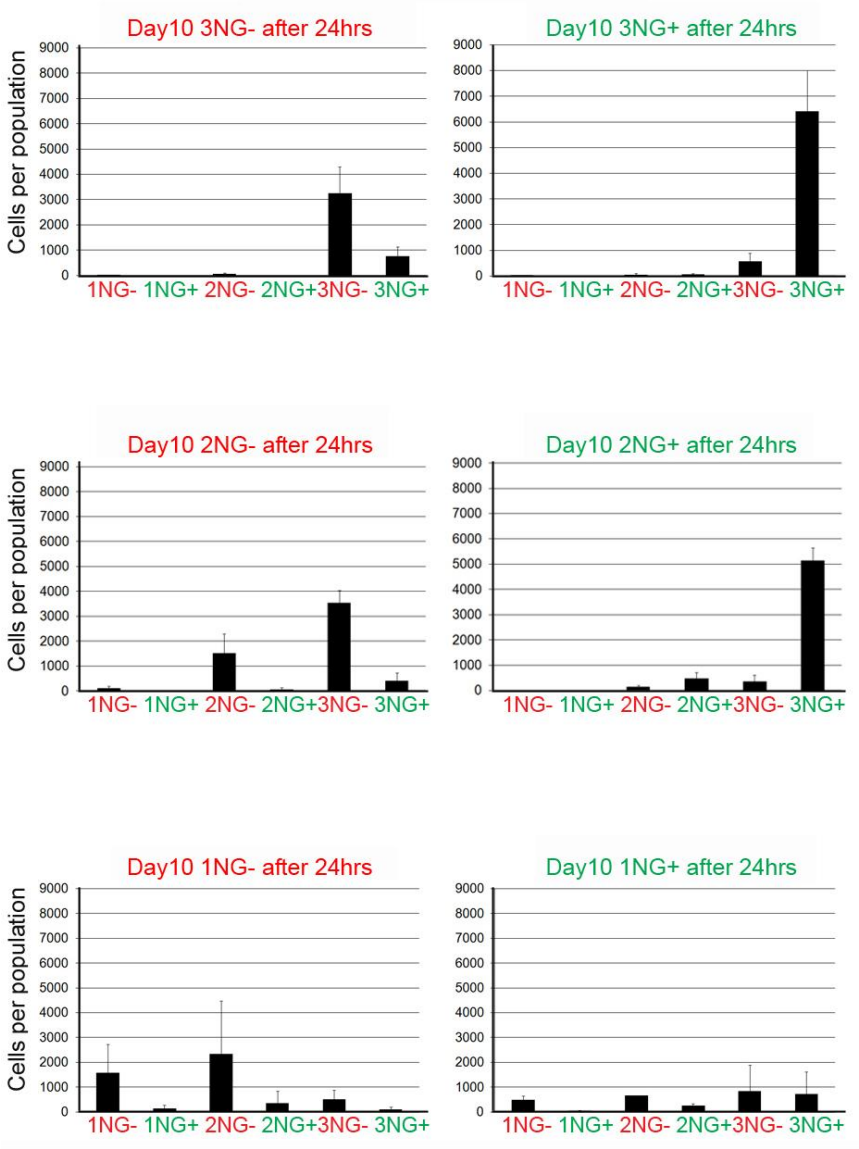


Figure 4.9: Total number of cells in each gate after 24 hours for day 10 sorted populations. Each reprogramming population was sorted at day 10 and cells cultured in reprogramming conditions for 24 hours. Cells were then harvested and CD44/ICAM-1/Nanog-GFP expression was re-analysed. Total cell numbers found in each gate were plotted. The error bars represent standard deviation. n=3.

population also shows a higher level of proliferation than 2NG+ cells. As observed, the 1NG+ cells have the potential to transition rapidly to the final 3NG+ state, however, they appear to be somewhat heterogeneous in their behaviour, generating similar numbers also of 2NG- and 3NG- cells. Finally, 1NG- cells display a more predictable pattern, with approximately half the total cells in a 2NG- state, with the remainder maintaining their sorted phenotype.

In order to investigate if the observed transition trend was a conserved characteristic of the cells in each identified gate, and was not specific to day 10 sub-populations, similar 24 hour analysis was carried out at day 8 of reprogramming (Figure 4.10). This revealed that although the proportions of each of the identified sub-populations varied at this time-point compared to Day 10, their behaviour was broadly similar.

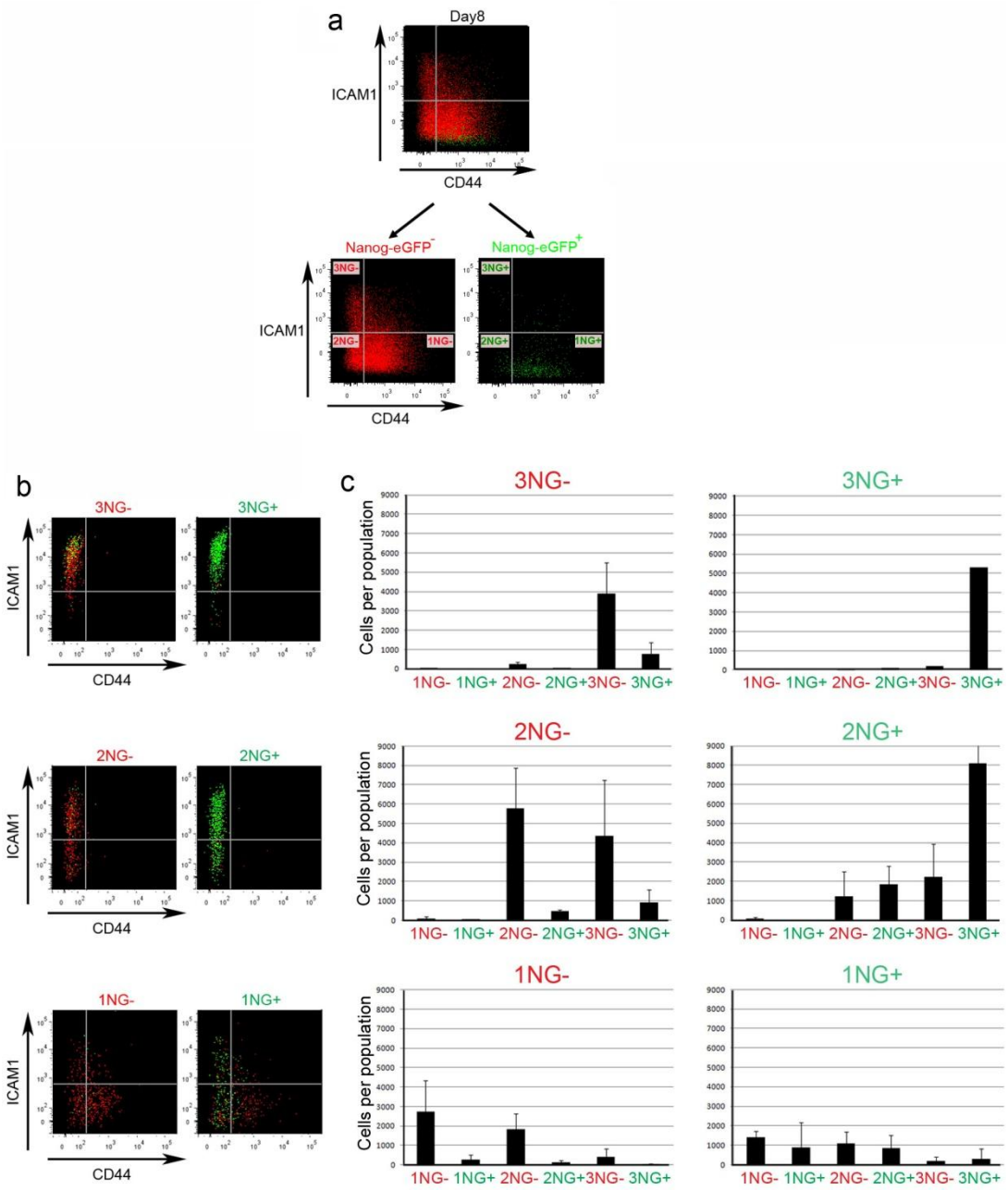


Figure 4.10: Behaviour of Day 8 sorted sub-populations are similar to that of day 10 sub-populations. **a.** Sorting strategy at day 8 of reprogramming. **b**. Each sub-population sorted at day 8 was re-plated in reprogramming conditions, and reanalyzed after 24 hours. **c.** Total cell numbers in each gate after 24 hour analysis for each sorted population. The error bars represent the standard deviation. n=3. Red; Nanog-GFP$^-$ cells, Green; Nanog-GFP$^+$ cells.

4.2.6 Validation and application of CD44/ICAM-1 reprogramming marker system.

Before progressing with further analysis using the CD44/ICAM-1 marker system and its associated sub-populations, it was important to validate the usefulness of these markers in other contexts. This would ensure that these observations were reproducible and not context specific.

The 6c cell line was generated using 4 separate *piggyBac* vectors to deliver the dox inducible reprogramming factors *c-Myc*, *Klf4*, *Oct4* and *Sox2* (Woltjen et al., 2009). Using secondary MEF isolated from 6c generated chimera, secondary reprogramming and FACS analysis was carried out (Figure 4.11a). Despite the absence of a pluripotency reporter, CD44 and ICAM-1 expression could be used to identify cell surface marker changes during the reprogramming process. It was observed that the changes in CD44/ICAM-1 expression recapitulated the observations made using the D6s4B5 cell line. The majority of cells were ICAM-1^{low} by day 6 of reprogramming, and by day 10 some cells had lost CD44 expression. Some cells expressed an ES cell-like ICAM-1^{+}/CD44^{-} profile, and this population was larger at a later time-point (day 14). This indicated that even in systems containing multiple, random insertions of the reprogramming factors these novel markers could be used to trace the fate of cells undergoing reprogramming.

Another potential issue identified with this marker system is its potential for use in lower efficiency, but vastly more common, primary reprogramming contexts. Primary reprogramming was carried out using the same vector used to generate the D6s4B5 cell line, pB TAP IRI MKOSimO (**MKOS**), and a modified version in which the polycistronic reprogramming cassette was replaced with factors in the order of *Oct4*, *Sox2*, *Klf4* and *c-Myc* (**OSKM**, prepared by Dr. Eleni Chantzoura)(Carey et al., 2011) (Figure 4.11b).

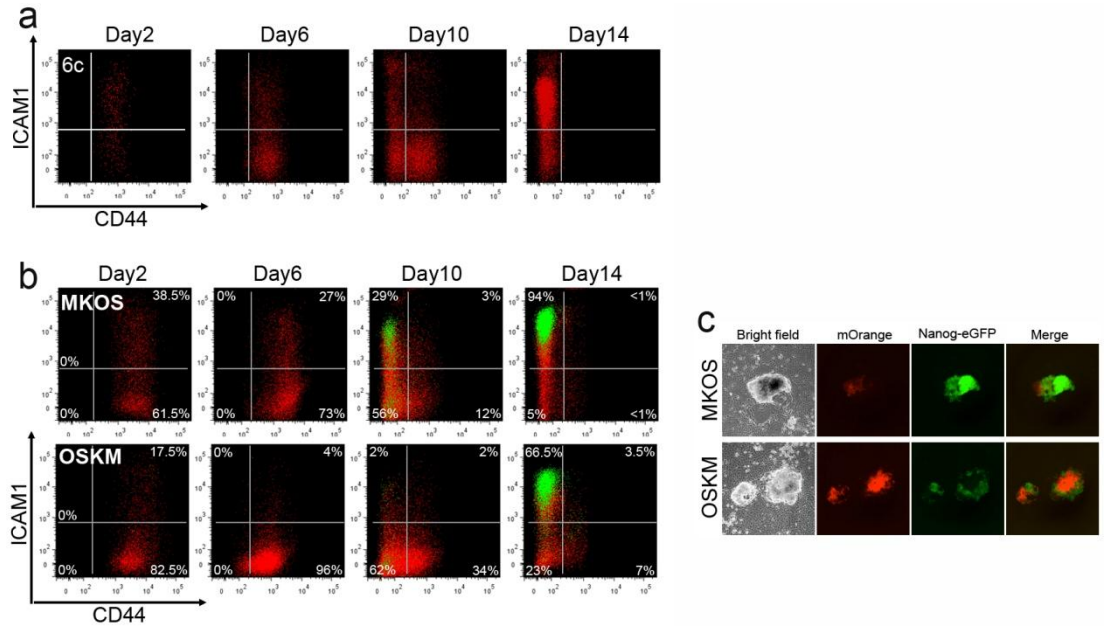


Figure 4.11: The reprogramming pathway is conserved in different reprogramming systems. a. Non-polycistronic PB 2^{o} reprogramming with the 6c cell line. **b.** Primary PB reprogramming using both MKOS and OSKM polycistronic cassettes. **c.** Typical colonies arising from MKOS, OKMS primary PB reprogramming. Red; Nanog-GFP^{-} cells, Green; Nanog-GFP^{+} cells.

It had been observed that the colonies generated using the MKOS and OSKM vectors differed somewhat, with some OSKM-derived colonies failing to down-regulate mOrange expression and similarly fail to acquire Nanog-GFP expression (Figure 4.11c). Further differences became apparent in the course of the primary reprogramming experiments; OSKM cultures appeared to down-regulate ICAM-1 expression more rapidly than MKOS cells (Day 6), however this apparent advantage was lost at later stages of reprogramming with MKOS cultures reaching 3NG+ more quickly and in greater numbers (Day 10). This experiment highlighted how differences observed in the kinetics of reprogramming using this marker system could accurately reflect events occurring on the heterogeneous, colony level.

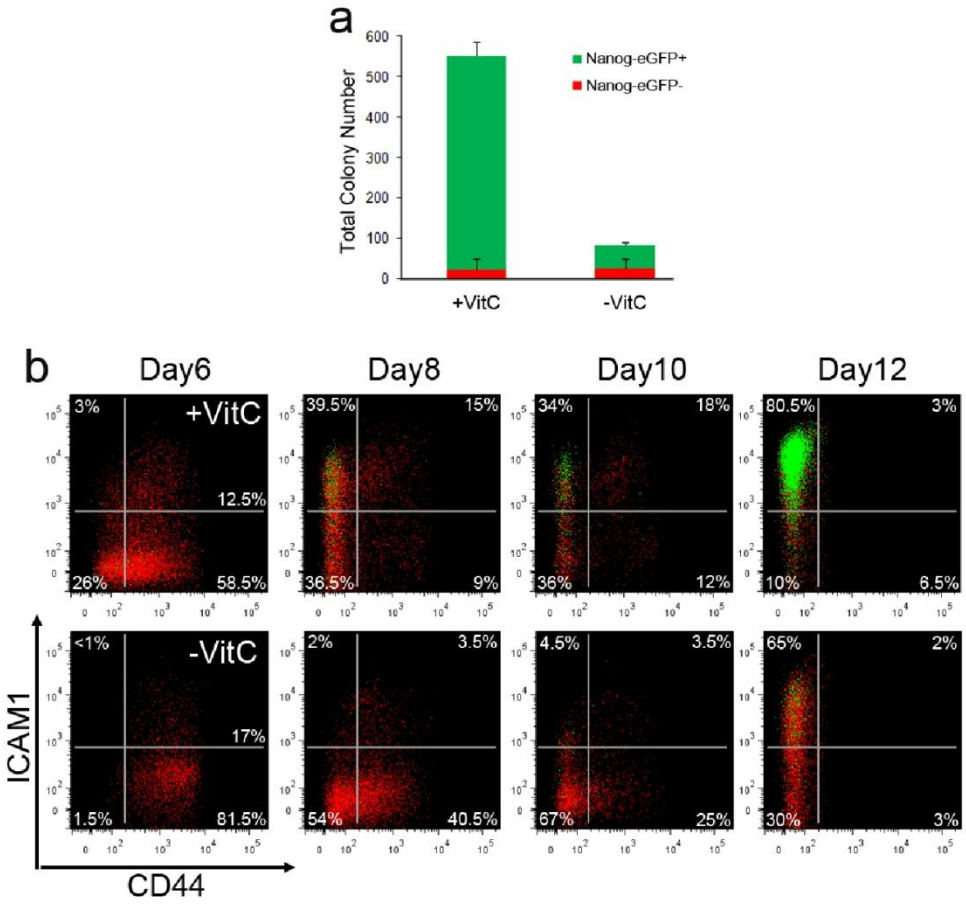


Figure 4.12: Secondary reprogramming in the presence of dox, ALKi either with or without VitaminC (+ or -VitC). a. Both total and Nanog-GFP$^+$ colony numbers decreased in the absence of VitC. **b.** Cells reprogrammed in the absence of VitC displayed delayed reprogramming kinetics. Red; Nanog-GFP$^-$ cells, Green; Nanog-GFP$^+$ cells.

As previously described, secondary reprogramming from MEF in the presence of both ALKi and VitC greatly increased colony number (Figure 4.1c). In order to investigate if this marker system could respond to changes in reprogramming efficiency mediated by small molecules, cells were reprogrammed in the absence of VitC, but in the presence of dox and ALKi (Figure 4.12). As expected, there was a decrease both in total colony number, and also in the number of Nanog-GFP$^+$ colonies generated in the absence of VitC (Figure 4.12a). FACS

analysis of cells undergoing reprogramming in these conditions revealed a delay in down-regulation of CD44 expression at day 6 compared to the +VitC cultures (Figure 4.12b). At day 8, cells also appeared to be stalled at the 2NG- stage ($CD44^-$/$ICAM\text{-}1^-$/$Nanog\text{-}GFP^-$), with this population seemingly delayed in its ability to increase ICAM-1 expression (Day 10). In addition, the transition to 3NG+ appeared to be somewhat hindered, with an accumulation of cells in the 3NG- stage (Day 12). These results confirmed that this marker system could be used to identify those populations of cells which were specifically affected by the action of small molecules in culture.

4.3 Discussion

4.3.1 Efficient reprogramming from secondary MEF

In order to accurately study the intermediate stages of reprogramming, a highly efficient reprogramming system was required. In the identification of the single integration D6s4B5 clone and subsequent cell line, much of the variability of reprogramming factor expression was addressed as the reprogramming factors were expressed in all cells simultaneously from an identical locus. The presence of insulator and replicator sequences in the reprogramming construct served two purposes; to help maintain high expression of the reprogramming factors during initial primary and subsequent secondary reprogramming, and to prevent *in vivo* differentiation-mediated modifications to the transgene-containing locus which may prevent re-expression of the reprogramming factors (Fu et al., 2006; Gaszner and Felsenfeld, 2006). While the presence of these genetic elements slightly reduced the average number of transgene insertions compared to constructs without (data not shown), it is unclear what effect they had upon transgene expression during secondary reprogramming.

The D6s4B5 reprogramming system displayed a reprogramming efficiency of ~3%, which is comparable to that described for MEF in the original secondary reprogramming system established in Rudolf Jaenisch's lab, (Wernig et al., 2008). This study posited insufficient transgene expression as a potential reason for this lower than expected efficiency, however, as illustrated in Figure 4.1d, many cells which do not successfully complete the reprogramming process appear to express high levels of the reprogramming factors. More likely is the inherent heterogeneity of the starting MEF population which can be observed at the cell surface-marker level (Figure 4.2b, 4.4b), and probably also at the epigenetic level, as outlined in another study from the Jaenisch lab (Meissner et al., 2007). Indeed, many small molecules which modulate histone-modifying enzymes have been reported to

positively influence reprogramming from MEF, given appropriate reprogramming conditions (see section 1.4.2).

While retrovirus silencing has long been associated with pluripotent cells, the down-regulation of reprogramming factors in this study's non-viral *piggyBac* transposon system suggests that this event may be important in ensuring successful reprogramming, and not simply a consequence of virus-mediate delivery of reprogramming factors (Hotta and Ellis, 2008). Supporting this, a recent report identified that continued dox-mediated expression of reprogramming factors at later stages (day 17) of reprogramming actually prevented up-regulation of a number of genes associated with the pluripotency network including endogenous Sox2, Utf1 and Dppa4 (Golipour et al., 2012).

4.3.2 Conventional markers are limited in their use during reprogramming

One of the most important findings from these experiments was the discovery that many commonly used markers of reprogramming were simply unsuitable for discriminating between cells which can successfully complete reprogramming and those which fail. Identification of ICAM-1 and CD44 enabled isolation of defined populations of cells which appeared to display specific reprogramming potentials and could be used to trace the progress of the MEF to iPS cell process.

THY-1 expression was rapidly down-regulated, as previously reported (Stadtfeld et al., 2008a). This confirms that the use of such early markers is not suitable to investigate the latter events of reprogramming such as pluripotency gene up-regulation.

It has previously been reported that **SSEA-1** expression is heterogeneous in ES cells, and sorting experiments have indicated that purified SSEA-1^{+} and SSEA-1^{-} sorted populations recreate pre-sort heterogeneity in an

identical manner within three days of re-plating (Cui et al., 2004). This is similar to my finding that expression of SSEA-1 remains relatively heterogeneous and constant after only 6 days of reprogramming (Figure 4.2d). This data along with the other findings outlined above suggests that SSEA-1 is not a reliable marker with which to isolate cells undergoing reprogramming, as at later stages SSEA-1$^-$ populations are just as likely to be undergoing successful reprogramming (Figure 4.3).

The **E-CADHERIN** expression data generated in this reprogramming system also casts doubt on another "cornerstone" of reprogramming (Figure 4.2c). This marker was found to display rapid kinetics, with almost all cells positive for this marker after 4 days of reprogramming factor expression in this reprogramming system. This finding questions the importance of the MET which has been the focus of much interest since it was first implicated in reprogramming (Li et al., 2010). That cells undergo this transition so readily suggests that it does not actually represent a major barrier to the reprogramming process in this system. Further to this, data from our lab has shown that even in the absence of ALKi, cells undergo MET just as efficiently as their inhibitor-treated counterparts (Tyson Ruetz, in preparation). Further study of this mechanism in the context of iPS cell generation may not actually yield useful information about the molecular requirements of successful reprogramming.

4.3.3 ICAM-1 and CD44 are superior markers for tracing the reprogramming process

Murine **ICAM-1** was first isolated using the rat monoclonal antibody YN1/1 and identified to have high similarity to the human Icam-1 (Horley et al., 1989). A member of the immunoglobulin (Ig) supergene family, it is one of 5 intercellular adhesion molecules identified, so named due to their Ig-like domains expressed on the extracellular portion of the cellular membrane (Hayflick et al., 1998). ICAM-1 is

expressed on vascular endothelial cells, lymphocytes and monocytes, but typically expressed only at low levels (Rothlein et al., 1986). ICAM-1 has been reported to be the ligand for the $\beta 2$ integrin molecules LFA-1 (CD11a/CD18 = $\alpha_1\beta_2$ integrin) and MAC1 (CD11b/CD18 = $\alpha_M\beta_2$ integrin) present on leucocytes (Staunton et al., 1990). Elevated levels of ICAM-1 are induced upon stimulation of cells with inflammatory cytokines (Dustin et al., 1986). This was abolished in ICAM-1 homozygous mutant mice, which, despite displaying retarded lymphocyte migration and decreased inflammatory response, were otherwise viable (Sligh et al., 1993).

ICAM-1 was reported to be expressed at homogeneously high levels in mouse ES cells, however its specific receptors (LFA-1 and MAC1) were found to be absent, suggesting this marker does not play a functional role in these cells, despite maintaining lymphocyte-binding ability, and upon differentiation, a response to IFN-γ stimulation (Tian et al., 1997). Up-regulation of Icam-1 expression has been associated with activation of the LIF signalling target Stat3 in glioblastoma cells treated with ionising radiation, suggesting a potential link between these factors (Kesanakurti et al., 2012) However, this up-regulation was dependent on the presence of activated nuclear factor-κB (NF-κB), and Tian et al. (1997) reported that treatment of ES cells with the NF-κB-activator tumour necrosis factor alpha (TNFα), failed to increase ICAM-1 expression levels. These results indicate that knockdown or blocking of ICAM-1 would more than likely not be detrimental to the reprogramming process, and as shown in Figure 4.5, the initial ICAM-1 expression in the starting MEF population appeared to have little influence over the reprogramming pathway.

CD44 was identified in the early 1980s as a cell-surface molecule on lymphocytes (Gallatin et al., 1983). It was soon linked to cancer-initiating cells (CICs) isolated from many different tissues, and discovered to be a member of the cartilage link protein family (Goldstein et al., 1989; Stamenkovic et al., 1989; Naor et al., 2008). Encoded by a single gene, many variants of CD44 have been described which arise through differences in N-glycosylation and O-Glycosylation patterns

and from various splicing products found in the extracellular domain of the protein (Screaton et al., 1992). Expression of CD44 is dependant of the specific variant, with the most common or standard being CD44s which is expressed on most vertebrate cells (Naor et al., 2008). The isoform found on epithelial cells, CD44v, is also found during early embryonic development and is associated with lymphocyte maturation and activation (Ruiz et al., 1995). Oligomerization of CD44 proteins at the cellular membrane can occur mediated via the transmembrane region, and results in the formation of glycolipid-enriched domains which can play an important role in mediating signal transduction (Oliferenko et al., 1999). Finally, the cytoplasmic tail of the CD44 proteins can interact with cytoskeletal proteins which can influence cell migration and shape (Fehon et al., 2010).

CD44 has been reported to be associated with certain signalling pathways found in both stem cells and transformed cancerous cells. For example, over-expression of CD44 in colorectal adenoma-carcinoma appears to be a direct result of aberrant Wnt signalling associated with the cancerous state (Wielenga et al., 1999). In addition, hyaluronan (HA) binding to CD44 has been described to result in protein kinase Cε (PKCε) activation in breast, head and neck tumour cells, which results in NANOG phosphorylation and translocation to the nucleus (Bourguignon et al., 2009; Bourguignon et al., 2012). However as CD44 is absent in ES cells, this function is probably highly context dependent. Intriguingly there appear to be numerous associations between the expression of CD44 and the epithelial-to-mesenchymal transition (EMT), which is in direct opposition to the required MET during reprogramming (see section 1.3.4 for a fuller discussion of MET and iPS cell generation). In breast cancer cells, over-expression of SLUG, a key protein required for EMT, appears to result in a shift to a $CD44^{+}$ phenotype (Bhat-Nakshatri et al., 2010). The TGF signalling pathway is responsible for driving EMT during development and there is evidence that TGFβ-receptor I can be activated by HA/CD44 interaction which leads to EMT-associated cellular reorganization (Bourguignon et al., 2002; Takahashi et al., 2010). The Takahashi study also utilised

CD44 knockout mice and found that EMT failed to occur. While this suggests that CD44/HA signalling in MEF may actually function to maintain the mesenchymal phenotype and interference with this signalling pathway may positively influence reprogramming, my results clearly showed CD44 expression was maintained after a gain in E-CADHERIN expression (Figures 4.2b, 4.4a). This suggests CD44 has little control over the progression of MET during reprogramming.

The above data suggests the functional activity of CD44 and ICAM-1 is most likely unimportant for iPS cell generation, and bears little impact on interpretation of results obtained using this reprogramming system. Combined staining for ICAM-1 and CD44 expression during the reprogramming process was capable, however, of tracing the movement of cells from one stage of reprogramming to the next. In addition, the ability to monitor, via mOrange expression, only those cells expressing the reprogramming factors greatly increased the accuracy of this study's analysis. Loss of ICAM-1 expression appears to be the first major transition using this marker system. The subsequent down-regulation of CD44 may indicate a significant change in cellular character, as observed in the *in vitro* reprogramming system when loss of its expression appeared more indicative of successful reprogramming compared to SSEA-1 (Chapter 3, Figure 3.5). The subsequent gain of ICAM-1 expression and acquisition of an ES cell-like phenotype provides, for the first time, an accurate profile with which iPS and iPS-like cells can be isolated from MEF undergoing reprogramming.

These markers also highlighted the heterogeneity of MEF cultures undergoing reprogramming. This heterogeneity was actually found to occur within the cells that comprise individual colonies, as illustrated by immunofluorescence staining (Figure 4.13). This further emphasises the requirement of suitable markers during reprogramming in order to isolate and group together cells from different colonies, but at similar stages for accurate analysis of key molecular mechanisms.

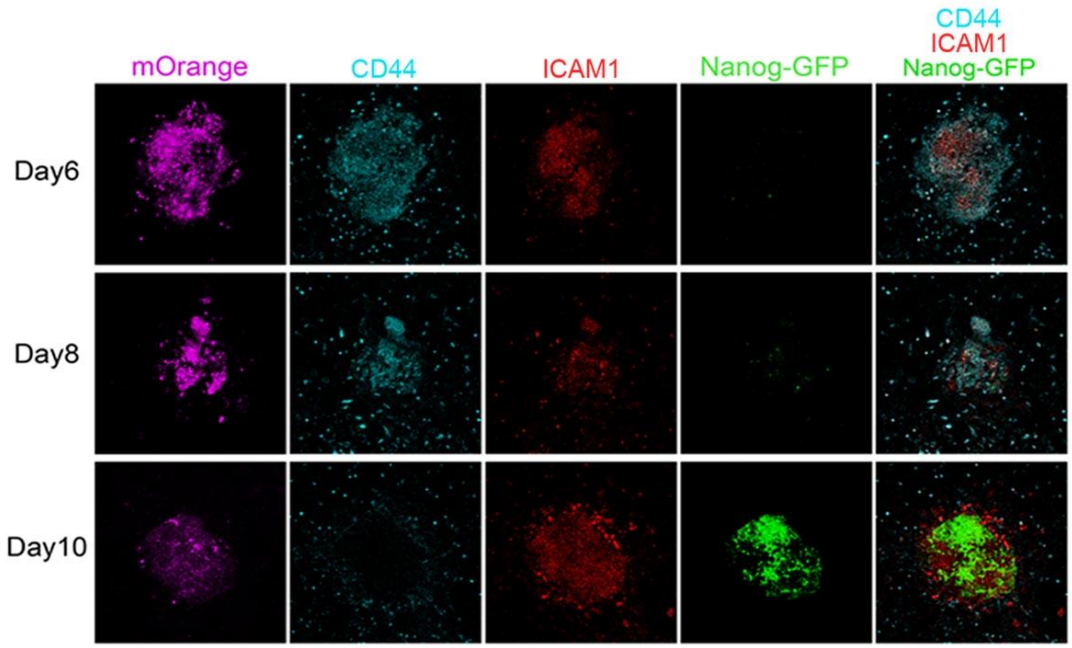


Figure 4.13: Immunofluorescence for CD44 and ICAM-1 at day 6, 8 and 10 after reprogramming initiation. Cells in a single colony have distinct CD44, ICAM-1, Nanog-eGFP expression, indicating clonal analysis is not sufficient to isolate cells in similar stages. Note expression of mOrange tends to be low in colonies with Nanog-eGFP^{+} cells consistent with this study's flow cytometry data. (Image provided by Tyson Ruetz).

4.3.4 *Nanog* has a key influence on reprogramming from MEF

Investigation of the three ICAM-1/CD44 identified populations was further enhanced through the use of the Nanog-GFP reporter. Interestingly, Nanog-GFP expression was observed in all sub-populations, even those not displaying an ES cell-like profile (Figure 4.4c, GFP^{+} panels). It should be noted however, that Nanog-GFP expression appeared to still be closely linked to completion of reprogramming as the overall representation of these cells was correlated with a movement towards ES cell-like ICAM-1/CD44 profile, reflecting the importance of this molecule in establishing pluripotency (Silva et al., 2009). In the past, reporters of *Nanog* expression have been used as strict indicators of completed reprogramming (Maherali et al., 2007; Okita et al., 2007; Wernig et al., 2007). Probably due to the

consistent use of selective agents for *Nanog* in all three cases, these studies reported homogenous *Nanog* expression in iPS cells generated from MEF. This approach does not take into account the possibility that *Nanog* expression may occur asynchronously during reprogramming, and indeed in all cases, delayed administration of the selective agent resulted in an increase in *Nanog* positive iPS cell colonies. As illustrated in Figure 4.13, the D6s4B5 secondary reprogramming system generated colonies that displayed heterogeneous activation of the Nanog-GFP reporter during reprogramming. This suggests that selective approaches may aberrantly target cells which are in fact undergoing successful reprogramming, but have not yet activated the Nanog promoter. This therefore decreases the quality and quantity of data obtained from these studies regarding the reprogramming process. In addition, the use of *egfp* cDNA in the original cell line from which D6s4B5 was derived may have resulted in a more sensitive reporter than the *gfp* used in the previous studies, which could have influenced the selection of lines for expansion/further study etc (Hatano et al., 2005; Maherali et al., 2007).

Isolation of the ICAM-1/CD44 populations and their further subdivision based on Nanog-GFP expression highlighted the importance of using a reporter of pluripotent gene activity, as this marker was capable of identifying distinct sub-populations of cells undergoing reprogramming (Figure 4.6). For each sorted sub-population, the presence of the Nanog-GFP reporter appeared to benefit the reprogramming process. The three day analysis of the sorted populations (Figure 4.6b) revealed two interesting aspects of this reprogramming system; 1) Sorted Nanog-GFP$^-$ cells, despite progressing through ICAM-1/CD44 profile changes, did not readily up-regulate Nanog-GFP, and 2) in contrast to this, a significant population of 1NG+ cells appeared to lose GFP expression but still showed enhanced progression of the reprogramming process compared to 1NG- cells. In contrast, 2NG+ cells did not lose GFP expression, suggesting a possible stabilisation of Nanog-GFP expression, which may also go some way towards explaining the proportional increase of GFP$^+$ cells observed (1NG+<2NG+<3NG+) at the time of

sorting (Figure 4.6a). This may be associated with increased expression of accessory, complementary pluripotency factors, or an epigenetic-mediated modification of the locus enabling enhanced *Nanog* promoter activity as cells progress towards a more reprogrammed cell-state (Buganim et al., 2012; Fidalgo et al., 2012).

The apparent differences in 1NG+/- reprogramming kinetics were also observed to influence their cfp (Figure 4.6). However, in contrast, 2NG+ cells produced a higher cfp than 2NG-, despite the similar rapid gain in ICAM-1 expression displayed by both populations. Interestingly, despite the heterogeneous nature of *Nanog* expression in ES cells and fully reprogrammed iPS cells, 3NG- produced fewer Nanog-GFP colonies than 3NG+ even though both populations displayed an iPS cell-like ICAM-1/CD44 profile. These results highlight the limitations of using these cell-surface markers in isolation, and the importance of integrating the Nanog-GFP reporter into this reprogramming strategy.

4.3.5 Reprogramming from MEF occurs in a step-wise, reproducible pattern

Analysis of the isolated sub-populations after 24 hours provided some of the most clear information about the possible routes available to cells undergoing reprogramming in this system (Figure 4.8). The differences observed between this data and the 72 hour analysis highlight one of the major issues with many reprogramming studies; the expansion of more/fully reprogrammed populations over others present in the culture. It is clear from the cell counting data that the sorted 3NG+ population rapidly expands, even within the two ES cell-cycles theoretically possible in 24 hours. All of the other sorted populations show generation of some 3NG+ cells, and it is the expansion of this population that is detected in the 72 hour analysis.

Closer examination of the subsequent transitions of each of the sorted populations revealed that Nanog-GFP$^-$ populations appeared to prefer a direct transition to the next NG- population, with a low number of cells acquiring GFP

expression after 24 hours. While 3NG- cells were more similar in terms of ICAM-1 expression, the transition to 3NG+ was relatively low. (Note: Nanog-GFP expression has been reported to fluctuate in ES cells, however after 24 hours only 6% of Nanog-GFP$^-$ cells become Nanog-GFP$^+$. This is in contrast to the ~25% of 3NG- cells which transition to 3NG+ after 24 hours in this study. This suggests the 3NG- population does not simply represent the *Nanog* low cells found in LIF/serum ES cell culture, but is in fact a distinct reprogramming intermediate (Chambers et al., 2007)). Many of the NG- populations also showed some propensity to maintain their sorted profile, a feature most pronounced with 1NG- cells. This suggests that some of these populations may act as sinks or reservoirs for cells whose transition to the next stage of reprogramming is prevented by as yet unknown hurdles which must be overcome for efficient iPS cell generation.

In contrast, apart from 3NG+, which appeared to represent a population similar to fully reprogrammed iPS cells, the NG+ populations did not display much, if any, self-renewal. The contrast between the rapid, almost homogenous transition of 2NG+ cells to 3NG+ and the very heterogeneous behaviour of the 1NG+ population, despite its Nanog-GFP expression, is striking. I speculate that at the single cell level, while both 1NG+ and 2NG+ cells have short lives in culture, the actual cellular characteristics of the 2NG+ group is more stably defined. As cells approach a more reprogrammed state, as with 2NG+ cells, it is conceivable that ES cell-associated factors, including pluripotency genes and epigenetic regulators, become increasingly active and regulated, with alternative behaviour, i.e. intermediate characteristics, becoming less favourable (Buganim et al., 2012; Golipour et al., 2012). On the other hand, 1NG+ cells may be the result of fortunate, but rare and temporal activation of reprogramming associated factors, the expression of which can neither be sustained nor is sufficient to establish a stable pluripotency network centred on the activation of the core factors *Nanog, Oct-4* and *Sox2* (Chambers and Tomlinson, 2009). In addition, as previously illustrated (Figure 4.1d), in comparison to cells at the later stages, early stage reprogramming

populations appear to express much higher levels of the reprogramming factors. These transcription factors have numerous targets and their very use in reprogramming is due to their ability to drive cells towards a pluripotent state by aberrantly activating their targets in non-physiological contexts, therefore these factors may also contribute to the premature appearance of the 1NG Nanog-GFP$^+$ population (Soufi et al., 2012).

The observation that each of the sub-populations displayed distinct, reproducible behaviour has enabled the construction of a "route map" from MEF to iPS cells (Figure 4.14). This analysis focused on the major transitions that cells undergo from one stage of the reprogramming process to the next. This visualization of the data attempts to highlight that in the case of this secondary reprogramming system, cells appear to have more than one option in order to achieve successful reprogramming: 1) Cells that fail to up-regulate Nanog-GFP expression at the early stages of reprogramming (1NG-) continue on a relatively low efficiency route with late stage activation of Nanog-GFP (1NG- → 2NG- → 3NG- → 3NG+); and 2) Those few cells which both activate and maintain Nanog-GFP expression early in the reprogramming process (1NG+) appear to use a more efficient route, transitioning to the final reprogrammed state with some ease (1NG+ → 2NG+ → 3NG+).

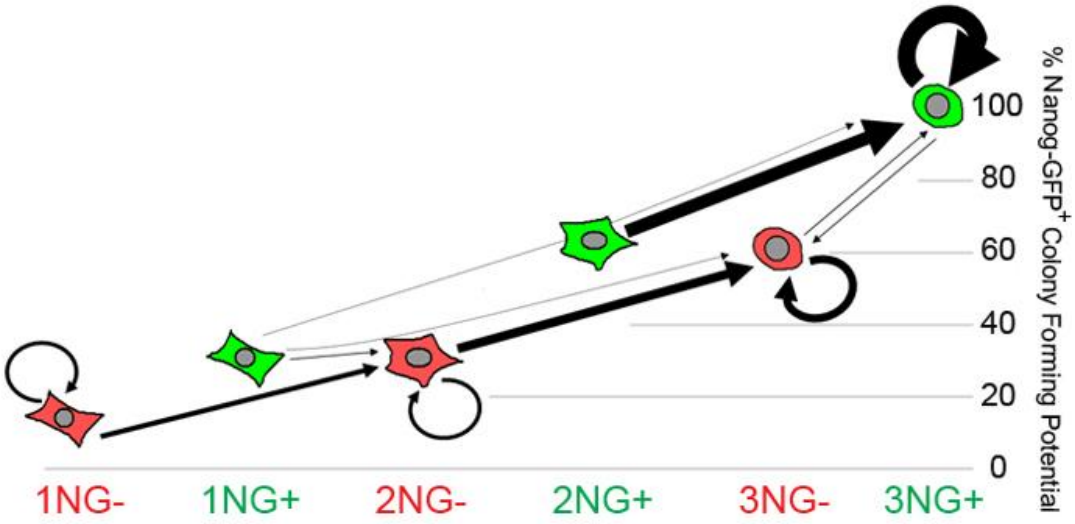


Figure 4.14: Major transitions (>500 cells) of each population within 24 hours at day 10 of secondary reprogramming. Y axis indicates relative colony formation potential after an additional 10 days. Arrow size reflects relative cell numbers.

This data indicates that although reprogramming is a stepwise event with cells passing through distinct stages, there is more than one route via which this process can occur. These findings also highlight the strength of using an accurate marker and reporter system. As shown in Figure 4.11b and 4.12, this system is robust enough for use even in lower efficiency primary reprogramming and in the context of altered reprogramming conditions, as the intermediate stages of reprogramming as identified by ICAM-1, CD44 and Nanog-GFP expression were largely conserved, even if the reprogramming kinetics may be altered. This data also shows that cells within individual colonies undergo reprogramming in a heterogeneous manner, and highlights the importance of identifying these different populations of cells. This enables much more accurate and thorough understanding of how reprogramming truly happens as different culture conditions, reprogramming systems and combinations of reprogramming factors can be compared to each other. For example, it has recently been shown that VitC acts to enhance reprogramming by reducing the levels of H3K9 methylation (Chen et al., 2013). Using this marker system VitC treated and untreated populations could be

isolated, their behaviour monitored, and individual populations analysed via H3K9me3 ChIP-seq to identify the most important sites of methylation loss. This type of approach is currently being applied in our lab to investigate the role of ALKi in reprogramming, and to compare the differences that appear during reprogramming with MKOS and OSKM reprogramming polycistronic cassettes. However, in all cases, identification of the different population dynamics represents merely the first step in understanding how a gene or molecule acts to affect reprogramming. RNA-sequencing analysis of the different stages of reprogramming identified in this chapter is described in Chapter 5.

CHAPTER 5 - High resolution analysis of sub-populations provides novel insights into the molecular mechanisms of reprogramming

5.1 Introduction

Many studies of cells undergoing the reprogramming process utilise next-generation sequencing technologies in order to probe the mechanism of iPS cell generation (Samavarchi-Tehrani et al., 2010; Golipour et al., 2012; Soufi et al., 2012). While these methods can provide information on the general progression of the reprogramming process there are a number of issues with regards the resulting data from these studies. The resolution of the resulting data is low as most strategies utilise bulk cultures of cells, a minority of which will successfully complete reprogramming. In addition, as demonstrated by the data in Chapter 4, even in secondary systems, cells undergoing reprogramming do not behave homogeneously, with different sub-populations and expansion of more-/fully-reprogrammed cells in culture. This means that sampling in bulk gives the average of cells at different stages of reprogramming rather than reflecting the different discrete stages of the reprogramming process.

Reduction in the background "contaminating" signal of cells which fail to become iPS cells is crucial, and this can only be achieved through the use of high-efficiency reprogramming systems. However, these systems have yet to be combined with accurate population analysis due to a lack of suitable markers to identify intermediate stages of the reprogramming process. In addition, many methods require some pre-existing knowledge or expectation regarding the analysis output, for example the selection of primers and probes to detect known factors in single-cell QPCR analysis, which limits discovery of novel mediators of the reprogramming process (Buganim et al., 2012). These technical limitations have

obscured correct analysis and understanding of the reprogramming process and must be addressed for the best use of these technologies.

5.1.1 Aims of this chapter

The aim of this chapter is to outline the results of the RNA sequencing (RNAseq) analysis carried out on the identified reprogramming sub-populations. Analysis of this data identified differentially expressed genes between each sub-population which identified five distinct gene expression patterns occurring during reprogramming from MEF. This analysis revealed that pluripotency gene up-regulation was identified to occur in two waves, which were classified as "Early" and "Late" relative to each-other. In addition, more un-expected patterns of gene expression were identified with transient up- and down-regulation of certain factors found to occur during reprogramming, providing a unique insight into the mechanism of iPS cell generation. Further investigation of transiently up-regulated genes revealed that during reprogramming from MEF, cells express epidermis-associated genes, and this was confirmed by immunofluorescence and single cell QPCR analysis. Finally, another reprogramming system which displayed lower reprogramming efficiency compared to the D6s4B5 cell line was also subjected to RNA-sequencing of the same sub-populations. Comparison of the gene expression profile of this cell line to D6s4B5 was used to identify novel transcription factors which enhanced the efficiency of primary reprogramming.

Note 1: All bioinformatic processing of data, including GeneProf processing of raw data, identification of differentially expressed genes (DEGs), clustering analysis, principal component analysis (PCA), gene ontology (GO) enrichment, and other applications of R were carried out by Dr. Stavroula Skylaki. Interpretation of some of this data was carried out by Dr. Keisuke Kaji.

Note 2: All single-cell QPCR experiments and data processing was carried out by Dr. Kumiko A. Iwabuchi.

5.2 Results

5.2.1 Sample collection for RNA sequencing analysis

In order to more fully investigate the nature of the identified reprogramming sub-populations, RNAseq analysis was carried out in collaboration with the Linnarsson group, which has developed a multiplexed RNA sequencing system as outlined in section 2.3.3 (Islam et al., 2011; Islam et al., 2012). Briefly, day 10 sorted sub-populations were FACS-sorted from three different experiments and RNA isolated. RNA quality was measured and at least two samples from each population were used for analysis. Control samples included the D6s4B5 iPS cell and E14 ES cell lines, as well as both wild-type MEF and D6s4B5 transgenic MEF isolated at E12.5. In addition, it had been observed that while the 3NG+ population at day 10 had a similar colony forming potential (cfp) as fully reprogrammed iPS cells, this was only in the continued culture of dox-containing media. 3NG+ cells isolated at day 10 of reprogramming were cultured at clonal density in the absence of dox, and it was observed that the cfp of this population was roughly 60% that of D6s4B5 iPS cells (Figure 5.1, 3NG+). However, when 3NG+ cells isolated at day 15 of reprogramming were cultured in the absence of dox, the cfp was now found to be similar to D6s4B5 iPS cells (Figure 5.1, 3NG+ D15). This indicated that a further five days of culture in reprogramming conditions was sufficient to enable all 3NG+ cells to reach a dox-independent state. This population was also included for analysis as comparison of this population to 3NG+ at day 10 could yield important information about the stabilisation of the exogenous-factor-independent reprogrammed state. However, it may be the case that this event does not represent a major barrier for successful reprogramming in this system, as all 3NG+ cells appeared to undergo this transition between day 10 and day 15.

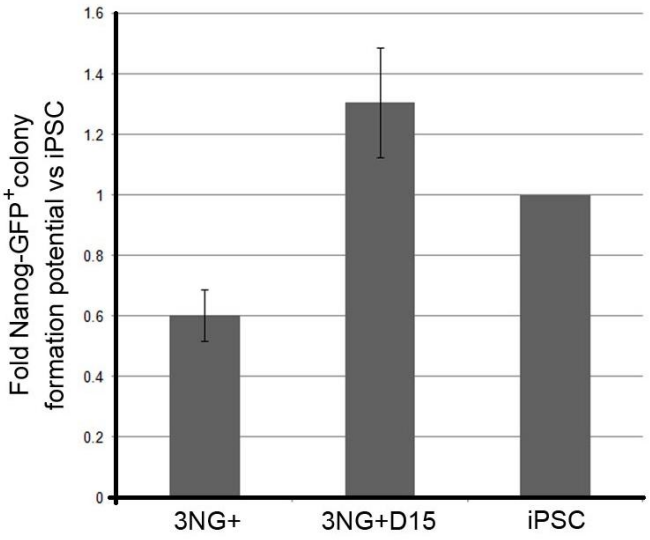


Figure 5.1: Nanog-GFP Colony formation potential of 3NG+ cells in the absence of dox. 3NG+ cells sorted at day 10 (3NG+) and day 15 (3NG+D15) of reprogramming were plated at clonal density in the absence of dox. The number of Nanog-GFP$^+$ colonies was counted after 10 days. Error bars represent standard deviation. n=3.

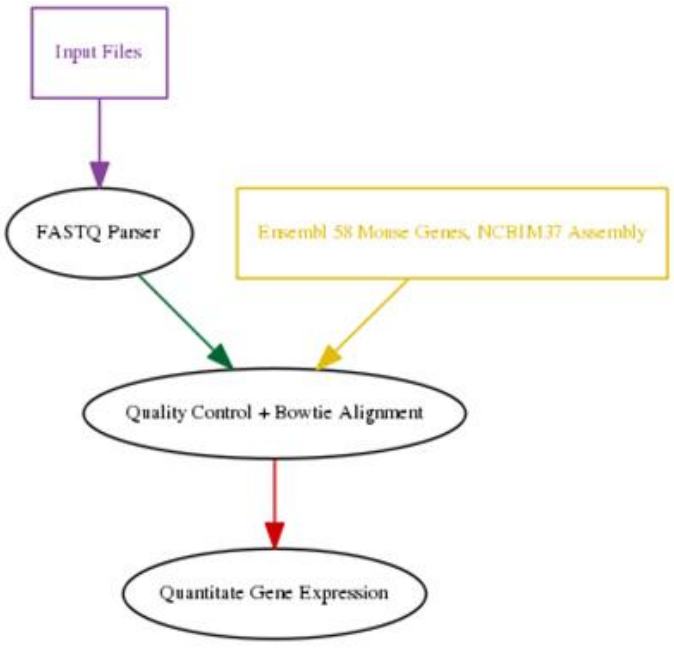


Figure 5.2: Workflow of GeneProf analysis. Reproduced with permission from www.GeneProf.net

5.2.2 Identification of differentially expressed genes

The raw data obtained from the RNAseq analysis was processed using GeneProf which allowed quality control and alignment of the obtained reads to the genome (Figure 5.2, see section 2.3.4 for methods used) (Halbritter et al., 2012).

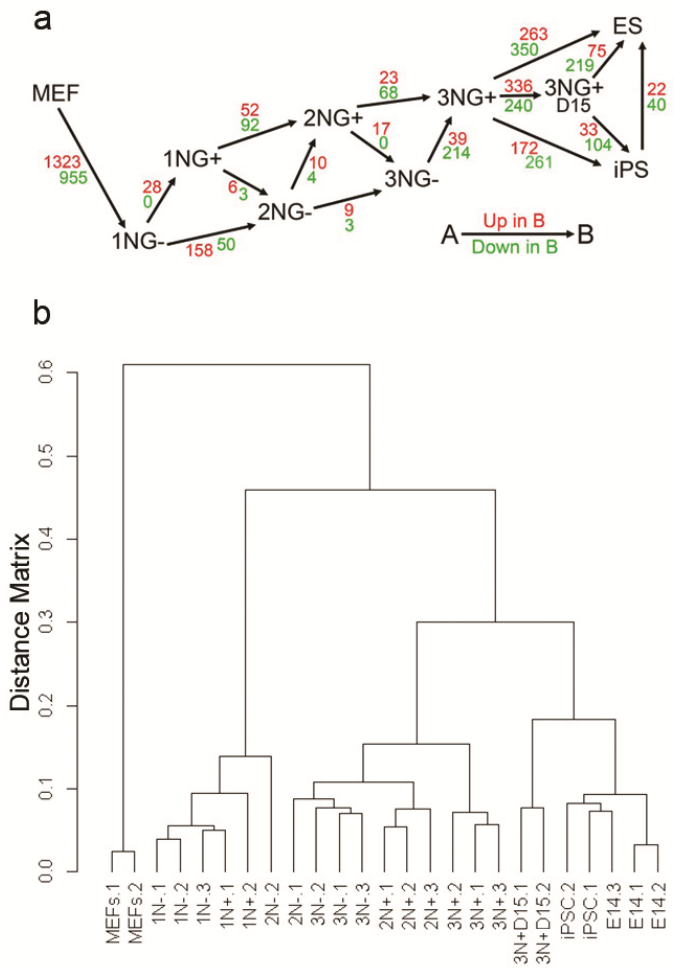


Figure 5.3: Hierarchical clustering of sub-populations based on DEG expression. a. Number of DEGs between samples identified via both edgeR and DESeq are indicated with arrows as shown. **b.** Agglomerative hierarchical clustering based on Pearson's Correlation Coefficient with complete linkage. P-values were adjusted using a threshold for false discovery rate (FDR) ≤ 0.05. Green and red colours represent up- or down-regulated genes identified by both edgeR and DESeq.

The total number of reads per sample, per gene, was quantified, and this allowed identification of the differentially expressed genes (DEGs, FDR ≤ 0.05) between each of the sub-populations (Figure 5.3a). This list of genes was then used to carry out hierarchical clustering of the sub-populations and revealed four major branches: [MEF], [1NG-/+], [2NG-/+ and 3NG-/+], and [3NG+D15, iPSCs and ESCs] (Figure 5.3b).

This data highlighted the relatively high reproducibility among the three experiments used to collect the RNA for this analysis. In addition, 3NG+D15 cells were identified to be more similar to iPS and ES cells than to day 10 3NG+ cells (Figure 5.4a). PCA analysis of the DEGs identified a distinct difference between 2NG+ and 3NG- sub-populations (Figure 5.4b).

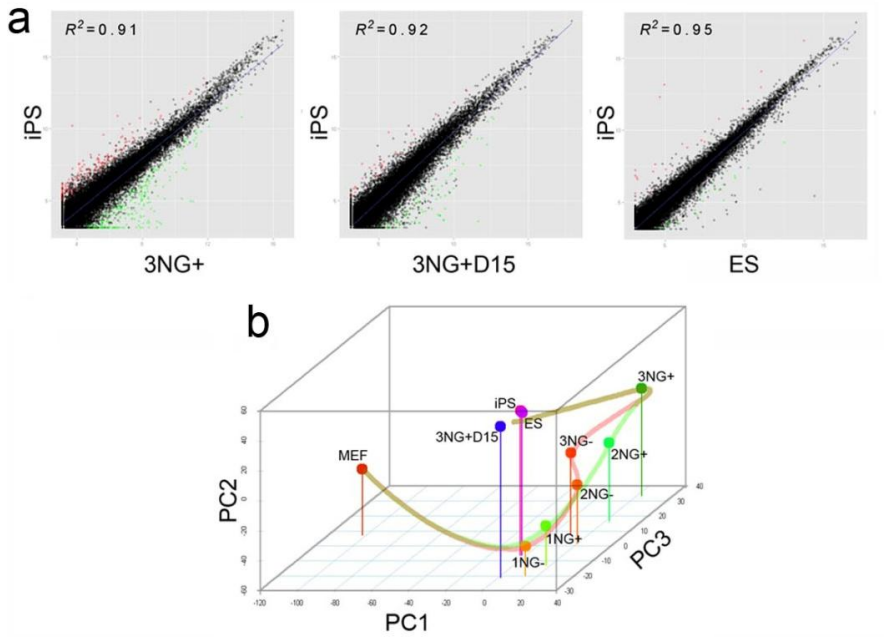


Figure 5.4: Analysis of reprogramming intermediates and fully reprogrammed iPS cells. a. Comparison of gene expression profiles between iPSCs and 3NG+, 3NG+D15 and ES cells. **b.** PCA analysis of sub-populations using DEGs. Green and Red lines indicate pathways identified from FACS analysis of sorted sub-populations.

This appeared to reflect the two previously identified routes to the fully reprogrammed state, as observed from FACS analysis of the sub-populations, reflecting their different probabilities to successfully transition to the 3NG+ state (Figure 5.4b, red and green paths).

5.2.3 DEG expression patterns reveal processes of reprogramming

The DEGs could be further categorized based on their expression pattern across the clustered sub-populations (Figure 5.5). This analysis revealed 5 major gene expression patterns among the DEGs during reprogramming (A-E). Some of the expression patterns were expected, such as **Group A** which indicated genes expressed in MEF but down-regulated in reprogramming intermediates and pluripotent cells, **Group D** which showed late-reprogramming-stage up-regulation of genes expressed in pluripotent cells but not in MEF, and **Group C** contained genes up-regulated early in, and maintained throughout, the reprogramming process and in iPS/ES cells. More unexpected was the identification of genes expressed at a similar level in both MEF and iPS/ES cells, but which showed transient up- (**Group B**) or down-regulation (**Group E**) exclusively in the reprogramming intermediate populations.

Gene ontology (GO) analysis was carried out using DAVID (see Supplementary Table 3 in Appendix)(Huang da et al., 2009). **Group A** was enriched for genes associated with diverse functions, but especially enriched for cell adhesion, extracellular matrix organisation and signal transduction-associated factors, as previously reported from microarray analysis of MEF (Tanaka et al., 2002). **Group D** contained genes associated with pluripotency. Interestingly, pluripotency genes were also found in **Group C**, suggesting that there may be a number of genes important for establishing iPS cells that are more easily re-activated than others (Buganim et al., 2012; Golipour et al., 2012).

Further investigation of this data was carried out via comparison to other published datasets. A recent report identified genes down-regulated in MEF very rapidly (within 3 cell-cycles) after reprogramming factor expression (Koche et al., 2011). In my data, a number of these genes were similarly down-regulated in all day 10 sub-populations and ES/iPS cells compared to MEF, indicating that Group A most likely contained genes down-regulated at the early stages of reprogramming (Figure 5.6a). Similarly, the expression pattern of 22 pluripotency-related genes in my data set was examined (Kim et al., 2008a; Xu et al., 2010). Including endogenous *Oct4* (*Pou5f1*), 8 pluripotency genes were already up-regulated at the 1NG+/2NG- stages to the level found in 3NG+ cells (Figure 5.6b, Group C), while 14 pluripotency genes were more gradually up-regulated in the later stage reprogramming populations (Figure 5.6b, Group D).

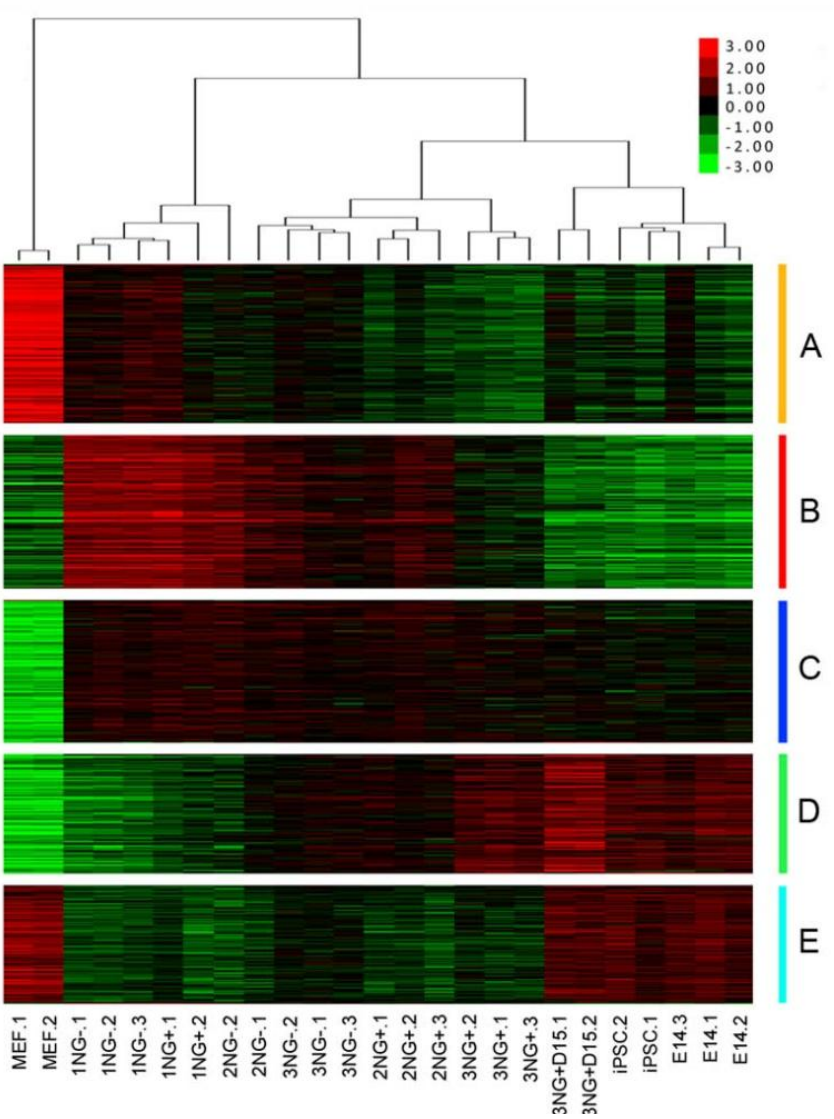


Figure 5.5: Global gene expression changes during stage transitions. Hierarchical clustering of samples with DEGs and expression heat map. Groups A-E represent different expression patterns. Key indicates log fold change.

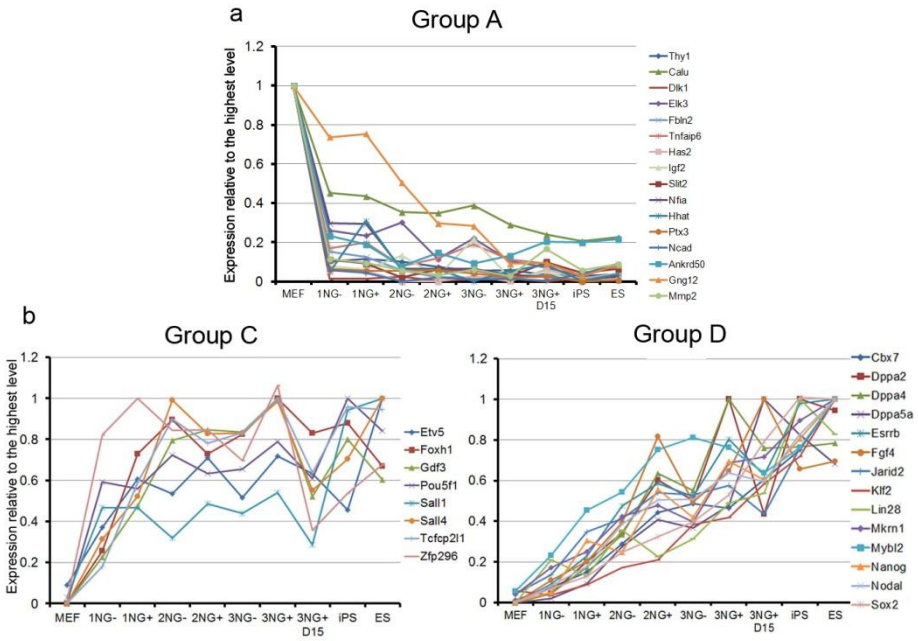


Figure 5.6: Gene expression changes during reprogramming. a. Down-regulation of genes expressed in reprogramming factor-un-induced MEF populations. **b.** Early and late up-regulation of pluripotency-related genes.

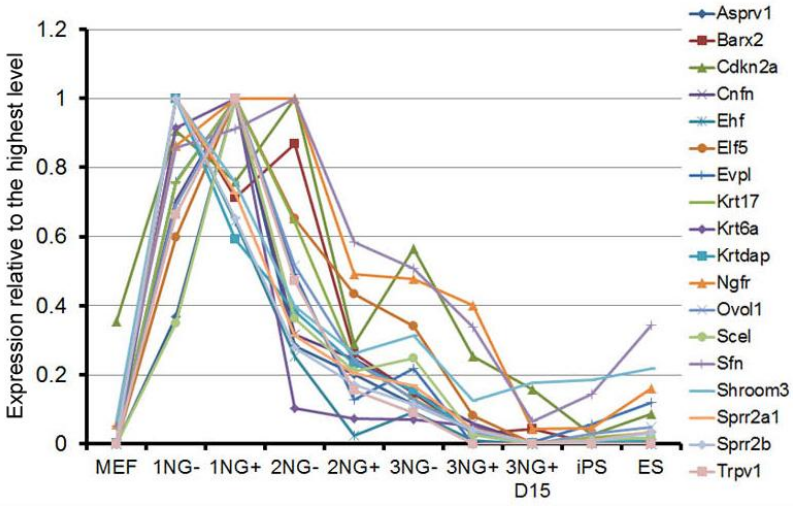


Figure 5.7: Epidermis gene expression during secondary reprogramming. Transient up-regulation of epidermis and keratinocyte genes is rapidly lost in 2NG+ and 3NG+/- populations.

5.2.4 Epidermis genes are up-regulated during reprogramming

While a loss of fibroblast character and gain of pluripotent cell properties is consistent with previous findings (Brambrink et al., 2008; Stadtfeld et al., 2008a; Li et al., 2010; Samavarchi-Tehrani et al., 2010), the identification of **Groups B** and **E**, with their transient expression pattern found exclusively in the intermediate stages of reprogramming, warranted further investigation, as this finding suggested that reprogramming from MEFs to iPSCs was not simply the loss of MEF genes and gain of ES cell genes.

GO analysis of **Group E** produced a small number of terms, with heart and tube development comprising the majority of hits. The significance of the transient down-regulation of these developmental processes is unclear, however this finding suggested that the transition from MEF to iPS cells may involve unexpected factors and processes.

Genes related to ectoderm/epidermis development and keratinocyte differentiation were highly enriched ($P \leq 0.000274$) in **Group B**. Down-regulation of those epidermis genes appeared to coincide with the second, later wave of pluripotency gene up-regulation, suggesting MEFs expressed multiple epidermis-related genes before reaching a pluripotent state (Figure 5.7). In addition, analysis of expressed sequence tag (EST) data of different developmental stages and tissues in the mouse failed to identify co-expression of these epidermis genes in any one single tissue, and significantly nor at any stage of embryonic development (Supplementary Table 5, Appendix). This suggested that during reprogramming cells do not simply reverse developmental processes.

The RNAseq analysis was carried out on populations of cells which displayed heterogeneous behaviour (Figure 4.7, 4.8). Therefore the observed transient up-regulation of epidermis associated genes may be the result of a minor population expressing high levels of these genes. Epidermis gene expression was identified by single-cell qPCR for Ehf and Ovol1, detected in ~65% and 70% of 1NG-

cells (Figure 5.8d). This analysis also confirmed the co-expression of epidermis genes with the earlier up-regulated pluripotency genes, but not the later set.

Single cell qPCR is limited by the number of cells used for analysis, and in order to investigate how the RNAseq analysis correlated with protein expression, immunofluorescence staining was carried out for two epidermis markers, SFN and KRT17 (Figure 5.8). These proteins were barely detectable in MEFs and iPSCs, while transient up-regulation was observed in the intermediate stages of reprogramming (Figure 5.8a,b). By day 9 of dox treatment both SFN and KRT17 were expressed throughout colonies undergoing reprogramming. SFN protein appeared to be down-regulated before the onset of Nanog-GFP expression. KRT17 protein was detectable in some Nanog-GFP^{+} cells, however KRT17 staining was highest in mOrange^{high}, Nanog-GFP^{-} cells (Figure 5.8c). In summary this data, in addition to the RNAseq single cell QPCR analysis, indicated that during reprogramming from MEF most cells express epidermis-associated genes. This expression pattern was greatest at the earlier stages of reprogramming, during an initial wave of pluripotency gene expression, but was down-regulated concurrent with a second wave of pluripotency gene up-regulation.

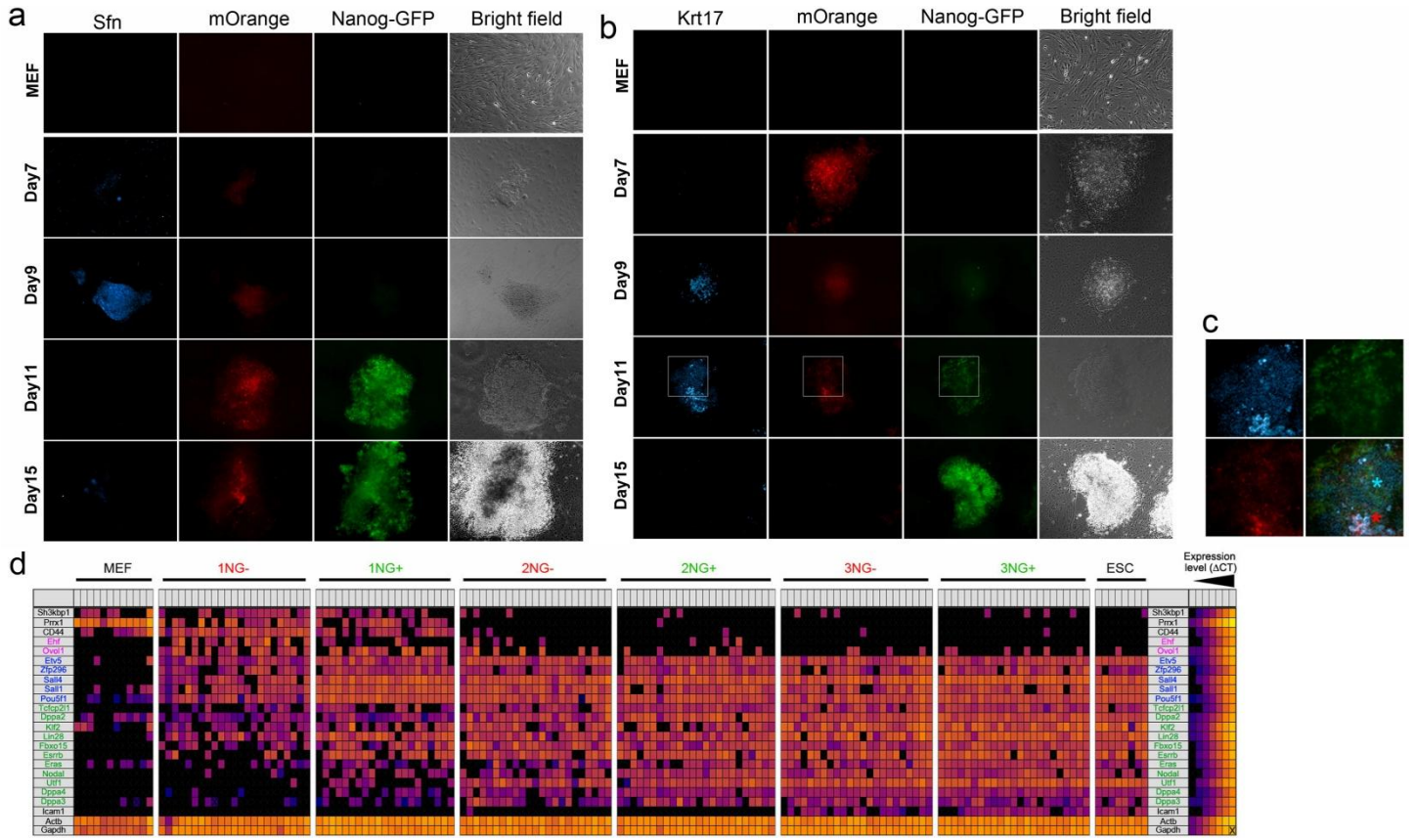


Figure 5.8: Transient up-regulation of epidermis genes during reprogramming. Staining was carried out on MEF undergoing secondary reprogramming for both SFN (**a**) and KRT17 (**b**). **c.** Inset of day 11 staining for KRT17. Blue asterisk = Nanog-GFP^{+}/KRT17^{low}, Red asterisk = mOrange^{high}/KRT17^{high}. **d.** Single-cell gene expression analysis. Each square represents one reaction chamber from one cell. Colour corresponds to ΔCT value as shown in legend at right.

5.2.5 Identification of candidate genes to improve reprogramming efficiency

Splinkerette PCR was used to identify the site containing the reprogramming vector in D6s4B5 iPS cells (data not shown, carried out by Laurence Lemier). The vector was found to be integrated into the *Sp3* locus, and the suitability of this locus for efficient reprogramming was investigated. $Nanog^{+/GFP}$, $Rosa^{+/rtTA}$ (RTANG) ES cells were targeted with PB-TAP IRI 2LMKOSimO at the Sp3 locus to produce $Sp3^{+/PB\text{-}TAP\ IRI\ 2LMKOSimO}$ ES cells (TANGO ES cells, all work carried out by Dr. Eleni Chantzoura and Dr. Keisuke Kaji). This ES cell line was used to generate chimera from which MEF were isolated for secondary reprogramming (TANGO MEF).

Despite expressing the reprogramming factors from the same locus as D6s4B5 MEF, TANGO cells appeared to be somewhat hindered in undergoing the reprogramming process (Figure 5.9a). While 3NG+ cells could eventually be detected, this appeared to be due to the expansion of a small number of cells, with the majority failing to undergo reprogramming. Cells were sorted in each of the six NG populations at day 10 of reprogramming (Figure 5.9b). The cfp of each of the sorted TANGO populations was almost 50% lower than that of the corresponding D6s4B5 populations (Figure 5.9c). Surprisingly, the 3NG+ TANGO population had a similar cfp as D6s4B5 cells. This suggested that although reprogramming was hindered at earlier stages, the few cells that successfully completed this process were *bone fide* iPS cells.

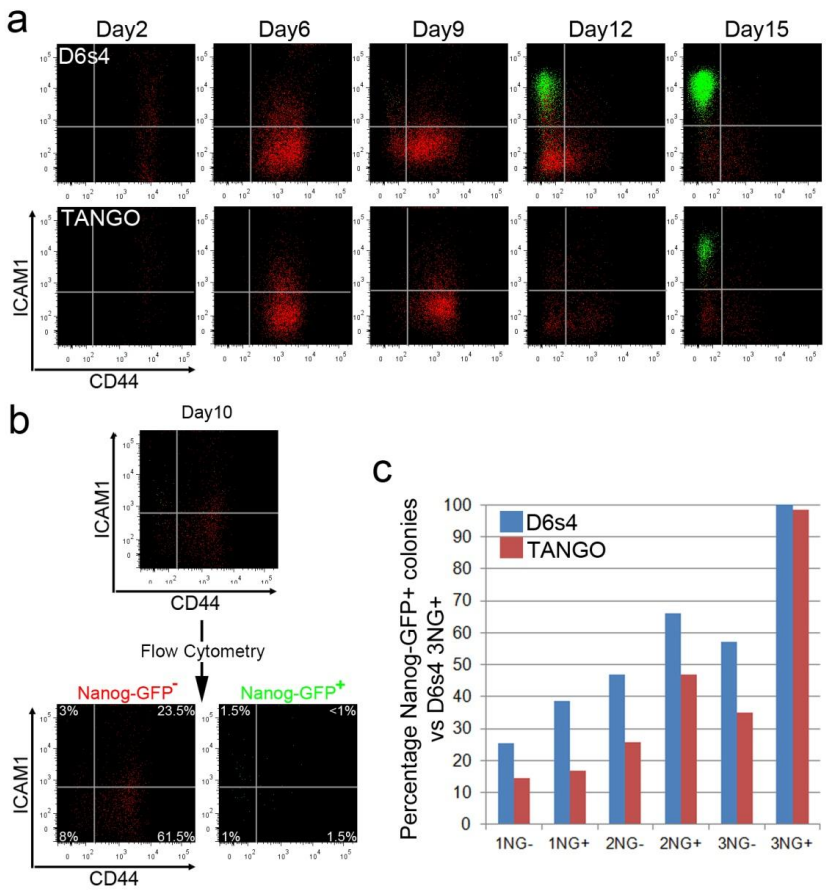


Figure 5.9: TANGO undergoes reprogramming less efficiently than D6s4B5. a. FACS analysis of secondary MEF undergoing reprogramming. **b.** TANGO MEF sorting at day 10 of reprogramming. **c.** Nanog-GFP cfp of sorted TANGO populations compared to D6s4B5 cells from the same time-point.

In order to identify factors which may increase the efficiency of reprogramming, a number of screening steps were carried out to compare the RNAseq analysis of D6s4B5 and TANGO sub-populations. Sufficient numbers of TANGO 3NG+ cells could not be isolated at day 10 and these were not included in the analysis. Briefly, DNA binding factors were identified from the D6s4B5 dataset using the MGI GO database (Hill et al., 2001), within this group genes identified to increase expression proportionally to their cfp were filtered, and those whose expression differed greatly in TANGO were selected for further study (Figure 5.10a). Genes not previously described in reprogramming from somatic cells were

identified using these criteria (Figure 5.11). Five genes from this analysis were cloned from cDNA isolated from E14 ES cells using primers which introduced SalI and BstZ17I sites. After a TOPO vector intermediate stage, these sites were used to ligate the cDNA of interest into a site in a pENTR vector upstream of an ires-linked human *CD2* (h*CD2*) gene which was to be used as a reporter for future isolation of novel cDNA-containing cells if desired. Gateway cloning was then used to introduce the cDNA sequence to the dox-inducible *piggyBac* vector pB TAP IRI to generate PB TAP IRI cDNA-ires-h*CD2*. Restriction sites introduced were selected based on their low frequency of occurrence in the mouse genome (www.tools.neb.com), and alternative cloning sites BstBI and AgeI could also be used to subsequently introduce newly cloned cDNA directly into the pB TAP IRI vector after the introduction of the *CD2* reporter to the vector (Figure 5.10b). Each over-expression vector was introduced alongside the PB-TAP IRI 2LMKOSimO construct in primary reprogramming of $Nanog^{+/GFP}$, $Rosa^{+/rtTA}$ (RTANG) MEF, and total Nanog-GFP^{+} colonies were quantified on day 15 of reprogramming (Figure 5.12). In order to maximise potential differences in reprogramming efficiency induced by the novel factors, cells were reprogrammed in the absence of ALKi which increased the efficiency of reprogramming from MEF in the secondary system (Figure 4.1c). Eras, one of the top hits in the comparative analysis between D6s4B5 and TANGO cell lines, has been reported to increase the efficiency of reprogramming, and was included as a positive control(Polo et al., 2012). However, this factor was not found to enhance reprogramming compared to the empty vector control. Of the tested genes, the helicase component Mcm3 and primordial germ cell specification-factor Prdm14 displayed a five- and six-fold increase in Nanog-GFP^{+} colony number respectively, compared to the empty PB TAP IRI control (Figure 5.12).

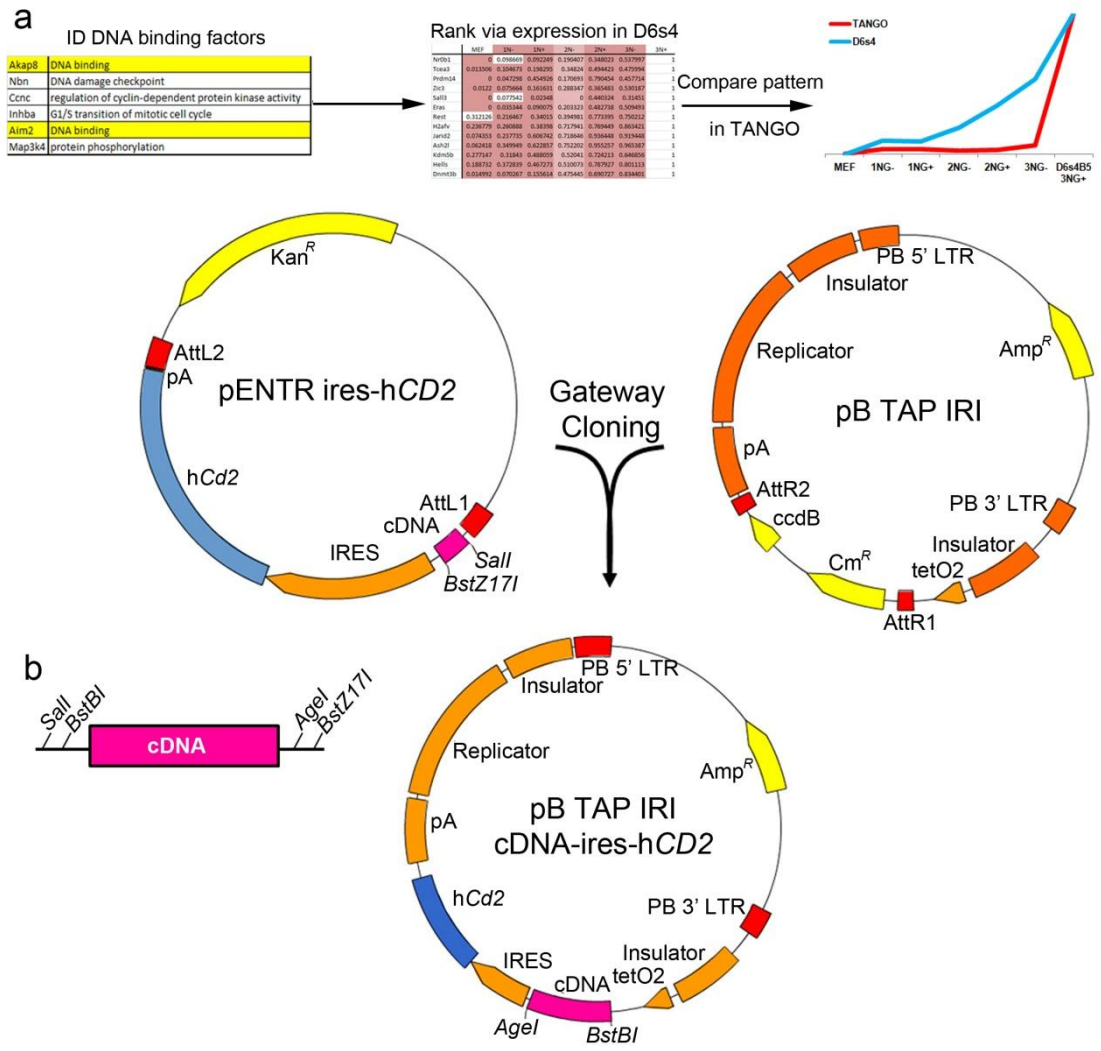


Figure 5.10: Identification and over-expression of novel reprogramming factors. a. Overview of method to identify novel factors differentially expressed between D6s4B5 and TANGO. **b.** Cloning strategy used to introduce cDNA of interest into pB TAP IRI vector for over-expression.

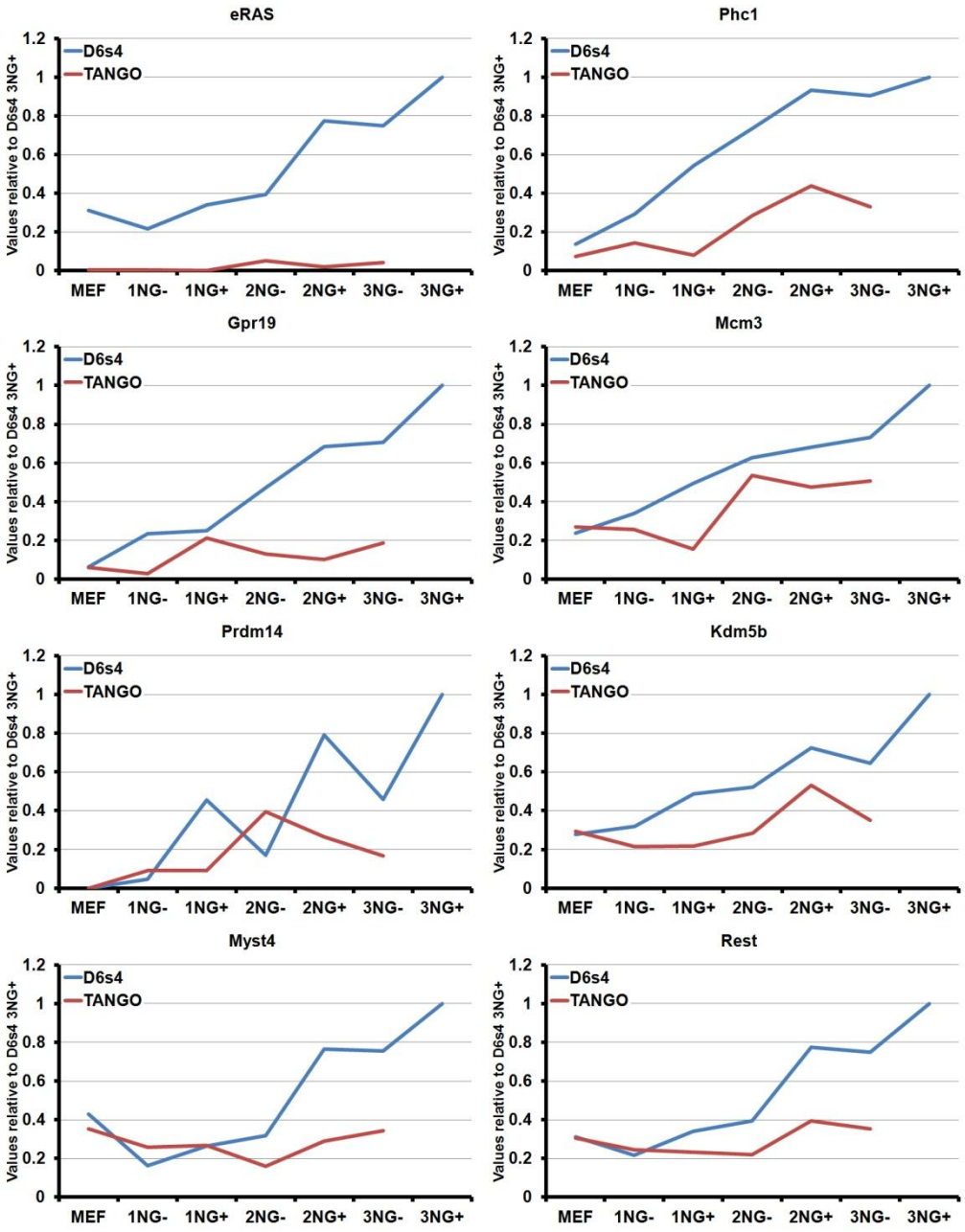


Figure 5.11: RNAseq analysis comparison of TANGO and D6s4B5 subpopulations. Total reads were normalised to D6s4B5 3NG+.

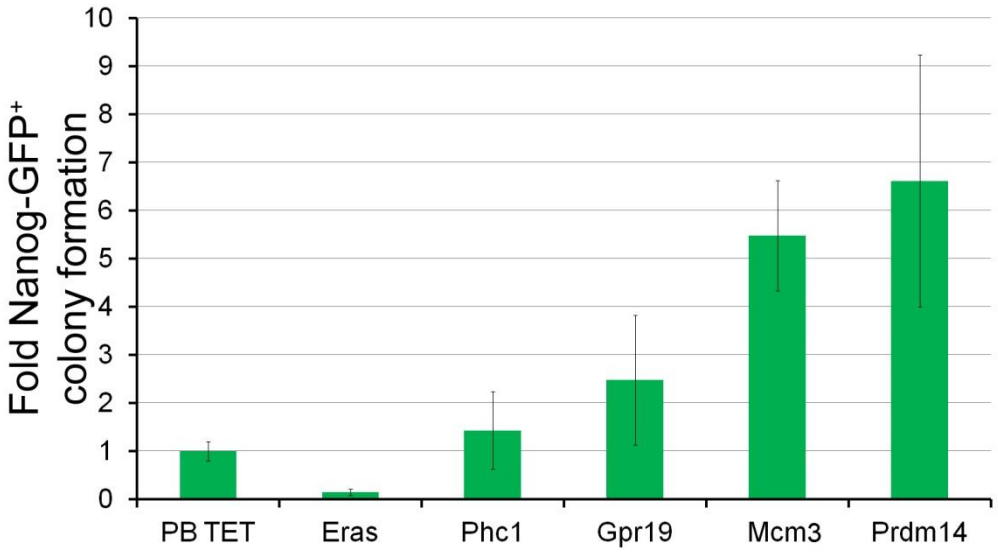


Figure 5.12: Over-expression of novel factors increases the colony formation potential in primary reprogramming. Values are fold increase over PB TET empty vector. Error bars represent standard deviation (*n* = 2).

This data indicated that the reprogramming system identified using ICAM-1 and CD44 markers and the Nanog-GFP reporter was suitable for not only identification of novel routes to reprogramming, but also factors which greatly increase the efficiency of iPS cell generation.

5.3 Discussion

5.3.1 RNAseq analysis reveals distinct properties of reprogramming populations

The functional differences evident between the identified reprogramming intermediate populations appeared to be reflected at the transcriptional level. Importantly, all the populations analysed were from the same time-point, and the data generated using this approach enables accurate transcriptional analysis of the reprogramming process independent of expansion of different reprogramming intermediates, and holds the potential to identify genes crucial for the success of reprogramming. This is unique among other marker-based approaches which have simply compared one reprogramming time-point against another (Polo et al., 2012).

This system also enables investigation of the stabilisation of the reprogrammed state via comparison of the gene expression differences of the 3NG+ sub-population which is transgene-dependent at day 10 and independent at day 15 (Figure 5.13). A number of the factors up-regulated at day 15 are expressed during embryonic development and in the inner cell mass, including the glycosphingolipid-associated enzyme β1,3gnt5 (B3gnt5), which has been shown to be essential for blastocyst formation (Biellmann et al., 2008). The functional relevance of these factors remains to be examined, but remains a source of further information about the establishment and maintenance of the reprogrammed state.

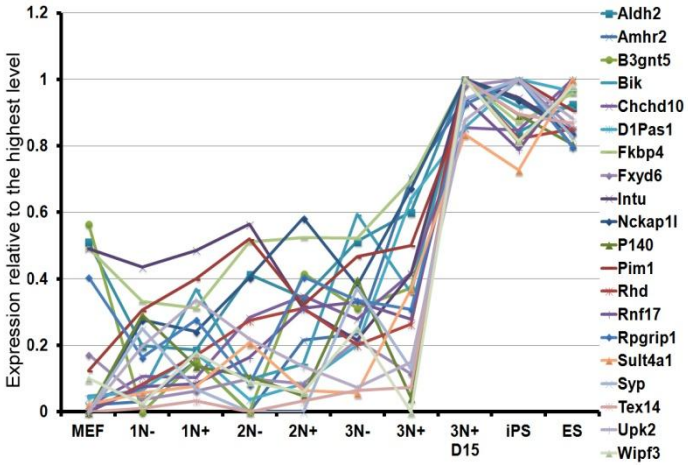


Figure 5.13: Genes up-regulated between day 10 and day 15 in 3NG+ cells. A number of factors are activated in this population to a level similar to that in fully reprogrammed iPS cells and ES cells.

Other differences between the reprogramming intermediates were revealed through the DEG analysis. This highlighted the series of transcriptional changes undertaken at each stage of the reprogramming process, and unsupervised clustering of the populations based on the total DEGs identified during reprogramming largely reflected the shifts in population behaviour, indicating a major phase of reprogramming requires the transition from the 1NG+/- and 2NG- populations, and subsequently the apparent differences in 2NG+ and 3NG- populations between each-other and the 3NG+ population. Another important transition appears to be the acquisition of a reprogramming factor-free, stably reprogrammed state, with day 15 3NG+ cells clustering with fully reprogrammed iPS cells and ES cells. These transitions could be observed at a high resolution, with the PCA analysis supporting the idea that cells have the potential to utilise different pathways to reach an iPS cell-state during reprogramming. This highlights the requirement for accurate identification of different sub-populations which arise, apparently even during the later stages of reprogramming; identification of pathways and processes which result in more efficient generation of iPS cells could potentially not only improve reprogramming technology, but also lend insights into the plasticity of cellular behaviour and how this is modulated at different stages of reprogramming thus resulting in the heterogeneity of this process.

5.3.2 Pluripotency gene up-regulation is a two-stage process during reprogramming

The unexpected identification of early- and late-stage up-regulation of pluripotency genes further supports the idea that reprogramming occurs, at least to some extent, in a step-by-step, predictable process, and requires appropriate conditions for successful progression. The presence of *Pou5f1* among the group of "early" genes is striking, considering the key role of OCT4 in ES cell maintenance (Niwa et al., 2000). Similarly, Sall4, a regulator of Pou5f1, was also identified as being up-regulated to an iPS/ES cell-level in all populations (Zhang et al., 2006). The

later set of genes included many factors with roles are important for successful reprogramming including Klf2 and Esrrb (Jiang et al., 2008; Festuccia et al., 2012).

A recent study also reported identification of two distinct groups of pluripotency-associated genes during reprogramming (Golipour et al., 2012). Factors were separately classified based on their contribution towards either the "maturation" or stabilisation of cells, based on changes in their expression upon exogenous reprogramming factor down-regulation during reprogramming. While the authors carried out a siRNA screen in ES cells and cells undergoing reprogramming to validate these classifications, the data was somewhat dubious as neither validation of gene knock-down nor rescue of knock-down phenotype were demonstrated, and therefore it is unclear if their separation of pluripotency factors is accurate. Another study also identified a two-phase pattern of pluripotency gene expression, and concluded that up-regulation of "early" wave genes, including Oct4 and Sall4 does not identify cells which will undergo successful reprogramming, however the "late" wave of genes can be markers of the fully reprogrammed state (Buganim et al., 2012). These findings have been extended by the RNAseq analysis carried out in my study. Buganim et al. (2012) puts forward a model where re-activation of endogenous SOX2 activity is important during reprogramming, with SOX2-activated factors in turn activating their targets, and thus leading to the establishment of a stabilised pluripotency network. The authors claim that therefore the latter stages of reprogramming occur in a hierarchical, ordered manner, however, this is described as occurring after a more stochastic phase in which chance activation of factors enables progression of reprogramming. However, this study was limited by the use of single cell qPCR analysis which is dependent on the selection of known, or expected factors, and it may be the case that the earlier "stochastic" phase also occurs in a similarly ordered manner as the later "hierarchical" phase, but that at present our understanding of the events of early reprogramming are poorly understood. Gene expression at earlier stages of reprogramming was examined in a study which sampled THY-1 and SSEA-1 sorted

cells at early time-points (Polo et al., 2012). This study also reported earlier up-regulation of *Sall4* and *Pou5f1* and later activation of *Sox2*, suggesting a conserved phenomenon, at least in MEF undergoing reprogramming, however as previously discussed, THY-1 and SSEA-1 are not highly accurate markers with which to identify populations of cells undergoing successful reprogramming (see Section 4.3.2).

As outlined previously (see Section 4.3.4), there appeared to be stabilisation of Nanog-GFP expression as cells progressed from one stage of reprogramming to the next (1NG→2NG/3NG), which correlated with an increase in cfp among the Nanog-GFP^{+} sorted populations themselves. I speculated that this increased stability of Nanog promoter activity in 2NG+ and 3NG+ is due to the presence of a pluripotency transcription factor network, or epigenetic changes which enabled greater access to this locus. The RNAseq analysis revealed *Nanog* was also expressed among the Nanog-GFP^{-} sorted populations, being detected at higher levels than the preceding NG- sub-population, but at lower levels than their NG+ equivalent (1NG-<2NG-≈1NG+<3NG-<2NG+<3NG+). Notably, *Nanog* was detected among the second wave of pluripotency genes. Importantly, a recent report claimed *Nanog* mRNA is expressed in a mono-allelic fashion in ES cells cultured in serum and LIF, with bi-allelic expression only occurring when cells were cultured in their naïve state in 2i conditions or on feeders (Miyanari and Torres-Padilla, 2012). While this could have explained the inconsistency observed between Nanog-GFP and *Nanog* transcription, a report recently demonstrated detection of NANOG protein as a result of transcription from both alleles in ES cells, and suggested poor reporter choice, genetic background or transcriptional bursting as reasons for the original findings (Faddah et al., 2013; Filipczyk et al., 2013). However, a further study also identified a few, rare cells which expressed *Nanog* in a mono-allelic manner in both serum/LIF and 2i conditions, and ruled out transcriptional bursting as an explanation for this observation (Hansen and van Oudenaarden, 2013). Why 3NG- cells, which express higher levels of *Nanog* than 1NG+ do not activate expression of the reporter is

unclear, however this discrepancy does not lessen the importance of the functional and transcriptional characteristics of the sub-populations identified by Nanog-GFP expression.

5.3.3 Epidermis gene expression occurs during reprogramming from MEF

The most striking finding from the RNAseq analysis of the intermediate populations was the identification of the transient gene expression patterns, in particular the transiently up-regulated group of epidermis-associated genes (Group B). Expression of epidermis genes in partially reprogrammed cells from B-cells had previously been noted, and was thought to be a result of the off-target effects of Sox2 and Klf4 activity (Mikkelsen et al., 2008). Consistent with my data, analysis of three published microarray data sets incorporating partially reprogrammed iPS cells (piPSCs), a time course experiment and a sub-population analysis confirmed transient epidermal gene expression during reprogramming (Sridharan et al., 2009; Samavarchi-Tehrani et al., 2010; Polo et al., 2012) (Figure 5.14). This data indicates this pattern of gene expression is conserved during reprogramming, indicating cells undergoing reprogramming experience apparently more complex gene expression changes than the simple down-regulation of MEF genes followed by up-regulation of ES cell genes. Notably, it has also been shown that pre-iPS cells derived from different starting materials and by different laboratories have similar characteristics, and the observed maintenance of epidermis gene expression in the study by Sridharan et al. (2009) confirms these genes are found in other pre-iPS cell lines, suggesting this may be a conserved feature of reprogramming from other cell-types, not just MEF (Polo et al., 2012). The timing of up-regulation and down-regulation of each individual epidermis gene is not synchronized in the time course and THY1/SSEA1/Oct4-GFP datasets (Figure 5.14). Interestingly, the RNAseq data generated from the ICAM-1/CD44/Nanog-GFP sub-population analysis indicated that this wave of transient gene expression is down-regulated at a point when the later phase of pluripotent gene expression is activated, which is only apparent in

sufficiently homogeneous systems, and therefore suggests why it may have been overlooked in other studies.

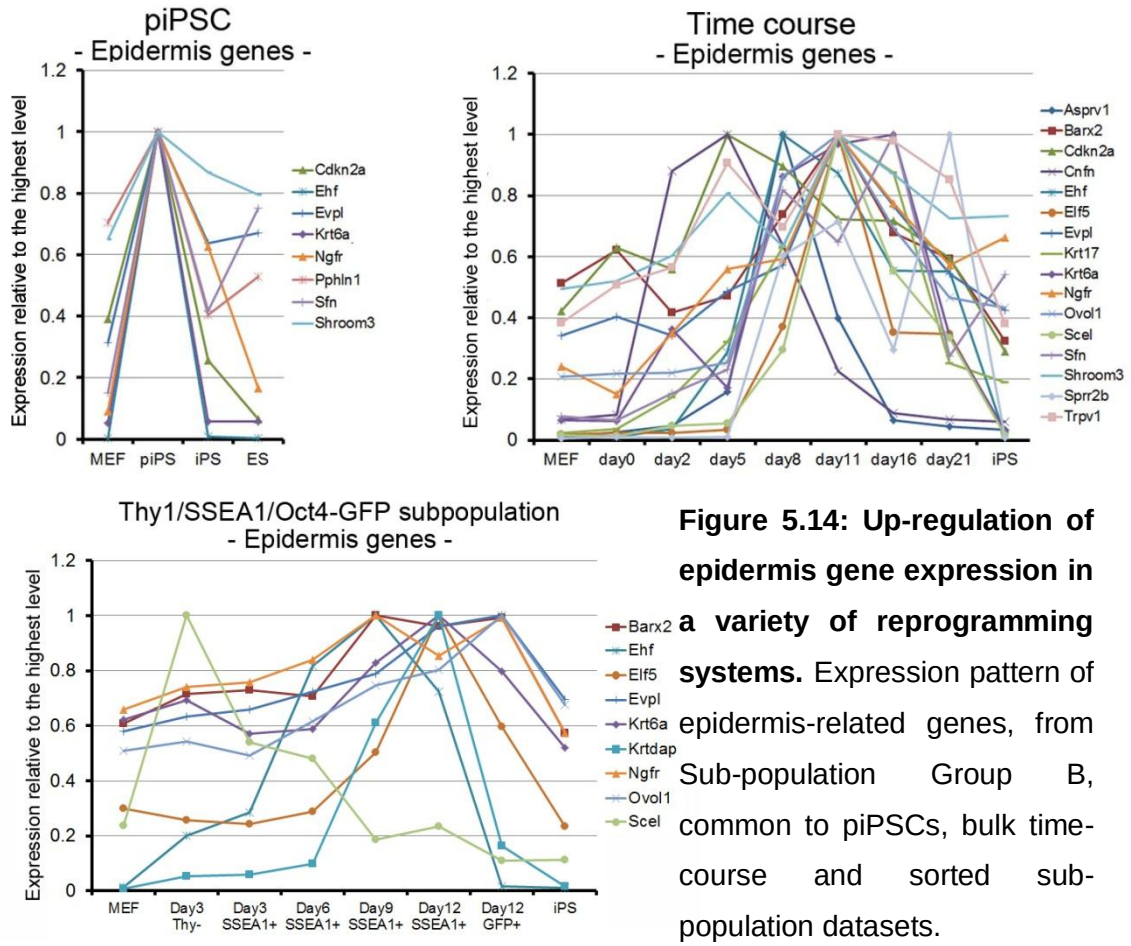


Figure 5.14: Up-regulation of epidermis gene expression in a variety of reprogramming systems. Expression pattern of epidermis-related genes, from Sub-population Group B, common to piPSCs, bulk time-course and sorted sub-population datasets.

While there is an apparent inverse correlation between epidermis and late stage pluripotency genes, it remains unclear if the former group play a functional role during reprogramming. Further to this, epidermis gene up-regulation was detected during Klf4 and Oct4, but not Sox2 and Oct4, based reprogramming of MEF (Nemajerova et al., 2012). This may suggest the observed up-regulation of epidermis genes may simply be a by-product of Klf4 over-expression during reprogramming. Reprogramming human fibroblasts in the absence of exogenous Klf4 has been demonstrated, and also in MEFs, both in cells lacking p53 activity and

by the inclusion of the kinase inhibitor Kenpaullone, with the latter strategy also shown to be effective in reprogramming mouse neural stem (NS) cells (Huangfu et al., 2008b; Kawamura et al., 2009; Lyssiotis et al., 2009). However, in all the above studies, the efficiency of reprogramming in the absence of Klf4 was severely affected and was not totally restored in any instance, indicating additional, as yet unknown functions for this factor during reprogramming. In addition, this suggests that even if the transient expression of these genes is the result of Klf4 activity, this is an important stage in efficient reprogramming.

5.3.4 RNAseq analysis reveals genes which improve reprogramming efficiency

As a proof of principle, preliminary results suggest that this reprogramming system can be used to identify factors which improve reprogramming from MEF. Prdm14 showed the greatest effect on colony formation, and recent reports have highlighted its important roles both in reprogramming epiblast stem cells to a more naïve state and in ES cell maintenance (Gillich et al., 2012; Grabole et al., 2013). Interestingly, Prdm14 activity in ES cells involves the repression of lineage specifiers in part by mediating PRC2-complex-member Jarid2 target binding, and was shown to co-localize with Nanog and Esrrb, all three of which are up-regulated as part of the second wave of pluripotency genes in the RNAseq dataset (Yamaji et al., 2013) (Figure 5.6b).

Mcm3 is a member of the minichromosome maintenance proteins (Mcm2-7) which constitute the core of the replicative DNA helicase complex which is recruited to origins of replication in early G_1, in a process termed "origin licensing" (Sclafani and Holzen, 2007). Abrogation of Mcm2 and Mcm4 activity in mice has previously been linked to chromosomal instability and an increased disposition to developing cancer (Pruitt et al., 2007; Shima et al., 2007). The complex relationship between members of the Mcm family was demonstrated by the finding that the efficiency of reprogramming from *Mcm4*$^{-/-}$, and *Mcm4*$^{-/-}$ *Mcm2*$^{-/+}$ MEF could be increased by

decreasing levels of Mcm3 expression (Chuang et al., 2010). This data indicated the levels of these family members relative to each other are important in reprogramming. However, it is unclear why altering the levels of individual members of a hetero-hexameric complex such as the one formed by the Mcm family members should have a rescue effect as described by this study. This may indicate theses proteins can function in other processes. Interestingly, in D6s4B5 cells, similarly to Mcm3, other members of the Mcm2-7 complex also increase during reprogramming (Figure 5.15, D-Mcm). Expression of these factors was largely similar in TANGO with the exception of Mcm3 and Mcm2 which were expressed at substantially lower levels than in D6s4B5 reprogramming (Figure 5.15, T-Mcm). Knockdown of Mcm3 and Mcm2 in human cells was found to result in hypersensitivity to replication stress with cells accumulating aberrant chromosome rearrangements, breaks and gaps (Ibarra et al., 2008). Similarly, reprogramming has been shown to induce expression of the tumour suppressor protein p53 which greatly decreases the efficiency of iPS cell generation, while inhibition of its activity generates iPS cells which carry DNA damage and chromosomal aberrations (Marion et al., 2009). It may be the case that during reprogramming expression of the Mcm complex members, especially Mcm3 and Mcm2, helps to maintain chromosomal integrity, thus reducing p53 activation and therefore overcoming this barrier to reprogramming.

Future investigation of factors identified through comparison with other cell lines, other reprogramming methods, different reprogramming culture conditions and using additional or alternative reprogramming factors can be carried out accurately using this reprogramming system, thus allowing further dissection of the molecular mechanism of reprogramming.

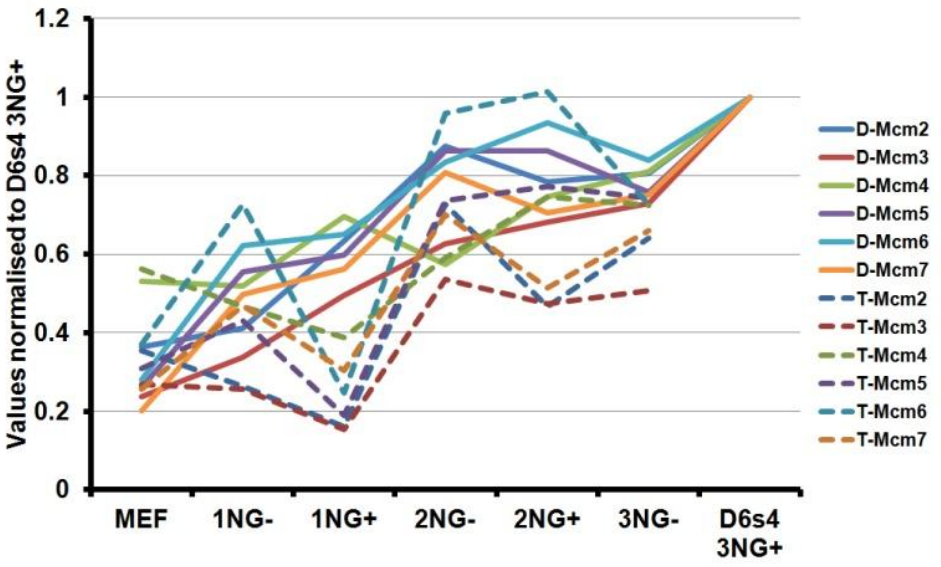


Figure 5.15: Mcm complex member expression during reprogramming. Sub-population expression levels for D6s4B5 (D-Mcm, solid lines) and TANGO (T-Mcm, dashed lines) cell lines. Expression is normalised to D6s4B5 3NG+.

CHAPTER 6 - General discussion and summary

6.1 Overview of results from this study

Starting with Gurdon's work in the 1960's, it had been shown that "resetting" of a differentiated cell to a pluripotent-like state was possible through cell-fusion or nuclear transfer, the exact factors required for this process to occur remained largely a mystery. Therefore the discovery that pluripotency could be induced in differentiated cells by over-expression of a defined set of transcription factors revolutionised the field of stem cell research. While this technology appeared to herald an era of huge potential for both investigation of pluripotency, differentiation and therapeutic applications of this knowledge, there were a number of obstacles to be overcome before these ideas would become reality.

This first set of induced pluripotent stem (iPS) cells was unable to contribute to viable chimeric mice, despite displaying many of the attributes of embryonic stem (ES) cells. It was later discovered that in cells undergoing the reprogramming process, more stringent selection for the expression of key pluripotency genes, such as Nanog and Oct4, produced iPS cells which could contribute to chimeric mice. These iPS cells were found to display an epigenetic profile which was much more similar to ES cells than those originally isolated by Takahashi and Yamanaka. Further advances were made through identification of additional factors, small molecules and culture conditions which were found to enhance the reprogramming process. However, despite these advances, the efficiency and heterogeneous of generating iPS cells in the vast majority of reprogramming systems prevented accurate high-resolution analysis of the requirements of successful reprogramming.

The aim of my project was to identify novel markers that could be used to isolate defined populations of cells undergoing reprogramming which could be

used to investigate the molecular mechanism of iPS cell generation from MEFs. By using a relatively efficient secondary reprogramming system, I discovered that markers widely used in reprogramming studies such as SSEA-1, THY-1 and E-CAD were insufficient to accurately identify the diverse populations of cells that arise during the reprogramming process. In contrast, I discovered that during reprogramming the dynamic expression patterns of novel markers ICAM-1 and CD44, when coupled with a Nanog-GFP reporter, could be used to isolate populations of cells which displayed distinct potentials to generate iPS cells. Uniquely, this demonstrated that even at a single time-point cultures and colonies of cells undergoing reprogramming are heterogeneous, highlighting the importance of this marker system's ability to isolate each of these sub-populations.

Further profiling of these populations by FACS analysis revealed distinct and reproducible kinetics during the transition from one population to the next. This demonstrated that the heterogeneity that arises during reprogramming can be attributed not only to differences in the timing of the reprogramming process but also to differences in the routes cells take to achieve a reprogrammed state. This was illustrated by the identification, for the first time, of two major routes available to cells undergoing reprogramming from MEF; one activating Nanog-GFP expression before loss of CD44, and gain of ICAM-1 expression, and the other doing so only after achieving an ES cell-like ICAM-1^{+}/CD44^{-} phenotype. Interestingly, while one route appeared to generate iPS cells more efficiently than the other, the ability of cells to take this "better" route was low.

In order to further investigate this observation and dissect the key characteristics of each of the sub-populations, RNAseq analysis was carried out. This method revealed that although isolated from a single time-point, each sub-population did indeed represent a distinct population of cells, with distinct transcriptional profiles, and each of the differentially expressed genes (DEGs) expressed in each were identified. A number of important results were obtained by using RNAseq analysis, including the identification of additional factors involved in

each of two waves of pluripotency gene up-regulation which had previously been reported in a study limited by single cell qPCR (sc-qPCR). In addition, analysis of DEGs between the sub-populations revealed two unexpected expression patterns; the transient up- and also down-regulation of genes during reprogramming. Gene ontology (GO) analysis identified a number of genes associated with epidermis development among the set of transiently up-regulated genes, and expression of these genes was confirmed by sc-qPCR of cells in each of the sub-populations and immunofluorescence staining of colonies undergoing reprogramming. This clearly demonstrated that during reprogramming from MEF cells enter an intermediate stage associated with the expression of epidermis-related genes. In addition, this transient gene expression pattern was identified in other reprogramming studies and in pre-iPS cells generated from B-cells, suggesting it may be a conserved characteristic of the reprogramming process (Sridharan et al., 2009; Samavarchi-Tehrani et al., 2010; Polo et al., 2012).

Finally, it was demonstrated that by applying this marker and high-resolution analysis other, less efficient, reprogramming systems could be investigated to identify novel factors that enhance the reprogramming process. Using this technique, two novel genes, Prdm14 and Mcm3, were identified to increase the efficiency of generating Nanog-GFP^{+} iPS cell during primary reprogramming.

These results demonstrated the ability of this system to accurately investigate and reveal novel insights into the molecular mechanism of reprogramming.

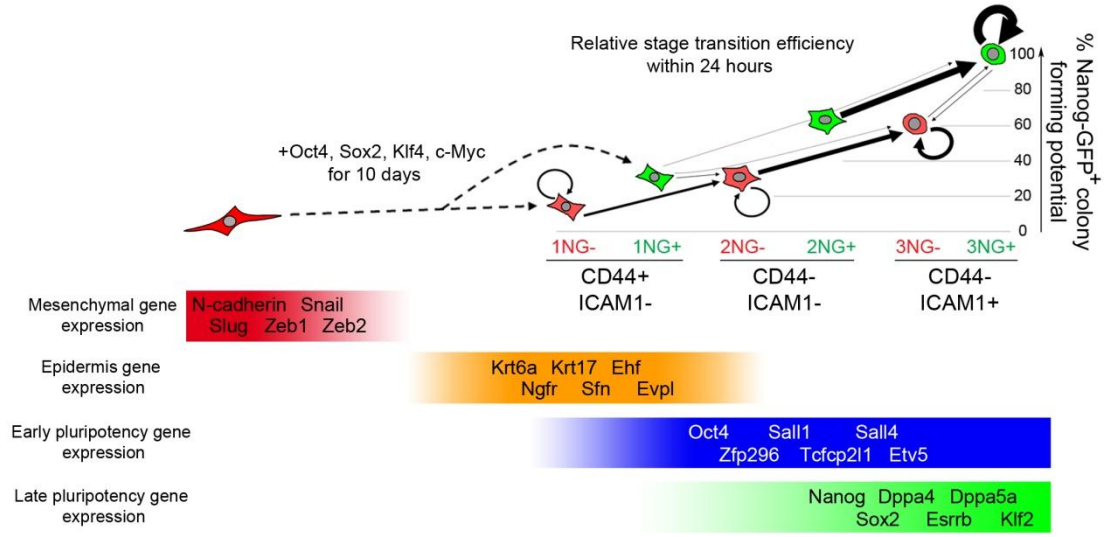


Figure 6.1: Overview of study findings. This study identified novel reprogramming cell surface markers ICAM-1 and CD44 which, in association with a Nanog-GFP reporter, illuminated the pathway(s) to the fully reprogrammed state through a number of distinct intermediate subpopulations. Transcriptional profiling revealed gene expression patterns during reprogramming including previously un-identified transient up-regulation of Epidermis genes.

6.2 Future directions

6.2.1 Other novel markers of the reprogramming process

While ICAM-1 and CD44 were sufficient to identify populations at the latter stages of reprogramming, other markers could be integrated in order to further increase the range and resolution of this system. For example, at earlier stages these markers may be unsuitable to correctly discriminate between the potential of populations to undergo reprogramming as illustrated by the almost identical reprogramming profile and kinetics illustrated by ICAM-1 sorted MEF cells (Figure 4.5). However, one caveat of this type of analysis however, is the heterogeneity of the reprogramming process at early stages of reprogramming, with a large number of genes responding to reprogramming factor expression, the majority of which may not play a role in promoting iPS cell generation (Soufi et al., 2012). One approach could be to sort $CD44^{+}$ cells at various time-points after reprogramming factor expression and determine a change in iPS cell potential. Changes in gene expression between this and earlier populations may reveal factors important for successful reprogramming, with the ultimate aim being identification of a very early up-regulated factor or factors which can be used to identify cells correlated with reprogramming success, for example a marker of a day 10 1NG+ precursor. However, as outlined in this study, even at the latter stages of reprogramming more than one route is available for cells to undergo reprogramming, and identification of such a marker may be difficult.

It may be more appropriate to apply the type of analysis carried out with ICAM-1 and CD44 to non-MEF reprogramming systems. While these exact markers may not be suitable, other differentiated cell types have well characterised cell surface markers, and the dynamics of their expression changes during reprogramming may enable similar analysis as carried out in this study. It has been reported that pre-iPS cells cluster separately from SSEA-1 cells undergoing reprogramming from MEF, suggesting they may take a different route to the

reprogrammed state (Polo et al., 2012). Interestingly, pre-iPS cells demonstrate an ICAM-1^{-}/CD44^{+} profile (see Appendix, Supplementary Figure 15), suggesting these markers could be used to investigate the mechanism of reprogramming from these intermediates.

6.2.2 Investigation of the epigenetic state of identified sub-populations

While the RNAseq analysis carried out in this study provides a wealth of information regarding the transcriptional capability of each of the sub-populations, the epigenetic state of these cells remains unknown. A large number of studies have emphasised the importance of epigenetic regulators in ensuring successful reprogramming, and the system generated in my study provides a unique opportunity to investigate the dynamics of epigenetic change during reprogramming. A factor which may prove difficult to overcome is the quantity of cells required for thorough analysis of epigenetic features such as Chromatin-immunoprecipitation coupled with high-throughput DNA sequencing (ChIP-seq), which may not be possible for the lesser-occurring sub-populations such as 1NG+ and 2NG+. Of course, this technical obstacle may be overcome by either increasing the starting numbers of MEFs used for analysis, or by sampling at different time-points which may be more enriched for the sub-population required, and this would require functional validation, as carried out for the day 10 and day 8 sorted cells (Figure 4.9, 4.10). In addition, ChIP-seq protocols using relatively low ($1X10^{5}$) number of cells have been recently been reported, although this reduction in cell number was correlated to an increase in unmapped sequencing reads (~30% of total), and the quality of the antibody used for immunoprecipitation introduces another potentially detrimental variable (Gilfillan et al., 2012).

Another approach to epigenetic investigation of the reprogramming process is to analyses the effect of specific factors. For example, one molecule associated with regulation of epigenetic factors, and shown to drastically affect the

reprogramming process in this study is VitaminC (VitC, Figure 4.12). As illustrated, the removal of this molecule reduced the efficiency of reprogramming and altered the transition kinetics of each sub-population. It would be interesting to investigate if the activity of reported VitC-dependent factors, and their associated epigenetic modifications (see Section 1.4.3) are altered compared to control sub-populations. This type of analysis could reveal the individual contribution of individual epigenetic factors to the series of events required to re-establish an ES cell-like epigenetic state during reprogramming.

6.2.3 The role of transient epidermis gene up-regulation during reprogramming

While it appears from the results of this study that transient up-regulation of epidermis-associated genes during reprogramming of MEF is conserved across different systems, the functional role of these genes, if any, was not investigated. Introduction of siRNA knock-down vectors against these epidermis genes during reprogramming could reveal the importance of these factors during reprogramming. Cultures could be monitored for differences in the kinetics of iPS cell colony formation and changes in the dynamic of ICAM-1/CD44 expression during reprogramming compared to control cells. This could illuminate if this transient state is required for, or hinders efficient reprogramming.

As previously discussed (see Section 5.3.3), a number of studies suggest that Klf4 plays a role in the up-regulation of these genes during reprogramming, and reprogramming in the absence of Klf4 could also address this question.

6.2.4 Further characterisation and identification of novel enhancers of the reprogramming process

In this study I demonstrated the identification of two novel factors which could increase the efficiency of primary reprogramming in co-operation with the

widely-used Yamanaka factors (Figure 5.12). Further investigation of these factors could lend further insights into their role in reprogramming. As discussed it would be interesting to monitor if Prdm14-associated targets are similarly up-regulated at earlier stages in reprogramming and are responsible for the observed increase in reprogramming efficiency upon over-expression of this factor. Similarly, investigation of factors associated with senescence and DNA damage such as activated p53 may lead to greater understanding of the role of Mcm3 in enhancing reprogramming.

In addition to these factors, a number of other factors differentially expressed between D6s4B5 and TANGO reprogramming systems remain to be investigated, in a similar manner to the genes above and these may also provide information about the mechanism of reprogramming.

6.3 Future perspectives for iPS cell technology

Although understanding of the mechanism of reprogramming is far from complete, and the system developed in this study can help to further our knowledge of this process, iPS cells have already been applied in therapeutic contexts. Recently, the Japanese Ministry of Health approved a plan to carry out the world's first clinical trial involving transplantation of retinal pigment epithelial cells generated from iPS cells derived from patients with age-related macular degeneration. This is a tremendously exciting development and is the product of intensive research in this field, and although it remains to be seen if this trial will be successful, it is likely that the promise of iPS cells is finally coming to fruition.

Much debate has focused on attempting to understand if the mechanism of reprogramming is a stochastic or deterministic process (Hanna et al., 2009b; Yamanaka, 2009; Buganim et al., 2012). In 2009, Hanna and colleagues reported that if given enough time (i.e. number of cell divisions) almost all clonally derived B-cells can undergo reprogramming (Hanna et al., 2009b). This indicated that reprogramming was a stochastic process, with the acquisition of the fully reprogrammed state varying from cell to cell (variable latency). More recently however, Hanna and colleagues (Rais et al., 2013) reported knock-down of a component of the nucleosome remodelling and deacetylation (NuRD) complex, methyl-CpG binding domain protein 3 (*Mbd3*), was sufficient to abolish the stochastic nature of reprogramming. Dramatically, this report demonstrated that in a secondary reprogramming system driven by dox-mediated expression of a polycistronic reprogramming vector in single-cell sorted *Mbd3*$^{-/-}$ MEF, 100% of cells expressed an Oct4-GFP reporter by day 8 of reprogramming. Unusually, cells underwent reprogramming in 2i/LIF under physiological oxygen (5% O_2) conditions. In addition to MEF, a number of other cell types were reprogrammed at a similar efficiency including pro- and mature B-cells, adult tail tip fibroblasts, hematopoietic stem cells and neural precursors. ChIP-seq and sc-qPCR analysis revealed *Mbd3*$^{-/-}$ cells lost DNA methylation and up-regulated pluripotency genes

between days 4 and 8 of reprogramming. In contrast to another report investigating Mbd3 in reprogramming, this report did not observe up-regulation of endogenous pluripotency-associated genes in the absence of Mbd3 in MEF, and rather reported a direct physical interaction between the reprogramming factors and Mbd3 itself (Luo et al., 2013). In wt cells, this interaction was found to recruit Mbd3 and another NuRD complex component, Chd4, to common target genes which displayed higher expression in *Mbd3*$^{-/-}$ cells. This lead to the conclusion that in the absence of *Mbd3*, reprogramming is a deterministic process, with a loss of the variable latency of individual cells achieving a reprogrammed state. Interestingly, it was reported in human fibroblasts that while the Yamanaka factors demonstrate quite promiscuous binding upon over-expression during reprogramming, large regions which contain reprogramming factor-bound targets in ES cells remain unbound and methylated, and it would be interesting to determine if there is a correlation between these two observations, especially as Hanna et al. also demonstrated *Mbd3*$^{-/-}$ enhanced reprogramming of differentiated human cells (Soufi et al., 2012). The 6 days required for reprogramming in *Mbd3*$^{-/-}$ cells was only reported in 2i/LIF medium which is not commonly used throughout the whole reprogramming time-course, and lower reprogramming efficiencies were reported for more usual reprogramming medium. This discrepancy could be investigated to understand the role of the 2i components in reprogramming, and indeed, the addition of other factors which may further accelerate the onset of reprogramming.

The fact that the previous report which demonstrated Mbd3 knockdown did not observe such a dramatic increase in reprogramming efficiency is surprising (Luo et al., 2013). In addition, it had previously been reported that in reprogramming from primary human cells Mbd3 knockdown negatively affected colony formation, in contrast to the reported increase in reprogramming efficiency observed in Hanna and colleague's report, although this could only be achieved by the addition of *LIN28* to the reprogramming cocktail (Onder et al., 2012). Of course, these experiments were carried out under different culture conditions, and the exact

experiments carried out by Hannah and colleagues must be repeated by other groups. Assuming these results can be independently verified, the study of reprogramming may not remain a mystery for much longer if totally homogeneous, successful cultures of cells can be readily generated.

REFERENCES

AL-HAJJ, M., WICHA, M. S., BENITO-HERNANDEZ, A., MORRISON, S. J. & CLARKE, M. F. 2003. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A,* 100**,** 3983-8.

ALBERIO, R., CROXALL, N. & ALLEGRUCCI, C. 2010. Pig epiblast stem cells depend on activin/nodal signaling for pluripotency and self-renewal. *Stem Cells Dev,* 19**,** 1627-36.

AMBASUDHAN, R., TALANTOVA, M., COLEMAN, R., YUAN, X., ZHU, S., LIPTON, S. A. & DING, S. 2011. Direct reprogramming of adult human fibroblasts to functional neurons under defined conditions. *Cell Stem Cell,* 9**,** 113-8.

AMIT, M., CARPENTER, M. K., INOKUMA, M. S., CHIU, C. P., HARRIS, C. P., WAKNITZ, M. A., ITSKOVITZ-ELDOR, J. & THOMSON, J. A. 2000. Clonally derived human embryonic stem cell lines maintain pluripotency and proliferative potential for prolonged periods of culture. *Dev Biol,* 227**,** 271-8.

AMIT, M., SHARIKI, C., MARGULETS, V. & ITSKOVITZ-ELDOR, J. 2004. Feeder layer- and serum-free culture of human embryonic stem cells. *Biol Reprod,* 70**,** 837-45.

ANDERS, S. & HUBER, W. 2010. Differential expression analysis for sequence count data. *Genome Biol,* 11**,** R106.

ANG, Y. S., TSAI, S. Y., LEE, D. F., MONK, J., SU, J., RATNAKUMAR, K., DING, J., GE, Y., DARR, H., CHANG, B., WANG, J., RENDL, M., BERNSTEIN, E., SCHANIEL, C. & LEMISCHKA, I. R. 2011. Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell,* 145**,** 183-97.

ANOKYE-DANSO, F., TRIVEDI, C. M., JUHR, D., GUPTA, M., CUI, Z., TIAN, Y., ZHANG, Y., YANG, W., GRUBER, P. J., EPSTEIN, J. A. & MORRISEY, E. E. 2011. Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell Stem Cell,* 8**,** 376-88.

BAN, H., NISHISHITA, N., FUSAKI, N., TABATA, T., SAEKI, K., SHIKAMURA, M., TAKADA, N., INOUE, M., HASEGAWA, M., KAWAMATA, S. & NISHIKAWA, S. 2011. Efficient generation of transgene-free human induced pluripotent stem cells (iPSCs) by temperature-sensitive Sendai virus vectors. *Proc Natl Acad Sci U S A,* 108**,** 14234-9.

BERNSTEIN, B. E., MIKKELSEN, T. S., XIE, X., KAMAL, M., HUEBERT, D. J., CUFF, J., FRY, B., MEISSNER, A., WERNIG, M., PLATH, K., JAENISCH, R.,

WAGSCHAL, A., FEIL, R., SCHREIBER, S. L. & LANDER, E. S. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell,* 125**,** 315-26.

BHAT-NAKSHATRI, P., APPAIAH, H., BALLAS, C., PICK-FRANKE, P., GOULET, R., JR., BADVE, S., SROUR, E. F. & NAKSHATRI, H. 2010. SLUG/SNAI2 and tumor necrosis factor generate breast cells with CD44+/CD24- phenotype. *BMC Cancer,* 10**,** 411.

BHUTANI, N., BRADY, J. J., DAMIAN, M., SACCO, A., CORBEL, S. Y. & BLAU, H. M. 2010. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature,* 463**,** 1042-7.

BIELLMANN, F., HULSMEIER, A. J., ZHOU, D., CINELLI, P. & HENNET, T. 2008. The Lc3-synthase gene B3gnt5 is essential to pre-implantation development of the murine embryo. *BMC Dev Biol,* 8**,** 109.

BLELLOCH, R., VENERE, M., YEN, J. & RAMALHO-SANTOS, M. 2007. Generation of induced pluripotent stem cells in the absence of drug selection. *Cell Stem Cell,* 1**,** 245-7.

BOTCHKINA, I. L., ROWEHL, R. A., RIVADENEIRA, D. E., KARPEH, M. S., JR., CRAWFORD, H., DUFOUR, A., JU, J., WANG, Y., LEYFMAN, Y. & BOTCHKINA, G. I. 2009. Phenotypic subpopulations of metastatic colon cancer stem cells: genomic analysis. *Cancer Genomics Proteomics,* 6**,** 19-29.

BOURGUIGNON, L. Y., EARLE, C., WONG, G., SPEVAK, C. C. & KRUEGER, K. 2012. Stem cell marker (Nanog) and Stat-3 signaling promote MicroRNA-21 expression and chemoresistance in hyaluronan/CD44-activated head and neck squamous cell carcinoma cells. *Oncogene,* 31**,** 149-60.

BOURGUIGNON, L. Y., SINGLETON, P. A., ZHU, H. & ZHOU, B. 2002. Hyaluronan promotes signaling interaction between CD44 and the transforming growth factor beta receptor I in metastatic breast tumor cells. *J Biol Chem,* 277**,** 39703-12.

BOURGUIGNON, L. Y., SPEVAK, C. C., WONG, G., XIA, W. & GILAD, E. 2009. Hyaluronan-CD44 interaction with protein kinase C(epsilon) promotes oncogenic signaling by the stem cell marker Nanog and the Production of microRNA-21, leading to down-regulation of the tumor suppressor protein PDCD4, anti-apoptosis, and chemotherapy resistance in breast tumor cells. *J Biol Chem,* 284**,** 26533-46.

BOYER, L. A., LEE, T. I., COLE, M. F., JOHNSTONE, S. E., LEVINE, S. S., ZUCKER, J. P., GUENTHER, M. G., KUMAR, R. M., MURRAY, H. L., JENNER, R. G., GIFFORD, D. K., MELTON, D. A., JAENISCH, R. & YOUNG, R. A. 2005.

Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell,* 122**,** 947-56.

BRADLEY, A., EVANS, M., KAUFMAN, M. H. & ROBERTSON, E. 1984. Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines. *Nature,* 309**,** 255-6.

BRAMBRINK, T., FOREMAN, R., WELSTEAD, G. G., LENGNER, C. J., WERNIG, M., SUH, H. & JAENISCH, R. 2008. Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell,* 2**,** 151-9.

BRONS, I. G., SMITHERS, L. E., TROTTER, M. W., RUGG-GUNN, P., SUN, B., CHUVA DE SOUSA LOPES, S. M., HOWLETT, S. K., CLARKSON, A., AHRLUND-RICHTER, L., PEDERSEN, R. A. & VALLIER, L. 2007. Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature,* 448**,** 191-5.

BUGANIM, Y., FADDAH, D. A., CHENG, A. W., ITSKOVICH, E., MARKOULAKI, S., GANZ, K., KLEMM, S. L., VAN OUDENAARDEN, A. & JAENISCH, R. 2012. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell,* 150**,** 1209-22.

CAIAZZO, M., DELL'ANNO, M. T., DVORETSKOVA, E., LAZAREVIC, D., TAVERNA, S., LEO, D., SOTNIKOVA, T. D., MENEGON, A., RONCAGLIA, P., COLCIAGO, G., RUSSO, G., CARNINCI, P., PEZZOLI, G., GAINETDINOV, R. R., GUSTINCICH, S., DITYATEV, A. & BROCCOLI, V. 2011. Direct generation of functional dopaminergic neurons from mouse and human fibroblasts. *Nature,* 476**,** 224-7.

CAREY, B. W., MARKOULAKI, S., BEARD, C., HANNA, J. & JAENISCH, R. 2010. Single-gene transgenic mouse strains for reprogramming adult somatic cells. *Nat Methods,* 7**,** 56-9.

CAREY, B. W., MARKOULAKI, S., HANNA, J., SAHA, K., GAO, Q., MITALIPOVA, M. & JAENISCH, R. 2009. Reprogramming of murine and human somatic cells using a single polycistronic vector. *Proc Natl Acad Sci U S A,* 106**,** 157-62.

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., WELSTEAD, G. G., GAO, Q. & JAENISCH, R. 2011. Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells. *Cell Stem Cell,* 9**,** 588-98.

CARY, L. C., GOEBEL, M., CORSARO, B. G., WANG, H. G., ROSEN, E. & FRASER, M. J. 1989. Transposon mutagenesis of baculoviruses: analysis of

Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology,* 172**,** 156-69.

CHAMBERS, I., COLBY, D., ROBERTSON, M., NICHOLS, J., LEE, S., TWEEDIE, S. & SMITH, A. 2003. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell,* 113**,** 643-55.

CHAMBERS, I., SILVA, J., COLBY, D., NICHOLS, J., NIJMEIJER, B., ROBERTSON, M., VRANA, J., JONES, K., GROTEWOLD, L. & SMITH, A. 2007. Nanog safeguards pluripotency and mediates germline development. *Nature,* 450**,** 1230-4.

CHAMBERS, I. & TOMLINSON, S. R. 2009. The transcriptional foundation of pluripotency. *Development,* 136**,** 2311-22.

CHAN, E. M., RATANASIRINTRAWOOT, S., PARK, I. H., MANOS, P. D., LOH, Y. H., HUO, H., MILLER, J. D., HARTUNG, O., RHO, J., INCE, T. A., DALEY, G. Q. & SCHLAEGER, T. M. 2009. Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nat Biotechnol,* 27**,** 1033-7.

CHANG, C. W., LAI, Y. S., PAWLIK, K. M., LIU, K., SUN, C. W., LI, C., SCHOEB, T. R. & TOWNES, T. M. 2009. Polycistronic lentiviral vector for "hit and run" reprogramming of adult skin fibroblasts to induced pluripotent stem cells. *Stem Cells,* 27**,** 1042-9.

CHEN, H. F., CHUANG, C. Y., LEE, W. C., HUANG, H. P., WU, H. C., HO, H. N., CHEN, Y. J. & KUO, H. C. 2011. Surface marker epithelial cell adhesion molecule and E-cadherin facilitate the identification and selection of induced pluripotent stem cells. *Stem Cell Rev,* 7**,** 722-35.

CHEN, J., LIU, H., LIU, J., QI, J., WEI, B., YANG, J., LIANG, H., CHEN, Y., WU, Y., GUO, L., ZHU, J., ZHAO, X., PENG, T., ZHANG, Y., CHEN, S., LI, X., LI, D., WANG, T. & PEI, D. 2013. H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nat Genet,* 45**,** 34-42.

CHEN, X., XU, H., YUAN, P., FANG, F., HUSS, M., VEGA, V. B., WONG, E., ORLOV, Y. L., ZHANG, W., JIANG, J., LOH, Y. H., YEO, H. C., YEO, Z. X., NARANG, V., GOVINDARAJAN, K. R., LEONG, B., SHAHAB, A., RUAN, Y., BOURQUE, G., SUNG, W. K., CLARKE, N. D., WEI, C. L. & NG, H. H. 2008. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell,* 133**,** 1106-17.

CHEVALIER-MARIETTE, C., HENRY, I., MONTFORT, L., CAPGRAS, S., FORLANI, S., MUSCHLER, J. & NICOLAS, J. F. 2003. CpG content affects gene silencing in mice: evidence from novel transgenes. *Genome Biol,* 4**,** R53.

CHEW, J. L., LOH, Y. H., ZHANG, W., CHEN, X., TAM, W. L., YEAP, L. S., LI, P., ANG, Y. S., LIM, B., ROBSON, P. & NG, H. H. 2005. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol Cell Biol,* 25**,** 6031-46.

CHO, H. J., LEE, C. S., KWON, Y. W., PAEK, J. S., LEE, S. H., HUR, J., LEE, E. J., ROH, T. Y., CHU, I. S., LEEM, S. H., KIM, Y., KANG, H. J., PARK, Y. B. & KIM, H. S. 2010. Induction of pluripotent stem cells from adult somatic cells by protein-based reprogramming without genetic manipulation. *Blood,* 116**,** 386-95.

CHUANG, C. H., WALLACE, M. D., ABRATTE, C., SOUTHARD, T. & SCHIMENTI, J. C. 2010. Incremental genetic perturbations to MCM2-7 expression and subcellular distribution reveal exquisite sensitivity of mice to DNA replication stress. *PLoS Genet,* 6**,** e1001110.

CICCONE, D. N., SU, H., HEVI, S., GAY, F., LEI, H., BAJKO, J., XU, G., LI, E. & CHEN, T. 2009. KDM1B is a histone H3K4 demethylase required to establish maternal genomic imprints. *Nature,* 461**,** 415-8.

CLOUAIRE, T., WEBB, S., SKENE, P., ILLINGWORTH, R., KERR, A., ANDREWS, R., LEE, J. H., SKALNIK, D. & BIRD, A. 2012. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev,* 26**,** 1714-28.

COSTA, Y., DING, J., THEUNISSEN, T. W., FAIOLA, F., HORE, T. A., SHLIAHA, P. V., FIDALGO, M., SAUNDERS, A., LAWRENCE, M., DIETMANN, S., DAS, S., LEVASSEUR, D. N., LI, Z., XU, M., REIK, W., SILVA, J. C. & WANG, J. 2013. NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature,* 495**,** 370-4.

COWAN, C. A., ATIENZA, J., MELTON, D. A. & EGGAN, K. 2005. Nuclear reprogramming of somatic cells after fusion with human embryonic stem cells. *Science,* 309**,** 1369-73.

CUI, L., JOHKURA, K., YUE, F., OGIWARA, N., OKOUCHI, Y., ASANUMA, K. & SASAKI, K. 2004. Spatial distribution and initial changes of SSEA-1 and other cell adhesion-related molecules on mouse embryonic stem cells before and during differentiation. *J Histochem Cytochem,* 52**,** 1447-57.

DAHERON, L., OPITZ, S. L., ZAEHRES, H., LENSCH, M. W., ANDREWS, P. W., ITSKOVITZ-ELDOR, J. & DALEY, G. Q. 2004. LIF/STAT3 signaling fails to maintain self-renewal of human embryonic stem cells. *Stem Cells,* 22**,** 770-8.

DAVIS, R. L., WEINTRAUB, H. & LASSAR, A. B. 1987. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell,* 51**,** 987-1000.

DE HOON, M. J., IMOTO, S., NOLAN, J. & MIYANO, S. 2004. Open source clustering software. *Bioinformatics,* 20**,** 1453-4.

DING, S., WU, X., LI, G., HAN, M., ZHUANG, Y. & XU, T. 2005. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell,* 122**,** 473-83.

DOEGE, C. A., INOUE, K., YAMASHITA, T., RHEE, D. B., TRAVIS, S., FUJITA, R., GUARNIERI, P., BHAGAT, G., VANTI, W. B., SHIH, A., LEVINE, R. L., NIK, S., CHEN, E. I. & ABELIOVICH, A. 2012. Early-stage epigenetic modification during somatic cell reprogramming by Parp1 and Tet2. *Nature,* 488**,** 652-5.

DONNELLY, M. L., LUKE, G., MEHROTRA, A., LI, X., HUGHES, L. E., GANI, D. & RYAN, M. D. 2001. Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: a putative ribosomal 'skip'. *J Gen Virol,* 82**,** 1013-25.

DUSTIN, M. L., ROTHLEIN, R., BHAN, A. K., DINARELLO, C. A. & SPRINGER, T. A. 1986. Induction by IL 1 and interferon-gamma: tissue distribution, biochemistry, and function of a natural adherence molecule (ICAM-1). *J Immunol,* 137**,** 245-54.

EFE, J. A., HILCOVE, S., KIM, J., ZHOU, H., OUYANG, K., WANG, G., CHEN, J. & DING, S. 2011. Conversion of mouse fibroblasts into cardiomyocytes using a direct reprogramming strategy. *Nat Cell Biol,* 13**,** 215-22.

EMINLI, S., FOUDI, A., STADTFELD, M., MAHERALI, N., AHFELDT, T., MOSTOSLAVSKY, G., HOCK, H. & HOCHEDLINGER, K. 2009. Differentiation stage determines potential of hematopoietic cells for reprogramming into induced pluripotent stem cells. *Nat Genet,* 41**,** 968-976.

ESTEBAN, M. A., WANG, T., QIN, B., YANG, J., QIN, D., CAI, J., LI, W., WENG, Z., CHEN, J., NI, S., CHEN, K., LI, Y., LIU, X., XU, J., ZHANG, S., LI, F., HE, W., LABUDA, K., SONG, Y., PETERBAUER, A., WOLBANK, S., REDL, H., ZHONG, M., CAI, D., ZENG, L. & PEI, D. 2010. Vitamin C enhances the generation of mouse and human induced pluripotent stem cells. *Cell Stem Cell,* 6**,** 71-9.

EVANS, M. 2011. Discovering pluripotency: 30 years of mouse embryonic stem cells. *Nat Rev Mol Cell Biol,* 12**,** 680-6.

EVANS, M. J. & KAUFMAN, M. H. 1981. Establishment in culture of pluripotential cells from mouse embryos. *Nature,* 292**,** 154-6.

FADDAH, D. A., WANG, H., CHENG, A. W., KATZ, Y., BUGANIM, Y. & JAENISCH, R. 2013. Single-Cell Analysis Reveals that Expression of Nanog

Is Biallelic and Equally Variable as that of Other Pluripotency Factors in Mouse ESCs. *Cell Stem Cell,* 13**,** 23-9.

FEHON, R. G., MCCLATCHEY, A. I. & BRETSCHER, A. 2010. Organizing the cell cortex: the role of ERM proteins. *Nat Rev Mol Cell Biol,* 11**,** 276-87.

FENG, B., JIANG, J., KRAUS, P., NG, J. H., HENG, J. C., CHAN, Y. S., YAW, L. P., ZHANG, W., LOH, Y. H., HAN, J., VEGA, V. B., CACHEUX-RATABOUL, V., LIM, B., LUFKIN, T. & NG, H. H. 2009a. Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol,* 11**,** 197-203.

FENG, B., NG, J. H., HENG, J. C. & NG, H. H. 2009b. Molecules that promote or enhance reprogramming of somatic cells to induced pluripotent stem cells. *Cell Stem Cell,* 4**,** 301-12.

FENG, R., DESBORDES, S. C., XIE, H., TILLO, E. S., PIXLEY, F., STANLEY, E. R. & GRAF, T. 2008. PU.1 and C/EBPalpha/beta convert fibroblasts into macrophage-like cells. *Proc Natl Acad Sci U S A,* 105**,** 6057-62.

FESTUCCIA, N., OSORNO, R., HALBRITTER, F., KARWACKI-NEISIUS, V., NAVARRO, P., COLBY, D., WONG, F., YATES, A., TOMLINSON, S. R. & CHAMBERS, I. 2012. Esrrb is a direct Nanog target gene that can substitute for Nanog function in pluripotent cells. *Cell Stem Cell,* 11**,** 477-90.

FIDALGO, M., FAIOLA, F., PEREIRA, C. F., DING, J., SAUNDERS, A., GINGOLD, J., SCHANIEL, C., LEMISCHKA, I. R., SILVA, J. C. & WANG, J. 2012. Zfp281 mediates Nanog autorepression through recruitment of the NuRD complex and inhibits somatic cell reprogramming. *Proc Natl Acad Sci U S A,* 109**,** 16202-7.

FILIPCZYK, A., GKATZIS, K., FU, J., HOPPE, P. S., LICKERT, H., ANASTASSIADIS, K. & SCHROEDER, T. 2013. Biallelic expression of nanog protein in mouse embryonic stem cells. *Cell Stem Cell,* 13**,** 12-3.

FRASER, M. J., CISZCZON, T., ELICK, T. & BAUSER, C. 1996. Precise excision of TTAA-specific lepidopteran transposons piggyBac (IFP2) and tagalong (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. *Insect Mol Biol,* 5**,** 141-51.

FU, H., WANG, L., LIN, C. M., SINGHANIA, S., BOUHASSIRA, E. E. & ALADJEM, M. I. 2006. Preventing gene silencing with human replicators. *Nat Biotechnol,* 24**,** 572-6.

FUSAKI, N., BAN, H., NISHIYAMA, A., SAEKI, K. & HASEGAWA, M. 2009. Efficient induction of transgene-free human pluripotent stem cells using a

vector based on Sendai virus, an RNA virus that does not integrate into the host genome. *Proc Jpn Acad Ser B Phys Biol Sci,* 85**,** 348-62.

GABUT, M., SAMAVARCHI-TEHRANI, P., WANG, X., SLOBODENIUC, V., O'HANLON, D., SUNG, H. K., ALVAREZ, M., TALUKDER, S., PAN, Q., MAZZONI, E. O., NEDELEC, S., WICHTERLE, H., WOLTJEN, K., HUGHES, T. R., ZANDSTRA, P. W., NAGY, A., WRANA, J. L. & BLENCOWE, B. J. 2011. An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell,* 147**,** 132-46.

GAGLIARDI, A., MULLIN, N. P., YING TAN, Z., COLBY, D., KOUSA, A. I., HALBRITTER, F., WEISS, J. T., FELKER, A., BEZSTAROSTI, K., FAVARO, R., DEMMERS, J., NICOLIS, S. K., TOMLINSON, S. R., POOT, R. A. & CHAMBERS, I. 2013. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J,* 32**,** 2231-47.

GALLATIN, W. M., WEISSMAN, I. L. & BUTCHER, E. C. 1983. A cell-surface molecule involved in organ-specific homing of lymphocytes. *Nature,* 304**,** 30-4.

GARDNER, R. L. 1982. Investigation of cell lineage and differentiation in the extraembryonic endoderm of the mouse embryo. *J Embryol Exp Morphol,* 68**,** 175-98.

GASZNER, M. & FELSENFELD, G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet,* 7**,** 703-13.

GAUTIER, L. C., L.; BOLSTAD, BM.; IRIZARRY, RA. 2004. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics,* 20**,** 307-15.

GENTLEMAN, R. C., CAREY, V. J., BATES, D. M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A. J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J. Y. & ZHANG, J. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol,* 5**,** R80.

GILBERT, D. M. 2002. Replication timing and metazoan evolution. *Nat Genet,* 32**,** 336-7.

GILFILLAN, G. D., HUGHES, T., SHENG, Y., HJORTHAUG, H. S., STRAUB, T., GERVIN, K., HARRIS, J. R., UNDLIEN, D. E. & LYLE, R. 2012. Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics,* 13**,** 645.

GILLICH, A., BAO, S., GRABOLE, N., HAYASHI, K., TROTTER, M. W., PASQUE, V., MAGNUSDOTTIR, E. & SURANI, M. A. 2012. Epiblast stem cell-based

system reveals reprogramming synergy of germline factors. *Cell Stem Cell,* 10**,** 425-39.

GIORGETTI, A., MARCHETTO, M. C., LI, M., YU, D., FAZZINA, R., MU, Y., ADAMO, A., PARAMONOV, I., CARDOSO, J. C., MONASTERIO, M. B., BARDY, C., CASSIANI-INGONI, R., LIU, G. H., GAGE, F. H. & IZPISUA BELMONTE, J. C. 2012. Cord blood-derived neuronal cells by ectopic expression of Sox2 and c-Myc. *Proc Natl Acad Sci U S A,* 109**,** 12556-61.

GOLDSTEIN, L. A., ZHOU, D. F., PICKER, L. J., MINTY, C. N., BARGATZE, R. F., DING, J. F. & BUTCHER, E. C. 1989. A human lymphocyte homing receptor, the hermes antigen, is related to cartilage proteoglycan core and link proteins. *Cell,* 56**,** 1063-72.

GOLIPOUR, A., DAVID, L., LIU, Y., JAYAKUMARAN, G., HIRSCH, C. L., TRCKA, D. & WRANA, J. L. 2012. A late transition in somatic cell reprogramming requires regulators distinct from the pluripotency network. *Cell Stem Cell,* 11**,** 769-82.

GONZALEZ, F., BARRAGAN MONASTERIO, M., TISCORNIA, G., MONTSERRAT PULIDO, N., VASSENA, R., BATLLE MORERA, L., RODRIGUEZ PIZA, I. & IZPISUA BELMONTE, J. C. 2009. Generation of mouse-induced pluripotent stem cells by transient expression of a single nonviral polycistronic vector. *Proc Natl Acad Sci U S A,* 106**,** 8918-22.

GORDON, J. & MANLEY, N. R. 2011. Mechanisms of thymus organogenesis and morphogenesis. *Development,* 138**,** 3865-78.

GOSSEN, M. & BUJARD, H. 1992. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc Natl Acad Sci U S A,* 89**,** 5547-51.

GRABOLE, N., TISCHLER, J., HACKETT, J. A., KIM, S., TANG, F., LEITCH, H. G., MAGNUSDOTTIR, E. & SURANI, M. A. 2013. Prdm14 promotes germline fate and naive pluripotency by repressing FGF signalling and DNA methylation. *EMBO Rep,* 14**,** 629-37.

GUNDRY, R. L., RIORDON, D. R., TARASOVA, Y., CHUPPA, S., BHATTACHARYA, S., JUHASZ, O., WIEDEMEIER, O., MILANOVICH, S., NOTO, F. K., TCHERNYSHYOV, I., RAGINSKI, K., BAUSCH-FLUCK, D., TAE, H. J., MARSHALL, S., DUNCAN, S. A., WOLLSCHEID, B., WERSTO, R. P., RAO, S., VAN EYK, J. E. & BOHELER, K. R. 2012. A cell surfaceome map for immunophenotyping and sorting pluripotent stem cells. *Mol Cell Proteomics,* 11**,** 303-16.

GUO, G. & SMITH, A. 2010. A genome-wide screen in EpiSCs identifies Nr5a nuclear receptors as potent inducers of ground state pluripotency. *Development,* 137**,** 3185-92.

GUO, G., YANG, J., NICHOLS, J., HALL, J. S., EYRES, I., MANSFIELD, W. & SMITH, A. 2009. Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development,* 136**,** 1063-9.

GURDON, J. B. 1962. The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *J Embryol Exp Morphol,* 10**,** 622-40.

HALBRITTER, F., VAIDYA, H. J. & TOMLINSON, S. R. 2012. GeneProf: analysis of high-throughput sequencing experiments. *Nat Methods,* 9**,** 7-8.

HAN, D. W., TAPIA, N., HERMANN, A., HEMMER, K., HOING, S., ARAUZO-BRAVO, M. J., ZAEHRES, H., WU, G., FRANK, S., MORITZ, S., GREBER, B., YANG, J. H., LEE, H. T., SCHWAMBORN, J. C., STORCH, A. & SCHOLER, H. R. 2012. Direct reprogramming of fibroblasts into neural stem cells by defined factors. *Cell Stem Cell,* 10**,** 465-72.

HANNA, J., CHENG, A. W., SAHA, K., KIM, J., LENGNER, C. J., SOLDNER, F., CASSADY, J. P., MUFFAT, J., CAREY, B. W. & JAENISCH, R. 2010. Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proc Natl Acad Sci U S A,* 107**,** 9222-7.

HANNA, J., MARKOULAKI, S., MITALIPOVA, M., CHENG, A. W., CASSADY, J. P., STAERK, J., CAREY, B. W., LENGNER, C. J., FOREMAN, R., LOVE, J., GAO, Q., KIM, J. & JAENISCH, R. 2009a. Metastable pluripotent states in NOD-mouse-derived ESCs. *Cell Stem Cell,* 4**,** 513-24.

HANNA, J., MARKOULAKI, S., SCHORDERET, P., CAREY, B. W., BEARD, C., WERNIG, M., CREYGHTON, M. P., STEINE, E. J., CASSADY, J. P., FOREMAN, R., LENGNER, C. J., DAUSMAN, J. A. & JAENISCH, R. 2008. Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell,* 133**,** 250-64.

HANNA, J., SAHA, K., PANDO, B., VAN ZON, J., LENGNER, C. J., CREYGHTON, M. P., VAN OUDENAARDEN, A. & JAENISCH, R. 2009b. Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature,* 462**,** 595-601.

HANSEN, C. H. & VAN OUDENAARDEN, A. 2013. Allele-specific detection of single mRNA molecules in situ. *Nat Methods*.

HATANO, S. Y., TADA, M., KIMURA, H., YAMAGUCHI, S., KONO, T., NAKANO, T., SUEMORI, H., NAKATSUJI, N. & TADA, T. 2005. Pluripotential competence of cells associated with Nanog activity. *Mech Dev,* 122**,** 67-79.

HAYFLICK, J. S., KILGANNON, P. & GALLATIN, W. M. 1998. The intercellular adhesion molecule (ICAM) family of proteins. New members and novel functions. *Immunol Res,* 17**,** 313-27.

HENG, J. C., FENG, B., HAN, J., JIANG, J., KRAUS, P., NG, J. H., ORLOV, Y. L., HUSS, M., YANG, L., LUFKIN, T., LIM, B. & NG, H. H. 2010. The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. *Cell Stem Cell,* 6**,** 167-74.

HILL, D. P., DAVIS, A. P., RICHARDSON, J. E., CORRADI, J. P., RINGWALD, M., EPPIG, J. T. & BLAKE, J. A. 2001. Program description: Strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics. *Genomics,* 74**,** 121-8.

HOCHEDLINGER, K. & JAENISCH, R. 2002. Monoclonal mice generated by nuclear transfer from mature B and T donor cells. *Nature,* 415**,** 1035-8.

HOCKEMEYER, D., SOLDNER, F., COOK, E. G., GAO, Q., MITALIPOVA, M. & JAENISCH, R. 2008. A drug-inducible system for direct reprogramming of human somatic cells to pluripotency. *Cell Stem Cell,* 3**,** 346-53.

HONG, H., TAKAHASHI, K., ICHISAKA, T., AOI, T., KANAGAWA, O., NAKAGAWA, M., OKITA, K. & YAMANAKA, S. 2009. Suppression of induced pluripotent stem cell generation by the p53-p21 pathway. *Nature,* 460**,** 1132-5.

HOOPER, M., HARDY, K., HANDYSIDE, A., HUNTER, S. & MONK, M. 1987. HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature,* 326**,** 292-5.

HORLEY, K. J., CARPENITO, C., BAKER, B. & TAKEI, F. 1989. Molecular cloning of murine intercellular adhesion molecule (ICAM-1). *EMBO J,* 8**,** 2889-96.

HOTTA, A. & ELLIS, J. 2008. Retroviral vector silencing during iPS cell induction: an epigenetic beacon that signals distinct pluripotent states. *J Cell Biochem,* 105**,** 940-8.

HUANG DA, W., SHERMAN, B. T. & LEMPICKI, R. A. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc,* 4**,** 44-57.

HUANG, H. P., CHEN, P. H., YU, C. Y., CHUANG, C. Y., STONE, L., HSIAO, W. C., LI, C. L., TSAI, S. C., CHEN, K. Y., CHEN, H. F., HO, H. N. & KUO, H. C. 2011. Epithelial cell adhesion molecule (EpCAM) complex proteins promote transcription factor-mediated pluripotency reprogramming. *J Biol Chem,* 286**,** 33520-32.

HUANGFU, D., MAEHR, R., GUO, W., EIJKELENBOOM, A., SNITOW, M., CHEN, A. E. & MELTON, D. A. 2008a. Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nat Biotechnol,* 26**,** 795-7.

HUANGFU, D., OSAFUNE, K., MAEHR, R., GUO, W., EIJKELENBOOM, A., CHEN, S., MUHLESTEIN, W. & MELTON, D. A. 2008b. Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nat Biotechnol,* 26**,** 1269-75.

IBARRA, A., SCHWOB, E. & MENDEZ, J. 2008. Excess MCM proteins protect human cells from replicative stress by licensing backup origins of replication. *Proc Natl Acad Sci U S A,* 105**,** 8956-61.

ICHIDA, J. K., BLANCHARD, J., LAM, K., SON, E. Y., CHUNG, J. E., EGLI, D., LOH, K. M., CARTER, A. C., DI GIORGIO, F. P., KOSZKA, K., HUANGFU, D., AKUTSU, H., LIU, D. R., RUBIN, L. L. & EGGAN, K. 2009. A small-molecule inhibitor of tgf-Beta signaling replaces sox2 in reprogramming by inducing nanog. *Cell Stem Cell,* 5**,** 491-503.

IEDA, M., FU, J. D., DELGADO-OLGUIN, P., VEDANTHAM, V., HAYASHI, Y., BRUNEAU, B. G. & SRIVASTAVA, D. 2010. Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell,* 142**,** 375-86.

IRIZARRY, R. A., BOLSTAD, B. M., COLLIN, F., COPE, L. M., HOBBS, B. & SPEED, T. P. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res,* 31**,** e15.

ISLAM, S., KJALLQUIST, U., MOLINER, A., ZAJAC, P., FAN, J. B., LONNERBERG, P. & LINNARSSON, S. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res,* 21**,** 1160-7.

ISLAM, S., KJALLQUIST, U., MOLINER, A., ZAJAC, P., FAN, J. B., LONNERBERG, P. & LINNARSSON, S. 2012. Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat Protoc,* 7**,** 813-28.

ITO, S., D'ALESSIO, A. C., TARANOVA, O. V., HONG, K., SOWERS, L. C. & ZHANG, Y. 2010. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature,* 466**,** 1129-33.

JAYAWARDENA, T. M., EGEMNAZAROV, B., FINCH, E. A., ZHANG, L., PAYNE, J. A., PANDYA, K., ZHANG, Z., ROSENBERG, P., MIROTSOU, M. & DZAU, V. J. 2012. MicroRNA-mediated in vitro and in vivo direct

reprogramming of cardiac fibroblasts to cardiomyocytes. *Circ Res,* 110**,** 1465-73.

JIA, F., WILSON, K. D., SUN, N., GUPTA, D. M., HUANG, M., LI, Z., PANETTA, N. J., CHEN, Z. Y., ROBBINS, R. C., KAY, M. A., LONGAKER, M. T. & WU, J. C. 2010. A nonviral minicircle vector for deriving human iPS cells. *Nat Methods,* 7**,** 197-9.

JIANG, H., SHUKLA, A., WANG, X., CHEN, W. Y., BERNSTEIN, B. E. & ROEDER, R. G. 2011. Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalent domains. *Cell,* 144**,** 513-25.

JIANG, J., CHAN, Y. S., LOH, Y. H., CAI, J., TONG, G. Q., LIM, C. A., ROBSON, P., ZHONG, S. & NG, H. H. 2008. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol,* 10**,** 353-60.

JUDSON, R. L., BABIARZ, J. E., VENERE, M. & BLELLOCH, R. 2009. Embryonic stem cell-specific microRNAs promote induced pluripotency. *Nat Biotechnol,* 27**,** 459-61.

KAJI, K., CABALLERO, I. M., MACLEOD, R., NICHOLS, J., WILSON, V. A. & HENDRICH, B. 2006. The NuRD component Mbd3 is required for pluripotency of embryonic stem cells. *Nat Cell Biol,* 8**,** 285-92.

KAJI, K., NICHOLS, J. & HENDRICH, B. 2007. Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells. *Development,* 134**,** 1123-32.

KAJI, K., NORRBY, K., PACA, A., MILEIKOVSKY, M., MOHSENI, P. & WOLTJEN, K. 2009. Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature,* 458**,** 771-5.

KAN, H. Y., GEORGOPOULOS, S., ZANNI, M., SHKODRANI, A., TZATSOS, A., XIE, H. X. & ZANNIS, V. I. 2004. Contribution of the hormone-response elements of the proximal ApoA-I promoter, ApoCIII enhancer, and C/EBP binding site of the proximal ApoA-I promoter to the hepatic and intestinal expression of the ApoA-I and ApoCIII genes in transgenic mice. *Biochemistry,* 43**,** 5084-93.

KAWAMURA, T., SUZUKI, J., WANG, Y. V., MENENDEZ, S., MORERA, L. B., RAYA, A., WAHL, G. M. & IZPISUA BELMONTE, J. C. 2009. Linking the p53 tumour suppressor pathway to somatic cell reprogramming. *Nature,* 460**,** 1140-4.

KELLER, G. 2005. Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes Dev,* 19**,** 1129-55.

KESANAKURTI, D., CHETTY, C., RAJASEKHAR MADDIRELA, D., GUJRATI, M. & RAO, J. S. 2012. Essential role of cooperative NF-kappaB and Stat3 recruitment to ICAM-1 intronic consensus elements in the regulation of radiation-induced invasion and migration in glioma. *Oncogene*.

KIM, D., KIM, C. H., MOON, J. I., CHUNG, Y. G., CHANG, M. Y., HAN, B. S., KO, S., YANG, E., CHA, K. Y., LANZA, R. & KIM, K. S. 2009a. Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. *Cell Stem Cell,* 4**,** 472-6.

KIM, J., CHU, J., SHEN, X., WANG, J. & ORKIN, S. H. 2008a. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell,* 132**,** 1049-61.

KIM, J., EFE, J. A., ZHU, S., TALANTOVA, M., YUAN, X., WANG, S., LIPTON, S. A., ZHANG, K. & DING, S. 2011. Direct reprogramming of mouse fibroblasts to neural progenitors. *Proc Natl Acad Sci U S A,* 108**,** 7838-43.

KIM, J., ZAEHRES, H., WU, G., GENTILE, L., KO, K., SEBASTIANO, V., ARAÚZO-BRAVO, M., RUAU, D., HAN, D., ZENKE, M. & SCHÖLER, H. 2008b. Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors. *Nature,* 454**,** 646-650.

KIM, J. B., GREBER, B., ARAUZO-BRAVO, M. J., MEYER, J., PARK, K. I., ZAEHRES, H. & SCHOLER, H. R. 2009b. Direct reprogramming of human neural stem cells by OCT4. *Nature,* 461**,** 649-3.

KIM, J. B., SEBASTIANO, V., WU, G., ARAUZO-BRAVO, M. J., SASSE, P., GENTILE, L., KO, K., RUAU, D., EHRICH, M., VAN DEN BOOM, D., MEYER, J., HUBNER, K., BERNEMANN, C., ORTMEIER, C., ZENKE, M., FLEISCHMANN, B. K., ZAEHRES, H. & SCHOLER, H. R. 2009c. Oct4-induced pluripotency in adult neural stem cells. *Cell,* 136**,** 411-9.

KOCHE, R. P., SMITH, Z. D., ADLI, M., GU, H., KU, M., GNIRKE, A., BERNSTEIN, B. E. & MEISSNER, A. 2011. Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell Stem Cell,* 8**,** 96-105.

KULESSA, H., FRAMPTON, J. & GRAF, T. 1995. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev,* 9**,** 1250-62.

LASSAR, A. B., PATERSON, B. M. & WEINTRAUB, H. 1986. Transfection of a DNA locus that mediates the conversion of 10T1/2 fibroblasts to myoblasts. *Cell,* 47**,** 649-56.

LEIGHT, E. R. & SUGDEN, B. 2001. Establishment of an oriP replicon is dependent upon an infrequent, epigenetic event. *Mol Cell Biol,* 21**,** 4149-61.

LI, H., COLLADO, M., VILLASANTE, A., STRATI, K., ORTEGA, S., CANAMERO, M., BLASCO, M. A. & SERRANO, M. 2009a. The Ink4/Arf locus is a barrier for iPS cell reprogramming. *Nature,* 460**,** 1136-9.

LI, R., LIANG, J., NI, S., ZHOU, T., QING, X., LI, H., HE, W., CHEN, J., LI, F., ZHUANG, Q., QIN, B., XU, J., LI, W., YANG, J., GAN, Y., QIN, D., FENG, S., SONG, H., YANG, D., ZHANG, B., ZENG, L., LAI, L., ESTEBAN, M. A. & PEI, D. 2010. A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell,* 7**,** 51-63.

LI, W., WEI, W., ZHU, S., ZHU, J., SHI, Y., LIN, T., HAO, E., HAYEK, A., DENG, H. & DING, S. 2009b. Generation of rat and human induced pluripotent stem cells by combining genetic reprogramming and chemical inhibitors. *Cell Stem Cell,* 4**,** 16-9.

LIANG, G., TARANOVA, O., XIA, K. & ZHANG, Y. 2010. Butyrate promotes induced pluripotent stem cell generation. *J Biol Chem,* 285**,** 25516-21.

LIGGES, U. M., M. 2003. scatterplot3d - An R Package for Visualizing Multivariate Data. *Journal of Statistical Software, American Statistical Association,* 8.

LIN, K. K. & GOODELL, M. A. 2011. Detection of hematopoietic stem cells by flow cytometry. *Methods Cell Biol,* 103**,** 21-30.

LOH, Y. H., WU, Q., CHEW, J. L., VEGA, V. B., ZHANG, W., CHEN, X., BOURQUE, G., GEORGE, J., LEONG, B., LIU, J., WONG, K. Y., SUNG, K. W., LEE, C. W., ZHAO, X. D., CHIU, K. P., LIPOVICH, L., KUZNETSOV, V. A., ROBSON, P., STANTON, L. W., WEI, C. L., RUAN, Y., LIM, B. & NG, H. H. 2006. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet,* 38**,** 431-40.

LUJAN, E., CHANDA, S., AHLENIUS, H., SUDHOF, T. C. & WERNIG, M. 2012. Direct conversion of mouse fibroblasts to self-renewing, tripotent neural precursor cells. *Proc Natl Acad Sci U S A,* 109**,** 2527-32.

LUO, M., LING, T., XIE, W., SUN, H., ZHOU, Y., ZHU, Q., SHEN, M., ZONG, L., LYU, G., ZHAO, Y., YE, T., GU, J., TAO, W., LU, Z. & GRUMMT, I. 2013. NuRD blocks reprogramming of mouse somatic cells into pluripotent stem cells. *Stem Cells,* 31**,** 1278-86.

LYSSIOTIS, C. A., FOREMAN, R. K., STAERK, J., GARCIA, M., MATHUR, D., MARKOULAKI, S., HANNA, J., LAIRSON, L. L., CHARETTE, B. D., BOUCHEZ, L. C., BOLLONG, M., KUNICK, C., BRINKER, A., CHO, C. Y., SCHULTZ, P. G. & JAENISCH, R. 2009. Reprogramming of murine fibroblasts to induced pluripotent stem cells with chemical complementation of Klf4. *Proc Natl Acad Sci U S A,* 106**,** 8912-7.

MAEKAWA, M., YAMAGUCHI, K., NAKAMURA, T., SHIBUKAWA, R., KODANAKA, I., ICHISAKA, T., KAWAMURA, Y., MOCHIZUKI, H., GOSHIMA, N. & YAMANAKA, S. 2011. Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1. *Nature,* 474**,** 225-9.

MAHERALI, N., AHFELDT, T., RIGAMONTI, A., UTIKAL, J., COWAN, C. & HOCHEDLINGER, K. 2008. A high-efficiency system for the generation and study of human induced pluripotent stem cells. *Cell Stem Cell,* 3**,** 340-5.

MAHERALI, N. & HOCHEDLINGER, K. 2009. Tgfbeta signal inhibition cooperates in the induction of iPSCs and replaces Sox2 and cMyc. *Curr Biol,* 19**,** 1718-23.

MAHERALI, N., SRIDHARAN, R., XIE, W., UTIKAL, J., EMINLI, S., ARNOLD, K., STADTFELD, M., YACHECHKO, R., TCHIEU, J., JAENISCH, R., PLATH, K. & HOCHEDLINGER, K. 2007. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell,* 1**,** 55-70.

MALI, P., CHOU, B. K., YEN, J., YE, Z., ZOU, J., DOWEY, S., BRODSKY, R. A., OHM, J. E., YU, W., BAYLIN, S. B., YUSA, K., BRADLEY, A., MEYERS, D. J., MUKHERJEE, C., COLE, P. A. & CHENG, L. 2010. Butyrate greatly enhances derivation of human induced pluripotent stem cells by promoting epigenetic remodeling and the expression of pluripotency-associated genes. *Stem Cells,* 28**,** 713-20.

MALI, P., YE, Z., HOMMOND, H. H., YU, X., LIN, J., CHEN, G., ZOU, J. & CHENG, L. 2008. Improved efficiency and pace of generating induced pluripotent stem cells from human adult and fetal fibroblasts. *Stem Cells,* 26**,** 1998-2005.

MANSOUR, A. A., GAFNI, O., WEINBERGER, L., ZVIRAN, A., AYYASH, M., RAIS, Y., KRUPALNIK, V., ZERBIB, M., AMANN-ZALCENSTEIN, D., MAZA, I., GEULA, S., VIUKOV, S., HOLTZMAN, L., PRIBLUDA, A., CANAANI, E., HORN-SABAN, S., AMIT, I., NOVERSHTERN, N. & HANNA, J. H. 2012. The H3K27 demethylase Utx regulates somatic and germ cell epigenetic reprogramming. *Nature,* 488**,** 409-13.

MARION, R. M., STRATI, K., LI, H., MURGA, M., BLANCO, R., ORTEGA, S., FERNANDEZ-CAPETILLO, O., SERRANO, M. & BLASCO, M. A. 2009. A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity. *Nature,* 460**,** 1149-53.

MARTIN, G. R. 1981. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci U S A,* 78**,** 7634-8.

MASUI, S., NAKATAKE, Y., TOYOOKA, Y., SHIMOSATO, D., YAGI, R., TAKAHASHI, K., OKOCHI, H., OKUDA, A., MATOBA, R., SHAROV, A. A., KO, M. S. & NIWA, H. 2007. Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat Cell Biol,* 9**,** 625-35.

MCCLINTOCK, B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A,* 36**,** 344-55.

MEISSNER, A., WERNIG, M. & JAENISCH, R. 2007. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat Biotechnol,* 25**,** 1177-81.

MIKKELSEN, T. S., HANNA, J., ZHANG, X., KU, M., WERNIG, M., SCHORDERET, P., BERNSTEIN, B. E., JAENISCH, R., LANDER, E. S. & MEISSNER, A. 2008. Dissecting direct reprogramming through integrative genomic analysis. *Nature,* 454**,** 49-55.

MITSUI, K., TOKUZAWA, Y., ITOH, H., SEGAWA, K., MURAKAMI, M., TAKAHASHI, K., MARUYAMA, M., MAEDA, M. & YAMANAKA, S. 2003. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell,* 113**,** 631-42.

MIYANARI, Y. & TORRES-PADILLA, M. E. 2012. Control of ground-state pluripotency by allelic regulation of Nanog. *Nature,* 483**,** 470-3.

MIYOSHI, N., ISHII, H., NAGANO, H., HARAGUCHI, N., DEWI, D. L., KANO, Y., NISHIKAWA, S., TANEMURA, M., MIMORI, K., TANAKA, F., SAITO, T., NISHIMURA, J., TAKEMASA, I., MIZUSHIMA, T., IKEDA, M., YAMAMOTO, H., SEKIMOTO, M., DOKI, Y. & MORI, M. 2011. Reprogramming of mouse and human cells to pluripotency using mature microRNAs. *Cell Stem Cell,* 8**,** 633-8.

NAJM, F. J., CHENOWETH, J. G., ANDERSON, P. D., NADEAU, J. H., REDLINE, R. W., MCKAY, R. D. & TESAR, P. J. 2011. Isolation of epiblast stem cells from preimplantation mouse embryos. *Cell Stem Cell,* 8**,** 318-25.

NAKAGAWA, M., KOYANAGI, M., TANABE, K., TAKAHASHI, K., ICHISAKA, T., AOI, T., OKITA, K., MOCHIDUKI, Y., TAKIZAWA, N. & YAMANAKA, S. 2008. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol,* 26**,** 101-6.

NANBO, A., SUGDEN, A. & SUGDEN, B. 2007. The coupling of synthesis and partitioning of EBV's plasmid replicon is revealed in live cells. *EMBO J,* 26**,** 4252-62.

NAOR, D., WALLACH-DAYAN, S. B., ZAHALKA, M. A. & SIONOV, R. V. 2008. Involvement of CD44, a molecule with a thousand faces, in cancer dissemination. *Semin Cancer Biol,* 18**,** 260-7.

NAVARRO, P., CHAMBERS, I., KARWACKI-NEISIUS, V., CHUREAU, C., MOREY, C., ROUGEULLE, C. & AVNER, P. 2008. Molecular coupling of Xist regulation and pluripotency. *Science,* 321**,** 1693-5.

NAVARRO, P., FESTUCCIA, N., COLBY, D., GAGLIARDI, A., MULLIN, N. P., ZHANG, W., KARWACKI-NEISIUS, V., OSORNO, R., KELLY, D., ROBERTSON, M. & CHAMBERS, I. 2012. OCT4/SOX2-independent Nanog autorepression modulates heterogeneous Nanog gene expression in mouse ES cells. *EMBO J,* 31**,** 4547-62.

NAVARRO, P., OLDFIELD, A., LEGOUPI, J., FESTUCCIA, N., DUBOIS, A., ATTIA, M., SCHOORLEMMER, J., ROUGEULLE, C., CHAMBERS, I. & AVNER, P. 2010. Molecular coupling of Tsix regulation and pluripotency. *Nature,* 468**,** 457-60.

NEMAJEROVA, A., KIM, S. Y., PETRENKO, O. & MOLL, U. M. 2012. Two-factor reprogramming of somatic cells to pluripotent stem cells reveals partial functional redundancy of Sox2 and Klf4. *Cell Death Differ,* 19**,** 1268-76.

NICHOLS, J., JONES, K., PHILLIPS, J. M., NEWLAND, S. A., ROODE, M., MANSFIELD, W., SMITH, A. & COOKE, A. 2009a. Validated germline-competent embryonic stem cell lines from nonobese diabetic mice. *Nat Med,* 15**,** 814-8.

NICHOLS, J., SILVA, J., ROODE, M. & SMITH, A. 2009b. Suppression of Erk signalling promotes ground state pluripotency in the mouse embryo. *Development,* 136**,** 3215-22.

NISHIMURA, K., SANO, M., OHTAKA, M., FURUTA, B., UMEMURA, Y., NAKAJIMA, Y., IKEHARA, Y., KOBAYASHI, T., SEGAWA, H., TAKAYASU, S., SATO, H., MOTOMURA, K., UCHIDA, E., KANAYASU-TOYODA, T., ASASHIMA, M., NAKAUCHI, H., YAMAGUCHI, T. & NAKANISHI, M. 2011. Development of defective and persistent Sendai virus vector: a unique gene delivery/expression system ideal for cell reprogramming. *J Biol Chem,* 286**,** 4760-71.

NIWA, H., BURDON, T., CHAMBERS, I. & SMITH, A. 1998. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev,* 12**,** 2048-60.

NIWA, H., MIYAZAKI, J. & SMITH, A. G. 2000. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet,* 24**,** 372-6.

O'MALLEY, J., SKYLAKI, S., IWABUCHI, K. A., CHANTZOURA, E., RUETZ, T., JOHNSSON, A., TOMLINSON, S. R., LINNARSSON, S. & KAJI, K. 2013. High-resolution analysis with novel cell-surface markers identifies routes to iPS cells. *Nature,* 499**,** 88-91.

O'MALLEY, J., WOLTJEN, K. & KAJI, K. 2009. New strategies to generate induced pluripotent stem cells. *Curr Opin Biotechnol,* 20**,** 516-21.

OKITA, K., ICHISAKA, T. & YAMANAKA, S. 2007. Generation of germline-competent induced pluripotent stem cells. *Nature,* 448**,** 313-7.

OKITA, K., MATSUMURA, Y., SATO, Y., OKADA, A., MORIZANE, A., OKAMOTO, S., HONG, H., NAKAGAWA, M., TANABE, K., TEZUKA, K., SHIBATA, T., KUNISADA, T., TAKAHASHI, M., TAKAHASHI, J., SAJI, H. & YAMANAKA, S. 2011. A more efficient method to generate integration-free human iPS cells. *Nat Methods,* 8**,** 409-12.

OKUMURA-NAKANISHI, S., SAITO, M., NIWA, H. & ISHIKAWA, F. 2005. Oct-3/4 and Sox2 regulate Oct-3/4 gene in embryonic stem cells. *J Biol Chem,* 280**,** 5307-17.

OLIFERENKO, S., PAIHA, K., HARDER, T., GERKE, V., SCHWARZLER, C., SCHWARZ, H., BEUG, H., GUNTHERT, U. & HUBER, L. A. 1999. Analysis of CD44-containing lipid rafts: Recruitment of annexin II and stabilization by the actin cytoskeleton. *J Cell Biol,* 146**,** 843-54.

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., LANDER, E. S., ARMSTRONG, S. A. & DALEY, G. Q. 2012. Chromatin-modifying enzymes as modulators of reprogramming. *Nature,* 483**,** 598-602.

PAPAPETROU, E. P. & SADELAIN, M. 2011. Generation of transgene-free human induced pluripotent stem cells with an excisable single polycistronic vector. *Nat Protoc,* 6**,** 1251-73.

PAPAPETROU, E. P., TOMISHIMA, M. J., CHAMBERS, S. M., MICA, Y., REED, E., MENON, J., TABAR, V., MO, Q., STUDER, L. & SADELAIN, M. 2009. Stoichiometric and temporal requirements of Oct4, Sox2, Klf4, and c-Myc expression for efficient human iPSC induction and differentiation. *Proc Natl Acad Sci U S A,* 106**,** 12759-64.

PARK, I. H., ZHAO, R., WEST, J. A., YABUUCHI, A., HUO, H., INCE, T. A., LEROU, P. H., LENSCH, M. W. & DALEY, G. Q. 2008. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature,* 451**,** 141-6.

PAWLAK, M. & JAENISCH, R. 2011. De novo DNA methylation by Dnmt3a and Dnmt3b is dispensable for nuclear reprogramming of somatic cells to a pluripotent state. *Genes Dev,* 25**,** 1035-40.

PEREIRA, C. F., TERRANOVA, R., RYAN, N. K., SANTOS, J., MORRIS, K. J., CUI, W., MERKENSCHLAGER, M. & FISHER, A. G. 2008. Heterokaryon-based reprogramming of human B lymphocytes for pluripotency requires Oct4 but not Sox2. *PLoS Genet,* 4**,** e1000170.

PFISTERER, U., KIRKEBY, A., TORPER, O., WOOD, J., NELANDER, J., DUFOUR, A., BJORKLUND, A., LINDVALL, O., JAKOBSSON, J. & PARMAR, M. 2011. Direct conversion of human fibroblasts to dopaminergic neurons. *Proc Natl Acad Sci U S A,* 108**,** 10343-8.

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., BAR-NUR, O., CHELOUFI, S., STADTFELD, M., FIGUEROA, M. E., ROBINTON, D., NATESAN, S., MELNICK, A., ZHU, J., RAMASWAMY, S. & HOCHEDLINGER, K. 2012. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell,* 151**,** 1617-32.

PRUITT, S. C., BAILEY, K. J. & FREELAND, A. 2007. Reduced Mcm2 expression results in severe stem/progenitor cell deficiency and cancer. *Stem Cells,* 25**,** 3121-32.

QIAN, L., HUANG, Y., SPENCER, C. I., FOLEY, A., VEDANTHAM, V., LIU, L., CONWAY, S. J., FU, J. D. & SRIVASTAVA, D. 2012. In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature,* 485**,** 593-8.

QIANG, L., FUJITA, R., YAMASHITA, T., ANGULO, S., RHINN, H., RHEE, D., DOEGE, C., CHAU, L., AUBRY, L., VANTI, W. B., MORENO, H. & ABELIOVICH, A. 2011. Directed conversion of Alzheimer's disease patient skin fibroblasts into functional neurons. *Cell,* 146**,** 359-71.

RAIS, Y., ZVIRAN, A., GEULA, S., GAFNI, O., CHOMSKY, E., VIUKOV, S., MANSOUR, A. A., CASPI, I., KRUPALNIK, V., ZERBIB, M., MAZA, I., MOR, N., BARAN, D., WEINBERGER, L., JAITIN, D. A., LARA-ASTIASO, D., BLECHER-GONEN, R., SHIPONY, Z., MUKAMEL, Z., HAGAI, T., GILAD, S., AMANN-ZALCENSTEIN, D., TANAY, A., AMIT, I., NOVERSHTERN, N. & HANNA, J. H. 2013. Deterministic direct reprogramming of somatic cells to pluripotency. *Nature,* 502**,** 65-70.

REUBINOFF, B. E., PERA, M. F., FONG, C. Y., TROUNSON, A. & BONGSO, A. 2000. Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. *Nat Biotechnol,* 18**,** 399-404.

RING, K. L., TONG, L. M., BALESTRA, M. E., JAVIER, R., ANDREWS-ZWILLING, Y., LI, G., WALKER, D., ZHANG, W. R., KREITZER, A. C. & HUANG, Y. 2012. Direct reprogramming of mouse and human fibroblasts into multipotent neural stem cells with a single factor. *Cell Stem Cell,* 11**,** 100-9.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-40.

RODDA, D. J., CHEW, J. L., LIM, L. H., LOH, Y. H., WANG, B., NG, H. H. & ROBSON, P. 2005. Transcriptional regulation of nanog by OCT4 and SOX2. *J Biol Chem,* 280**,** 24731-7.

ROTHLEIN, R., DUSTIN, M. L., MARLIN, S. D. & SPRINGER, T. A. 1986. A human intercellular adhesion molecule (ICAM-1) distinct from LFA-1. *J Immunol,* 137**,** 1270-4.

RUIZ, P., SCHWARZLER, C. & GUNTHERT, U. 1995. CD44 isoforms during differentiation and development. *Bioessays,* 17**,** 17-24.

SALDANHA, A. J. 2004. Java Treeview--extensible visualization of microarray data. *Bioinformatics,* 20**,** 3246-8.

SAMAVARCHI-TEHRANI, P., GOLIPOUR, A., DAVID, L., SUNG, H. K., BEYER, T. A., DATTI, A., WOLTJEN, K., NAGY, A. & 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.

SCLAFANI, R. A. & HOLZEN, T. M. 2007. Cell cycle regulation of DNA replication. *Annu Rev Genet,* 41**,** 237-80.

SCREATON, G. R., BELL, M. V., JACKSON, D. G., CORNELIS, F. B., GERTH, U. & BELL, J. I. 1992. Genomic structure of DNA encoding the lymphocyte homing receptor CD44 reveals at least 12 alternatively spliced exons. *Proc Natl Acad Sci U S A,* 89**,** 12160-4.

SEKI, T., YUASA, S., ODA, M., EGASHIRA, T., YAE, K., KUSUMOTO, D., NAKATA, H., TOHYAMA, S., HASHIMOTO, H., KODAIRA, M., OKADA, Y., SEIMIYA, H., FUSAKI, N., HASEGAWA, M. & FUKUDA, K. 2010. Generation of induced pluripotent stem cells from human terminally differentiated circulating T cells. *Cell Stem Cell,* 7**,** 11-4.

SHAO, L., FENG, W., SUN, Y., BAI, H., LIU, J., CURRIE, C., KIM, J., GAMA, R., WANG, Z., QIAN, Z., LIAW, L. & WU, W. S. 2009. Generation of iPS cells using defined factors linked via the self-cleaving 2A sequences in a single open reading frame. *Cell Res,* 19**,** 296-306.

SHI, Y., DESPONTS, C., DO, J. T., HAHM, H. S., SCHOLER, H. R. & DING, S. 2008a. Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds. *Cell Stem Cell,* 3**,** 568-74.

SHI, Y., DO, J. T., DESPONTS, C., HAHM, H. S., SCHOLER, H. R. & DING, S. 2008b. A combined chemical and genetic approach for the generation of induced pluripotent stem cells. *Cell Stem Cell,* 2**,** 525-8.

SHIMA, N., ALCARAZ, A., LIACHKO, I., BUSKE, T. R., ANDREWS, C. A., MUNROE, R. J., HARTFORD, S. A., TYE, B. K. & SCHIMENTI, J. C. 2007. A viable allele of Mcm4 causes chromosome instability and mammary adenocarcinomas in mice. *Nat Genet,* 39**,** 93-8.

SILVA, J., BARRANDON, O., NICHOLS, J., KAWAGUCHI, J., THEUNISSEN, T. W. & SMITH, A. 2008. Promotion of reprogramming to ground state pluripotency by signal inhibition. *PLoS Biol,* 6**,** e253.

SILVA, J., CHAMBERS, I., POLLARD, S. & SMITH, A. 2006. Nanog promotes transfer of pluripotency after cell fusion. *Nature,* 441**,** 997-1001.

SILVA, J., NICHOLS, J., THEUNISSEN, T. W., GUO, G., VAN OOSTEN, A. L., BARRANDON, O., WRAY, J., YAMANAKA, S., CHAMBERS, I. & SMITH, A. 2009. Nanog is the gateway to the pluripotent ground state. *Cell,* 138**,** 722-37.

SINGHAL, N., GRAUMANN, J., WU, G., ARAUZO-BRAVO, M. J., HAN, D. W., GREBER, B., GENTILE, L., MANN, M. & SCHOLER, H. R. 2010. Chromatin-Remodeling Components of the BAF Complex Facilitate Reprogramming. *Cell,* 141**,** 943-55.

SLIGH, J. E., JR., BALLANTYNE, C. M., RICH, S. S., HAWKINS, H. K., SMITH, C. W., BRADLEY, A. & BEAUDET, A. L. 1993. Inflammatory and immune responses are impaired in mice deficient in intercellular adhesion molecule 1. *Proc Natl Acad Sci U S A,* 90**,** 8529-33.

SMITH, A. G. 1991. Culture and differentiation of embryonic stem cells. *Methods in Cell Science,* 13**,** 89-94.

SMITH, A. G., HEATH, J. K., DONALDSON, D. D., WONG, G. G., MOREAU, J., STAHL, M. & ROGERS, D. 1988. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature,* 336**,** 688-90.

SOLDNER, F., HOCKEMEYER, D., BEARD, C., GAO, Q., BELL, G. W., COOK, E. G., HARGUS, G., BLAK, A., COOPER, O., MITALIPOVA, M., ISACSON, O. & JAENISCH, R. 2009. Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell,* 136**,** 964-77.

SOLTER, D. & KNOWLES, B. B. 1978. Monoclonal antibody defining a stage-specific mouse embryonic antigen (SSEA-1). *Proc Natl Acad Sci U S A,* 75**,** 5565-9.

SOMERS, A., JEAN, J. C., SOMMER, C. A., OMARI, A., FORD, C. C., MILLS, J. A., YING, L., SOMMER, A. G., JEAN, J. M., SMITH, B. W., LAFYATIS, R., DEMIERRE, M. F., WEISS, D. J., FRENCH, D. L., GADUE, P., MURPHY, G. J., MOSTOSLAVSKY, G. & KOTTON, D. N. 2010. Generation of transgene-free lung disease-specific human induced pluripotent stem cells using a single excisable lentiviral stem cell cassette. *Stem Cells,* 28**,** 1728-40.

SOMMER, C. A. & MOSTOSLAVSKY, G. 2010. Experimental approaches for the generation of induced pluripotent stem cells. *Stem Cell Res Ther,* 1**,** 26.

SOMMER, C. A., STADTFELD, M., MURPHY, G. J., HOCHEDLINGER, K., KOTTON, D. N. & MOSTOSLAVSKY, G. 2009. Induced pluripotent stem cell generation using a single lentiviral stem cell cassette. *Stem Cells,* 27**,** 543-9.

SON, E. Y., ICHIDA, J. K., WAINGER, B. J., TOMA, J. S., RAFUSE, V. F., WOOLF, C. J. & EGGAN, K. 2011. Conversion of mouse and human fibroblasts into functional spinal motor neurons. *Cell Stem Cell,* 9**,** 205-18.

SONG, K., NAM, Y. J., LUO, X., QI, X., TAN, W., HUANG, G. N., ACHARYA, A., SMITH, C. L., TALLQUIST, M. D., NEILSON, E. G., HILL, J. A., BASSEL-DUBY, R. & OLSON, E. N. 2012. Heart repair by reprogramming non-myocytes with cardiac transcription factors. *Nature,* 485**,** 599-604.

SOUFI, A., DONAHUE, G. & ZARET, K. S. 2012. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell,* 151**,** 994-1004.

SRIDHARAN, R., TCHIEU, J., MASON, M. J., YACHECHKO, R., KUOY, E., HORVATH, S., ZHOU, Q. & PLATH, K. 2009. Role of the murine reprogramming factors in the induction of pluripotency. *Cell,* 136**,** 364-77.

STADTFELD, M., APOSTOLOU, E., AKUTSU, H., FUKUDA, A., FOLLETT, P., NATESAN, S., KONO, T., SHIODA, T. & HOCHEDLINGER, K. 2010a. Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature,* 465**,** 175-81.

STADTFELD, M., APOSTOLOU, E., FERRARI, F., CHOI, J., WALSH, R. M., CHEN, T., OOI, S. S., KIM, S. Y., BESTOR, T. H., SHIODA, T., PARK, P. J. & HOCHEDLINGER, K. 2012. Ascorbic acid prevents loss of Dlk1-Dio3

imprinting and facilitates generation of all-iPS cell mice from terminally differentiated B cells. *Nat Genet,* 44**,** 398-405, S1-2.

STADTFELD, M., MAHERALI, N., BORKENT, M. & HOCHEDLINGER, K. 2010b. A reprogrammable mouse strain from gene-targeted embryonic stem cells. *Nat Methods,* 7**,** 53-5.

STADTFELD, M., MAHERALI, N., BREAULT, D. T. & HOCHEDLINGER, K. 2008a. Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse. *Cell Stem Cell,* 2**,** 230-40.

STADTFELD, M., NAGAYA, M., UTIKAL, J., WEIR, G. & HOCHEDLINGER, K. 2008b. Induced pluripotent stem cells generated without viral integration. *Science,* 322**,** 945-9.

STAMENKOVIC, I., AMIOT, M., PESANDO, J. M. & SEED, B. 1989. A lymphocyte molecule implicated in lymph node homing is a member of the cartilage link protein family. *Cell,* 56**,** 1057-62.

STAUNTON, D. E., DUSTIN, M. L., ERICKSON, H. P. & SPRINGER, T. A. 1990. The arrangement of the immunoglobulin-like domains of ICAM-1 and the binding sites for LFA-1 and rhinovirus. *Cell,* 61**,** 243-54.

SUBRAMANYAM, D., LAMOUILLE, S., JUDSON, R. L., LIU, J. Y., BUCAY, N., DERYNCK, R. & BLELLOCH, R. 2011. Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells. *Nat Biotechnol,* 29**,** 443-8.

SZABO, E., RAMPALLI, S., RISUENO, R. M., SCHNERCH, A., MITCHELL, R., FIEBIG-COMYN, A., LEVADOUX-MARTIN, M. & BHATIA, M. 2010. Direct conversion of human fibroblasts to multilineage blood progenitors. *Nature,* 468**,** 521-6.

TADA, M., TAKAHAMA, Y., ABE, K., NAKATSUJI, N. & TADA, T. 2001. Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. *Curr Biol,* 11**,** 1553-8.

TAKAHASHI, E., NAGANO, O., ISHIMOTO, T., YAE, T., SUZUKI, Y., SHINODA, T., NAKAMURA, S., NIWA, S., IKEDA, S., KOGA, H., TANIHARA, H. & SAYA, H. 2010. Tumor necrosis factor-alpha regulates transforming growth factor-beta-dependent epithelial-mesenchymal transition by promoting hyaluronan-CD44-moesin interaction. *J Biol Chem,* 285**,** 4060-73.

TAKAHASHI, K., TANABE, K., OHNUKI, M., NARITA, M., ICHISAKA, T., TOMODA, K. & YAMANAKA, S. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell,* 131**,** 861-72.

TAKAHASHI, K. & YAMANAKA, S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell,* 126**,** 663-76.

TANAKA, T. S., KUNATH, T., KIMBER, W. L., JARADAT, S. A., STAGG, C. A., USUDA, M., YOKOTA, T., NIWA, H., ROSSANT, J. & KO, M. S. 2002. Gene expression profiling of embryo-derived stem cells reveals candidate genes associated with pluripotency and lineage specificity. *Genome Res,* 12**,** 1921-8.

TESAR, P. J., CHENOWETH, J. G., BROOK, F. A., DAVIES, T. J., EVANS, E. P., MACK, D. L., GARDNER, R. L. & MCKAY, R. D. 2007. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature,* 448**,** 196-9.

THEUNISSEN, T. W., VAN OOSTEN, A. L., CASTELO-BRANCO, G., HALL, J., SMITH, A. & SILVA, J. C. 2011. Nanog overcomes reprogramming barriers and induces pluripotency in minimal conditions. *Curr Biol,* 21**,** 65-71.

THOMSON, J. A., ITSKOVITZ-ELDOR, J., SHAPIRO, S. S., WAKNITZ, M. A., SWIERGIEL, J. J., MARSHALL, V. S. & JONES, J. M. 1998. Embryonic stem cell lines derived from human blastocysts. *Science,* 282**,** 1145-7.

TIAN, L., CATT, J. W., O'NEILL, C. & KING, N. J. 1997. Expression of immunoglobulin superfamily cell adhesion molecules on murine embryonic stem cells. *Biol Reprod,* 57**,** 561-8.

TZATSOS, A., PFAU, R., KAMPRANIS, S. C. & TSICHLIS, P. N. 2009. Ndy1/KDM2B immortalizes mouse embryonic fibroblasts by repressing the Ink4a/Arf locus. *Proc Natl Acad Sci U S A,* 106**,** 2641-6.

UTIKAL, J., POLO, J. M., STADTFELD, M., MAHERALI, N., KULALERT, W., WALSH, R. M., KHALIL, A., RHEINWALD, J. G. & HOCHEDLINGER, K. 2009. Immortalization eliminates a roadblock during cellular reprogramming into iPS cells. *Nature,* 460**,** 1145-8.

VIERBUCHEN, T., OSTERMEIER, A., PANG, Z. P., KOKUBU, Y., SUDHOF, T. C. & WERNIG, M. 2010. Direct conversion of fibroblasts to functional neurons by defined factors. *Nature,* 463**,** 1035-41.

VISVADER, J. E. & LINDEMAN, G. J. 2008. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. *Nat Rev Cancer,* 8**,** 755-68.

WABL, M. R., BRUN, R. B. & DU PASQUIER, L. 1975. Lymphocytes of the toad Xenopus laevis have the gene set for promoting tadpole development. *Science,* 190**,** 1310-2.

WANG, S., SHEN, Y., YUAN, X., CHEN, K., GUO, X., CHEN, Y., NIU, Y., LI, J., XU, R. H., YAN, X., ZHOU, Q. & JI, W. 2008a. Dissecting signaling pathways that govern self-renewal of rabbit embryonic stem cells. *J Biol Chem,* 283**,** 35929-40.

WANG, T., CHEN, K., ZENG, X., YANG, J., WU, Y., SHI, X., QIN, B., ZENG, L., ESTEBAN, M. A., PAN, G. & PEI, D. 2011a. The histone demethylases Jhdm1a/1b enhance somatic cell reprogramming in a vitamin-C-dependent manner. *Cell Stem Cell,* 9**,** 575-87.

WANG, W., LIN, C., LU, D., NING, Z., COX, T., MELVIN, D., WANG, X., BRADLEY, A. & LIU, P. 2008b. Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc Natl Acad Sci U S A,* 105**,** 9290-5.

WANG, W., YANG, J., LIU, H., LU, D., CHEN, X., ZENONOS, Z., CAMPOS, L. S., RAD, R., GUO, G., ZHANG, S., BRADLEY, A. & LIU, P. 2011b. Rapid and efficient reprogramming of somatic cells to induced pluripotent stem cells by retinoic acid receptor gamma and liver receptor homolog 1. *Proc Natl Acad Sci U S A,* 108**,** 18283-8.

WARREN, L., MANOS, P. D., AHFELDT, T., LOH, Y. H., LI, H., LAU, F., EBINA, W., MANDAL, P. K., SMITH, Z. D., MEISSNER, A., DALEY, G. Q., BRACK, A. S., COLLINS, J. J., COWAN, C., SCHLAEGER, T. M. & ROSSI, D. J. 2010. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell,* 7**,** 618-30.

WERNIG, M., LENGNER, C. J., HANNA, J., LODATO, M. A., STEINE, E., FOREMAN, R., STAERK, J., MARKOULAKI, S. & JAENISCH, R. 2008. A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. *Nat Biotechnol,* 26**,** 916-24.

WERNIG, M., MEISSNER, A., FOREMAN, R., BRAMBRINK, T., KU, M., HOCHEDLINGER, K., BERNSTEIN, B. E. & JAENISCH, R. 2007. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature,* 448**,** 318-24.

WICKER, L. S., TODD, J. A. & PETERSON, L. B. 1995. Genetic control of autoimmune diabetes in the NOD mouse. *Annu Rev Immunol,* 13**,** 179-200.

WIELENGA, V. J., SMITS, R., KORINEK, V., SMIT, L., KIELMAN, M., FODDE, R., CLEVERS, H. & PALS, S. T. 1999. Expression of CD44 in Apc and Tcf mutant mice implies regulation by the WNT pathway. *Am J Pathol,* 154**,** 515-23.

WILLIAMS, R. L., HILTON, D. J., PEASE, S., WILLSON, T. A., STEWART, C. L., GEARING, D. P., WAGNER, E. F., METCALF, D., NICOLA, N. A. &

GOUGH, N. M. 1988. Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature,* 336**,** 684-7.

WILMUT, I., SCHNIEKE, A. E., MCWHIR, J., KIND, A. J. & CAMPBELL, K. H. 1997. Viable offspring derived from fetal and adult mammalian cells. *Nature,* 385**,** 810-3.

WOLTJEN, K., MICHAEL, I. P., MOHSENI, P., DESAI, R., MILEIKOVSKY, M., HAMALAINEN, R., COWLING, R., WANG, W., LIU, P., GERTSENSTEIN, M., KAJI, K., SUNG, H. K. & NAGY, A. 2009. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature,* 458**,** 766-70.

WU, S., YING, G., WU, Q. & CAPECCHI, M. R. 2007. Toward simpler and faster genome-wide mutagenesis in mice. *Nat Genet,* 39**,** 922-30.

XU, H., LEMISCHKA, I. R. & MA'AYAN, A. 2010. SVM classifier to predict genes important for self-renewal and pluripotency of mouse embryonic stem cells. *BMC Syst Biol,* 4**,** 173.

YAMAJI, M., UEDA, J., HAYASHI, K., OHTA, H., YABUTA, Y., KURIMOTO, K., NAKATO, R., YAMADA, Y., SHIRAHIGE, K. & SAITOU, M. 2013. PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells. *Cell Stem Cell,* 12**,** 368-82.

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

YE, Z., ZHAN, H., MALI, P., DOWEY, S., WILLIAMS, D. M., JANG, Y. Y., DANG, C. V., SPIVAK, J. L., MOLITERNO, A. R. & CHENG, L. 2009. Human-induced pluripotent stem cells from blood cells of healthy donors and patients with acquired blood disorders. *Blood,* 114**,** 5473-80.

YING, Q. L., NICHOLS, J., CHAMBERS, I. & SMITH, A. 2003. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell,* 115**,** 281-92.

YING, Q. L., WRAY, J., NICHOLS, J., BATLLE-MORERA, L., DOBLE, B., WOODGETT, J., COHEN, P. & SMITH, A. 2008. The ground state of embryonic stem cell self-renewal. *Nature,* 453**,** 519-23.

YU, J., HU, K., SMUGA-OTTO, K., TIAN, S., STEWART, R., SLUKVIN, II & THOMSON, J. A. 2009. Human induced pluripotent stem cells free of vector and transgene sequences. *Science,* 324**,** 797-801.

YU, J. & THOMSON, J. A. 2008. Pluripotent stem cell lines. *Genes Dev,* 22**,** 1987-97.

YU, J., VODYANIK, M. A., SMUGA-OTTO, K., ANTOSIEWICZ-BOURGET, J., FRANE, J. L., TIAN, S., NIE, J., JONSDOTTIR, G. A., RUOTTI, V., STEWART, R., SLUKVIN, II & THOMSON, J. A. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science,* 318**,** 1917-20.

YUSA, K., RAD, R., TAKEDA, J. & BRADLEY, A. 2009. Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nat Methods,* 6**,** 363-9.

ZHANG, J., TAM, W. L., TONG, G. Q., WU, Q., CHAN, H. Y., SOH, B. S., LOU, Y., YANG, J., MA, Y., CHAI, L., NG, H. H., LUFKIN, T., ROBSON, P. & LIM, B. 2006. Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1. *Nat Cell Biol,* 8**,** 1114-23.

ZHAO, Y., YIN, X., QIN, H., ZHU, F., LIU, H., YANG, W., ZHANG, Q., XIANG, C., HOU, P., SONG, Z., LIU, Y., YONG, J., ZHANG, P., CAI, J., LIU, M., LI, H., LI, Y., QU, X., CUI, K., ZHANG, W., XIANG, T., WU, Y., LIU, C., YU, C., YUAN, K., LOU, J., DING, M. & DENG, H. 2008. Two supporting factors greatly improve the efficiency of human iPSC generation. *Cell Stem Cell,* 3**,** 475-9.

ZHOU, H., WU, S., JOO, J. Y., ZHU, S., HAN, D. W., LIN, T., TRAUGER, S., BIEN, G., YAO, S., ZHU, Y., SIUZDAK, G., SCHOLER, H. R., DUAN, L. & DING, S. 2009. Generation of induced pluripotent stem cells using recombinant proteins. *Cell Stem Cell,* 4**,** 381-4.

ZHOU, Q., BROWN, J., KANAREK, A., RAJAGOPAL, J. & MELTON, D. A. 2008. In vivo reprogramming of adult pancreatic exocrine cells to beta-cells. *Nature,* 455**,** 627-32.

ZHOU, W. & FREED, C. R. 2009. Adenoviral gene delivery can reprogram human fibroblasts to induced pluripotent stem cells. *Stem Cells,* 27**,** 2667-74.

Appendix: Relevant Publication

High resolution analysis with novel cell-surface markers identifies routes to iPS cells.

LETTER

doi:10.1038/nature12243

High-resolution analysis with novel cell-surface markers identifies routes to iPS cells

James O'Malley[1], Stavroula Skylaki[2], Kumiko A. Iwabuchi[1], Eleni Chantzoura[1], Tyson Ruetz[1], Anna Johnsson[3], Simon R. Tomlinson[1], Sten Linnarsson[3] & Keisuke Kaji[1]

The generation of induced pluripotent stem (iPS) cells presents a challenge to normal developmental processes. The low efficiency and heterogeneity of most methods have hindered understanding of the precise molecular mechanisms promoting, and roadblocks preventing, efficient reprogramming. Although several intermediate populations have been described[1–7], it has proved difficult to characterize the rare, asynchronous transition from these intermediate stages to iPS cells. The rapid expansion of minor reprogrammed cells in the heterogeneous population can also obscure investigation of relevant transition processes. Understanding the biological mechanisms essential for successful iPS cell generation requires both accurate capture of cells undergoing the reprogramming process and identification of the associated global gene expression changes. Here we demonstrate that in mouse embryonic fibroblasts, reprogramming follows an orderly sequence of stage transitions, marked by changes in the cell-surface markers CD44 and ICAM1, and a Nanog–enhanced green fluorescent protein (Nanog–eGFP) reporter. RNA-sequencing analysis of these populations demonstrates two waves of pluripotency gene upregulation, and unexpectedly, transient upregulation of several epidermis-related genes, demonstrating that reprogramming is not simply the reversal of the normal developmental processes. This novel high-resolution analysis enables the construction of a detailed reprogramming route map, and the improved understanding of the reprogramming process will lead to new reprogramming strategies.

Several reports have suggested that reprogramming progresses in an ordered manner[3,5,6,8–10]. To identify markers whose expression changed concurrent with pluripotency gene expression, we performed time course microarray analysis using a piggyBac transposon-based secondary reprogramming system[3,11] (Supplementary Fig. 2a). Of a number of candidate cell-surface markers, *Cd44* and *Icam1* (also known as *CD54*) demonstrated the most dynamic expression changes throughout secondary mouse embryonic fibroblast (MEF) reprogramming (Supplementary Fig. 2b). For further investigation, we generated an efficient secondary reprogramming system in which doxycycline-mediated induction of the reprogramming factors could be monitored by an mOrange reporter placed after the 2A-peptide-linked reprogramming cassette *c-Myc-Klf4-Oct4-Sox2* (MKOS)[12], and endogenous *Nanog* promoter activation could be followed by expression of enhanced green fluorescent protein (eGFP)[13] (Supplementary Fig 3). Reprogramming cultures were supplemented with vitamin C and an Alk inhibitor, both of which enhance reprogramming efficiency[10,14,15]. In this secondary reprogramming system, Nanog–eGFP$^+$ cells appeared as early as day 6, and >60% of mOrange$^+$ transgene-expressing cells were found to be Nanog–eGFP$^+$ by day 12 (Supplementary Figs 4 and 5a). Most mOrange$^+$ transgene-expressing cells lost expression of Thy1 (also known as CD90) and gained E-cadherin (also known as Cdh1) expression by day 4 (Supplementary Fig. 5b, c). Expression of stage-specific embryonic antigen 1 (SSEA-1, also known as Fut4) barely changed after day 8, with a gradual gain of Nanog–eGFP$^+$ cells in both SSEA-1$^+$ and SSEA-1$^-$ cell populations (Supplementary Fig. 5d). Consistent with heterogeneous expression of SSEA1 in iPS and embryonic stem (ES) cells, it was not possible to delineate the reprogramming process accurately using SSEA-1 (Supplementary Fig. 6). By contrast, the appearance of CD44$^-$ and ICAM1$^+$ cells at later time points closely correlated with Nanog–eGFP expression (Supplementary Fig. 5e, f). Double staining for CD44 and ICAM1 revealed that a distinct series of population changes occur during reprogramming (Fig. 1). Initially, MEFs displayed high CD44 and broad ICAM1 expression, with most becoming ICAM1$^-$ by day 6, along with the appearance of a minor CD44$^-$ ICAM1$^-$ cell population. By day 8, CD44$^-$ populations appeared enriched, and at day 12 almost all cells displayed an iPS/ES-cell-like CD44$^-$ ICAM1$^+$ profile, of which more than 60% expressed Nanog–eGFP. Consistent with the observation that Nanog expression is not necessarily a sign of completed reprogramming[16], Nanog–eGFP$^+$ cells were observed even before cells obtained this iPS/ES-cell-like phenotype (CD44$^-$ ICAM1$^+$). Both ICAM1$^+$- and ICAM1$^-$-sorted MEFs demonstrated similar fluorescence-activated cell sorting (FACS) profile changes during reprogramming (Supplementary Fig. 7). Immunofluorescence for CD44 and ICAM1 revealed that reprogramming is not synchronized even within individual colonies (Supplementary Fig. 8). Secondary reprogramming of the non-polycistronic iPS cell line 6c (refs 3, 11) and primary reprogramming using MKOS and *Oct4-P2A-Sox2-T2A-Klf4-E2A-cMyc* (OSKM)[17] piggyBac transposons resulted in similar ICAM1 and CD44 profile changes, indicating their suitability for use in other systems and contexts (Supplementary Fig. 9). These findings demonstrated the asynchronous but stepwise manner of reprogramming, and highlighted the potential usefulness of CD44 and ICAM1 to isolate intermediate reprogramming subpopulations.

Next, we aimed to confirm that the observed CD44/ICAM1 profile changes reflected the transition of individual cells from one stage to the next, and not merely the loss of one major population and expansion of another minor population. CD44$^+$ ICAM1$^-$ (gate 1), CD44$^-$ ICAM1$^-$ (gate 2) and CD44$^-$ ICAM1$^+$ (gate 3) cell populations, either Nanog–eGFP$^+$ (that is, 1NG$^+$, 2 NG$^+$ and 3NG$^+$) or Nanog–eGFP$^-$ (1NG$^-$, 2NG$^-$ and 3NG$^-$), were isolated by cell-sorting at day 10 of reprogramming and re-plated in reprogramming conditions (Fig. 2a). After 3 days, both NG$^+$ and NG$^-$ cells progressed in the order of gates 1 to 2 to 3 (Fig. 2b). This progression correlated well with increased Nanog–eGFP$^+$ colony-forming potential (c.f.p.), with 3NG$^+$ cells displaying similar clonogenicity to fully reprogrammed iPS cells (Fig. 2c). Of cells with the same CD44/ICAM1 profile, Nanog–eGFP expression correlated with a higher c.f.p. (for example, 1NG$^-$ versus 1NG$^+$).

To examine the progression of the reprogramming process more accurately, cells from each gate were sorted, and their expression of CD44/ICAM1/Nanog–eGFP was re-analysed after 24 h (Fig. 2d). On the basis of total cell numbers in each gate after 24 h (Supplementary Fig. 10), we generated a reprogramming route map representing differences in the

[1]MRC Centre for Regenerative Medicine, University of Edinburgh, Edinburgh BioQuarter, 5 Little France Drive, Edinburgh EH16 4UU, UK. [2]Stem Cell Dynamics Research Unit, Helmholtz Center Munich, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany. [3]Laboratory for Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institute, Scheeles väg 1, SE-171 77 Stockholm, Sweden.

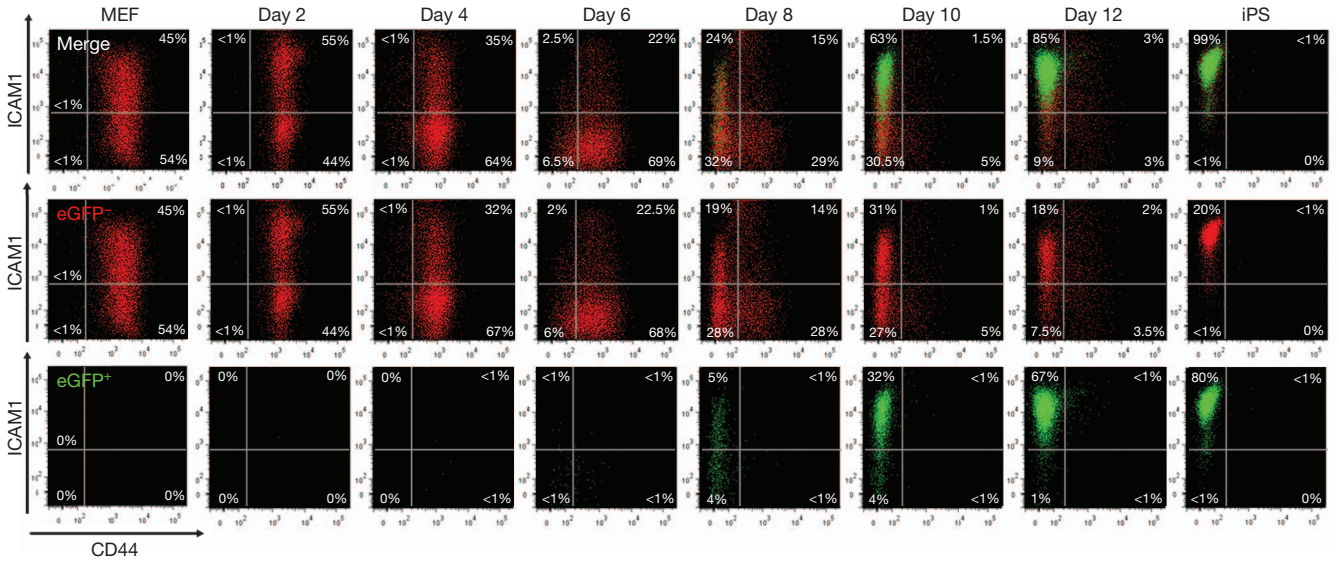


Figure 1 | FACS analysis during secondary reprogramming of MEFs with CD44/ICAM1 double staining. Loss of CD44 expression was rapidly followed by ICAM1 upregulation and Nanog–eGFP expression. By day 12, most cells displayed an ICAM$^+$/CD44$^-$ ES-cell-like profile. Red denotes Nanog–eGFP$^-$ cells; green denotes Nanog–eGFP$^+$ cells.

efficiency of these stage transitions and in Nanog–eGFP$^+$ c.f.p. (Fig. 2e). Similar results were obtained when each subpopulation was sorted at day 8 (Supplementary Fig. 11). This analysis revealed that reaching a Nanog–eGFP$^+$ state is a rate-limiting step—as few cells overcame this barrier in the 24 h assay—and those that do so reprogram more efficiently than their Nanog–eGFP$^-$ counterparts, consistent with the role of *Nanog* as an accelerator of reprogramming and the gateway to pluripotency[18,19].

To determine global gene expression changes during these stage transitions, we carried out RNA-sequencing analysis using a highly multiplexed sample bar-coding system[20–26] (see Methods and Supplementary Table 1). Hierarchical clustering using the complete list of differentially expressed genes (DEGs) revealed four major branches: (1) MEFs; (2) 1NG$^{-/+}$ and 2NG$^-$; (3) 2NG$^{-/+}$ and 3NG$^{-/+}$; and (4) 3NG$^+$ sorted at day 15 (3NG$^+$D15), iPS and ES cells (Fig. 3a). There was a prominent gene expression difference between 3NG$^+$ and 3NG$^+$D15 cells, with the latter being more similar to iPS and ES cells (Fig. 3a and Supplementary Fig. 12), possibly reflecting the observed difference in the c.f.p. in the absence of doxycycline (Supplementary

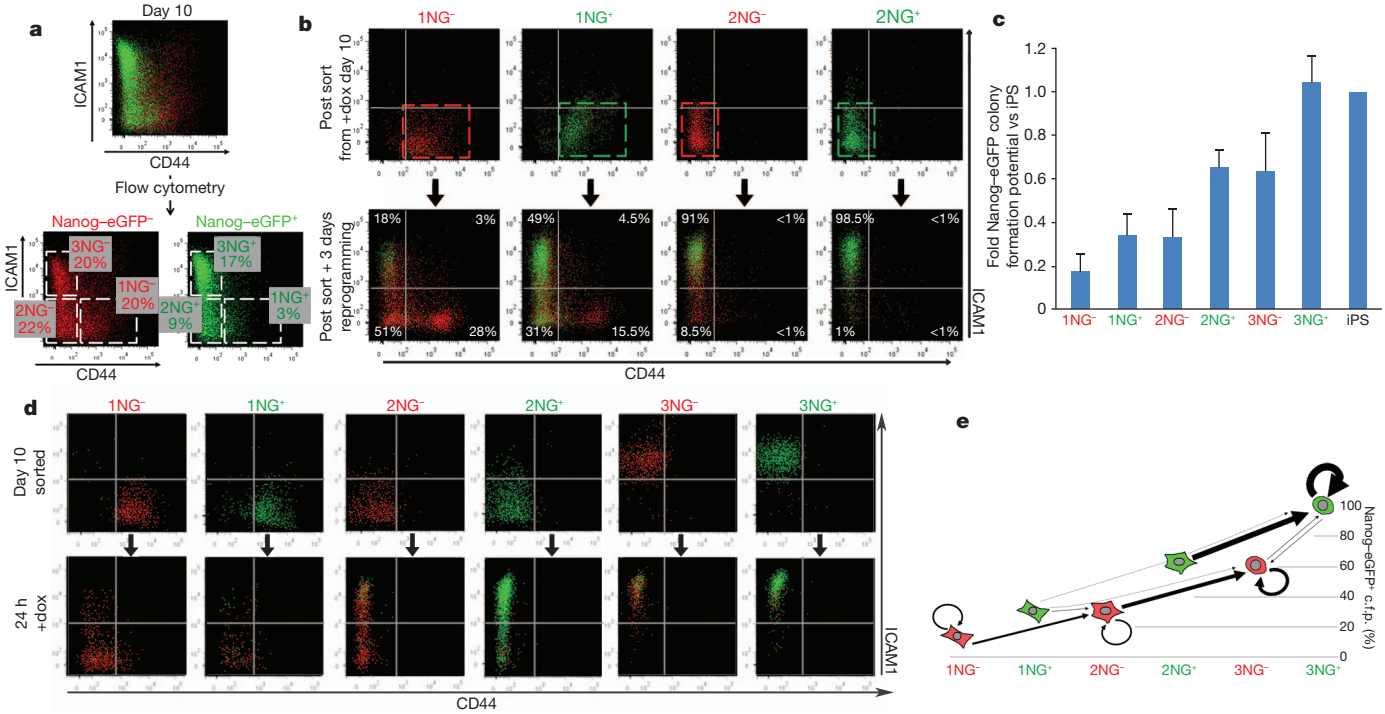


Figure 2 | CD44/ICAM1 subpopulations represent distinct stages of reprogramming. **a**, Nanog–eGFP$^+$ (NG$^+$) and Nanog–eGFP$^-$ (NG$^-$) cells were subdivided into CD44$^+$ ICAM1$^-$ (gate 1), CD44$^-$ ICAM1$^-$ (gate 2) and CD44$^-$ ICAM1$^+$ (gate 3) populations at day 10 of reprogramming. **b**, FACS analysis of sorted subpopulations after a 3-day culture in the presence of doxycycline (dox). **c**, Relative probability to generate Nanog–eGFP$^+$ iPS cell colonies from each subpopulation compared to fully reprogrammed iPS cells. Error bars represent s.d., $n = 3$. **d**, Expression of CD44, ICAM1 and Nanog–eGFP was re-analysed 24 h after sorting. **e**, Major transitions (>500 cells) of each population within 24 h. The *y* axis indicates relative c.f.p. after a further 10 days. Arrow size reflects relative cell numbers.

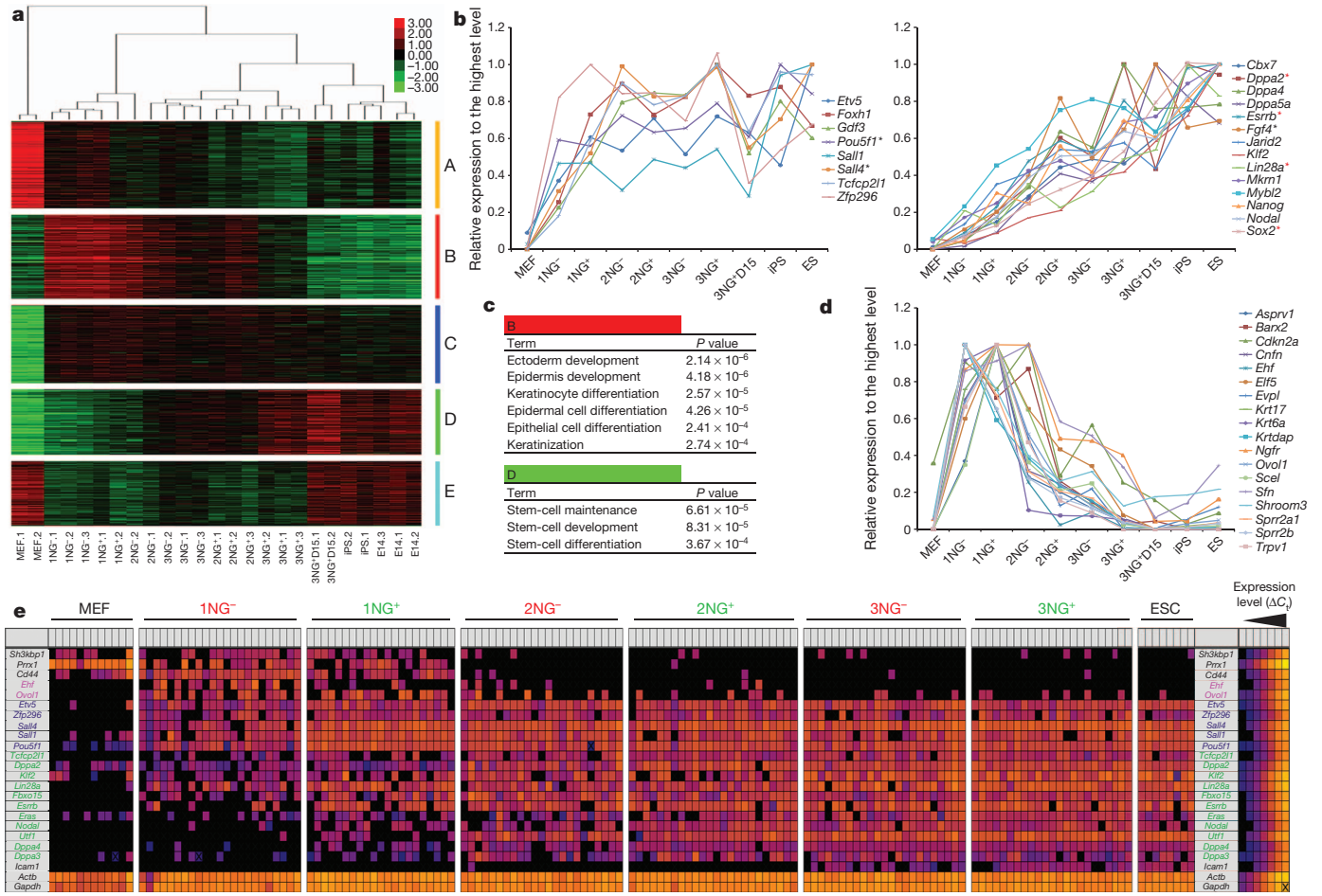


Figure 3 | Global gene expression changes during the stage transition. **a**, Hierarchical clustering of samples with DEGs and expression heat map. Groups A–E represent different expression patterns. **b**, Early (left) and late (right) upregulation of pluripotency-related genes. Black and red asterisks indicate early and late pluripotency genes, respectively, previously identified by single-cell quantitative PCR (qPCR)[5]. **c**, Epidermal and stem-cell gene enrichment in gene list B and D, respectively. **d**, Transient upregulation of 18 epidermis/keratinocyte-related genes during reprogramming. **e**, Single-cell gene expression analysis. Each square represents one reaction chamber from one cell. Colour corresponds to ΔC_t value, as shown in the legend.

Fig. 13). The DEGs between these two populations may be involved in the establishment of an exogenous-factor-independent self-renewal state. Principal component analysis clearly distinguished $2NG^+$ from $3NG^-$ cells, consistent with the higher probability of the former to reach the $3NG^+$ state within 24 h (Supplementary Figs 10 and 12b). DEGs could be classified into five distinct expression pattern groups (A–E) (Fig. 3a and Supplementary Tables 2 and 3). Group A contained readily downregulated fibroblast-related genes. Group D comprised factors gradually upregulated towards iPS cells, in which ES cell genes were highly enriched ($P \leq 0.000367$) (Fig. 3c). However group C, which contained genes upregulated at early stages and maintained throughout reprogramming, also included some pluripotency-related factors. To extend this finding, we examined the expression pattern of 22 pluripotency-related genes in our data set[27,28]. Interestingly, 8 pluripotency genes, including endogenous *Oct4* (also known as *Pou5f1*), were already upregulated at the $1NG^+/2NG^-$ stages to the level found in $3NG^+$ cells (Fig. 3b, left), whereas 14 pluripotency genes were more gradually upregulated in the later stage reprogramming populations (Fig. 3b, right, and Supplementary Table 4). This early and late pluripotency gene upregulation was confirmed at the single cell level[5] (Fig. 3e), highlighting the high resolution of the CD44/ICAM1 sorting system.

We also identified two additional gene expression patterns displaying transient upregulation (group B) or downregulation (group E) exclusively in the intermediate stages of reprogramming. This finding indicates that reprogramming from MEFs to iPS cells is not simply the loss of MEF genes and gain of ES cell genes. Gene Ontology analysis revealed that genes related to ectoderm/epidermis development and keratinocyte differentiation were highly enriched in group B ($P \leq 0.000274$) (Fig. 3c, d and Supplementary Tables 3–5). Although SFN and KRT17 were barely detectable by immunofluorescence in MEFs and iPS cells, transient upregulation was observed in the intermediate stages of reprogramming (Supplementary Fig. 14). Single-cell PCR confirmed the co-expression of epidermis genes (*Ehf* and *Ovol1*) with early pluripotency genes in the $1NG^{-/+}$ stage (Fig. 3e). Consistent with our data, analysis of three published microarray data sets incorporating partially reprogrammed iPS cells[1], a time course experiment[3] and a subpopulation analysis with Thy1, SSEA-1 and Oct4–eGFP (ref. 6) confirmed transient epidermal gene expression during reprogramming (Supplementary Figs 15–17 and Supplementary Tables 6–8). Partially reprogrammed cells from B cells also displayed similar epidermis gene expression[4], whereas two factor-reprogramming (Oct4 and Sox2) of MEFs did not[29]. Therefore, this intermediate state could be a consequence of the use of Klf4 that is important for efficient reprogramming, and demonstrates that the reprogramming process is not simply a reversion of normal differentiation (summarized in Supplementary Fig. 1). It would be intriguing to investigate whether similar transient gene expression changes can be seen in reprogramming of ectoderm or endoderm lineages. Down-regulation of these epidermis genes coincided with upregulation of 'late' pluripotency genes. Future examination of this rapid switch in gene expression may provide a new insight into the molecular mechanism of reprogramming.

The integrative data analysis described above demonstrated that this CD44/ICAM1/Nanog–eGFP marker system could uniquely provide high-resolution information during late pluripotency gene upregulation, enabling the discrimination of 'reprogramming' from 'expansion of reprogrammed cells' (Fig. 3b and Supplementary Figs 16b and 17f). This system also refines investigation of the kinetics of reprogramming. It has recently been shown that vitamin C increases reprogramming efficiency by facilitating histone 3 Lys 9 (H3K9) demethylation[7], and that reprogramming factors fail to bind trimethylated H3K9-rich regions in the initial stages of reprogramming[30]. We carried out reprogramming in the absence of vitamin C and observed not only a decrease in the iPS cell colony number, but also a marked delay in the transition from one stage of reprogramming to the next (Supplementary Fig. 18). Similar analyses can be performed using our marker system to investigate the mechanism of action of other factors that alter reprogramming efficiency. Isolation and analysis of subpopulations affected by these factors could reveal the downstream genes specifically involved in, and required for, successful reprogramming. Further studies using this high-resolution analysis system have the potential to make a considerable contribution towards revealing the molecular mechanisms of reprogramming.

METHODS SUMMARY

The vector PB-TAP IRI 2LMKOSimO, a modified version of polycistronic reprogramming vector pCAG2LMKOSimO (ref. 12), containing insulator and replicator sequences and driven by the $tetO_2$ promoter, was constructed as described in the Methods. This vector was used to generate iPS cell line D6s4B5 from reverse tetracycline transactivator (rtTA)-expressing MEFs carrying a Nanog–eGFP reporter[13]. D6s4B5 iPS cells were used to generate chimaeric embryos from which MEFs were isolated at embryonic day 12.5. Transgenic MEFs were cultured in doxycycline (300 ng ml^{-1}), vitamin C (10 μg ml^{-1}) and Alk inhibitor (500 nM), and collected for flow cytometry analysis (BD Fortessa), carried out using antibodies for CD44 and ICAM1 every 2–3 days. Cells were sorted (BD FACS Aria II) at day 10 or 15, and replated on gelatin for analysis at 24 h, or at clonal density on irradiated MEFs for Nanog–eGFP^{+} c.f.p. 10 days after cell sorting. All flow cytometry data were analysed using FlowJo (Tree Star). Immunofluorescence was carried out using confocal microscopy (Leica TSC SP2). RNA from sorted samples was extracted using Trizol (Invitrogen), and 10 ng total RNA was used for multiplexed RNA-sequencing[20,21]. Data were analysed using GeneProf[22], and DEGs were identified using edgeR and DESeq Bioconductor libraries[23–25]. Gene Ontology enrichment was calculated using DAVID[26].

Full Methods and any associated references are available in the online version of the paper.

Received 19 September 2012; accepted 3 May 2013.
Published online 2 June 2013.

1. Sridharan, R. *et al.* Role of the murine reprogramming factors in the induction of pluripotency. *Cell* **136,** 364–377 (2009).
2. 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).
3. 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).
4. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454,** 49–55 (2008).
5. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150,** 1209–1222 (2012).
6. Polo, J. M. *et al.* A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* **151,** 1617–1632 (2012).
7. Chen, J. *et al.* H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nature Genet.* **45,** 34–42 (2013).
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. Brambrink, T. *et al.* Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell* **2,** 151–159 (2008).
10. 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).
11. Woltjen, K. *et al.* piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **458,** 766–770 (2009).
12. Kaji, K. *et al.* Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature* **458,** 771–775 (2009).
13. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450,** 1230–1234 (2007).
14. 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).
15. Maherali, N. & Hochedlinger, K. Tgfβ signal inhibition cooperates in the induction of iPSCs and replaces Sox2 and cMyc. *Curr. Biol.* **19,** 1718–1723 (2009).
16. Chan, E. M. *et al.* Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells. *Nature Biotechnol.* **27,** 1033–1037 (2009).
17. Carey, B. W. *et al.* Reprogramming of murine and human somatic cells using a single polycistronic vector. *Proc. Natl Acad. Sci. USA* **106,** 157–162 (2009).
18. Hanna, J. *et al.* Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462,** 595–601 (2009).
19. Silva, J. *et al.* Nanog is the gateway to the pluripotent ground state. *Cell* **138,** 722–737 (2009).
20. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21,** 1160–1167 (2011).
21. Islam, S. *et al.* Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nature Protocols* **7,** 813–828 (2012).
22. Halbritter, F., Vaidya, H. J. & Tomlinson, S. R. GeneProf: analysis of high-throughput sequencing experiments. *Nature Methods* **9,** 7–8 (2012).
23. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
24. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
25. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
26. Huang, d. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).
27. Kim, J., Chu, J., Shen, X., Wang, J. & Orkin, S. H. An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132,** 1049–1061 (2008).
28. Xu, H., Lemischka, I. R. & Ma'ayan, A. SVM classifier to predict genes important for self-renewal and pluripotency of mouse embryonic stem cells. *BMC Syst. Biol.* **4,** 173 (2010).
29. Nemajerova, A., Kim, S. Y., Petrenko, O. & Moll, U. M. Two-factor reprogramming of somatic cells to pluripotent stem cells reveals partial functional redundancy of Sox2 and Klf4. *Cell Death Differ.* **19,** 1268–1276 (2012).
30. 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).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A. Nagy and K. Woltjen for providing the 6c iPS cell line, I. Chambers for providing TNG mice, S. Monard and O. Rodrigues for assistance with flow cytometry, and T. Kunath, T. Burdon, S. Lowell and N. Festuccia for discussions and comments on the manuscript. We also thank L. Robertson for technical assistance, and Biomed unit staff for mouse husbandry. This work was supported by ERC grants ROADTOIPS (no. 261075) and BRAINCELL (no. 261063), and the Anne Rowling Regenerative Neurology Clinic. J.O.'M. and T.R. are funded by an MRC PhD Studentship and a Darwin Trust of Edinburgh Scholarship, respectively.

Author Contributions J.O.'M. designed and performed flow cytometry analysis and sorting experiments, prepared RNA for sequencing, carried out immunofluorescence imaging, and collected, analysed and interpreted data, and wrote the manuscript. S.S. analysed RNA-sequencing and published microarray data sets. K.A.I. carried out single-cell PCR analysis. E.C. performed primary reprogramming and FACS analysis. T.R. carried out immunofluorescence and confocal imaging. S.R.T. performed microarray analysis to identify cell-surface marker candidates. A.J. and S.L. performed multiplexed RNA-sequencing and collected data. K.K. conceived the study, identified the surface markers, generated the D6s4B5 iPS cell line, analysed RNA-sequencing data, supervised experiment design and data interpretation, and wrote the manuscript.

Author Information RNA-sequencing data are deposited in the ArrayExpress under accession number E-MTAB-1654. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.K. (keisuke.kaji@ed.ac.uk).

METHODS

Vector construction. The piggyBac transposon PB-TAP containing the $tetO_2$ promoter, an attR1R2 Gateway cloning cassette (Invitrogen) and rabbit β-globin poly A signal, was provided by A. Nagy. To minimize silencing of the reprogramming vector, a chicken β-globin insulator[31] was inserted into the PacI site between the piggyBac 3′-terminal repeat (3′-TR) and the $tetO_2$ promoter, and a human lamin B2 (LMB2) replicator[32] plus another chicken β-globin insulator were inserted into the EcoRV site between the rabbit β-globin poly A signal and the piggyBac 5′-TR, to generate PB-TAP IRI. The BamHI fragment containing *loxP*-flanked MKOS reprogramming cassette followed by ires-mOrange (2LMKOSimO) from pCAG2LMKOSimO (ref. 12) was inserted into a Gateway entry vector pENTR 2B (Life Technologies), to generate attP2LMKOSimO pENTR. Finally the attP2LMKOSimO cassette was Gateway-cloned into the PB-TAP IRI to yield reprogramming piggyBac transposon PB-TAP IRI attP2LMKOSimO. Similarly, reprogramming piggyBac transposon PB-TAP IRI 2LOSKMimO was generated after transferring the OSKM reprogramming cassette[17] into attP2LMKOSimO pENTR replacing the MKOS cassette. Plasmid sequences are available on request.

Generation of a primary iPS cell line D6s4B5. Embryos at 12.5 days post coitum (d.p.c.) were obtained from $Rosa^{rtTA/rtTA}$, $Nanog^{eGFP/+}$, $Col1a1^{+/+}$ mice, which were derived by crossing TNG mice[13] and B6;129-$Gt(ROSA)26Sor^{tm1(rtTA*M2)Jae}$ $Col1a1^{tm2(tetO\text{-}Pou5f1)Jae}$/J (Jackson Laboratory). The embryos were decapitated, eviscerated, dissociated with 0.25% trypsin and 0.1% EDTA, and plated in MEF medium (GMEM, 10% FBS, penicillin–streptomycin, 1× non-essential amino acids (Invitrogen), 1 mM sodium pyruvate and 0.05 mM 2-mercaptoethanol). The PB-TAP IRI attP2LMKOSimO (500 ng) and pCyL43 piggyBac transposase expression vector[33] (2 μg) were introduced into the MEFs by nucleofection (Amaxa) as before[12], and cells were cultured in ES cell medium (MEF medium supplemented with 1,000 U ml^{-1} leukaemia inhibiting factor (LIF)) in the presence of 1.0 μg ml^{-1} doxycycline (Sigma) for an initial 8 days, and thereafter 0.5 μg ml^{-1} doxycycline. Pluripotency of a clonal iPS cell line D6 was confirmed by teratoma formation, and a subclone D6s4B5 was used for secondary reprogramming. To compare CD44 and ICAM1 profiles of primary reprogramming with PB-TAP IRI attP2LMKOSimO and PB-TAP IRI 2LOSKMimO vectors, MEFs were nucleofected as above and cultured in the presence of 1.0 μg ml^{-1} doxycycline, 10 μg ml^{-1} vitamin C (Sigma) and 500 nM Alk inhibitor A 83-01 (TOCRIS Bioscience).

Secondary reprogramming. Each chimaeric embryo was collected at 12.5 d.p.c., dissociated and cultured in MEF medium. One-twentieth of the dissociated cells were exposed to doxycycline (300 ng ml^{-1}) for 2 days, and the proportion of transgenic MEFs was measured by FACS analysis of mOrange expression. For FACS time course and colony counting experiments, secondary transgenic MEFs were diluted to 5% and 30% by addition of 129 wild-type MEFs and plated in a gelatinized 6-well-plate at 1×10^5 cells per well (5,000 and 30,000 transgenic MEFs per well, respectively). For sorting experiments, MEFs were plated at 2×10^5 cells per gelatinized 100 mm plate (1×10^4 transgenic MEFs per plate). Cells were cultured in reprogramming medium, which is ES cell medium supplemented with 300 ng ml^{-1} doxycycline, 10 μg ml^{-1} vitamin C and 500 nM Alk inhibitor. Medium was changed every 2 days.

Flow cytometry and cell sorting. Cell-surface marker analysis was performed with the following eBioscience antibodies: ICAM-1-biotin (13-0541; 1/100), CD44-biotin (17-0441; 1/100), CD44- allophycocyanin (APC) (17-0441; 1/300), streptavidin-phycoerythrin (PE)-Cy7 (25-4317-82; 1/1500), SSEA-1-647 (51-8813; 1/50), E-cadherin-biotin (13-3249; 1/100), Thy1-APC (17-0902, 1/300) and CD2-biotin (13-0029; 1/100). For sorting experiments, dead cells were excluded using 4′,6-diamidino-2-phenylindole (DAPI) nucleic acid stain (Invitrogen) (0.5 ng ml^{-1}). Cells were incubated in 0.25% trypsin and 1 mM EDTA (Life Technologies) for 1–2 min at 37 °C, collected in GMEM media containing 10% FCS and counted. Staining was carried out in FACS buffer (2% FCS in PBS) at $\sim 1 \times 10^6$ cells ml^{-1} for 15–30 min at 4 °C, and followed by washing with FACS buffer, sorting and/or analysis with FACSAriaII and LSRFortessa (both BD Biosciences), respectively. Excitation laser lines and filters used for each fluorophore are summarized in Supplementary Table 9. Data were analysed using FlowJo software (Tree Star). Intact cells were identified based on forward and side light scatter, and subsequently analysed for fluorescence intensity. Additional gating was carried out as outlined in Supplementary Fig. 2. For colony formation assays, sorted cells were plated on γ-irradiated MEFs in 12-well plates at 3.5×10^3 cells per well. Nanog–$eGFP^+$ colonies were quantified 10 days after sorting. For 24 h or time-course analysis, sorted cells were plated in gelatinized 48-well plate at 1×10^4 cells per well. In both cases, cells were cultured in reprogramming medium after sorting.

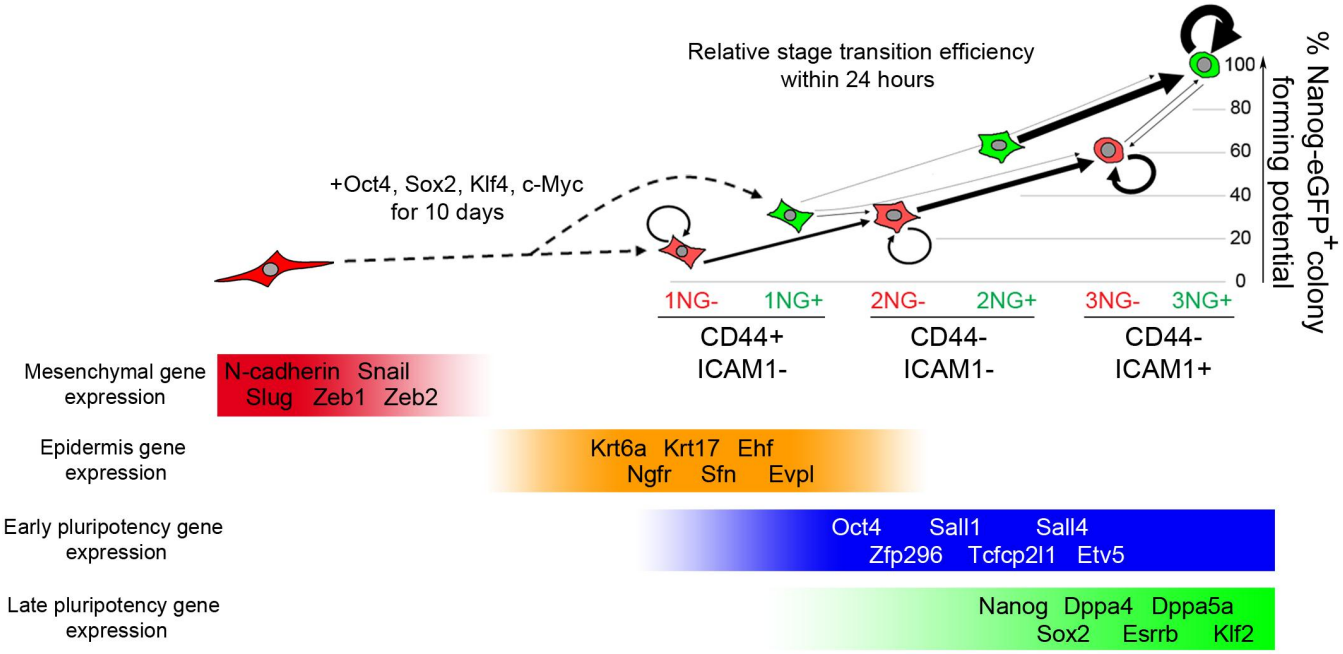
Immunofluorescenceand confocal microscopy imaging. Images of cells stained with ICAM-1-biotin (1/100), CD44-APC (1/300) and streptavidin-PE-Cy7 (1/1,500) antibodies described above were captured with a confocal microscope (Leica TSC SP2) and Leica confocal software. Cells stained with anti-Krt17 (LifeSpan BioSciences) and anti-Sfn (Sigma) antibodies and anti-Rabbit IgG CF633 secondary antibody (Sigma) were imaged with a fluorescence microscopy (Olympus).

Multiplexed RNA sequencing and data analysis. RNA was isolated with TRI reagent (Sigma) following the manufacturer's instructions. RNA quality and concentration was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies). Using 10 ng RNA, reverse transcription with bar-coded primers, complementary DNA amplification, and sequencing with Illumina HiSeq 2000 were performed as previously described[20,21]. Quality control of the obtained reads and alignment to the mouse reference genome (NCBI37/mm9) were performed using the GeneProf web-based analysis suite with default parameters[22]. Gene expression read counts were exported and analysed in R to identify DEGs, using the edgeR and DESeq Bioconductor libraries[23–25]. For both methods, low expression transcripts (less than 13 reads in all samples) were filtered out, and *P* values were adjusted using a threshold for false discovery rate (FDR) ≤ 0.05. Genes listed as DEGs by both methods in any two subpopulation comparison indicated in Supplementary Table 1 and Supplementary Fig. 12a (total 3,171) were used for further analysis. Hierarchical clustering and *K*-means clustering ($K = 5$) was performed using Cluster 3.0, and Java Treeview was used for visualization[34,35]. This multiplexed RNA-sequencing technology reads only the 5′ end of transcript, thus detecting only endogenous *Oct4* and *Sox2*. *Nanog* expression was detectable in Nanog–$eGFP^-$ populations owing to the reporter system. Principal components analysis was performed in R and plotted with the scatterplot3d library[36]. Gene Ontology enrichment was calculated using the DAVID functional annotation bioinformatics tool[26]. Gene Ontology term enrichment analysis was carried out with a modified Fisher exact *P* value. The three additional published studies[1,3,6] (GEO accession numbers GSE21757, GSE14012 and GSE42379) were analysed in a similar way. For the time course data, the analysis was performed as following: data were robust multi-array average (RMA)[37] normalized using the expression console from Affymetrix and, because no replicates were provided, fold changes between two samples were calculated in Excel. Genes with more than 1.5-fold changes were classified as DEGs. For the Plath and Polo data set, data were RMA-normalized using the 'affy' package[38] in R, and DEGs were identified using the 'limma' package[38] in R with fold change ≥ 1.5 and FDR ≤ 0.05, or fold change ≥ 1.5 where no replicates were available. Subsequently, *K*-means clustering of the identified DEGs was performed for all studies. Selected gene expression data are shown as the relative expression against the highest signal among the samples using an averaged signal value (reads per million) of duplicates/triplicates.

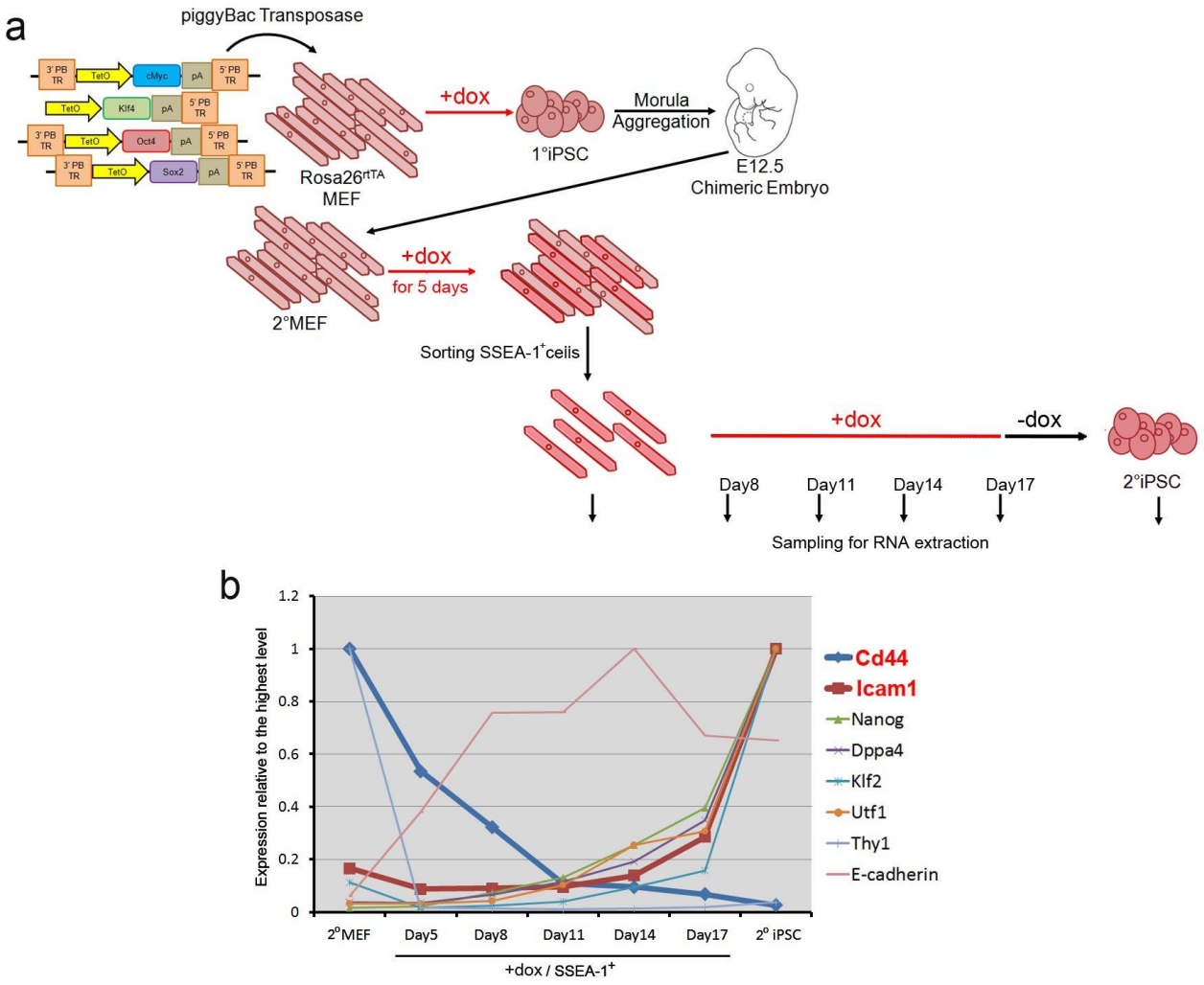
Single-cell gene expression analysis. Single-cell qPCR was performed as described previously[5] with slight modifications. In brief, 22 sets of TaqMan gene expression assays (Applied Biosystems; Supplementary Table 9) were pooled at a final concentration of 180 nM per primer set and 50 μM per probe. Individual cells were sorted directly into 10 μl RT-PreAmp Master Mix (5 μl of CellsDirect reaction mix (Invitrogen), 2.5 μl of pooled assays, 0.2 μl of SuperScript III (Invitrogen), 1.3 μl of water) using FACSAria II. Cell lysis and sequence-specific reverse transcription were performed at 50 °C for 15 min. Reverse transcriptase was inactivated by heating to 95 °C for 2 min. Subsequently, in the same tube, cDNA went through sequence-specific amplification by denaturing at 95 °C for 15 s, and annealing and amplification at 60 °C for 4 min for 22 cycles. Preamplified products were diluted fivefold with water and analysed in 48.48 dynamic arrays on a biomark system (Fluidigm) following the Fluidigm protocol. C_t values were calculated and visualized using BioMark real-time PCR analysis software (Fluidigm). Each assay was performed in replicate.

31. Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Rev. Genet.* **7,** 703–713 (2006).
32. Fu, H. *et al.* Preventing gene silencing with human replicators. *Nature Biotechnol.* **24,** 572–576 (2006).
33. Wang, W. *et al.* Chromosomal transposition of *PiggyBac* in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA* **105,** 9290–9295 (2008).
34. Saldanha, A. J. Java Treeview–extensible visualization of microarray data. *Bioinformatics* **20,** 3246–3248 (2004).
35. de Hoon, M. J., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20,** 1453–1454 (2004).
36. Ligges, U. & Maechler, M. scatterplot3d — an R package for visualizing multivariate data. *J. Stat. Softw.* **8,** 1–20 (2003).
37. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31,** e15 (2003).
38. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy–analysis of *Affymetrix GeneChip* data at the probe level. *Bioinformatics* **20,** 307–315 (2004).

doi:10.1038/nature12243

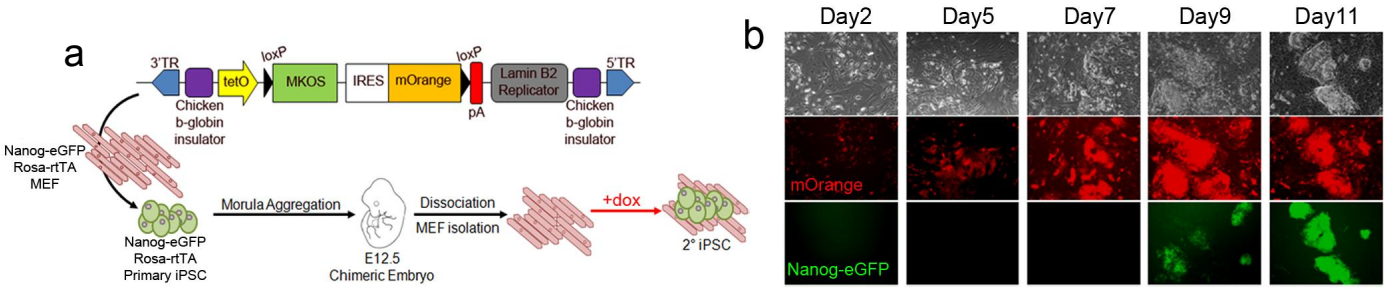


Supplementary Figure 1. A Route map to iPSCs defined by CD44 and ICAM1 expression change. Sequential changes of CD44 and ICAM1 expression allowed isolation of subpopulations at different stages of reprogramming, the progression of which was accompanied by a gradual increase of iPSC colony formation efficiency. Differences in the transition rate of each subpopulation from one stage to the next revealed preferential routes to iPSCs. Global gene expression profiling highlighted transient up-regulation of multiple epidermis genes. There were two groups of pluripotency genes; those displaying early initiation of expression which overlapped with epidermis gene expression (Early) and those which were up-regulated in parallel with down-regulation of epidermis genes (Late).

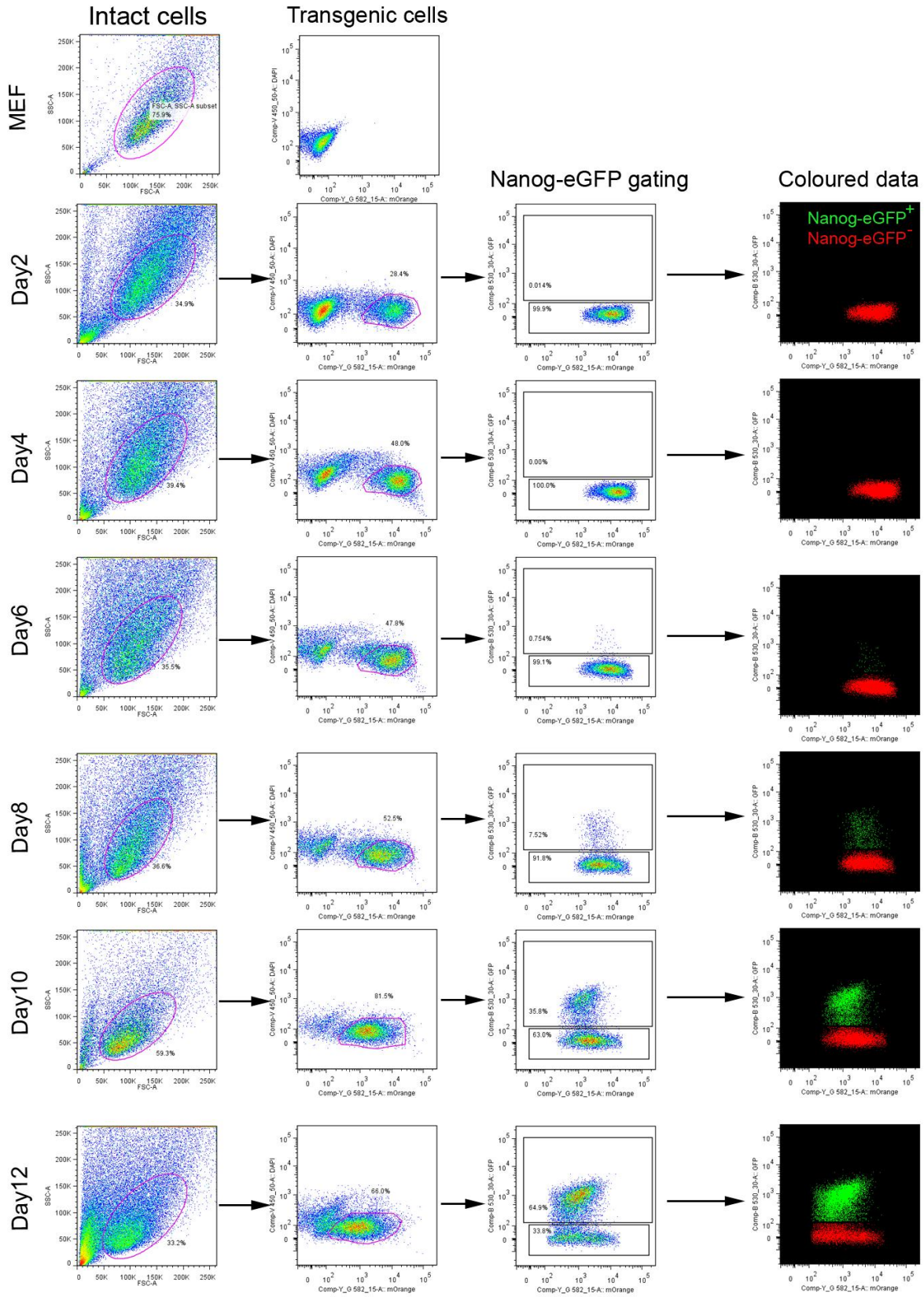


Supplementary Figure 2. Identification of novel reprogramming cell surface markers, Cd44 and Icam1. a. Secondary (2^o) reprogramming was carried out using the 6c cell line generated by 4 *piggyBac* (PB) transposons carrying reprogramming factors *c-Myc*, *Klf4*, *Oct4* and *Sox2*. Subsequently RNA was extracted from SSEA-1^+ cells at day5, 8, 11, 14, 17 as well as from 2^o iPSCs for microarray analysis to identify novel reprogramming cell surface markers. This experiment was performed in the absence of VitC and Alki. **b.** Microarray analysis of 6c 2^o reprogramming identified Cd44 and Icam1 as potential cell surface markers to dissect the reprogramming process. *Thy1* expression is already down-regulated in SSEA1^+ cells at day5 and *E-cadherin* expression plateaus at day8, suggesting previously identified markers are not suitable for investigation of the later stages of reprogramming.

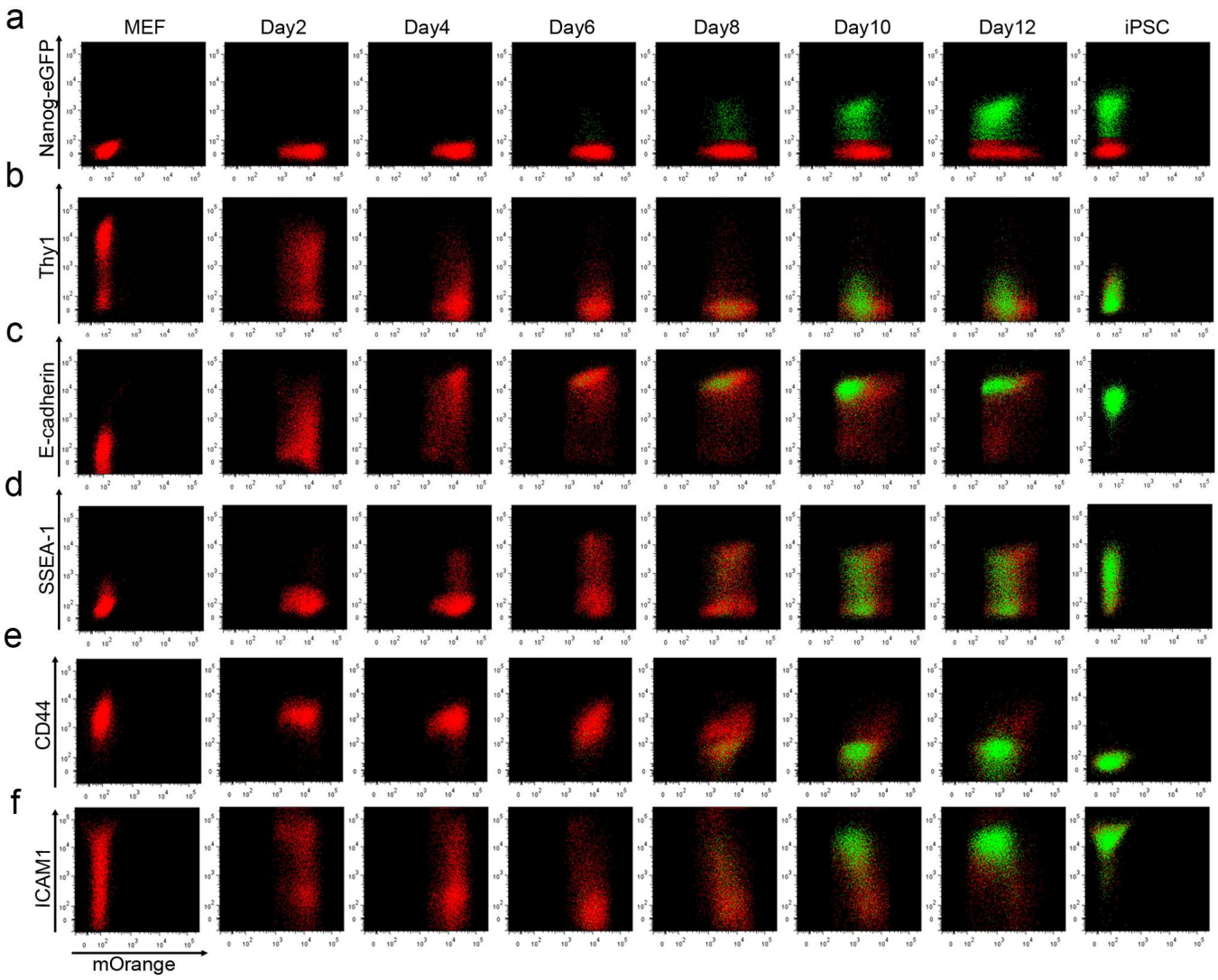
.



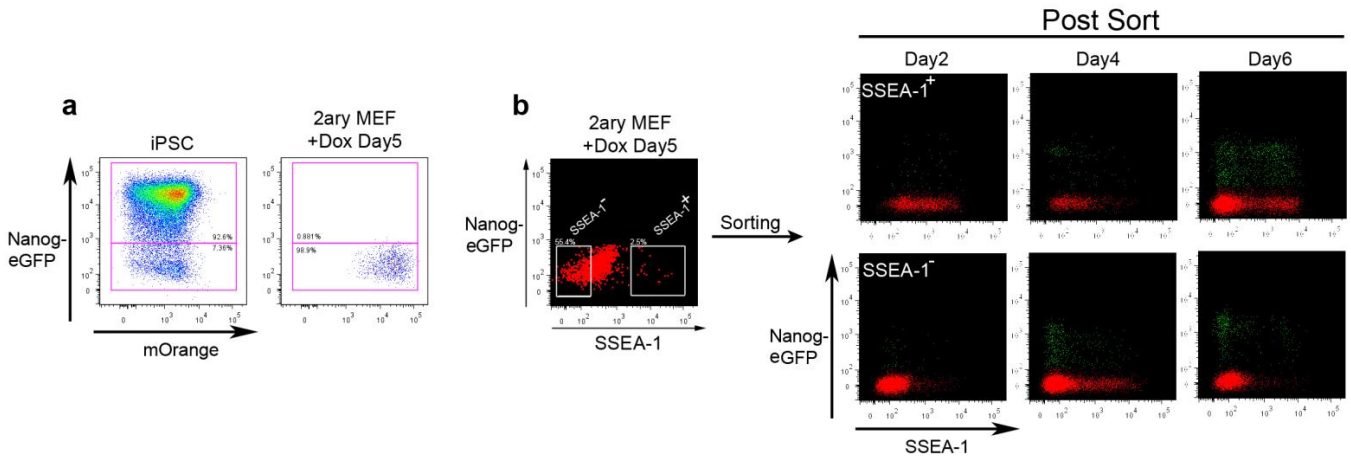
Supplementary Figure 3. A *piggyBac* (PB) secondary (2°) reprogramming system with 2A peptide-linked reprogramming cassette MKOS followed by ires mOrange. a. The reprogramming PB transposon with insulators and replicator introduced into Nanog-eGFP MEFs. **b.** Upon administration of dox, induction of reprogramming factors was observed as mOrange expression. Nanog-eGFP expression was observed at later time points.



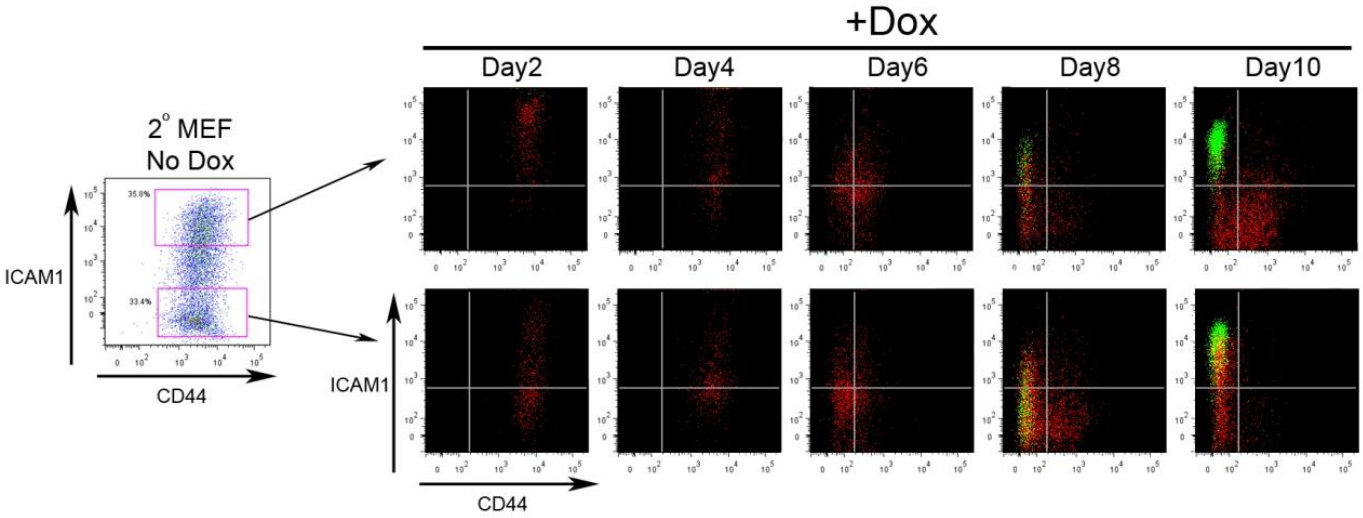
Supplementary Figure 4. Gating strategy for secondary reprogramming. Intact cells were gated using side and forward scatter (Intact cells). Transgenic cells were gated from wild-type cells using mOrange reporter (Transgenic cells). Nanog-eGFP^{+} and Nanog-eGFP^{-} cells were determined based on wild-type MEFs and coloured in green and red respectively.



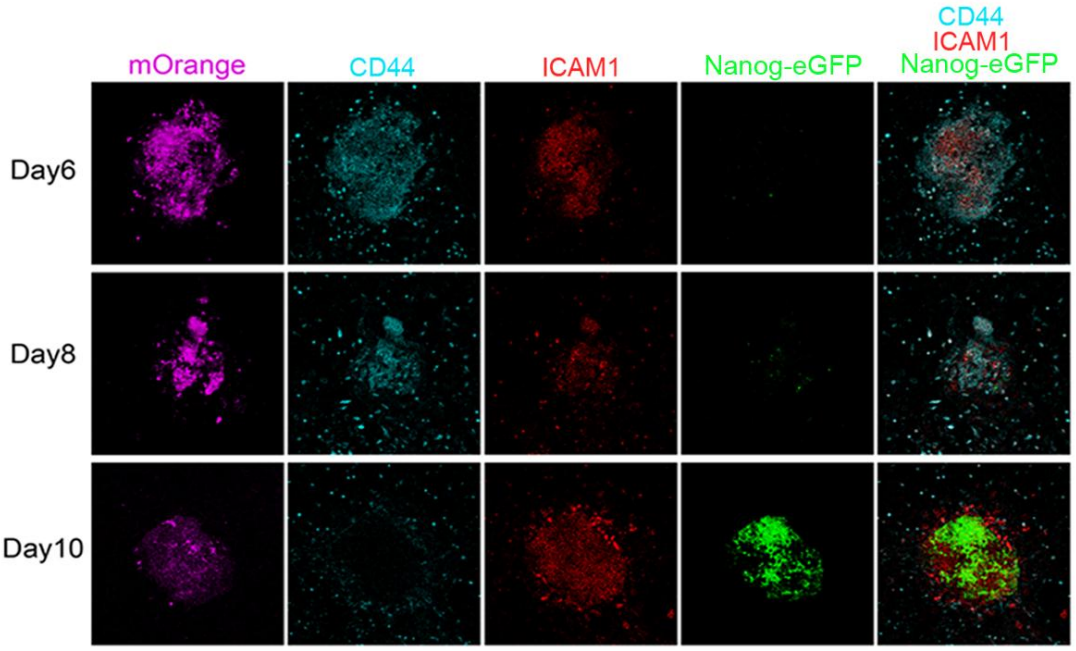
Supplementary Figure 5. Expression pattern of previously used and novel markers during 2° reprogramming. a. Nanog-eGFP expression during secondary reprogramming. Note appearance of eGFP$^+$ cells at day 6 from mOrange expressing cells. **b, c.** Thy1 negative population (b) and E-cadherin positive population (c) plateau by day 4 of 2^o reprogramming. **d.** SSEA-1 expression level in established iPSCs are heterogeneous and Nanog-eGFP$^+$ cells appear from both SSEA-1$^+$ and SSEA-1$^-$ cells. **e.** Downregulation of CD44 occurs later than that of Thy1, more closely correlating with the appearance of Nanog-eGFP$^+$ cells. **f.** ICAM1 expression is heterogeneous in MEFs, but the majority of cells become ICAM1$^-$ by around day 6. Re-upregulation of ICAM1 closely correlates with Nanog-eGFP expression.



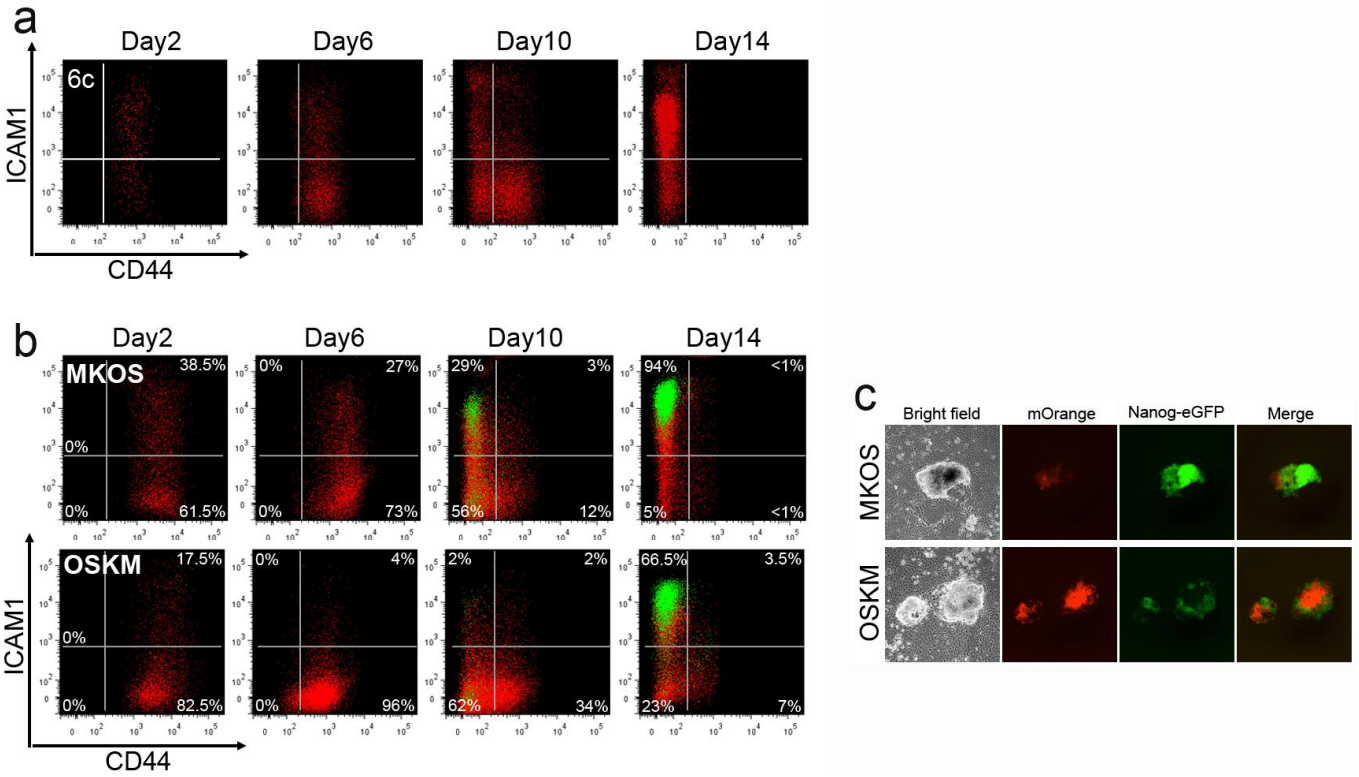
Supplementary Figure 6. SSEA-1 expression does not predict the appearance of Nanog-eGFP^{+} cells. a. Gating strategy for Nanog-eGFP^{-} cells at day 5 of reprogramming. **b.** Sorting strategy for Nanog-eGFP^{-}, SSEA-1$^{+/-}$ cells at day 5 of reprogramming. Cells were isolated, replated in reprogramming conditions and reanalyzed every 48hours. Red; Nanog-eGFP^{-} cells, Green; Nanog-eGFP^{+} cells.



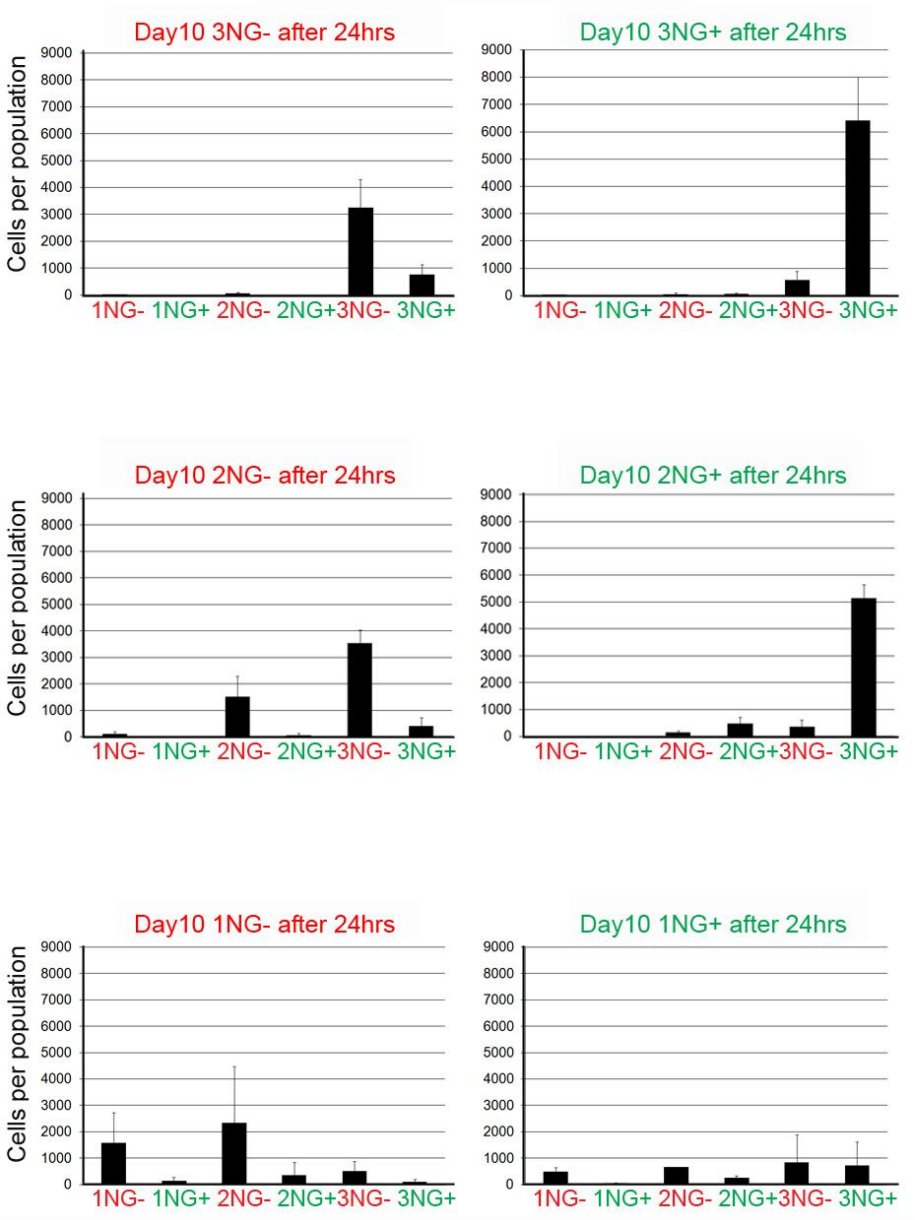
Supplementary Figure 7. Similar reprogramming kinetics with ICAM1$^{+/-}$ MEFs. ICAM1^{+} and ICAM1^{-} secondary MEF were sorted before initiating reprogramming. CD44, ICAM1, Nanog-eGFP expression was monitored every two days during reprogramming. Red; Nanog-eGFP^{-} cells, Green; Nanog-eGFP^{+} cells.



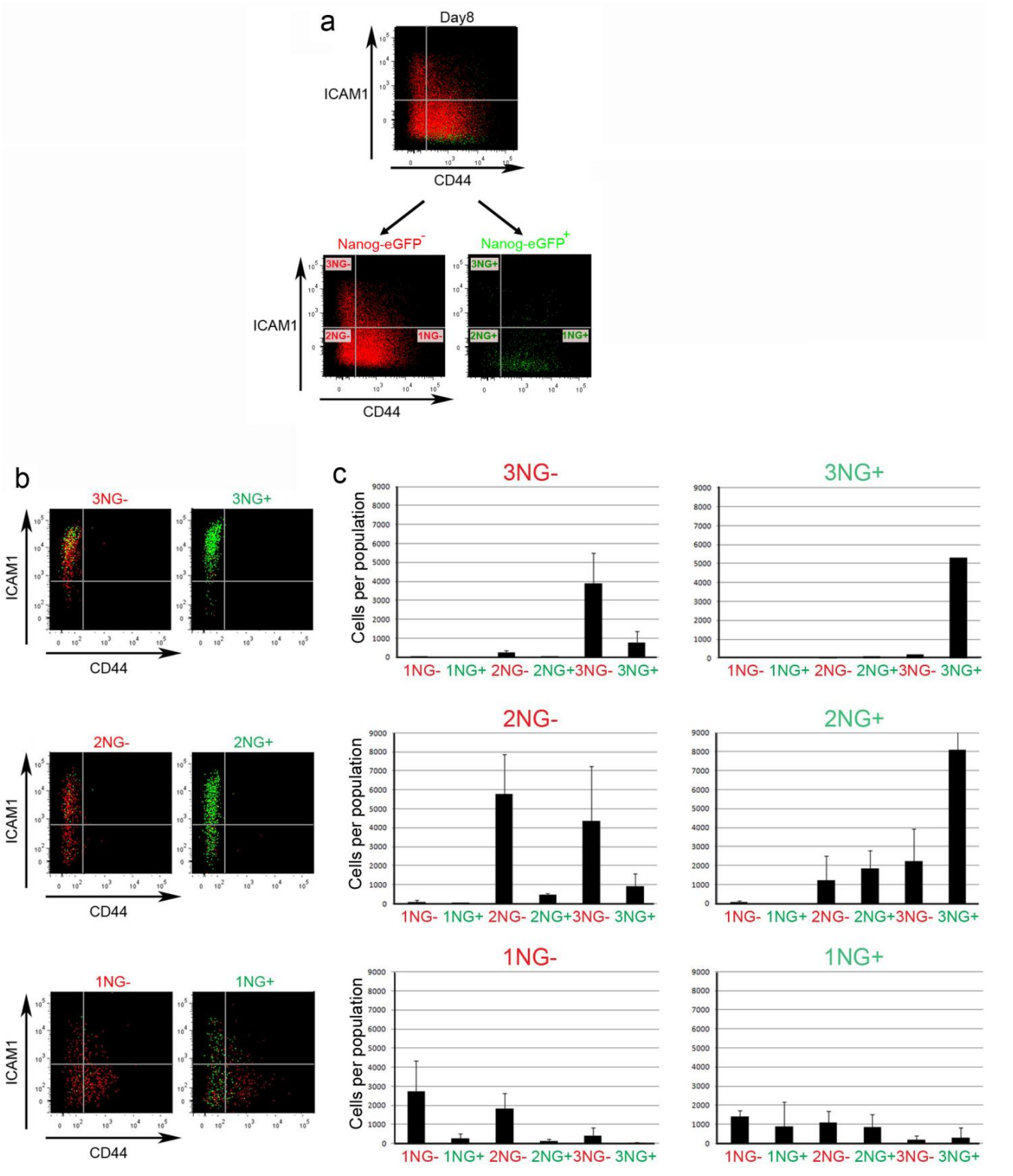
Supplementary Figure 8. Immunofluorescence for CD44 and ICAM1 at day 6, 8 and 10 after reprogramming initiation. Cells in a single colony have distinct CD44, ICAM1, Nanog-eGFP expression, indicating clonal analysis is not sufficient to isolate cells in similar stages. Note expression of mOrange tends to be low in colonies with Nanog-eGFP^{+} cells consistent with our flow cytometry data.



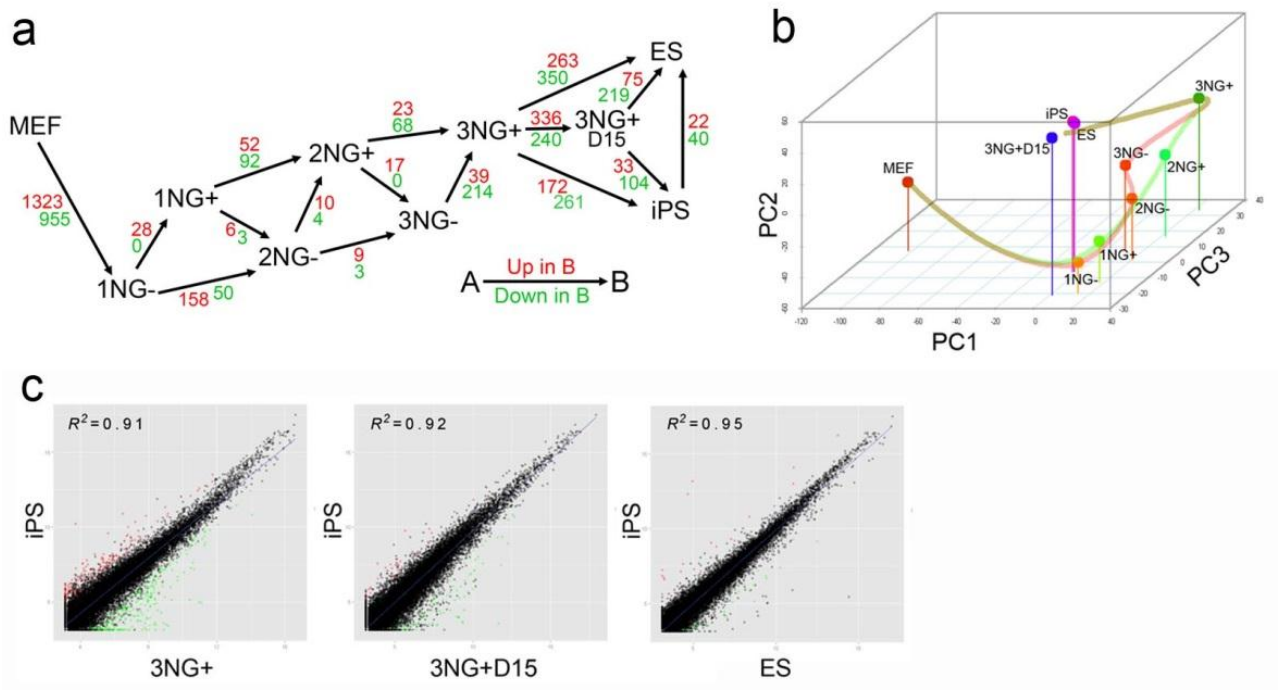
Supplementary Figure 9. The reprogramming pathway is conserved in different reprogramming systems. a, Non-polycistronic PB 2^{o} reprogramming with the 6c cell line. **b.** Primary PB reprogramming using both MKOS and OSKM polycistronic cassettes. **c.** Typical colonies arising from MKOS, OKMS primary PB reprogramming.



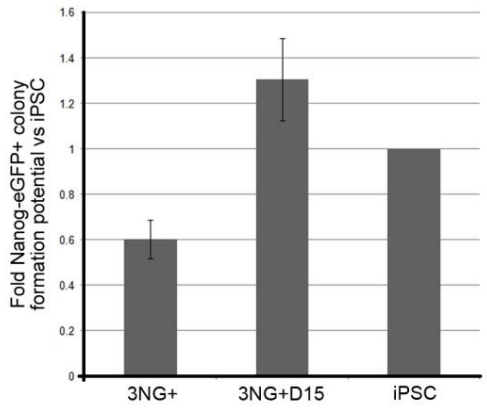
Supplementary Figure 10. Total number of cells in each gate after 24 hours for day 10 sorted populations. Each reprogramming population was sorted at day 10 and cells cultured in reprogramming conditions for 24 hours. Cells were then harvested and CD44/ICAM1/Nanog-eGFP expression was re-analysed. Total cell numbers found in each gate were plotted. This data highlighted the rapid expansion of cells once they entered the 3NG+ gate. The error bars represent the standard deviation of three independent experiments.



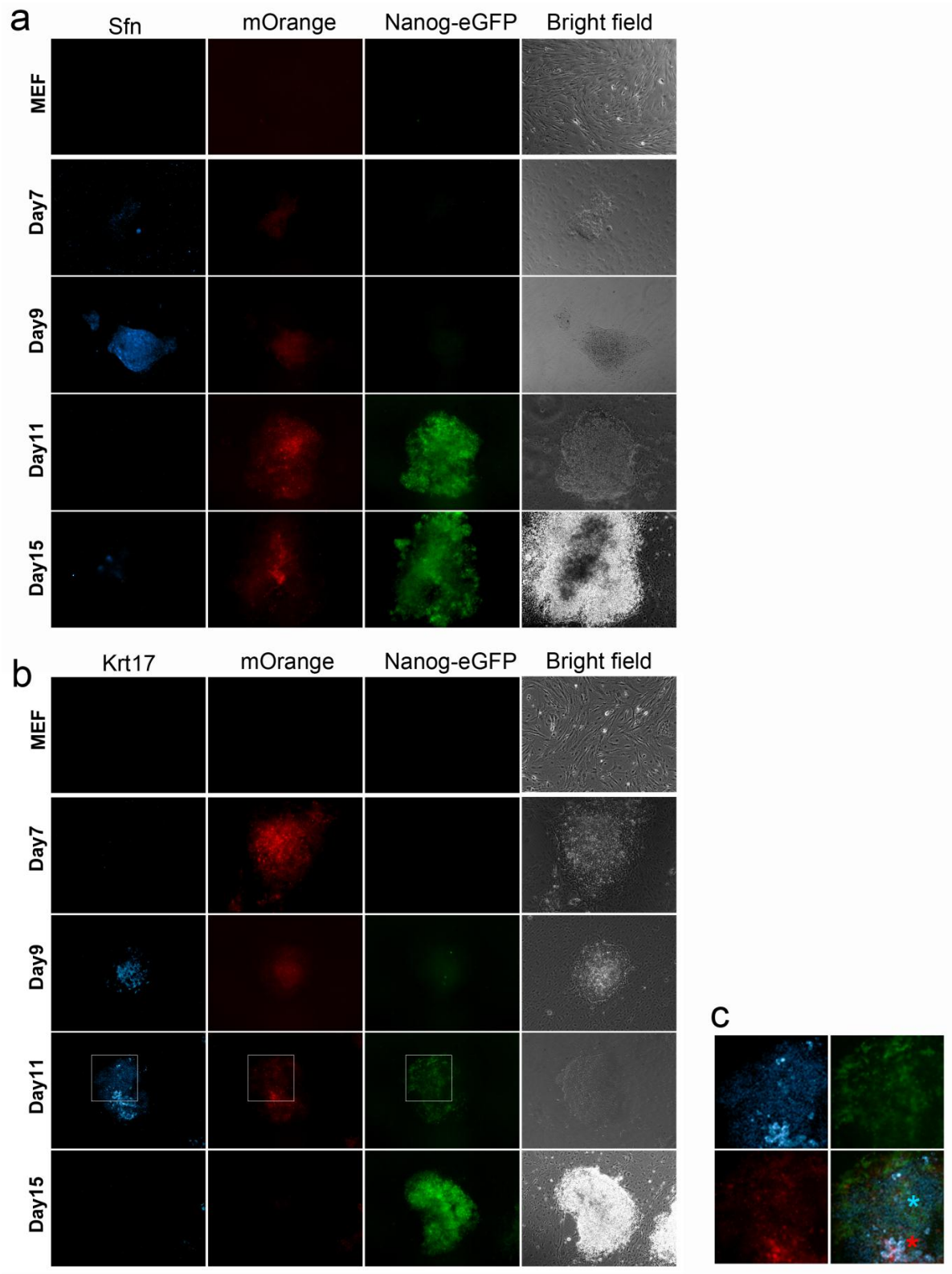
Supplementary Figure 11. Behavior of day8 sorted subpopulations are similar to that of day 10 subpopulations. a. Sorting strategy at day 8 of reprogramming. **b.** Each subpopulation sorted at day8 were replated in reprogramming conditions, and reanalyzed after 24 hours. **c.** Total cell numbers in each gate after 24 hour analysis for each sorted population. The error bars represent the standard deviation of three independent experiments.



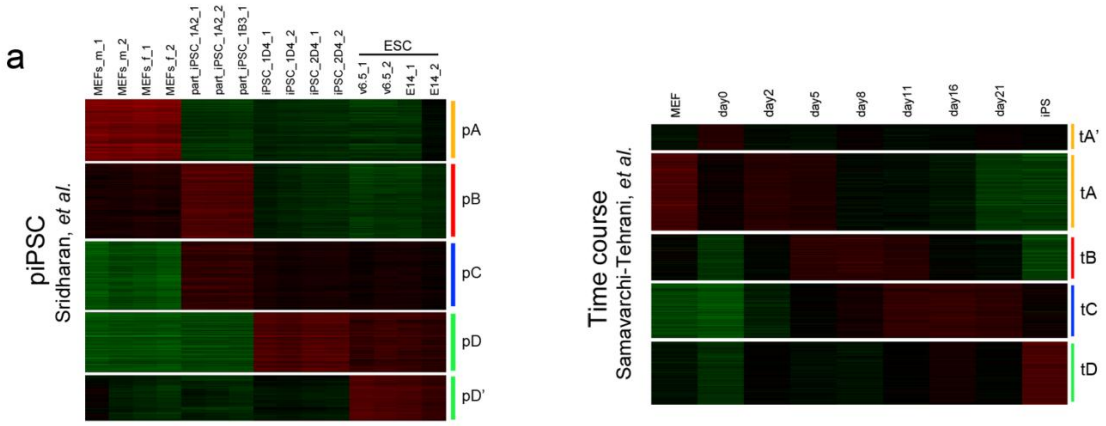
Supplementary Figure 12. Differentially expressed genes (DEGs) and Principal Component Analysis of reprogramming intermediates. a. Number of DEGs between samples identified via both edgeR and DESeq are indicated with arrows as shown. P-values were adjusted using a threshold for false discovery rate (FDR) ≤ 0.05. **b.** Using these DEGs (total 3,171), Principal Components Analysis (PCA) was performed. The green and red lines connecting the samples are based on the result that 3NG- cells were frequently produced by 2NG- and 1NG- cells (Figure 3d), while 2NG+ cells eventually appeared from 1NG+ cells (Figure 3b). **c.** Comparison of gene expression profiles between iPSCs and 3NG+, 3NG+D15, ESCs. Green and red color represents up- or down-regulated genes identified by both edgeR and DESeq.

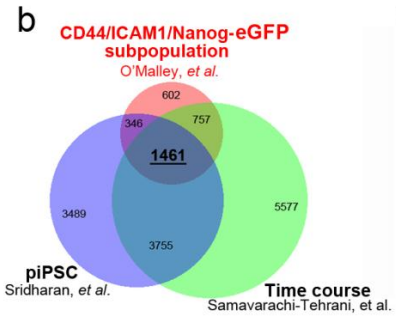


Supplementary Figure 13. Nanog-eGFP Colony formation potential of 3NG+ cells in the absence of dox. 3NG+ cells sorted at day10 (3NG+) and day15 (3NG+D15) of reprogramming were plated at clonal density on feeders in the absence of dox. The number of Nanog-eGFP$^+$ colonies was counted after 10 days. 3NG+ cells from day10 post transgene induction showed reduced ability (60%) to generate colonies compared to established iPSCs, while cells from day15 generated similar colony numbers to iPSCs. This suggests that about 40% of day10 3NG+ cells have not acquired exogenous reprogramming factor independent self-renewal capacity, but this trait can be acquired within an additional 5 days in the presence of dox. The error bars represent the standard deviation of three independent experiments.



Supplementary Figure 14. Transient up-regulation of Sfn and Krt17 during reprogramming. The highest protein level of both Sfn and Krt17 was observed around day9 post reprogramming (**a, b**). **c.** Higher magnification images of white squares in b. While Sfn protein was down-regulated before Nanog-eGFP expression (a), Krt protein was detectable in Nanog-eGFP expressing cells in the earlier stage, probably due to the protein stability (blue asterisk in c). Higher Krt17 expression was observed in mOrange^{high}, Nanog-eGFP^{-} cells (red asterisk in c).





c

CD44/ICAM1/Nanog-eGFP subpopulation vs piPSC

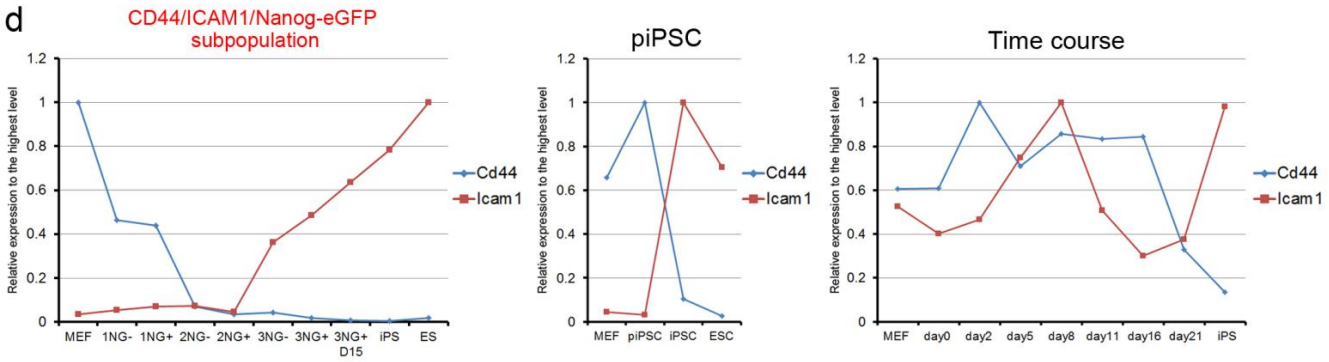
	A	B	C	D	E	Total
pA	**308**	28	1	4	97	438
pB	**109**	**153**	15	1	25	303
pC	8	67	**192**	47	19	333
pD	7	16	98	**184**	**60**	365
pD'	13	11	36	29	37	126
Total	445	275	342	265	238	1565

%	A	B	C	D	E	Total
pA	**70.3**	**6.4**	**0.2**	**0.9**	**22.1**	100
pB	**36.0**	**50.5**	**5.0**	**0.3**	**8.3**	100
pC	**2.4**	**20.1**	**57.7**	**14.1**	**5.7**	100
pD	**1.9**	**4.4**	26.8	**50.4**	**16.4**	100
pD'	**10.3**	**8.7**	28.6	**23.0**	29.4	100

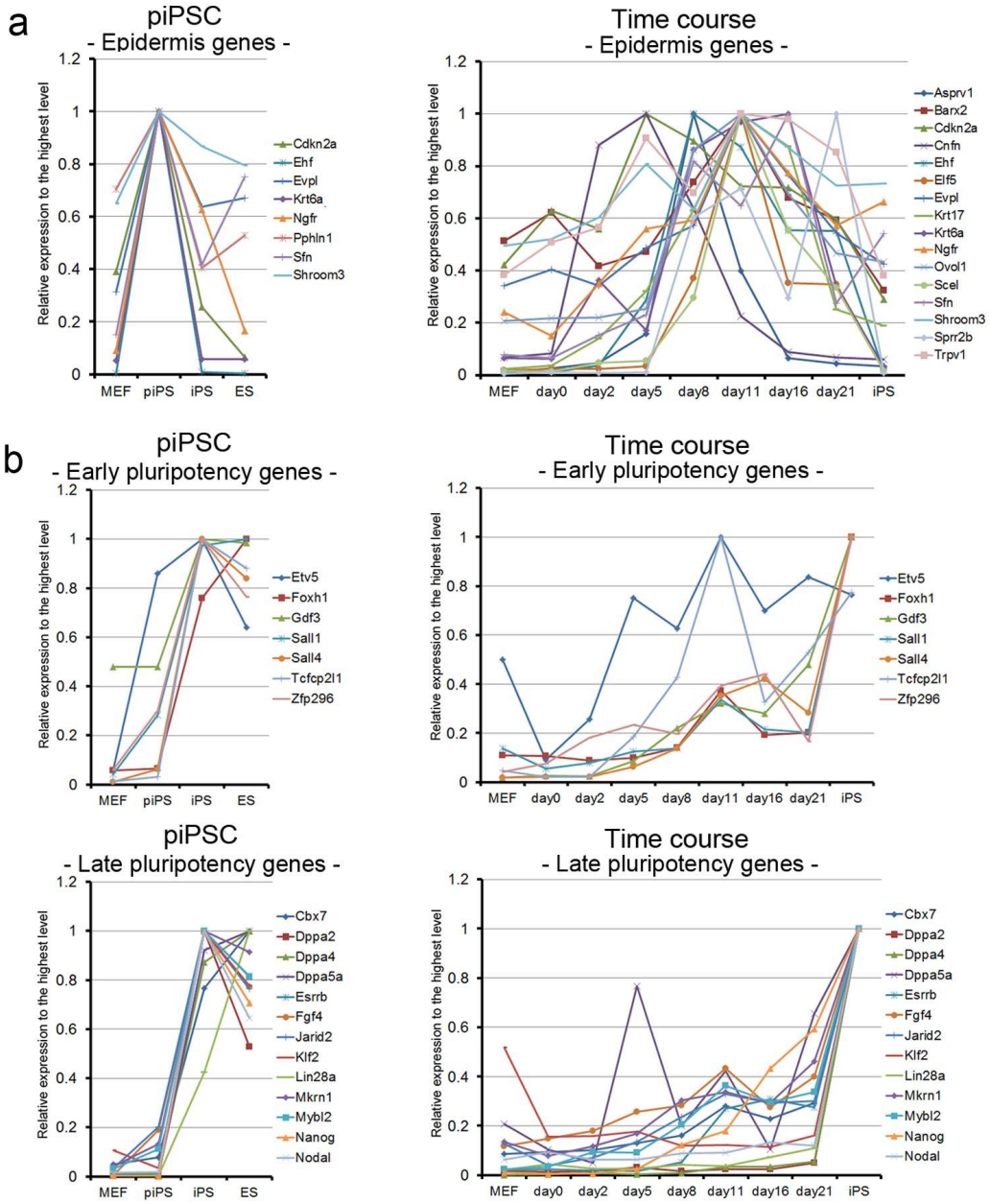
CD44/ICAM1/Nanog-eGFP subpopulation vs Time course

	A	B	C	D	E	Total
tA'	**10**	3	2	6	3	24
tA	**340**	42	2	4	125	513
tB	54	**109**	11	6	21	201
tC	9	96	**217**	64	29	415
tD	5	11	86	**165**	46	313
Total	418	261	318	245	224	1466

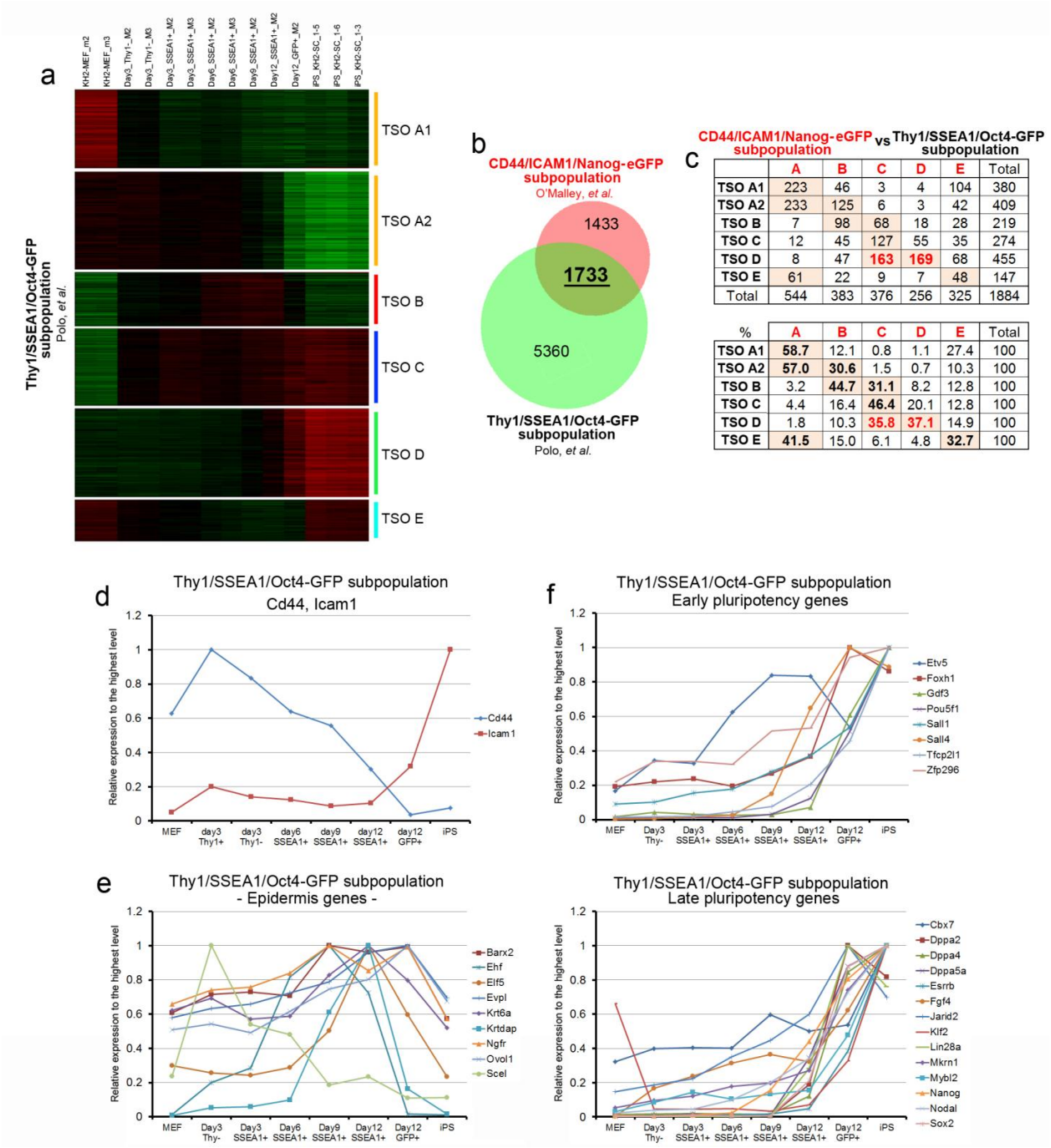
%	A	B	C	D	E	Total
tA'	**41.7**	12.5	8.3	25.0	12.5	100
tA	**66.3**	8.2	0.4	0.8	24.4	100
tB	26.9	**54.2**	5.5	3.0	10.4	100
tC	2.2	23.1	**52.3**	15.4	7.0	100
tD	1.6	3.5	27.5	**52.7**	14.7	100



Supplementary Figure 15. Comparative analysis using published reprogramming time course and partially reprogrammed cell data sets. a. Heat maps from time course[11] and partially reprogrammed iPSCs (piPSC)[27] data sets with differentially expressed genes (DEGs) with > 2.0 fold change and FDR ≤ 0.05, respectively. The DEGs were classified into 5 categories based on their expression pattern as identified from our dataset. A complete list of DEGs in each category is available in Supplementary Table 6 and 7. **b.** Venn diagram of all DEGs from this (Subpopulation), time course and piPSC data sets, highlighting 1461 DEGs common to all analyses. **c.** These 1461 genes were used to compare the overlap within each category against our subpopulation data set. Total number of genes belonging to each category is indicated in the upper tables, and their percentages against each Subpopulation A-E category are shown below. Overlap of more than 30% with Subpopulation A-E genes are highlighted in pink. Note that in the time course and piPSC microarrary datasets some genes are represented by multiple probe sets, resulting in several genes appearing in more than one category, with the total number of common genes standing at1466 and 1565, respectively. **d.** Cd44 and Icam1 expression profiles from Subpopulation, time course and piPSC data sets.



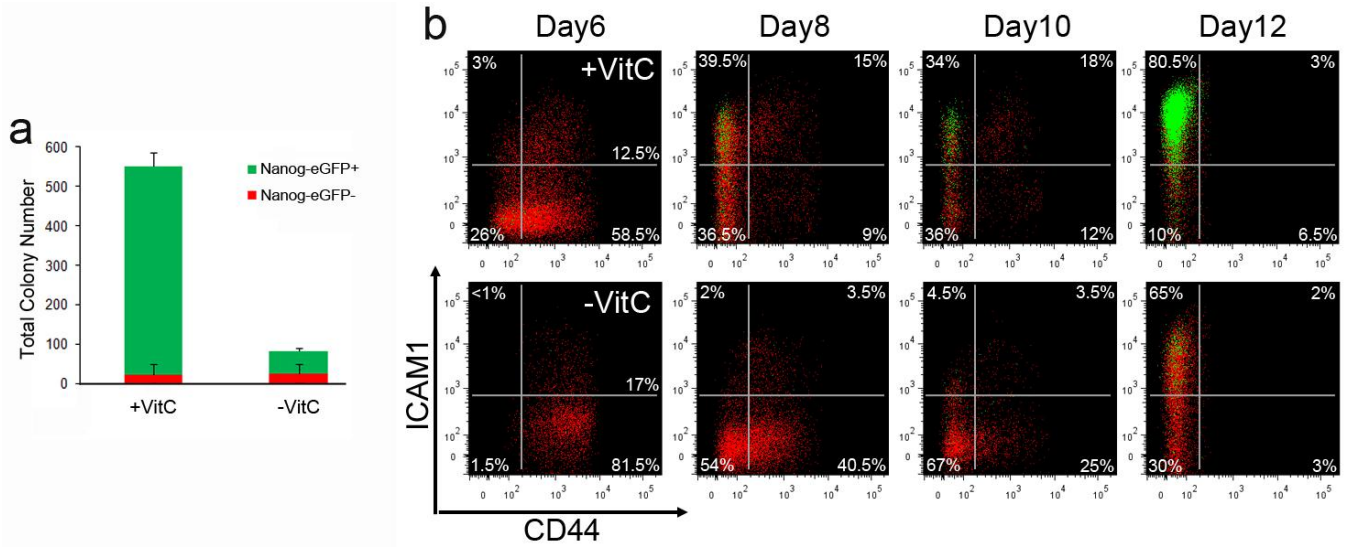
Supplementary Figure 16. Conserved transient epidermis gene up-regulation and limited resolution at the later stage of reprogramming without the sorting strategy. a. Expression pattern of epidermis-related genes, identified in Subpopulation Group B, common to time course and piPSC datasets. Data are shown as relative expression against the highest single value among the samples. Signal values are summarized in Supplementary Table 4. The data indicated transient up-regulation of epidermis-related genes is a common feature during reprogramming. **b.** Expression pattern of 19 pluripotency-related genes from time course and piPSC datasets. Genes are grouped according to their expression pattern from the Subpopulation dataset, Early (up-regulated in 1N populations) and Late (gradually up-regulated through later stages of reprogramming). It is notable that there is a large increase in the expression level of many pluripotency genes between the last time point (day21) and iPSCs in the Timecourse data, suggesting that proportionally there were few reprogrammed cells by day21. In general, such bulk population analysis may not be suitable to investigate how pluripotency genes are up-regulated. Consistent with the fact that piPSCs have a low potential to generate iPSCs, most pluripotency genes are not expressed in piPSCs.



	A	B	C	D	E	Total
TSO A1	223	46	3	4	104	380
TSO A2	233	125	6	3	42	409
TSO B	7	98	68	18	28	219
TSO C	12	45	127	55	35	274
TSO D	8	47	163	169	68	455
TSO E	61	22	9	7	48	147
Total	544	383	376	256	325	1884

%	A	B	C	D	E	Total
TSO A1	58.7	12.1	0.8	1.1	27.4	100
TSO A2	57.0	30.6	1.5	0.7	10.3	100
TSO B	3.2	44.7	31.1	8.2	12.8	100
TSO C	4.4	16.4	46.4	20.1	12.8	100
TSO D	1.8	10.3	35.8	37.1	14.9	100
TSO E	41.5	15.0	6.1	4.8	32.7	100

Supplementary Figure 17. Comparative analysis of another marker system (Polo et al. Cell, 2012). a. Heat maps from Thy1/SSEA1/Oct4-GFP (TSO) subpopulation datasets with differentially expressed genes (DEGs). A complete list of DEGs in each category is available in Supplementary Table 8. **b.** Venn diagram highlighting 1733 DEGs common to our CD44/ICAM1/Nanog-eGFP subpopulation dataset. **c.** These 1733 genes were used to compare the overlap within each category against our subpopulation dataset. Overlap of more than 30% with our A-E genes are highlighted in pink. Almost equal numbers of TSO D genes belong to our C and D groups (red), indicating CD44/ICAM1/Nanog-eGFP sorting strategy gives higher resolution at the late stage of reprogramming. Note that in the TSO microarrary dataset some genes are represented by multiple probe sets, resulting in several genes appearing in more than one category, with the total number of common genes standing at 1884. **d, e, f.** Expression pattern of Cd44 and Icam1, epidermis-related genes, early and late pluripotency genes from TSO Subpopulation dataset, respectively.



Supplementary Figure 18. Secondary reprogramming in the presence of dox, Alki either with or without VitaminC (+ or -VitC). a. Both total and Nanog-eGFP^{+} colony numbers decreased in the absence of VitC. **b.** Cells reprogrammed in the absence of VitC displayed delayed reprogramming kinetics.

Supplementary Table 3

A			
Term	Pop Hits	%	PValue
GO:0007155~cell adhesion	561	8.77	2.92.E-13
GO:0022610~biological adhesion	562	8.77	3.13.E-13
GO:0030198~extracellular matrix organization	101	3.51	1.58.E-12
GO:0043062~extracellular structure organization	149	3.95	3.56.E-11
GO:0030029~actin filament-based process	176	4.09	3.14.E-10
GO:0007167~enzyme linked receptor protein signaling pathway	273	5.12	5.09.E-10
GO:0001568~blood vessel development	244	4.68	1.85.E-09
GO:0030036~actin cytoskeleton organization	165	3.80	1.88.E-09
GO:0001944~vasculature development	250	4.68	3.35.E-09
GO:0001558~regulation of cell growth	92	2.78	5.83.E-09
GO:0001501~skeletal system development	285	4.97	5.95.E-09
GO:0035295~tube development	264	4.53	4.54.E-08
GO:0048729~tissue morphogenesis	238	4.24	6.13.E-08
GO:0048514~blood vessel morphogenesis	198	3.80	8.00.E-08
GO:0040008~regulation of growth	256	4.39	8.18.E-08
GO:0010810~regulation of cell-substrate adhesion	40	1.75	1.40.E-07
GO:0007169~transmembrane receptor protein tyrosine kinase signaling pathway	192	3.65	1.76.E-07
GO:0030199~collagen fibril organization	21	1.32	4.44.E-07
GO:0007010~cytoskeleton organization	326	4.82	4.85.E-07
GO:0042127~regulation of cell proliferation	538	6.43	1.64.E-06
GO:0001655~urogenital system development	146	2.92	1.84.E-06
GO:0040007~growth	193	3.36	2.76.E-06
GO:0010811~positive regulation of cell-substrate adhesion	27	1.32	3.98.E-06
GO:0030334~regulation of cell migration	92	2.19	6.73.E-06
GO:0001503~ossification	106	2.34	7.94.E-06
GO:0051270~regulation of cell motion	107	2.34	8.93.E-06
GO:0042325~regulation of phosphorylation	290	4.09	1.02.E-05
GO:0060537~muscle tissue development	136	2.63	1.10.E-05
GO:0001763~morphogenesis of a branching structure	125	2.49	1.47.E-05
GO:0006928~cell motion	367	4.68	1.64.E-05

GO:0014706~striated muscle tissue development	127	2.49	1.80.E-05
GO:0051174~regulation of phosphorus metabolic process	301	4.09	2.00.E-05
GO:0019220~regulation of phosphate metabolic process	301	4.09	2.00.E-05
GO:0045785~positive regulation of cell adhesion	43	1.46	2.23.E-05
GO:0007507~heart development	223	3.36	2.77.E-05
GO:0007517~muscle organ development	176	2.92	2.84.E-05
GO:0032535~regulation of cellular component size	161	2.78	2.87.E-05
GO:0060348~bone development	118	2.34	2.95.E-05
GO:0043549~regulation of kinase activity	192	3.07	2.97.E-05
GO:0001525~angiogenesis	133	2.49	3.22.E-05
GO:0048754~branching morphogenesis of a tube	93	2.05	3.63.E-05
GO:0001822~kidney development	107	2.19	3.91.E-05
GO:0030155~regulation of cell adhesion	94	2.05	4.07.E-05
GO:0008038~neuron recognition	12	0.88	4.35.E-05
GO:0032989~cellular component morphogenesis	351	4.39	4.54.E-05
GO:0051338~regulation of transferase activity	199	3.07	4.96.E-05
GO:0040012~regulation of locomotion	110	2.19	5.33.E-05
GO:0022604~regulation of cell morphogenesis	97	2.05	5.70.E-05
GO:0048589~developmental growth	100	2.05	7.86.E-05
GO:0017015~regulation of transforming growth factor beta receptor signaling pathway	30	1.17	8.86.E-05
GO:0048705~skeletal system morphogenesis	130	2.34	9.18.E-05
GO:0042692~muscle cell differentiation	117	2.19	1.05.E-04
GO:0007409~axonogenesis	163	2.63	1.14.E-04
GO:0051216~cartilage development	78	1.75	1.32.E-04
GO:0008361~regulation of cell size	108	2.05	1.74.E-04
GO:0045859~regulation of protein kinase activity	186	2.78	1.89.E-04
GO:0035239~tube morphogenesis	171	2.63	2.05.E-04
GO:0043009~chordate embryonic development	421	4.68	2.13.E-04
GO:0016477~cell migration	240	3.22	2.31.E-04
GO:0031032~actomyosin structure organization	25	1.02	2.43.E-04
GO:0042060~wound healing	112	2.05	2.51.E-04
GO:0009792~embryonic development ending in birth or egg hatching	425	4.68	2.52.E-04
GO:0001656~metanephros development	58	1.46	2.57.E-04

GO:0007411~axon guidance	98	1.90	2.61.E-04
GO:0048812~neuron projection morphogenesis	176	2.63	2.89.E-04
GO:0000904~cell morphogenesis involved in differentiation	212	2.92	3.36.E-04
GO:0030030~cell projection organization	319	3.80	3.41.E-04
GO:0044092~negative regulation of molecular function	132	2.19	3.79.E-04
GO:0048732~gland development	197	2.78	3.84.E-04
GO:0060415~muscle tissue morphogenesis	18	0.88	3.91.E-04
GO:0055008~cardiac muscle tissue morphogenesis	18	0.88	3.91.E-04
GO:0048667~cell morphogenesis involved in neuron differentiation	182	2.63	4.28.E-04
GO:0051146~striated muscle cell differentiation	89	1.75	4.30.E-04
GO:0045185~maintenance of protein location	28	1.02	4.69.E-04
GO:0007229~integrin-mediated signaling pathway	76	1.61	4.70.E-04
GO:0031175~neuron projection development	218	2.92	4.75.E-04
GO:0060429~epithelium development	271	3.36	4.77.E-04
GO:0016192~vesicle-mediated transport	466	4.82	5.73.E-04
GO:0001657~ureteric bud development	42	1.17	8.00.E-04
GO:0048598~embryonic morphogenesis	359	3.95	8.57.E-04
GO:0032990~cell part morphogenesis	212	2.78	9.00.E-04
B			
Term	Pop Hits	%	PValue
GO:0007398~ectoderm development	133	2.75	2.14.E-06
GO:0008544~epidermis development	125	2.60	4.18.E-06
GO:0030216~keratinocyte differentiation	48	1.53	2.57.E-05
GO:0009913~epidermal cell differentiation	51	1.53	4.26.E-05
GO:0009611~response to wounding	347	4.27	4.31.E-05
GO:0030855~epithelial cell differentiation	123	2.14	2.41.E-04
GO:0006952~defense response	448	4.73	2.49.E-04
GO:0009266~response to temperature stimulus	51	1.37	2.67.E-04
GO:0031424~keratinization	28	1.07	2.74.E-04
GO:0006954~inflammatory response	225	2.90	5.78.E-04
GO:0050727~regulation of inflammatory response	57	1.37	5.81.E-04

GO:0032101~regulation of response to external stimulus	103	1.83	6.55.E-04
GO:0002478~antigen processing and presentation of exogenous peptide antigen	23	0.92	8.41.E-04
C			
Term	Pop Hits	%	PValue
GO:0034660~ncRNA metabolic process	202	2.83	3.66.E-04
GO:0006790~sulfur metabolic process	94	1.83	4.81.E-04
D			
Term	Pop Hits	%	PValue
GO:0045934~negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	397	6.24	1.49.E-07
GO:0051172~negative regulation of nitrogen compound metabolic process	401	6.24	1.81.E-07
GO:0010558~negative regulation of macromolecule biosynthetic process	418	6.00	1.41.E-06
GO:0010605~negative regulation of macromolecule metabolic process	506	6.70	1.50.E-06
GO:0031327~negative regulation of cellular biosynthetic process	430	6.00	2.34.E-06
GO:0009890~negative regulation of biosynthetic process	434	6.00	2.78.E-06
GO:0010629~negative regulation of gene expression	410	5.77	3.37.E-06
GO:0016481~negative regulation of transcription	372	5.31	7.49.E-06
GO:0045449~regulation of transcription	2227	17.09	1.25.E-05
GO:0045892~negative regulation of transcription, DNA-dependent	308	4.62	1.76.E-05
GO:0051253~negative regulation of RNA metabolic process	310	4.62	1.93.E-05
GO:0019827~stem cell maintenance	22	1.39	6.61.E-05
GO:0048864~stem cell development	23	1.39	8.31.E-05
GO:0019953~sexual reproduction	386	4.85	1.23.E-04
GO:0000122~negative regulation of transcription from RNA polymerase II promoter	231	3.46	2.71.E-04
GO:0007548~sex differentiation	130	2.54	3.18.E-04
GO:0045596~negative regulation of cell differentiation	182	3.00	3.46.E-04
GO:0048863~stem cell differentiation	31	1.39	3.67.E-04
GO:0007276~gamete generation	331	4.16	4.42.E-04
GO:0051053~negative regulation of DNA metabolic process	20	1.15	6.24.E-04

GO:0048609~reproductive process in a multicellular organism	409	4.62	6.99.E-04
GO:0032504~multicellular organism reproduction	409	4.62	6.99.E-04
GO:0048610~reproductive cellular process	173	2.77	8.20.E-04
GO:0045165~cell fate commitment	147	2.54	8.38.E-04
E			
Term	Pop Hits	%	PValue
GO:0007507~heart development	223	4.27	1.09.E-06
GO:0042127~regulation of cell proliferation	538	6.10	1.18.E-04
GO:0035295~tube development	264	3.86	1.48.E-04
GO:0022604~regulation of cell morphogenesis	97	2.24	1.77.E-04
GO:0035239~tube morphogenesis	171	2.85	4.36.E-04
GO:0003007~heart morphogenesis	74	1.83	5.69.E-04

Supplementary Table 3. Gene Ontology from O'Malley sub-population data

Supplementary Table 5. Epidermis genes EST profile

Breakdown by Body Sites *-Transcripts per million*

	Asprv1	*Barx2*	*Cdkn2a*	*Cnfn*	*Ehf*	*Elf5*	*Evpl*	*Jun*	*Krt17*	*Krt6a*	*Krtdap*	*Ngfr*	*Ovol1*	*Pphln1*	*Scel*	*Sfn*	*Shroom3*	*Sprr2a1*	*Sprr2b*	*Trpv1*	*Tsg101*	total
tongue	2699	89	0	629	179	0	0	0	0	359	2429	0	89	0	0	359	0	1889	n.a.	0	269	8990
stomach	503	0	0	251	0	0	0	472	157	0	440	0	0	0	0	188	0	5383	n.a.	0	31	7425
vagina	613	0	0	0	766	0	153	0	306	766	0	0	153	0	0	0	0	0	n.a.	0	0	2757
intestine	0	0	0	0	172	0	0	241	11	0	0	0	46	34	0	126	57	1784	n.a.	0	80	2551
skin	201	0	0	92	58	33	100	193	353	67	109	25	67	117	100	311	8	8	n.a.	0	109	1951
turbinate	0	0	0	0	729	0	0	729	0	0	0	0	0	0	0	0	0	0	n.a.	0	0	1458
extraembryonic tissue	0	0	0	0	0	0	0	93	0	0	0	0	13	107	0	53	26	990	n.a.	0	174	1456
sympathetic ganglion	0	0	0	0	0	0	0	801	0	0	0	500	0	0	0	0	0	0	n.a.	0	0	1301
epididymis	0	0	0	0	645	0	0	0	0	0	0	0	0	0	645	0	0	0	n.a.	0	0	1290
mammary gland	16	13	9	0	270	260	13	178	82	9	0	9	26	89	0	42	42	0	n.a.	0	112	1170
lung	50	0	180	0	330	180	0	110	50	0	0	50	10	20	0	50	50	20	n.a.	0	70	1170
dorsal root ganglion	0	0	0	0	0	0	0	411	0	0	0	411	0	164	0	0	0	0	n.a.	164	0	1150
nasopharynx	0	0	0	0	0	0	0	125	251	125	0	0	0	251	0	0	125	0	n.a.	0	125	1002
bladder	0	0	0	0	368	0	0	61	0	0	0	61	0	0	0	0	61	368	n.a.	0	0	919
uterus	0	0	0	0	0	0	0	291	0	0	0	0	291	0	0	0	145	0	n.a.	0	145	872
molar	0	0	0	0	0	0	0	0	282	0	0	0	0	0	565	0	0	0	n.a.	0	0	847
embryonic tissue	75	1	1	4	8	1	13	51	19	13	47	23	16	82	10	67	36	67	n.a.	0	89	623
pancreas	0	0	0	0	28	0	9	103	0	0	0	28	0	37	18	9	169	150	n.a.	0	56	607
olfactory mucosa	0	0	0	0	0	0	0	591	0	0	0	0	0	0	0	0	0	0	n.a.	0	0	591
prostate	0	0	0	0	270	0	0	0	0	0	33	0	0	0	0	0	0	237	n.a.	0	33	573
thyroid	0	0	0	0	0	0	0	340	0	0	0	0	0	0	0	0	113	0	n.a.	0	113	566
eye	64	0	0	0	5	0	21	53	0	91	0	43	0	102	5	26	48	0	n.a.	0	102	560
bone	0	0	0	0	0	0	0	410	0	0	0	0	0	0	0	0	0	0	n.a.	0	58	468
ovary	0	0	0	0	91	0	0	0	0	0	0	0	18	36	18	0	127	0	n.a.	0	145	435
bone marrow	0	0	0	0	0	0	0	95	0	0	0	0	0	168	0	7	7	0	n.a.	0	139	416
connective tissue	0	0	151	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	n.a.	0	201	402
fertilized ovum	0	0	0	0	107	0	0	35	0	0	0	0	35	179	35	0	0	0	n.a.	0	0	391
thymus	0	0	0	0	0	0	0	33	16	16	0	0	8	206	0	0	0	0	n.a.	0	107	386
muscle	36	0	0	0	0	0	0	110	0	0	0	0	0	184	0	0	0	0	n.a.	0	36	366
spleen	10	0	21	0	32	0	0	32	0	0	0	43	10	0	21	0	10	21	n.a.	0	162	362
pituitary gland	0	0	0	0	0	0	0	110	0	0	0	0	0	55	0	0	0	0	n.a.	0	166	331
heart	0	0	0	18	0	0	0	73	0	0	0	0	0	128	0	18	0	0	n.a.	0	91	328
kidney	0	0	0	0	32	0	32	24	0	0	0	0	24	80	24	8	40	16	n.a.	0	40	320
salivary gland	0	0	0	0	51	51	0	0	51	0	0	0	0	0	0	0	0	0	n.a.	0	154	307
inner ear	0	0	0	0	0	0	0	106	0	0	0	26	0	80	0	0	0	0	n.a.	0	80	292
testis	0	0	0	0	0	0	0	90	0	8	0	0	16	49	8	0	32	0	n.a.	0	82	285
oviduct	0	0	0	0	261	0	0	0	0	0	0	0	0	0	0	0	0	0	n.a.	0	0	261
joint	0	0	117	0	0	0	0	117	0	0	0	0	0	0	0	0	0	0	n.a.	0	0	234
brain	0	2	0	0	0	0	4	69	0	0	0	8	0	29	0	0	48	0	n.a.	6	67	233
lymph node	0	0	0	0	0	0	0	0	0	0	0	0	0	68	0	0	0	0	n.a.	0	136	204
spinal cord	40	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	40	0	n.a.	0	80	200
liver	0	0	0	0	8	0	0	8	0	0	0	0	0	26	0	8	8	0	n.a.	0	62	120
blood	0	0	0	0	0	0	0	0	0	0	0	0	0	59	0	0	0	0	n.a.	0	59	118
adipose tissue	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n.a.	0	0	0
adrenal gland	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n.a.	0	0	0
pineal gland	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n.a.	0	0	0
vesicular gland	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n.a.	0	0	0

n.a. = not available

Breakdown by Developmental Stage *-Transcripts per million*

	Asprv1	*Barx2*	*Cdkn2a*	*Cnfn*	*Ehf*	*Elf5*	*Evpl*	*Jun*	*Krt17*	*Krt6a*	*Krtdap*	*Ngfr*	*Ovol1*	*Pphln1*	*Scel*	*Sfn*	*Shroom3*	*Sprr2a1*	*Sprr2b*	*Trpv1*	*Tsg101*	total
oocyte	0	0	0	0	206	0	0	0	0	0	0	0	0	51	0	0	0	0	n.a.	0	310	567
unfertilized ovum	0	0	0	0	49	0	0	0	0	0	0	0	0	0	49	0	0	0	n.a.	0	0	98
zygote	0	0	0	0	104	0	0	34	0	0	0	0	34	173	34	0	0	0	n.a.	0	0	379
cleavage	0	0	0	0	36	0	0	0	0	0	0	0	0	145	72	36	0	0	n.a.	0	181	470
morula	0	0	0	0	81	0	54	243	0	0	0	0	0	54	0	243	0	0	n.a.	0	27	702
blastocyst	0	0	0	0	0	14	0	0	0	0	0	0	43	87	0	204	0	0	n.a.	0	29	377
egg cylinder	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	n.a.	0	0	0
gastrula	0	0	0	0	0	0	0	33	0	0	0	67	0	135	0	67	0	0	n.a.	0	169	471
organogenesis	15	7	0	0	0	0	0	45	0	0	0	0	0	114	7	30	15	0	n.a.	0	99	332
fetus	99	4	0	19	7	0	8	96	34	23	71	51	11	75	5	32	99	103	n.a.	2	90	829
neonate	0	0	0	0	18	0	9	101	27	9	9	83	36	27	9	0	0	9	n.a.	0	18	355
juvenile	31	0	0	0	27	0	3	76	20	34	0	6	24	101	13	20	31	10	n.a.	0	59	455
adult	85	4	2	27	118	58	29	113	377	76	83	3	13	73	10	60	85	859	n.a.	0	76	2151

n.a. = not available