Title
Identification and Characterization of Klf4 Functional Domains in Somatic Cell Reprogramming

Author
Schmidt, Ryan

Publication Date
2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Identification and Characterization of Klf4 Functional Domains in Somatic Cell Reprogramming

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Neuroscience

by

Ryan Jason Schmidt

2012

ABSTRACT OF THE DISSERTATION

Identification and Characterization of Klf4 Functional Domains in Somatic Cell Reprogramming

by

Ryan Jason Schmidt

Doctor of Philosophy in Neuroscience

University of California, Los Angeles, 2012

Professor Kelsey Martin, Co-Chair

Professor Kathrin Plath, Co-Chair

Somatic cell reprogramming refers to the conversion of a differentiated cell with restricted developmental potential to a pluripotent cell through the collective action of defined factors. This method of resetting the epigenome to an embryonic-like state has tremendous medical implications since these cells can subsequently be converted into any cell type of the body for use in regenerative therapies. The reprogramming process is most often initiated by the expression of three transcription factors - Oct4, Sox2, and Klf4 - in a target cell. This work aims to better understand the mechanism of reprogramming by mapping functional domains in the Klf4 protein that are required for the induction of pluripotency. Within these important regions of Klf4, we characterize specific contact sites with DNA and cofactor proteins that determine its reprogramming activity. Finally, to further understand the properties within the DNA binding domain of Klf4 that enable its reprogramming activity, we determine the *in vitro* binding

preferences of DNA binding domains within the Klf family and demonstrate the molecular basis of their functional divergence in somatic cell reprogramming.

The dissertation of Ryan Jason Schmidt is approved.

Michael F. Carey

Stephen Smale

James Akira Wohlschlegel

Kelsey C. Martin, Committee Co-Chair

Kathrin Plath, Committee Co-Chair

University of California, Los Angeles

2012

I dedicate this work to my parents. Thank you for your love and support.

TABLE OF CONTENTS

FIGURES AND TABLES

Chapter 2

Chapter 3

Chapter 4

ACKNOWLEDGEMENTS

Thank you to the many colleagues and mentors that have trained and guided me throughout my scientific career. Working with you has been such a wonderful and enriching experience. Specifically, I would like to thank Kathrin Plath for her excellent mentorship. I am forever indebted to her for giving me a chance and showing me what it takes to be a professional scientist. Thank you to the members of my thesis committee. Their advice and support was truly invaluable. I would also like to recognize all of the individuals associated with the Neuroscience IDP, MSTP, and Department of Biological Chemistry for creating an outstanding training environment that is a privilege to be a part of.

Thank you to my family for supporting me in my scientific endeavors despite the many years and long hours. I love you all very much.

Chapter 2 is a version of a manuscript in preparation for publication authored by Ryan Schmidt and Kathrin Plath.

Chapter 3 is a version of a manuscript in preparation for publication authored by Ryan Schmidt, Bernadette Papp, Ajay Vashischt, Robin McKee, Xiaofen Chen, Alissa Minkovsky, Michael Carey, James Wohlschlegel, and Kathrin Plath.

Chapter 4 is a version of a manuscript in preparation for publication authored by Ryan Schmidt, Bernadette Papp, Anastasia Vedenko, Robin McKee, Reid Johnson, Martha Bulyk, and Kathrin Plath.

This work was supported by the NIH Director's Young Innovator Award (DP2OD001686) and a CIRM Young Investigator Award (RN1-00564) awarded to K.P.

VITA

Education

1999-2003	Bachelor of Arts in Biochemistry and Biology Master of Science in Chemistry University of Pennsylvania Philadelphia, PA

Honors and Awards

2006	Graduate Research and Education in Adaptive Biotechnology (GREAT) Individual Training Award from the UC Biotechnology & Research Program (UC BREP)
2010	UCLA Dissertation Year Fellowship

Publications

Yamamoto M, Watt CW, Schmidt RJ, Kuscuoglu U, Miesfeld RL, and Goldhammer DJ. Cloning and characterization of a novel MyoD enhancer-binding factor. Mech. Dev. 124(9-10):715-28, 2007.

CHAPTER 1

INTRODUCTION

Vertebrate embryonic development is composed of a remarkable series of events that rapidly transform a single zygote into a functional, patterned organism. As cells mature, they become increasingly restricted in their developmental potential and specialized in their function. The progressive loss of developmental potential that occurs during normal development is not typically reversible and is governed by a layer of epigenetic control that is imposed on the various cells of the organism, each of which contains identical genetic information in its nucleus.

Nuclear Reprogramming

Given that each cell carries the same genetic blueprint, it is possible through experimental manipulation to reverse the effects of development on a cell nucleus by resetting its epigenetic state. This was first carried out in amphibians by transferring a somatic cell nucleus into an enucleated oocyte [1]. This procedure led to the generation of sexually mature organisms derived from a nucleus which had previously been restricted in its developmental potential [2]. Somatic cell nuclear transfer (SCNT) was later used to generate viable sheep and mice from differentiated adult nuclei [3, 4]. These results indicate that oocyte cytoplasm from a variety of species contains an activity that is able to reprogram a cell nucleus.

Nuclei have also been reprogrammed by fusing somatic cells with embryonic stem (ES) or embryonic germ (EG) cells, which both exhibit pluripotency *in vitro* [5, 6]. When two cells are fused, the earlier developmental state acts in a dominant fashion to erase the epigenetic marks that restrict the potential of the differentiated nucleus. Nuclear reprogramming has also been achieved by permeabilizing a somatic cell and incubating it briefly in embryonic carcinoma (EC) or ES cell extract [7, 8]. Thus, the properties that control developmental potential can be transmitted by soluble factors that exist in these cell types.

Somatic Cell Reprogramming Using Defined Factors

Once the presence of nuclear reprogramming activity in pluripotent cells was appreciated, investigators sought to isolate the factors responsible. Transcription factors were good candidates given their known roles as master regulators of various cell fates. For example, the expression of the transcription factor, MyoD, had been shown to be sufficient to induce myogenesis in fibroblasts [9]. Conditions for culturing pluripotent cells from the inner cell mass of the blastocyst had been defined [10, 11], which subsequently allowed large amounts of these cells to be grown and analyzed for the presence of similar master regulators. Comparative gene expression analysis and other studies of ES cells identified candidate factors associated with the pluripotent state [12]. Remarkably, a cocktail of four transcription factors - Oct4, Sox2, Klf4, and c-Myc - expressed in mouse embryonic fibroblasts (MEFs) was sufficient to reprogram these cells to an ES-like, pluripotent state [12-14]. Cells created through this procedure are referred to as induced pluripotent stem (iPS) cells.

Follow-up experiments demonstrated that iPS cells could also be generated from human fibroblasts [15, 16]. Additionally, iPS cells were made from a variety of adult cell types, highlighting the broad utility of this reprogramming technique [17-20]. Finally, reprogramming was found to occur in the absence of c-Myc, albeit with reduced efficiency [21, 22].

Somatic cell reprogramming is relatively slow and inefficient when compared with SCNT. In a typical experiment, several days and multiple cell divisions are required before the reactivation of pluripotency markers, such as Nanog, is observed. Furthermore, reprogramming to the pluripotent state only occurs in a small fraction of the starting cells. This phenomenon is not simply attributable to variations in reprogramming factor expression between cells, since less than 2% of genetically identical MEFs carrying integrated, doxycycline-inducible reprogramming factors form iPS colonies [23]. In addition, reprogramming factor induction

must be sustained for at least 7 days in order to induce pluripotency [23]. Given that only a small minority of cells reach the pluripotent state in a standard reprogramming experiment, it was proposed that reprogrammable cells may represent an elite minority with increased developmental potential. However, this notion was dispelled by monitoring parallel reprogramming cultures derived from a single cell clone carrying inducible factors in an extended reprogramming experiment [24]. After 3-4 weeks, the length of most reprogramming experiments, only a small fraction of cultures contained Nanog+ cells [24]. However, after 18 weeks, 93% of wells had undergone reprogramming, proving that reprogramming is a stochastic process with variable latency [24]. Despite this observation that each cell in a culture is capable of being reprogrammed, cell types with increased developmental potential can be converted to iPS cells with greater efficiencies [25].

Comparison of ES and iPS Cells

Following the initial production of iPS cells, research was directed towards examining the degree of similarity between these cells and ES cells. ES cells differentiate into cells from all three germ layers both *in vitro* and when injected into nude mice to form teratomas. Additionally, mouse ES cells contribute to chimeric mice when injected into developing blastocysts.

At first glance, ES and iPS cells are virtually identical. They appear similar morphologically and grow under the same cell culture conditions [12, 13]. Initial functional assays demonstrated that iPS cells also form well-differentiated teratomas and contribute to chimeric mice [13]. Eventually, entire organisms were generated from iPS cells through tetraploid complementation [26-28]. Genomics techniques demonstrated that the patterns of gene expression and chromatin marks are roughly equivalent between ES and iPS cells [29-31].

These cells were even found to be indistinguishable when analyzed by infrared spectroscopy, which measures their internal chemical compositions [32].

Notwithstanding these results, some subtle differences that distinguish ES and iPS cells have been reported. Differences in gene expression signatures and copy number variation can be seen between ES and early passage iPS cells [33, 34]. Interestingly, these differences are lessened with extended passaging [33, 34]. Additionally, iPS cells contain regions with incompletely reprogrammed DNA methylation [35, 36]. These sites of epigenetic memory may be responsible for observed differences in differentiation potential between iPS cells lines associated with their cell type of origin [36, 37]. Finally, both reprogramming factor stoichiometry and the expression status of a single gene cluster have been associated with the ability of a given iPS cell line to form viable mice through tetraploid complementation [38, 39].

Utility of Somatic Cell Reprogramming

The generation of human iPS cells has opened the door for the widespread use of somatic cell reprogramming in the laboratory and clinical settings. iPS cells, like ES cells, are karyotypically normal cells that can be propagated indefinitely *in vitro*. Differentiation of either ES or iPS cells represents a useful means of obtaining a specific cell type of interest for scientific investigations. Previously, investigators relied on primary cell cultures, which can be difficult to obtain and expand, or transformed cells, which can be easily expanded but harbor significant genetic mutations. Oftentimes, researchers seek to create targeted genetic lesions in normal cells in culture in order to model the effects of a given disease. While this process is somewhat efficient in mouse ES cells, it has been much more difficult to carry out in human ES cells. Additionally, it is not possible to model the effects of complex or unknown genetic lesions using these techniques. Somatic cell reprogramming allows for a way around these technical barriers.

Disease-specific iPS cells can be generated directly from affected patients and then differentiated for use in basic research or high-throughput drug screening. This approach has already led to the successful generation of a large number of iPS cell-based disease models that accurately recapitulate relevant disease phenotypes.

Despite the widespread use of hematopoietic stem cell transplantation, cell-based therapies largely represent an untapped resource in the fight against human disease. Somatic cell reprogramming represents an important future weapon within this arsenal. Patient-specific iPS cells could be generated and converted into needed cell types that could then be reintroduced back into a patient for therapeutic purposes. This technique is well suited for the replacement of a single cell type that is absent or has been destroyed, as is observed in the case of autoimmune destruction of pancreatic β-cells in type I diabetes mellitus. The autologous transplantation strategy should not activate an immune response and thus would not require dangerous regimens of immunosuppression. The alleviation of the potential for complications due to immune rejection and immunosuppressive therapy alters the clinical risk-reward calculus, prospectively allowing these cells to be used as a treatment for a wide range of conditions.

However, regenerative therapies involving iPS cells come with their own set of risks. iPS cells have an incredible proliferative capacity and the ability to differentiate into any cell type in the body given the appropriate cues. Therefore, their accidental introduction in the undifferentiated state may lead to unpredictable consequences due to unchecked proliferation or disruption of normal tissue. Additionally, iPS cells retain a small number of somatic mutations that may increase the risk of neoplastic transformation of their differentiated progeny [40]. The currently used method of reprogramming factor expression depends on engineered viruses that integrate transgenes into the genome of a target cell. These random integration events may cause

insertional mutagenesis, which would also raise the risk of neoplasia in iPS-derived cells. Nucleic acid-free factor delivery based on protein transduction has been developed to overcome this concern, although reprogramming efficiency is greatly reduced [41, 42].

Molecular Mechanisms of Reprogramming Factor Activity

Once the therapeutic potential of somatic cell reprogramming became apparent, the reprogramming factor mechanism of action emerged as an important topic of investigation. Insights into the reprogramming mechanism may allow for improvement in the efficiency and fidelity of iPS cell generation. Moreover, these investigations may shed light on the large-scale reorganization of the epigenome that occurs during development in response to the action of transcription factors.

In this dissertation, chapter 2 reviews what is currently known regarding the molecular mechanisms of reprogramming factor activity. Then, attention will be focused exclusively on Klf4 in chapters 3 and 4. Work presented in chapter 3 characterizes regions of the Klf4 molecule that are important for its ability to induce pluripotency. Additionally, critical residues within its main transactivation domain are mapped and a novel cofactor protein that interacts with this region is identified. Chapter 4 reveals the molecular determinants within the Klf4 DNA binding domain that enable its reprogramming activity. This work also determines the DNA binding specificities of proteins within the Klf family and proposes a structural explanation for the observed differences. Finally, chapter 5 summarizes these results and outlines future directions in the study of the mechanism of induced pluripotency.

References

1. Briggs R, King TJ: **Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs.** *Proc Natl Acad Sci USA* 1952, **38:**455-463.
2. Gurdon JB, ELSDALE TR, FISCHBERG M: **Sexually mature individuals of Xenopus laevis from the transplantation of single somatic nuclei.** *Nature* 1958, **182:**64-65.
3. Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KH: **Viable offspring derived from fetal and adult mammalian cells.** *Nature* 1997, **385:**810-813.
4. Wakayama T, Perry AC, Zuccotti M, Johnson KR, Yanagimachi R: **Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei.** *Nature* 1998, **394:**369-374.
5. Tada M, Tada T, Lefebvre L, Barton SC, Surani MA: **Embryonic germ cells induce epigenetic reprogramming of somatic nucleus in hybrid cells.** *EMBO J* 1997, **16:**6510-6520.
6. Tada M, Takahama Y, Abe K, Nakatsuji N, Tada T: **Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells.** *Curr Biol* 2001, **11:**1553-1558.
7. Freberg CT, Dahl JA, Timoskainen S, Collas P: **Epigenetic reprogramming of OCT4 and NANOG regulatory regions by embryonal carcinoma cell extract.** *Molecular Biology of the Cell* 2007, **18:**1543-1553.
8. Taranger CK, Noer A, Sørensen AL, Håkelien A-M, Boquest AC, Collas P: **Induction of dedifferentiation, genomewide transcriptional programming, and epigenetic reprogramming by extracts of carcinoma and embryonic stem cells.** *Molecular Biology of the Cell* 2005, **16:**5719-5735.
9. Tapscott SJ, Davis RL, Thayer MJ, Cheng PF, Weintraub H, Lassar AB: **MyoD1: a nuclear phosphoprotein requiring a Myc homology region to convert fibroblasts to myoblasts.** *Science* 1988, **242:**405-411.
10. Evans MJ, Kaufman MH: **Establishment in culture of pluripotential cells from mouse embryos.** *Nature* 1981, **292:**154-156.
11. Martin GR: **Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells.** *Proc Natl Acad Sci USA* 1981, **78:**7634-7638.
12. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell* 2006, **126:**663-676.
13. Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R: **In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state.** *Nature* 2007, **448:**318-324.
14. Okita K, Ichisaka T, Yamanaka S: **Generation of germline-competent induced pluripotent stem cells.** *Nature* 2007, **448:**313-317.
15. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S: **Induction of pluripotent stem cells from adult human fibroblasts by defined factors.** *Cell* 2007, **131:**861-872.
16. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, et al: **Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells.** *Science* 2007, **318:**1917-1920.

17. Aoi T, Yae K, Nakagawa M, Ichisaka T, Okita K, Takahashi K, Chiba T, Yamanaka S: **Generation of pluripotent stem cells from adult mouse liver and stomach cells.** *Science* 2008, **321:**699-702.
18. Hanna J, Markoulaki S, Schorderet P, Carey BW, Beard C, Wernig M, Creyghton MP, Steine EJ, Cassady JP, Foreman R, et al: **Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency.** *Cell* 2008, **133:**250-264.
19. Stadtfeld M, Brennand K, Hochedlinger K: **Reprogramming of pancreatic beta cells into induced pluripotent stem cells.** *Curr Biol* 2008, **18:**890-894.
20. Utikal J, Maherali N, Kulalert W, Hochedlinger K: **Sox2 is dispensable for the reprogramming of melanocytes and melanoma cells into induced pluripotent stem cells.** *J Cell Sci* 2009, **122:**3502-3510.
21. Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochiduki Y, Takizawa N, Yamanaka S: **Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts.** *Nat Biotechnol* 2008, **26:**101-106.
22. Wernig M, Meissner A, Cassady JP, Jaenisch R: **c-Myc is dispensable for direct reprogramming of mouse fibroblasts.** *Cell Stem Cell* 2008, **2:**10-12.
23. Stadtfeld M, Maherali N, Borkent M, Hochedlinger K: **A reprogrammable mouse strain from gene-targeted embryonic stem cells.** *Nature Methods* 2010, **7:**53-55.
24. Hanna J, Saha K, Pando B, Zon Jv, Lengner CJ, Creyghton MP, Oudenaarden Av, Jaenisch R: **Direct cell reprogramming is a stochastic process amenable to acceleration.** *Nature* 2009, **462:**595-601.
25. Eminli S, Foudi A, Stadtfeld M, Maherali N, Ahfeldt T, Mostoslavsky G, Hock H, Hochedlinger K: **Differentiation stage determines potential of hematopoietic cells for reprogramming into induced pluripotent stem cells.** *Nat Genet* 2009, **41:**968-976.
26. Boland MJ, Hazen JL, Nazor KL, Rodriguez AR, Gifford W, Martin G, Kupriyanov S, Baldwin KK: **Adult mice generated from induced pluripotent stem cells.** *Nature* 2009, **461:**91-94.
27. Kang L, Wang J, Zhang Y, Kou Z, Gao S: **iPS Cells Can Support Full-Term Development of Tetraploid Blastocyst-Complemented Embryos.** *Stem Cell* 2009, **5:**135-138.
28. Zhao X-y, Li W, Lv Z, Liu L, Tong M, Hai T, Hao J, Guo C-l, Ma Q-w, Wang L, et al: **iPS cells produce viable mice through tetraploid complementation.** *Nature* 2009, **461:**86-90.
29. Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, Arnold K, Stadtfeld M, Yachechko R, Tchieu J, Jaenisch R, et al: **Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution.** *Cell Stem Cell* 2007, **1:**55-70.
30. Sridharan R, Tchieu J, Mason MJ, Yachechko R, Kuoy E, Horvath S, Zhou Q, Plath K: **Role of the murine reprogramming factors in the induction of pluripotency.** *Cell* 2008, **136:**364-377.
31. Guenther MG, Frampton GM, Soldner F, Hockemeyer D, Mitalipova M, Jaenisch R, Young RA: **Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells.** *Cell Stem Cell* 2010, **7:**249-257.
32. Sandt C, Féraud O, Oudrhiri N, Bonnet ML, Meunier MC, Valogne Y, Bertrand A, Raphaël M, Griscelli F, Turhan AG, et al: **Identification of spectral modifications occurring during reprogramming of somatic cells.** *PLoS ONE* 2012, **7:**e30743.

33. Chin MH, Mason MJ, Xie W, Volinia S, Singer M, Peterson C, Ambartsumyan G, Aimiuwu O, Richter L, Zhang J, et al: **Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures.** *Cell Stem Cell* 2009, **5:**111-123.
34. Hussein SM, Batada NN, Vuoristo S, Ching RW, Autio R, Närvä E, Ng S, Sourour M, Hämäläinen R, Olsson C, et al: **Copy number variation and selection during reprogramming to pluripotency.** *Nature* 2011, **471:**58-62.
35. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, et al: **Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells.** *Nature* 2011, **471:**68-73.
36. Kim K, Doi A, Wen B, Ng K, Zhao R, Cahan P, Kim J, Aryee MJ, Ji H, Ehrlich LIR, et al: **Epigenetic memory in induced pluripotent stem cells.** *Nature* 2010, **467:**285-290.
37. Polo JM, Liu S, Figueroa ME, Kulalert W, Eminli S, Tan KY, Apostolou E, Stadtfeld M, Li Y, Shioda T, et al: **Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells.** *Nat Biotechnol* 2010, **28:**848-855.
38. Carey BW, Markoulaki S, Hanna JH, Faddah DA, Buganim Y, Kim J, Ganz K, Steine EJ, Cassady JP, Creyghton MP, et al: **Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells.** *Cell Stem Cell* 2011, **9:**588-598.
39. Stadtfeld M, Apostolou E, Akutsu H, Fukuda A, Follett P, Natesan S, Kono T, Shioda T, Hochedlinger K: **Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells.** *Nature* 2010, **465:**175-181.
40. Gore A, Li Z, Fung H-L, Young JE, Agarwal S, Antosiewicz-Bourget J, Canto I, Giorgetti A, Israel MA, Kiskinis E, et al: **Somatic coding mutations in human induced pluripotent stem cells.** *Nature* 2011, **471:**63-67.
41. Cho H-J, Lee C-S, Kwon Y-W, Paek JS, Lee S-H, Hur J, Lee EJ, Roh T-Y, Chu I-S, Leem S-H, et al: **Induction of pluripotent stem cells from adult somatic cells by protein-based reprogramming without genetic manipulation.** *Blood* 2010, **116:**386-395.
42. Kim D, Kim C-H, Moon J-I, Chung Y-G, Chang M-Y, Han B-S, Ko S, Yang E, Cha KY, Lanza R, Kim K-S: **Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins.** *Cell Stem Cell* 2009, **4:**472-476.

CHAPTER 2

ROLES OF THE REPROGRAMMING FACTORS DURING IPS CELL GENERATION

Abstract

Somatic cell reprogramming by defined factors is a form of engineered reverse development carried out in an *in vitro* cell culture. This process, which is most often initiated by forced expression of three transcription factors (Oct4, Sox2, and Klf4), leads to a dramatic reorganization of the epigenome and concurrent change in gene expression that ultimately results in the induced pluripotent stem (iPS) cell fate. Recent investigation has begun to elucidate the molecular mechanisms whereby these factors function during reprogramming.

Introduction

Current reprogramming technology, pioneered by Takahashi and Yamanaka [1], was built on several seminal advances in the field of developmental biology. First, nuclear transfer experiments demonstrated that a somatic cell nucleus could be epigenetically reset to an early developmental state. Second, cell culture conditions were developed that allowed for the isolation and culture of pluripotent cells, termed embryonic stem (ES) cells, from the inner cell mass of the blastocyst. Finally, study of these cells led to the identification of candidate factors that were ultimately able to reprogram mouse embryonic fibroblasts (MEFs) to the iPS state [1].

Several groups rapidly followed up on the initial generation of iPS cells and demonstrated that these cells are functionally equivalent to ES cells in their ability to contribute to healthy adult mice and their offspring in addition to forming teratomas when injected into athymic mice [2-7]. In accordance with these results, the gene expression and chromatin states of iPS cells were also found to be strikingly similar to their ES counterparts [7-9].

Tremendous innovation has also occurred in the method of factor delivery to the somatic cells being reprogrammed. Initially, reprogramming factors were expressed from retroviral transgenes integrated into the genome. Subsequent advances have eliminated the requirement

for genomic insertion and viral infection altogether. Additionally, iPS cells have been generated from a variety of somatic cell types (reviewed in [10]). However, despite all of these advances, much remains to be learned about the reprogramming process itself. We believe that the MEF reprogramming paradigm still holds the most promise for answering these questions due to the ease of obtaining primary cells that are genetically tractable and easy to expand and reprogram. The next frontier for the reprogramming field will be a complete mechanistic understanding of how the factors cooperate to reshape the epigenome and gene expression profile of a target cell.

Enhancer and Replacement Factors

The core reprogramming cocktail, consisting of Oct4, Sox2, and Klf4 (OSK), can be augmented by the coexpression of factors that enhance the efficiency of iPS cell generation (Figure 2-1a). Most well known of these enhancer factors is c-Myc, which was added alongside OSK in the original reprogramming experiments but later shown to be dispensible [1, 7, 11, 12]. c-Myc, as well as family members N-Myc and L-Myc [11], are proto-oncogenes that act early in reprogramming to promote an active chromatin environment and enhance cell proliferation [9, 13]. In support of the notion that c-Myc acts mainly in early reprogramming stages, c-Myc greatly enhances the generation of trapped reprogramming intermediates (pre-iPS cells) when combined with OSK.

Several transcription factors normally expressed in the early stages of embryonic development enhance reprogramming. These include Glis1, Sall4, and Nanog [14-16]. This class of enhancer factors likely acts late in the reprogramming process to establish and stabilize the pluripotency transcription network. In contrast to c-Myc, Glis1 added to OSK enhances the generation of iPS colonies without producing Nanog-negative, putative pre-iPS colonies [15]. Remarkably, adding Glis1 and c-Myc together with OSK greatly enhances iPS colony formation

without the presence of Nanog-negative colonies, suggesting that Glis1 is able to coerce them to the fully reprogrammed state. Nanog overexpression in pre-iPS cells leads to their conversion to iPS cells, demonstrating its late-stage reprogramming activity [17, 18].

The ability of cells to pass through the cell cycle has also been shown to be an important determinant of reprogramming efficiency. Knockdown or gene deletion of p53, p21, or proteins expressed from the *Ink4/Arf* locus allows cells undergoing reprogramming to avoid the activation of cell cycle checkpoints and cellular senescence, leading to greater iPS formation [19]. Consequently, it is likely that any manipulation that accelerates the cell cycle would enhance reprogramming. Thus, reprogramming cultures should be monitored for alterations in their proliferation rate to determine whether the action of an enhancer factor can be attributed to changes in the cell cycle (Figure 2-1a).

In summary, the induction of pluripotency by OSK is a multistep progression whose course can be accelerated by enhancer factors. The generation of pre-iPS cells and the conversion of these cells to the fully reprogrammed state allows one to assay for enhancers of the early and late stages of reprogramming, respectively. It will be important to identify the subset of genes whose expression is changed by the introduction of each enhancer factor. Do these genes work alongside the core gene expression changes conferred by OSK or do they simply amplify the magnitude and kinetics of these changes?

Replacement factors possess the unique ability to substitute for Oct4, Sox2, or Klf4 in reprogramming (Figure 2-1b). Esrrb, an orphan nuclear receptor that is expressed highly in ES cells, has been reported to replace Klf4 [20]. Additionally, p53 knockdown has been shown to permit reprogramming in the absence of Klf4 [21].

High-throughput screens have been used successfully to identify small molecule replacement factors. Treatment of cells with kenpaullone allows reprogramming to occur without Klf4 [22], and several distinct classes of small molecules contribute to iPS cell generation in the absence of Sox2 [23-25].

Reprogramming enhancer and replacement factors are not necessarily mutually exclusive. Nr5a2, for instance, is capable of both enhancing reprogramming and replacing Oct4 [26]. In the human reprogramming system, Lin28 and Nanog, mentioned above as enhancer factors, combine to replace Klf4 [27].

Replacement factors, despite their substantial molecular and functional divergence, may provide important insights into the mechanism whereby OSK function in reprogramming. Future work will demonstrate whether these factors regulate the same key genes and pathways as the reprogramming factors that they replace or whether they help achieve the iPS end state via different means.

Gene Expression Changes During Reprogramming

The introduction of OSK brings about a dramatic change in the MEF transcriptional profile that eventually leads to induced pluripotency. Of the genes examined by Sridharan et al. ([9]; GSE14012) using expression microarrays, more than 6,000 change their expression by >2-fold between MEFs and iPS cells (Figure 2-2a; 3,562 up/3,239 down; median overall fold change=1.59). Expression changes in response to reprogramming factor induction begin immediately; however, the pluripotent state is not achieved until several days later. Hierarchical clustering of data obtained from a reprogramming timecourse demonstrates that reprogramming can be separated into 3 distinct gene expression phases [28].

The first of these phases includes downregulation of lineage-specific genes and activation of a genetic program that radically alters cell morphology [28]. This change, known as mesenchymal-to-epithelial transition (MET), is activated by BMP/Smad signaling and inhibited by the TGF-β pathway [24, 28, 29]. The difference in morphology that results from MET is not simply cosmetic. For example, knockdown of *Cdh1*, which encodes the epithelial protein E-cadherin, significantly reduces reprogramming efficiency and iPS cell contribution to chimeric mice [29].

The intermediates generated in a reprogramming culture do not appear to be stable when factor expression is lost before pluripotency is achieved [28, 30]. In this instance, cells revert back to a MEF-like gene expression pattern [28]. In contrast, pre-iPS cells are a stable intermediate with maintained OSK and c-Myc overexpression [8, 9]. These cells have successfully downregulated fibroblast genes and initiated MET, but have not activated the self-reinforcing network of transcription that characterizes the ES/iPS state [8, 13].

Fully reprogrammed cells exhibit indefinite self-renewal and the capacity to differentiate into any of the cell types that make up the developing organism. These unique properties are governed by a complex transcriptional program involving many transcription factors, including OSK expressed from their endogenous loci [13, 31]. Transcription factors within the pluripotency network appear to work cooperatively to regulate genes. Genome-wide chromatin immunoprecipitation (ChIP) experiments demonstrate cobinding among these factors at levels well beyond what would be expected by chance [9, 13, 31]. Additionally, the presence of multiple factors at a given locus is associated with increased levels of ES/iPS cell-specific gene expression [9, 13, 31].

Knockdown of any one of a number of transcription factors leads to loss of the pluripotent state, indicating the interconnected nature of the transcriptional network [32]. However, one factor - Nanog - seems to be of special importance. Overexpression of Nanog was able to rescue several of the aforementioned loss-of-function effects and allows ES cells to maintain pluripotency in the absence of the growth factor, LIF [32-34]. Furthermore, reprogramming of Nanog-deficient cells proceeds to the pre-iPS state but cannot transition to the iPS state due to impaired upregulation of the pluripotency network [17]. These data illustrate Nanog's central role in the establishment and maintenance of pluripotency and are consistent with its role as a late-stage enhancer of reprogramming.

Now that transcription factors within the pluripotency network have been largely identified, future research can determine their relative importance by performing similar gain- and loss-of-function assays to those described above involving Nanog. Are all pluripotency-associated factors enhancers of reprogramming? Why or why not?

In addition to the changes in specific gene programs mentioned above, reprogramming fundamentally alters the cell in several important ways. ES/iPS cells have an altered cell cycle with a shortened G1 phase [35]. Thus, reprogrammed cells have a reduced doubling time, and a greater fraction of these cells reside in the later phases of the cell cycle [35]. In order to protect genomic integrity during early development, ES/iPS cells have an enhanced capacity for DNA repair [36, 37]. Finally, pluripotent cells have an increased nuclear:cytoplasmic ratio when compared to differentiated cells as shown by electron microscopy [38].

In accordance with the reduction in cell membrane surface area and secretory function relative to MEFs, iPS cells generally express genes whose products function outside of the nucleus at lower levels. Signficantly enriched cellular compartment gene ontology (GO) terms

within the list of genes whose expression is reduced at least 2-fold from MEFs to iPS cells include: Golgi apparatus, endoplasmic reticulum, and extracellular matrix (Figure 2-2a). Conversely, genes whose expression is up at least 2-fold relative to MEFs act primarily within the nucleus and are enriched for GO terms such as nuclear lumen, chromosome, and chromatin (Figure 2-2a).

One important class of nuclear proteins whose gene expression is increased dramatically in ES/iPS cells relative to MEFs is chromatin modifying complexes (Figure 2-2b) [39]. These molecular machines modulate gene expression partly by covalent and non-covalent modification of nucleosomes. The expression levels of physically associated subunits within these complexes are largely coordinately regulated during reprogramming. For example, transcripts encoding the DNA methyltransferases (Dnmts) and the components of the PRC2 complex are highly upregulated as cells progress to the pluripotent state (Figure 2-2b). On the other hand, the TFIID and MLL/Set complexes are more moderately upregulated as a whole, yet they contain highly upregulated individual subunits, which play important roles in pluripotency and reprogramming (Figure 2-2b; Taf7, Dpy30, and Wdr5) [40-42]. Finally, expression switches within chromatin modifying complexes may affect the induction of pluripotency. Smarcc1 (BAF155) replaces Smarcc2 (BAF170) in the specific form of the BAF complex expressed in pluripotent cells (Figure 2-2b) [43].

The presence of increased levels of chromatin modifying complexes in ES/iPS cells may serve one of two purposes. First, these proteins may contribute to the maintenance of the self-renewing, undifferentiated state. Examples of this class, where loss-of-function disrupts self-renewal, include Brg, Chd1, and Wdr5 [40, 43, 44]. Second, while a given protein may not be required for normal growth of ES/iPS cells, its presence may be required for the proper execution

of subsequent developmental events. Thus, a loss-of-function phenotype will only be detected upon differentiation, as is seen for PRC2, G9a, TAF3, and the DNA methyltransferases - Dnmt1, Dnmt3a, and Dnmt3b [44-48].

In addition to their roles in ES/iPS cells, several components of the chromatin modifying machinery have been shown to affect reprogamming efficiency. Knockdown of LSD1, as well as chemical inhibition of DNA methyltransferases and histone deacetylases, leads to enhanced reprogramming [49]. Also, overexpression of Jhdm1a, Jhdm1b/Kdm2b, and the SWI/SNF complex components - Brg1 and Baf155 - increases the efficiency of iPS cell generation [50, 51]. In contrast, knockdown of Chd1 and Wdr5 inhibits reprogramming in a cell proliferation-independent manner [40, 44]. Knockdown of candidate chromatin modifying proteins during human reprogramming identified DOT1L and members of the PRC1 and PRC2 complexes as modulators of reprogramming activity [52].

Chromatin Changes During Reprogramming

Epigenetic changes during reprogramming, most frequently seen in the posttranslational modification status of histone tails, are likely to be both cause and consequence of the previously mentioned changes in gene expression. Differences in H3K4me2 and H3K27me3 are detected rapidly upon reprogramming factor induction and oftentimes precede transcriptional upregulation of the underlying loci [53]. Shifts in the balance of "active" and "inactive" chromatin marks at proximal gene regulatory elements are highly correlated with transcriptional changes during reprogramming. ChIP experiments demonstrate that the promoter regions of many genes with the greatest expression increases from MEFs to iPS cells lose H3K27me3 and gain H3K4me3 [7, 9]. Similar to what has been observed regarding changes in gene expression, the resetting of chromatin marks does not appear to occur all at once. In support of this notion,

pre-iPS cells display an intermediate chromatin modification pattern that lies between the MEF and iPS states [9, 54].

High-throughput sequencing coupled with ChIP has allowed for the identification of putative distal regulatory elements based on combinations of chromatin marks. These "enhancer" regions have been mainly defined by the presence of H3K4me1 and H3K4me2 at sites that lie at a distance from transcription start sites, which are frequently marked by H3K4me3 [53, 55]. Chromatin at these distal sites is reset to an ES-like state over the course of the reprogramming process [53, 55]. In addition to promoting the proper expression of pluripotency-related genes, these sites may contribute to the developmental potential of pluripotent cells by maintaining a poised state that allows for the upregulation of lineage-specific genes in response to the appropriate signals [55]. Future studies that analyze more histone marks and incorporate machine learning techniques will help to better characterize these regions as well as other important chromatin states during iPS cell generation.

Over the course of reprogramming, cells experience dramatic global increases in a variety of "active" histone acetylation and methylation marks while H3K27me3 levels remain unchanged [54]. The majority of these changes occur during the late stages of reprogramming - between the pre-iPS and fully reprogrammed states [54]. Additionally, the number of heterochromatin foci per cell, as marked by HP1α, is reduced in iPS cells when compared to MEFs [54]. In accordance with this observation, electron spectroscopic imaging demonstrates that lineage-committed cells have compacted blocks of chromatin near the nuclear envelope that are not seen in the pluripotent state [56, 57]. The specific increase in "active" chromatin is somewhat surprising given that the expression levels of chromatin modifying complexes associated with both the deposition of "active" and "inactive" marks increase as reprogramming

proceeds. Overall, changes in chromatin structure and histone marks coupled with increased transcription of repeat regions indicate that the pluripotent state may possess a unique, open chromatin architecture [39].

Another epigenetic modification, DNA methylation, plays an important role in silencing key pluripotency genes, including *Oct4* and *Nanog*, as cells undergo differentiation. During reprogramming, this repressive mark must be erased in order to allow for the establishment of induced pluripotency [2, 6-8]. Bisulfite sequencing demonstrates that removal of DNA methylation from pluripotency loci is a late event that can be placed between the pre-iPS and iPS states in the reprogramming continuum [8]. Furthermore, the enhancement in reprogramming efficiency in response to the DNA methyltransferase inhibitor, 5-aza-cytidine, is greatest when it is added in a brief window towards the end of the reprogramming process [8].

Molecular Mechanisms of Reprogramming Factor Activity

Over the course of reprogramming, OSK vary considerably in their DNA binding patterns. Eventually, however, they adopt an ES-like binding configuration upon reaching the iPS state [9]. Genes that exhibit the largest expression changes during reprogramming are frequently bound by all three reprogramming factors in ES and iPS cells [9]. Increased factor binding at gene promoters is associated with higher levels of transcription, indicating that OSK work together to regulate genes primarily as transcriptional activators [9].

Reprogramming factors must navigate a dynamic chromatin landscape at the various stages of iPS cell generation. While it is plausible that DNA binding differences may be due in part to changes in local chromatin accessibility, OSK are not blocked by the presence of the repressive mark, H3K27me3 [9]. Despite this finding, future work may identify specific chromatin signatures that enable or inhibit reprogramming factor binding.

While there is considerable overlap between the ChIP profiles of all three factors in ES cells, Oct4 and Sox2 are found together most frequently, whereas Klf4 binds to approximately twice as many sites genome-wide as either of the other factors [9, 13, 31]. Oct4 and Sox2 can bind cooperatively to composite sox-oct motifs that are frequently found within the regulatory elements of important pluripotency genes [58-60]. These genes include those that encode Oct4 and Sox2 themselves, indicating that these factors act within autoregulatory positive feedback loops that help to reinforce the pluripotent state [58, 59].

Reprogramming factors can sometimes be functionally replaced by paralogs within their respective families (Figure 2-3a). Comparison of OSK with their paralogs grouped in terms of functional redundancy may provide insight into their mechanisms of action during reprogramming. The binding pattern in ES cells and DNA binding specificity *in vitro* measured for Klf4 overlaps substantially with Klf2 and Klf5 [61]. Only triple knockdown of all three of these proteins together is sufficient to induce the loss of pluripotency [61]. However, each of these factors may also play more nuanced roles in maintaining self-renewal [62]. During reprogramming, Klf2, Klf5, and another close family member, Klf1, have been reported to replace Klf4 with varying degrees of efficiency (Figure 2-3a) [11]. Sox2, on the other hand, can be replaced by several diverse family members from across its phylogenetic tree, but not others (Figure 2-3a) [11]. Interestingly, reprogramming activity can be activated in Sox17, a reprogramming-incompetent paralog, by point mutation of two residues to the corresponding amino acids in Sox2 [63]. This change enables cooperative binding with Oct4 at a specific subset of sox-oct motifs [63]. Thus, the physical association between Sox2 and Oct4 when bound to DNA is likely to be critical for the induction of pluripotency. Oct4 cannot be replaced by Oct1 or Oct6 in reprogramming, suggesting that it may possess divergent activity not seen in

other family members (Figure 2-3a) [11]. This difference in reprogramming activity among the POU factors may not be simply due to differences in DNA binding. Oct1 and Oct4 both bind cooperatively to sox-oct elements in the *Fgf4* enhancer, but only Oct4 promotes transcriptional activation of the gene [60, 64].

Each reprogramming factor contains a highly conserved domain that functions primarily to bind DNA in a sequence-specific manner (Figure 2-3b). The OSK DNA binding domains each have distinct evolutionary origins with differing modes of interacting with the double helix. Klf4 binds DNA through 3 tandem C2H2 zinc fingers that wrap around the major groove [65]. Arginine and histidine side chains that project into the major groove and make contacts with the electronegative surface presented by guanine dictate Klf4's GC-rich DNA binding motif (Figure 2-3c) [65]. Sox2 binds an AT-rich motif (Figure 2-3c) though its HMG box which forms an L-shaped binding surface that exclusively contacts the minor groove [66]. This unique shape, along with amino acid side chains that intercalate between the DNA base pair stacks, creates a substantial bend in the DNA that is important for its ability to activate transcription [66, 67]. Oct4 interacts with DNA through two separate domains containing helix-turn-helix motifs that each contact half sites within its DNA binding motif (Figure 2-3c) in a cooperative manner [68].

Additional residues that lie outside of the highly conserved DNA binding domains in OSK are also important for their ability to activate transcription and mediate reprogramming (Figure 2-3b). Klf4 possesses an acidic transactivation domain (TAD) that interacts non-covalently with SUMO-1 [69]. Oct4 contains TADs both N- and C-terminal of its DNA binding domains, while Sox2 contains several regions with transactivation activity C-terminal of its HMG box [70]. Since these regions were characterized using assays from different

developmental contexts, future work is needed to determine which of these TADs function in reprogramming.

Reprogramming efficiency can be enhanced by fusing TADs from other proteins to the reprogramming factors. Addition of a TAD from VP16 to Oct4 or Sox2 increases reprogramming efficiency [71, 72]. Fusion of the MyoD TAD to either terminus of Oct4 accelerates and enhances the induction of pluripotency [73]. This enhancement activity is highly specific, since a variety of other known TADs were unable to accomplish the same feat [73]. Additionally, the MyoD TAD was unable to replace the transactivation regions within the Oct4 protein, indicating that these TADs are functionally distinct [73]. Further investigation is needed to elucidate the mechanism by which these TADs cooperate with the reprogramming factors to enhance reprogramming.

The reprogramming factors likely effect changes in transcription through interaction with protein cofactors that recruit the RNA polymerase machinery or modify the local chromatin structure. Several of these cofactors have been identified thus far. For instance, Sox2 and Oct4 have been reported to bind to a complex of XPC, RAD23B, and CENT2 to mediate the transactivation of *Nanog* [74]. Loss-of-function experiments demonstrated that these proteins are important for ES cell pluripotency and somatic cell reprogramming [74]. Additionally, several proteomic studies have identified a multitude of candidate interacting proteins that warrant further study [75-78].

Reprogramming factor activity can also be modulated by posttranslational modifications (PTMs). Oct4 phosphorylation at S229 within the POU homeodomain reduces its transactivation activity possibly by impairing DNA binding as a result of the disruption of a hydrogen bond with the DNA backbone [66, 79]. Reprogramming activity is completely abolished in a

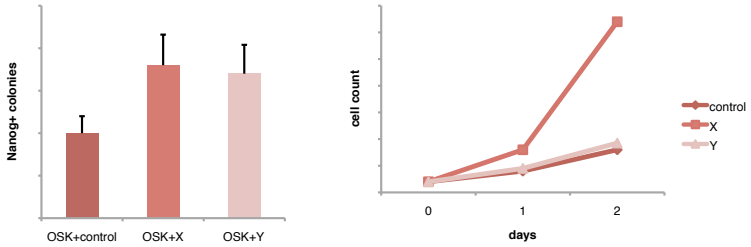
phosphomimetic mutant (S229D) protein [80]. Additionally, Oct4 can be O-GlcNAcylated at T228 [80]. Mutation of this residue to alanine substantially reduces reprogramming activity, indicating that this PTM may be important for the induction of pluripotency [80]. Given these results, it will be important to examine the effects of other known OSK PTMs during reprogramming.

The identification and study of defined reprogramming factors has helped to gain insight into the mechanism of induced pluripotency. Conversely, the process of reprogramming serves as a robust functional assay that allows us to advance our understanding of OSK. In a broad sense, knowledge gained through the study of somatic cell reprogramming may be applicable to other gene regulatory events that transform the epigenome and drive embryonic development.

Figure 2-1 Enhancer and replacement factors.

A) Example characterization of enhancer factors (X and Y). Enhancer factors may act through proliferation-dependent (X) or -independent mechanisms (Y). Example growth curves for MEFs infected with X, Y, or control retroviruses. **B**) Example characterization of a Sox2 replacement factor (Z).

A Enhancer Factors



B Reprogramming Factor Replacement

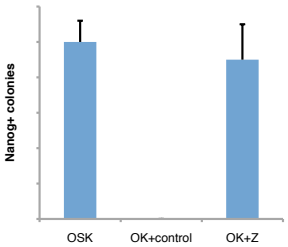
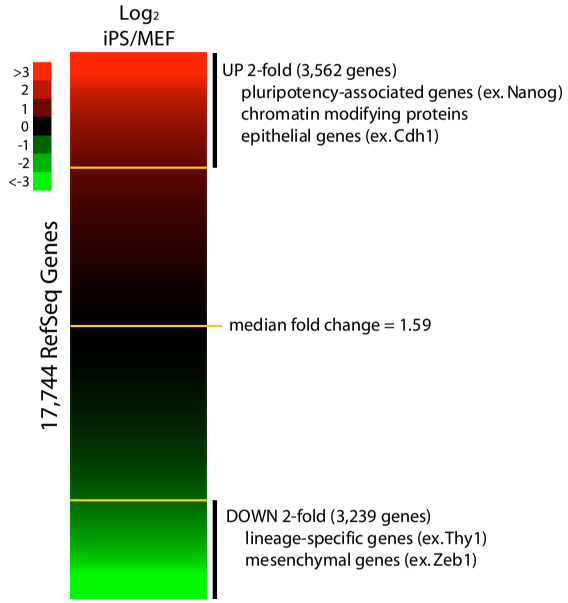


Figure 2-2 Gene expression changes during MEF reprogramming.

Data derived from Sridharan et al. [9]. **A**) Log_2 iPS/MEF expression ratios for each RefSeq gene ordered from highest to lowest. Selected enriched GO terms from genes with at least a 2-fold expression difference. **B**) Average log_2 expression ratios for selected chromatin modifying complexes. Red line indicates overall median expression change. Expression changes for individual complex subunits normalized to MEF value. Expression changes for Taf7 (green), Dpy30 (maroon), Wdr5 (purple), Smarcc1 (BAF155, red), and Smarcc2 (BAF170, blue) are highlighted.

A



UP 2-fold		Fold Enrichment(FDR)
GO:0031981	nuclear lumen	2.78(4.37E-82)
GO:0005694	chromosome	3.29(1.77E-51)
GO:0000785	chromatin	2.43(9.63E-08)

DOWN 2-fold		Fold Enrichment(FDR)
GO:0005794	Golgi apparatus	2.05(1.13E-25)
GO:0005783	endoplasmic reticulum	1.92(1.43E-25)
GO:0031012	extracellular matrix	2.48(1.06E-20)

B

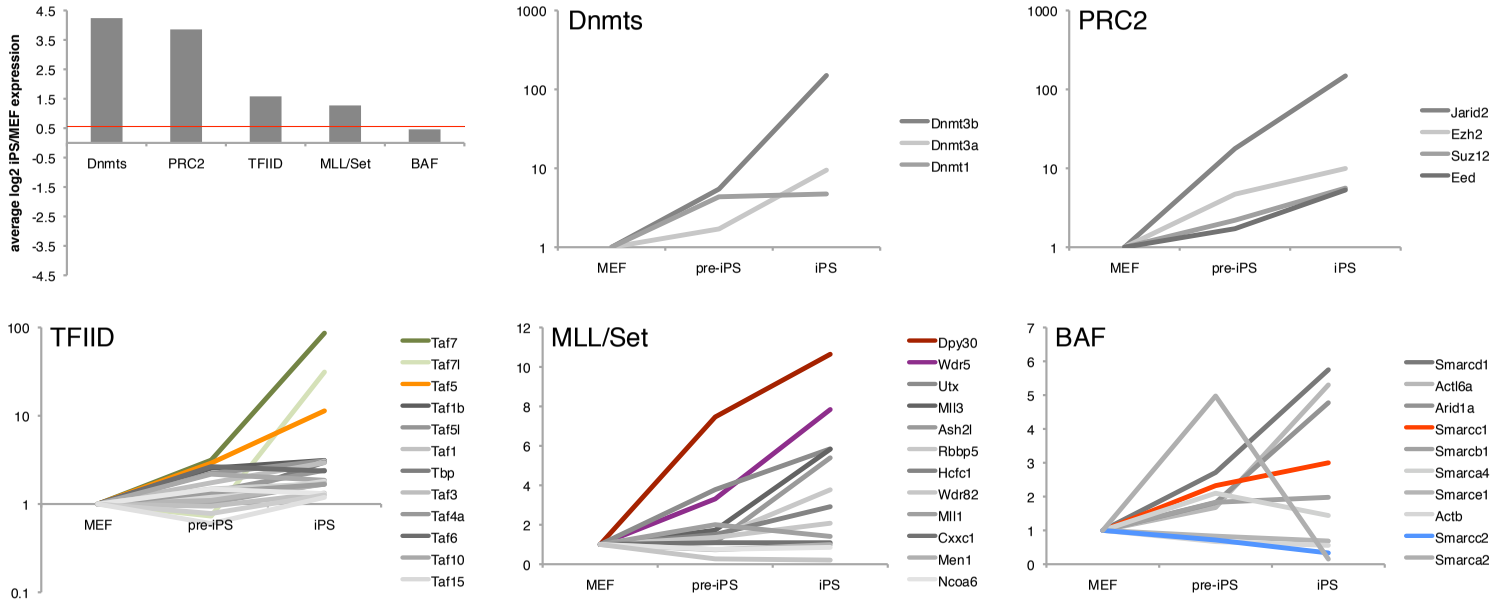
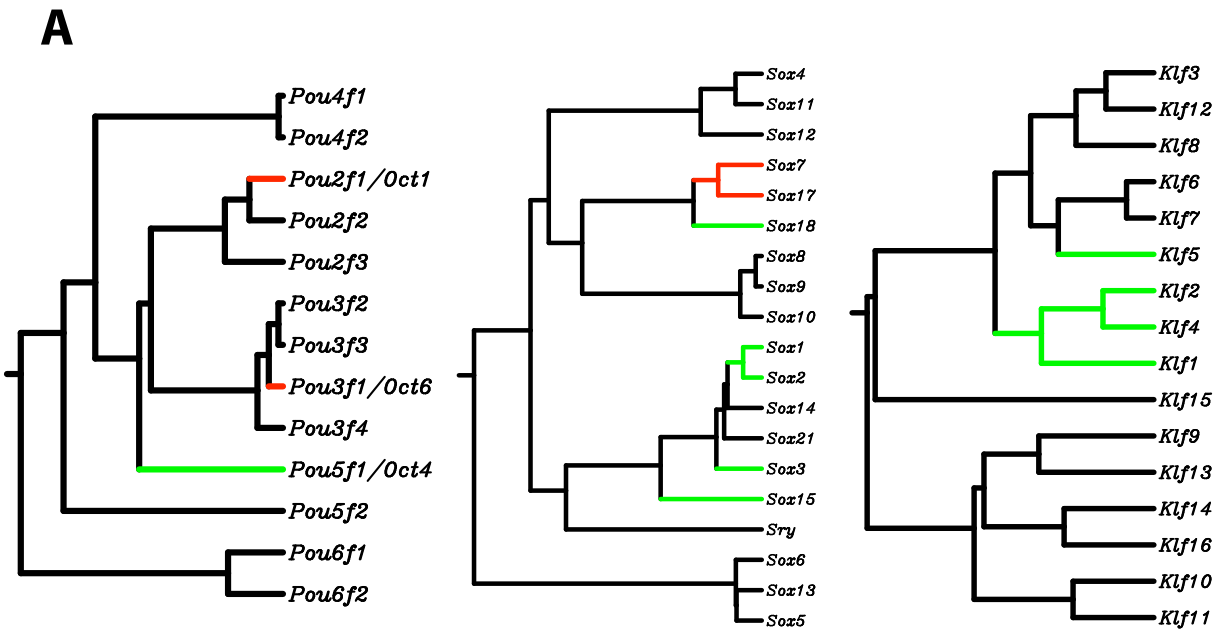
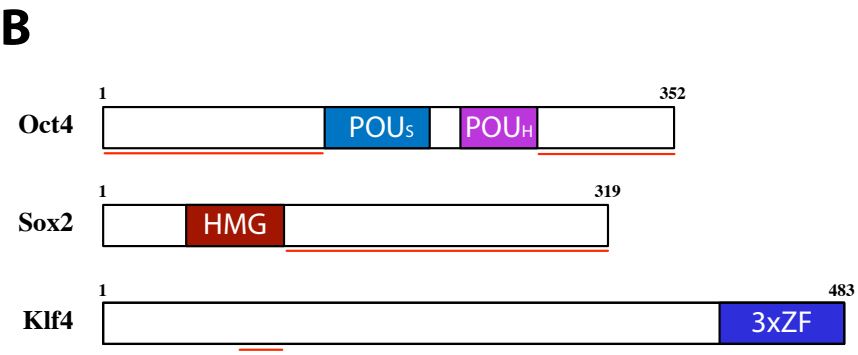
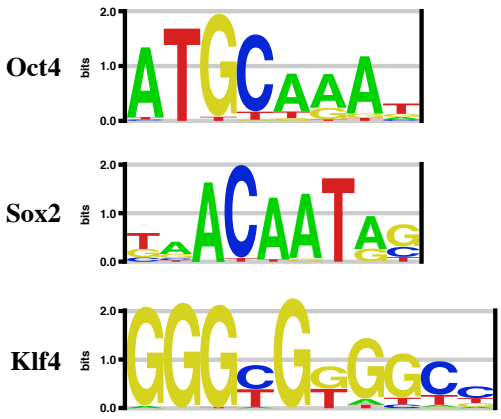


Figure 2-3 Reprogramming factors Oct4, Sox2, and Klf4.

A) Phylogenetic trees show evolutionary relationships to paralogs. Colors highlight family members that are able (green) or unable (red) to mediate reprogramming [11]. **B**) Schematic of each reprogramming factor with DNA binding domains indicated by colored boxes and transactivation domains underlined in red. **C**) Reprogramming factor DNA binding motifs determined by *de novo* motif discovery.


A
Pou4f1
Pou4f2
Pou2f1/Oct1
Pou2f2
Pou2f3
Pou3f2
Pou3f3
Pou3f1/Oct6
Pou3f4
Pou5f1/Oct4
Pou5f2
Pou6f1
Pou6f2
Sox4
Sox11
Sox12
Sox7
Sox17
Sox18
Sox8
Sox9
Sox10
Sox1
Sox2
Sox14
Sox21
Sox3
Sox15
Sry
Sox6
Sox13
Sox5
Klf3
Klf12
Klf8
Klf6
Klf7
Klf5
Klf2
Klf4
Klf1
Klf15
Klf9
Klf13
Klf14
Klf16
Klf10
Klf11


B
Oct4
1
352
POUS
POUH
Sox2
1
319
HMG
Klf4
1
483
3xZF


C
Oct4
Sox2
Klf4
bits
2.0
1.0
0.0

References

1. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell* 2006, **126:**663-676.
2. Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R: **In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state.** *Nature* 2007, **448:**318-324.
3. Zhao X-y, Li W, Lv Z, Liu L, Tong M, Hai T, Hao J, Guo C-l, Ma Q-w, Wang L, et al: **iPS cells produce viable mice through tetraploid complementation.** *Nature* 2009, **461:**86-90.
4. Kang L, Wang J, Zhang Y, Kou Z, Gao S: **iPS Cells Can Support Full-Term Development of Tetraploid Blastocyst-Complemented Embryos.** *Stem Cell* 2009, **5:**135-138.
5. Boland MJ, Hazen JL, Nazor KL, Rodriguez AR, Gifford W, Martin G, Kupriyanov S, Baldwin KK: **Adult mice generated from induced pluripotent stem cells.** *Nature* 2009, **461:**91-94.
6. Okita K, Ichisaka T, Yamanaka S: **Generation of germline-competent induced pluripotent stem cells.** *Nature* 2007, **448:**313-317.
7. Maherali N, Sridharan R, Xie W, Utikal J, Eminli S, Arnold K, Stadtfeld M, Yachechko R, Tchieu J, Jaenisch R, et al: **Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution.** *Cell Stem Cell* 2007, **1:**55-70.
8. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, Bernstein BE, Jaenisch R, Lander ES, Meissner A: **Dissecting direct reprogramming through integrative genomic analysis.** *Nature* 2008, **454:**49-55.
9. Sridharan R, Tchieu J, Mason MJ, Yachechko R, Kuoy E, Horvath S, Zhou Q, Plath K: **Role of the murine reprogramming factors in the induction of pluripotency.** *Cell* 2008, **136:**364-377.
10. Hochedlinger K, Plath K: **Epigenetic reprogramming and induced pluripotency.** *Development* 2009, **136:**509-523.
11. Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochiduki Y, Takizawa N, Yamanaka S: **Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts.** *Nat Biotechnol* 2008, **26:**101-106.
12. Wernig M, Meissner A, Cassady JP, Jaenisch R: **c-Myc is dispensable for direct reprogramming of mouse fibroblasts.** *Cell Stem Cell* 2008, **2:**10-12.
13. Kim J, Chu J, Shen X, Wang J, Orkin SH: **An extended transcriptional network for pluripotency of embryonic stem cells.** *Cell* 2008, **132:**1049-1061.
14. Tsubooka N, Ichisaka T, Okita K, Takahashi K, Nakagawa M, Yamanaka S: **Roles of Sall4 in the generation of pluripotent stem cells from blastocysts and fibroblasts.** *Genes Cells* 2009, **14:**683-694.
15. Maekawa M, Yamaguchi K, Nakamura T, Shibukawa R, Kodanaka I, Ichisaka T, Kawamura Y, Mochizuki H, Goshima N, Yamanaka S: **Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1.** *Nature* 2011, **474:**225-229.

16. Hanna J, Saha K, Pando B, Zon Jv, Lengner CJ, Creyghton MP, Oudenaarden Av, Jaenisch R: **Direct cell reprogramming is a stochastic process amenable to acceleration.** *Nature* 2009, **462:**595-601.
17. Silva J, Nichols J, Theunissen TW, Guo G, van Oosten AL, Barrandon O, Wray J, Yamanaka S, Chambers I, Smith A: **Nanog is the gateway to the pluripotent ground state.** *Cell* 2009, **138:**722-737.
18. Theunissen TW, van Oosten AL, Castelo-Branco G, Hall J, Smith A, Silva JC: **Nanog overcomes reprogramming barriers and induces pluripotency in minimal conditions.** *Curr Biol*, **21:**65-71.
19. Banito A, Rashid ST, Acosta JC, Li S, Pereira CF, Geti I, Pinho S, Silva JC, Azuara V, Walsh M, et al: **Senescence impairs successful reprogramming to pluripotent stem cells.** *Genes & Development* 2009, **23:**2134-2139.
20. Feng B, Jiang J, Kraus P, Ng J-H, Heng J-CD, Chan Y-S, Yaw L-P, Zhang W, Loh Y-H, Han J, et al: **Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb.** *Nat Cell Biol* 2009, **11:**197-203.
21. Kawamura T, Suzuki J, Wang YV, Menendez S, Morera LB, Raya A, Wahl GM, Izpisúa Belmonte JC: **Linking the p53 tumour suppressor pathway to somatic cell reprogramming.** *Nature* 2009, **460:**1140-1144.
22. Lyssiotis CA, Foreman RK, Staerk J, Garcia M, Mathur D, Markoulaki S, Hanna J, Lairson LL, Charette BD, Bouchez LC, et al: **Reprogramming of murine fibroblasts to induced pluripotent stem cells with chemical complementation of Klf4.** *Proc Natl Acad Sci USA* 2009, **106:**8912-8917.
23. Ichida JK, Blanchard J, Lam K, Son EY, Chung JE, Egli D, Loh KM, Carter AC, Di Giorgio FP, Koszka K, et al: **A small-molecule inhibitor of tgf-Beta signaling replaces sox2 in reprogramming by inducing nanog.** *Cell Stem Cell* 2009, **5:**491-503.
24. Maherali N, Hochedlinger K: **Tgfbeta signal inhibition cooperates in the induction of iPSCs and replaces Sox2 and cMyc.** *Curr Biol* 2009, **19:**1718-1723.
25. Shi Y, Desponts C, Do JT, Hahm HS, Schöler HR, Ding S: **Induction of pluripotent stem cells from mouse embryonic fibroblasts by Oct4 and Klf4 with small-molecule compounds.** *Cell Stem Cell* 2008, **3:**568-574.
26. Heng J-CD, Feng B, Han J, Jiang J, Kraus P, Ng J-H, Orlov YL, Huss M, Yang L, Lufkin T, et al: **The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells.** *Cell Stem Cell* 2010, **6:**167-174.
27. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, et al: **Induced pluripotent stem cell lines derived from human somatic cells.** *Science* 2007, **318:**1917-1920.
28. Samavarchi-Tehrani P, Golipour A, David L, Sung H-k, Beyer TA, Datti A, Woltjen K, Nagy A, Wrana JL: **Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming.** *Stem Cell* 2010, **7:**64-77.
29. Li R, Liang J, Ni S, Zhou T, Qing X, Li H, He W, Chen J, Li F, Zhuang Q, et al: **A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts.** *Cell Stem Cell* 2010, **7:**51-63.
30. Stadtfeld M, Maherali N, Breault DT, Hochedlinger K: **Defining molecular cornerstones during fibroblast to iPS cell reprogramming in mouse.** *Cell Stem Cell* 2008, **2:**230-240.

31. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al: **Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.** *Cell* 2008, **133:**1106-1117.
32. Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, DeCoste C, Schafer X, Lun Y, Lemischka IR: **Dissecting self-renewal in stem cells with RNA interference.** *Nature* 2006, **442:**533-538.
33. Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S, Smith A: **Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells.** *Cell* 2003, **113:**643-655.
34. Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, Maruyama M, Maeda M, Yamanaka S: **The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells.** *Cell* 2003, **113:**631-642.
35. White J, Dalton S: **Cell cycle control of embryonic stem cells.** *Stem Cell Rev* 2005, **1:**131-138.
36. Saretzki G, Armstrong L, Leake A, Lako M, von Zglinicki T: **Stress defense in murine embryonic stem cells is superior to that of various differentiated murine cells.** *Stem Cells* 2004, **22:**962-971.
37. Hong Y, Cervantes RB, Tichy E, Tischfield JA, Stambrook PJ: **Protecting genomic integrity in somatic cells and embryonic stem cells.** *Mutat Res* 2007, **614:**48-55.
38. Sampath P, Pritchard DK, Pabon L, Reinecke H, Schwartz SM, Morris DR, Murry CE: **A hierarchical network controls protein translation during murine embryonic stem cell self-renewal and differentiation.** *Cell Stem Cell* 2008, **2:**448-460.
39. Efroni S, Duttagupta R, Cheng J, Dehghani H, Hoeppner DJ, Dash C, Bazett-Jones DP, Le Grice S, McKay RD, Buetow KH, et al: **Global transcription in pluripotent embryonic stem cells.** *Cell Stem Cell* 2008, **2:**437-447.
40. Ang Y-S, Tsai S-Y, Lee D-F, Monk J, Su J, Ratnakumar K, Ding J, Ge Y, Darr H, Chang B, et al: **Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network.** *Cell* 2011, **145:**183-197.
41. Gegonne A, Tai X, Zhang J, Wu G, Zhu J, Yoshimoto A, Hanson J, Cultraro C, Chen Q-R, Guinter T, et al: **The general transcription factor TAF7 is essential for embryonic development but not essential for the survival or differentiation of mature T cells.** *Molecular and Cellular Biology* 2012, **32:**1984-1997.
42. Jiang H, Shukla A, Wang X, Chen W-y, Bernstein BE, Roeder RG: **Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalent domains.** *Cell* 2011, **144:**513-525.
43. Ho L, Ronan JL, Wu J, Staahl BT, Chen L, Kuo A, Lessard J, Nesvizhskii AI, Ranish J, Crabtree GR: **An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency.** *Proc Natl Acad Sci USA* 2009, **106:**5181-5186.
44. Tsumura A, Hayakawa T, Kumaki Y, Takebayashi S, Sakaue M, Matsuoka C, Shimotohno K, Ishikawa F, Li E, Ueda HR, et al: **Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b.** *Genes Cells* 2006, **11:**805-814.
45. Lei H, Oh SP, Okano M, Juttermann R, Goss KA, Jaenisch R, Li E: **De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells.** *Development* 1996, **122:**3195-3205.

46. Okano M, Bell DW, Haber DA, Li E: **DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.** *Cell* 1999, **99:**247-257.
47. Chamberlain SJ, Yee D, Magnuson T: **Polycomb repressive complex 2 is dispensable for maintenance of embryonic stem cell pluripotency.** *Stem Cells* 2008, **26:**1496-1505.
48. Tachibana M, Sugimoto K, Nozaki M, Ueda J, Ohta T, Ohki M, Fukuda M, Takeda N, Niida H, Kato H, Shinkai Y: **G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis.** *Genes & Development* 2002, **16:**1779-1791.
49. Huangfu D, Maehr R, Guo W, Eijkelenboom A, Snitow M, Chen AE, Melton DA: **Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds.** *Nat Biotechnol* 2008, **26:**795-797.
50. Liang G, He J, Zhang Y: **Kdm2b promotes induced pluripotent stem cell generation by facilitating gene activation early in reprogramming.** *Nat Cell Biol*.
51. Singhal N, Graumann J, Wu G, Araúzo-Bravo MJ, Han DW, Greber B, Gentile L, Mann M, Schöler HR: **Chromatin-Remodeling Components of the BAF Complex Facilitate Reprogramming.** *Cell* 2010, **141:**943-955.
52. Onder TT, Kara N, Cherry A, Sinha AU, Zhu N, Bernt KM, Cahan P, Mancarci OB, Unternaehrer J, Gupta PB, et al: **Chromatin-modifying enzymes as modulators of reprogramming.** *Nature* 2012.
53. Koche RP, Smith ZD, Adli M, Gu H, Ku M, Gnirke A, Bernstein BE, Meissner A: **Reprogramming factor expression initiates widespread targeted chromatin remodeling.** *Cell Stem Cell* 2011, **8:**96-105.
54. Mattout A, Biran A, Meshorer E: **Global epigenetic changes during somatic cell reprogramming to iPS cells.** *J Mol Cell Biol*, **3:**341-350.
55. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al: **Histone H3K27ac separates active from poised enhancers and predicts developmental state.** *Proc Natl Acad Sci U S A*, **107:**21931-21936.
56. Ahmed K, Dehghani H, Rugg-Gunn P, Fussner E, Rossant J, Bazett-Jones DP: **Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo.** *PLoS ONE* 2010, **5:**e10531.
57. Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, Papp B, Fussner E, Bazett-Jones DP, Plath K, Dalton S, et al: **Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis.** *Genome Res*, **20:**155-169.
58. Chew JL, Loh YH, Zhang W, Chen X, Tam WL, Yeap LS, Li P, Ang YS, Lim B, Robson P, Ng HH: **Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells.** *Mol Cell Biol* 2005, **25:**6031-6046.
59. Masui S, Nakatake Y, Toyooka Y, Shimosato D, Yagi R, Takahashi K, Okochi H, Okuda A, Matoba R, Sharov AA, et al: **Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells.** *Nat Cell Biol* 2007, **9:**625-635.
60. Ambrosetti DC, Basilico C, Dailey L: **Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites.** *Mol Cell Biol* 1997, **17:**6321-6329.

61. Jiang J, Chan Y-S, Loh Y-H, Cai J, Tong G-Q, Lim C-A, Robson P, Zhong S, Ng H-H: **A core Klf circuitry regulates self-renewal of embryonic stem cells.** *Nat Cell Biol* 2008, **10:**353-360.
62. Hall J, Guo G, Wray J, Eyres I, Nichols J, Grotewold L, Morfopoulou S, Humphreys P, Mansfield W, Walker R, et al: **Oct4 and LIF/Stat3 additively induce Kruppel factors to sustain embryonic stem cell self-renewal.** *Cell Stem Cell* 2009, **5:**597-609.
63. Jauch R, Aksoy I, Hutchins AP, Ng CKL, Tian XF, Chen J, Palasingam P, Robson P, Stanton LW, Kolatkar PR: **Conversion of Sox17 into a pluripotency reprogramming factor by reengineering its association with Oct4 on DNA.** *Stem Cells* 2011, **29:**940-951.
64. Yuan H, Corbi N, Basilico C, Dailey L: **Developmental-specific activity of the FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3.** *Genes & Development* 1995, **9:**2635-2645.
65. Schuetz A, Nana D, Rose C, Zocher G, Milanovic M, Koenigsmann J, Blasig R, Heinemann U, Carstanjen D: **The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation.** *Cell Mol Life Sci*, **68:**3121-3131.
66. Remenyi A, Lins K, Nissen LJ, Reinbold R, Scholer HR, Wilmanns M: **Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers.** *Genes Dev* 2003, **17:**2048-2059.
67. Scaffidi P, Bianchi ME: **Spatially precise DNA bending is an essential activity of the sox2 transcription factor.** *J Biol Chem* 2001, **276:**47296-47302.
68. Dailey L, Basilico C: **Coevolution of HMG domains and homeodomains and the generation of transcriptional regulation by Sox/POU complexes.** *J Cell Physiol* 2001, **186:**315-328.
69. Du JX, McConnell BB, Yang VW: **A small ubiquitin-related modifier-interacting motif functions as the transcriptional activation domain of Kruppel-like factor 4.** *J Biol Chem*, **285:**28298-28308.
70. Ambrosetti DC, Scholer HR, Dailey L, Basilico C: **Modulation of the activity of multiple transcriptional activation domains by the DNA binding domains mediates the synergistic action of Sox2 and Oct-3 on the fibroblast growth factor-4 enhancer.** *J Biol Chem* 2000, **275:**23387-23397.
71. Wang Y, Chen J, Hu J-L, Wei X-X, Qin D, Gao J, Zhang L, Jiang J, Li J-S, Liu J, et al: **Reprogramming of mouse and human somatic cells by high-performance engineered factors.** *EMBO Rep* 2011, **12:**373-378.
72. Hirai H, Katoku-Kikyo N, Karian P, Firpo M, Kikyo N: **Efficient iPS Cell Production with the MyoD Transactivation Domain in Serum-Free Culture.** *PLoS ONE* 2012, **7:**e34149.
73. Hirai H, Tani T, Katoku-Kikyo N, Kellner S, Karian P, Firpo M, Kikyo N: **Radical Acceleration of Nuclear Reprogramming by Chromatin Remodeling with the Transactivation Domain of MyoD.** *Stem Cells* 2011, **29:**1349-1361.
74. Fong YW, Inouye C, Yamaguchi T, Cattoglio C, Grubisic I, Tjian R: **A DNA repair complex functions as an Oct4/Sox2 coactivator in embryonic stem cells.** *Cell*, **147:**120-131.
75. Wang J, Rao S, Chu J, Shen X, Levasseur DN, Theunissen TW, Orkin SH: **A protein interaction network for pluripotency of embryonic stem cells.** *Nature* 2006, **444:**364-368.

76. van den Berg DL, Snoek T, Mullin NP, Yates A, Bezstarosti K, Demmers J, Chambers I, Poot RA: **An Oct4-centered protein interaction network in embryonic stem cells.** *Cell Stem Cell*, **6:**369-381.
77. Pardo M, Lang B, Yu L, Prosser H, Bradley A, Babu MM, Choudhary J: **An expanded Oct4 interaction network: implications for stem cell biology, development, and disease.** *Cell Stem Cell*, **6:**382-395.
78. Mallanna SK, Ormsbee BD, Iacovino M, Gilmore JM, Cox JL, Kyba M, Washburn MP, Rizzino A: **Proteomic analysis of Sox2-associated proteins during early stages of mouse embryonic stem cell differentiation identifies Sox21 as a novel regulator of stem cell fate.** *Stem Cells*, **28:**1715-1727.
79. Saxe JP, Tomilin A, Schöler HR, Plath K, Huang J: **Post-translational regulation of Oct4 transcriptional activity.** *PLoS ONE* 2009, **4:**e4467.
80. Jang H, Kim TW, Yoon S, Choi S-Y, Kang T-W, Kim S-Y, Kwon Y-W, Cho E-J, Youn H-D: **O-GlcNAc Regulates Pluripotency and Reprogramming by Directly Acting on Core Components of the Pluripotency Network.** *Cell Stem Cell* 2012.

CHAPTER 3

MAPPING OF KLF4 FUNCTIONAL DOMAINS IN REPROGRAMMING

Introduction

Somatic cell reprogramming refers to the conversion of a differentiated cell, such as a mouse embryonic fibroblast (MEF), into a pluripotent, embryonic stem cell (ESC)-like state. Reprogrammed cells, known as induced pluripotent stem cells (iPSCs), can be reliably generated through the ectopic expression of three transcription factors - Oct4, Sox2, and Klf4 - in a target cell [1, 2]. This procedure holds tremendous promise for regenerative therapies whereby patient-specific cells of a certain lineage could be generated following differentiation of their iPSCs.

Klf4 plays an essential role in the reprogramming process, although its mechanism of action is still widely under investigation. Cobinding of Klf4 with Oct4 and Sox2 during reprogramming occurs at the promoter regions of pluripotency-specific genes and is associated with their upregulation [3]. Knockdown of Klf4 in ESCs, along with its functionally redundant family members, leads to the loss of self-renewal [4]. On the contrary, the capacity of Klf4 to promote the pluripotent state is somewhat surprising given its anti-proliferative activity and role in the terminal differentiation of epithelial cells in the gut [5]. Klf4 has been described as both an oncogene as well as a tumor suppressor, depending upon the type of cancer under consideration [5]. Thus, it is plausible that expression of Klf4 leads to disparate effects in different cellular contexts.

Klf4 contains two well-characterized domains that may contribute to its ability to effect reprogramming. First, the highly conserved C-terminal DNA binding domain is known to bind sequence-specifically to a GC-rich motif [6, 7]. Second, a region near the N-terminus has been shown to possess potent transactivation activity, which is dependent upon its acidic residues [8]. However, the functional relevance of these and other domains in reprogramming has not yet been established.

Here, we use mutagenesis to probe the importance of various regions of Klf4 in reprogramming. We identify three domains that are essential for induced pluripotency. Furthermore, we characterize a novel transactivation domain (TAD) that functions in reprogramming in the absence of the well-studied acidic TAD. Within the acidic TAD, we find that reprogramming-specific transactivation is driven by a previously unidentified subregion in its C-terminal half containing critical hydrophobic residues. In a search for potential coactivators that associate with this region, we isolated clathrin heavy chain, which binds specifically through a consensus binding motif.

Results

Klf4 Deletion Mutagenesis Identifies Regions Required for Reprogramming

To identify regions of Klf4 that function in somatic cell reprogramming, we assayed deletion mutants for their ability to replace the full-length protein. Klf4 constructs were expressed in MEFs along with Oct4 and Sox2 using retroviruses on day 0 of the reprogramming procedure (Figure 3-1a). On day 5, the reprogramming culture was transitioned to media containing KSR (Figure 3-1a). Finally, each experiment was stopped on day 12 and reprogramming was quantified by counting Nanog+ colonies (Figure 3-1a).

Each construct was monitored for expression and subcellular localization defects that might confound the interpretation of a reprogramming result. Only mutants that exhibit similar expression levels, viral infection efficiencies, and subcellular localization to wild-type Klf4 are presented. These attributes were monitored by Western blotting and immunofluorescence against a FLAG epitope that was added to the N-terminus of each construct. Representative data from these control experiments is presented in Figure 3-2. Additionally, the presence of the FLAG epitope did not affect reprogramming efficiency (Figure 3-3b,c).

As a further control, we measured the effect of reduced viral titer on Nanog+ colony formation. Reduction in the amount of virus added led to a sub-linear reduction in reprogramming efficiency, indicating that the amount of virus added in our experiments approaches saturation of reprogramming efficiency (Figure 3-3a). Thus, large differences in reprogramming efficiencies cannot be explained by small fluctuations in viral titer.

Several separable regions within Klf4 are critical for its reprogramming activity. At the C-terminus, the highly conserved DNA binding domain, consisting of 3 tandem C2H2 zinc fingers, is necessary for iPSC colony formation (1-396, Figure 3-1b). Therefore, Klf4 likely must bind to target stretches of DNA in a sequence-specific manner to effect reprogramming.

Another region, which lies immediately N-terminal to the DNA binding domain and clearly outside of the zinc finger fold, is also required for reprogramming (Figure 3-1b). This region (residues 350-396) contains a basic stretch that has been previously identified as a nuclear localization signal (NLS, residues 383-396) [9]. However, Klf4 also contains another NLS within its DNA binding domain [9], and nuclear localization is unaffected by the Δ350-396 mutation (data not shown). It is notable that the basic residues contained in the 383-396 region are well-conserved in family members, Klf1 and Klf2, which can mediate reprogramming in place of Klf4 [1, 9]. Interestingly, a region containing several basic residues immediately N-terminal of the DNA binding domain within Sp1 was shown to be important for its ability to activate transcription [10].

Deletion of the first 89 amino acids from the N-terminus did not alter reprogramming efficiency. However, additional removal of the following 21 residues led to a dramatic, but not complete, reduction in the number of iPSC colonies observed (Figure 3-1b). This region (residues 90-110) contains a well-characterized acidic TAD [8, 11]. Deletion of this region or

mutation of its acidic residues leads to a complete loss of transactivation activity in CHO cells, HEK293T cells, and yeast [8, 11]. Furthermore, mutation of these acidic residues completely disrupts the growth-suppressing phenotype that results from wild-type Klf4 overexpression in Rat1a and COS-1 cells [8, 11]. In contrast to these prior observations from other assays, reprogramming activity remains in deletion mutants lacking residues 90-110, suggesting that residues 111-209 may also contain transactivation activity that functions specifically in the reprogramming context (Figure 3-1b). This putative activity is likely present in multiple places within residues 111-209 since the 170-483 construct still leads to iPSC colony formation, albeit at a lower efficiency than 111-483 (Figure 3-1b). Reprogramming activity was completely abolished in the 210-483 and 290-483 constructs, indicating that sequences within the 90-209 region are required to generate iPSCs (Figure 3-1b).

We sought to further assess the relative importance of portions of the 111-209 region by deleting them in the presence of the acidic TAD. Deletion of three subregions of 111-209 in this context had no effect on the generation of iPSC colonies (Figure 3-1b). Thus, these putative TADs are accessory to the acidic TAD (residues 90-110) and only impact reprogramming efficiency in its absence.

Klf4 has been previously shown to be post-translationally modified at several sites. Phosphorylation within a serine-rich region C-terminal to its acidic TAD has been reported to contribute to protein degradation and loss of ESC self-renewal [12]. Thus, it is tempting to hypothesize that deletion of these serines may lead to enhanced reprogramming. However, deletion of this region (residues 111-144) had no effect on reprogramming activity (Figure 3-1b). Additionally, Klf4 was shown to be SUMOylated at Lys 275 [11]. Mutation of this residue to arginine also did not affect reprogramming efficiency (Figure 3-4).

The Acidic TAD Exhibits Reprogramming-specific Function

We examined the ability of other well-studied TADs to rescue Klf4 deletion mutants lacking N-terminal sequences. Sp1 and VP16 contain powerful TADs that lead to robust transcriptional activation when tethered to the promoters of reporter genes [10, 13]. These domains are enriched for distinct classes of amino acids - Sp1 is glutamine-rich, while VP16 is acidic, analogous to Klf4 90-110. Previous experiments that fused the VP16 TAD to several reprogramming factors led to enhanced iPSC colony formation [14]. Additionally, fusion of the MyoD TAD to either terminus of Oct4 enhanced reprogramming [15]. However, the MyoD sequence could not substitute for the endogenous TADs within Oct4 [15]. Fusion of the Klf4, but not Sp1 or VP16, TAD to the reprogramming-deficient 210-483 construct resulted in a partial rescue of reprogramming activity (Figure 3-5a). Surprisingly, a C-terminal portion of the TAD consisting of residues 100-110 was also able to partially rescue the defect in iPSC colony formation (Figure 3-5a). This region lacks acidic residues that had previously been shown to be required for the transactivation and anti-proliferative functions of Klf4 [8, 11]. Thus, reprogramming seems to employ a specific subregion of the Klf4 TAD (residues 100-110).

Partial rescue may be due to the placement of the TAD adjacent to a region that had been previously shown to contain transcriptional repression activity [16]. Thus, we assayed the Klf4 90-110 and VP16 TADs for their ability to rescue the 170-483 mutant, which exhibits substantially reduced reprogramming activity relative to the full-length protein (Figure 3-5b). The Klf4 TAD fully restores reprogramming activity, while VP16 does not rescue the lost reprogramming function (Figure 3-5b).

Hydrophobic Residues within the Acidic TAD are Critical for Reprogramming

Given that the acidic TAD is important for reprogramming activity, we aimed to identify residues within this region that are required for reprogramming-specific transactivation. To do this, we made similar point mutations to those made by Du et al. [11] and tested them in the 90-483 deletion mutant, which reprograms MEFs with an efficiency similar to full-length Klf4 (1-483, Figure 3-6a). Previous work shows that mutation of three glutamate (Figure 3-6a, 3EA, red) or two aspartate residues (Figure 3-6a, 2DA, green) caused a complete loss of function in an anti-proliferation assay and multiple transactivation assays, whereas a double mutation of a leucine and an isoleucine residue (Figure 3-6a, LI, blue) led to partial loss of function [11]. Each of these mutations disrupted a non-covalent interaction with SUMO-1 that may be important for transcriptional activation [11]. Strikingly, mutation of the acidic residues causes little change in reprogramming efficiency, while the LI mutation leads to dramatically reduced reprogramming (Figure 3-6a). The level of reprogramming driven by 90-483 LI is similar to what was observed when the acidic TAD was deleted entirely (111-483, Figure 3-1b), suggesting that the leucine and isoleucine residues may be required for reprogramming-specific transactivation.

The LI mutation falls within the C-terminal subregion of the TAD that bears homology to putative acidic TADs in the N-terminal portions of several other reprogramming-competent Klf family members - Klf1, Klf2, and Klf5 [1, 17]. These regions contain a mixture of conserved acidic and hydrophobic residues. A 6-amino acid stretch is absolutely conserved in Klf2 (Figure 3-6a, underline). Klf2 is the most closely related paralog to Klf4 by sequence conservation and exhibits the strongest reprogramming activity of the other family members [1].

Our data indicate that Klf4 observes unique amino acid dependencies within its acidic TAD in the reprogramming context. We wondered if the decreased reprogramming activity of the 90-483 LI mutation can be explained by differences in its transactivation activity. To test

this hypothesis, we performed dual luciferase assays in two reprogramming-related cell types - MEFs and ESCs. ESCs were chosen because they are functionally equivalent to iPSCs. The 90-110 region functioned as a potent TAD in both MEFs (Figure 3-6b) and ESCs (Figure 3-6b), approaching the level of reporter activation seen from VP16. Mutation of the glutamate residues reduced transactivation ~10-fold in both cell types (Figure 3-6b,c, 90-110 3EA). However, this reduction was not nearly as dramatic as what was previously observed in similar assays in other cell types where transactivation activity was completely abolished [8, 11]. Also in contrast with previous findings, the LI mutation resulted in an almost complete loss of transactivation activity that greatly exceeded the reduction observed due to 3EA (Figure 3-6b,c) [11]. Therefore, the acidic TAD depends on different critical amino acids to activate transcription in different cell types as measured by reporter gene assays. The reprogramming activity attributable to the 90-110 region in the 90-483 LI construct correlates with its transactivation activity, suggesting that the LI mutation may lead to reduced reprogramming as a result of inefficient transcriptional activation. The 111-209 region, which is required to mediate reprogramming in the absence of the acidic TAD, possesses weak transactivation activity in MEFs and ESCs relative to residues 90-110 (Figure 3-6b,c). However, its reporter gene activation significantly exceeds 90-110 LI (Figure 3-6b,c). Thus, the residual reprogramming activity in the 111-483 and 90-483 LI constructs may be driven by weak transactivation activity contained in residues 111-209.

Clathrin Heavy Chain Binds to the Acidic TAD through a Clathrin-box Motif

We hypothesized that the leucine and/or isoleucine within the C-terminal portion of the Klf4 acidic activation domain altered by the LI mutation make specific contacts with a coactivator present in MEFs and ESCs to mediate reprogramming. Thus, we attempted to isolate proteins that bind specifically to the wild-type 90-110 sequence from ESC nuclear extract by

affinity purification. Silver staining of elution fractions separated on a polyacrylamide gel identified a high molecular weight protein in the 90-110 purification that was not found in 90-110 LI or GST only purifications (Figure 3-7a). Analysis of the first elution fraction from each purification suggests that this protein is clathrin heavy chain (Figure 3-7b). Searching within the primary sequence of the 90-110 region revealed the presence of a clathrin-box binding motif that is specifically disrupted by the leucine to alanine mutation in the LI mutant (Figure 3-7c). This motif perfectly matches the consensus of LϕXϕ[DE] (ϕ represents large hydrophobic residues), which recognizes a specific site in the N-terminal β-propeller domain of clathrin heavy chain [18].

Clathrin heavy chain is well known for its role in vesicle trafficking in the cytoplasm. However, recent work implicates this protein in transcriptional activation. A small fraction of the clathrin heavy chain protein in a cell was found to be present in the nucleus [19]. Additionally, clathrin heavy chain binds an acidic TAD within p53, and this interaction is important for the ability of p53 to activate transcription [19, 20]. Similar to what we have observed regarding Klf4, binding depends on hydrophobic residues within the acidic TAD, although these residues comprise a distinct motif that competes with clathrin light chain to make contact with a site within the C-terminal portion of clathrin heavy chain [19-21].

Discussion

Klf4 is a critical component of the reprogramming cocktail, originally identified by Takahashi and Yamanaka, which resets somatic cells to the pluripotent state [22]. In this study, we dissected the functional domains of Klf4 by mutagenesis to identify regions important for reprogramming. We determined that Klf4 contains three distinct regions that are required for its reprogramming activity - N-terminal TADs, the C-terminal DNA binding domain, and a C-

terminal region of unknown function. Follow-up experiments are necessary to determine whether this C-terminal region is required for the ability of Klf4 to activate transcription and whether this activity is related to the inability of the Δ350-396 construct to mediate reprogramming. The separable and modular nature of the transactivation and DNA binding domains indicates that Klf4 functions as a classic transcription factor in the model of proteins such as GAL4.

Within the N-terminal portion of the Klf4 protein, we identified a region (residues 111-209) containing weak transactivation activity that functions specifically in the reprogramming-relevant cell types used - MEFs and ESCs. The presence of this region allowed for iPSC generation in the absence of the much stronger acidic TAD (residues 90-110).

We found that a subregion of the TAD (residues 101-110) alone is sufficient to partially rescue the reprogramming function of a deletion mutant lacking the N-terminal TADs. Additionally, we identified a reprogramming-specific dependence on hydrophobic residues within the 90-110 TAD for reprogramming and transactivation activity. These results stand in stark contrast to previous findings regarding the functions of the residues within the 90-110 region using different assays in other cell types, suggesting that reprogramming may employ a distinct mechanism of transcriptional activation using this sequence.

We observed high levels of luciferase reporter activity in both MEFs and ESCs in response to the expression of the Klf4 90-110 fusion protein. Therefore, the coactivators that interact with this TAD during reprogramming are likely to be present throughout the entire process and would not need to be "unlocked" by a later reprogramming event. We also noticed a rough correlation between reprogramming efficiency and transactivation activity due to the N-terminal domains. For example, the defect in iPSC colony formation in the 90-483 LI mutant,

which had comparable reprogramming efficiency to 111-483, corresponds to the almost complete loss of transactivation activity observed from the 90-110 LI sequence in the luciferase reporter assay. Nevertheless, this putative relationship between reprogramming and transactivation activity is not perfectly linear. The reprogramming activity of the 90-483 3EA mutant is similar to the wild-type 90-483 construct, whereas the 3EA mutation results in a ~10-fold reduction in transactivation activity. However, despite this reduction, transactivation mediated by the both the wild-type and 3EA mutant acidic TADs is still quite strong. Thus, the transactivation function attributable to the three glutamate residues does not likely play an important role in reprogramming, while the transactivation activity disrupted by the LI mutation has a critical function in reprogramming-specific gene activation. Given the relationship observed between reprogramming and transactivation activity, we speculate that Klf4 may mediate reprogramming solely through transcriptional activation events.

The residues disrupted within Klf4 by the LI mutation likely interface with an important transcriptional coactivator during reprogramming. One candidate coactivator is Tfb1/p62, a subunit of the TFIIH complex that binds to the acidic TAD of Klf1 [17]. The binding interface consists of hydrophobic residues projecting from the extended TAD into hydrophobic pockets on the surface of Tfb1/p62 [17]. The leucine mutated in the LI construct aligns with a tryptophan residue in Klf1 that makes a specific contact with Tfb1/p62 and is critical for transactivation activity [17].

We took an open-ended approach to search for coactivators that bind to the acidic TAD of Klf4 via the leucine and/or isoleucine residues. To our surprise, we identified clathrin heavy chain as a specific interactor with the wild-type 90-110 TAD. This region contains a clathrin-box motif that is disrupted by the leucine to alanine mutation in the LI construct, explaining the

specific binding to the wild-type protein. The LI mutation completely abolished clathrin heavy chain binding as well as transactivation and reprogramming activity driven by the acidic TAD. These results suggest that clathrin heavy chain may serve as an important coactivator that allows Klf4 to activate reprogramming-specific transcriptional events. Due to its large size and the numerous interaction sites on its surface, clathrin heavy chain may serve as a nuclear scaffold that allows multiple transcriptional regulators to assemble on its surface.

It is important to note that the 2DA mutation includes the D104A mutation, which also disrupts the clathrin-box motif. This mutation had a minimal effect on reprogramming efficiency and was not tested in the transactivation assay. Despite the disruption of the clathrin-box, it is possible that the negatively-charged aspartate residue in the last position is not critical for clathrin heavy chain binding since alternative clathrin-interacting motifs have been identified containing hydrophobic residues at this site [23]. Thus, it will be important to determine the ability of this mutant to activate transcription and bind to clathrin heavy chain.

Our data demonstrate that the Klf4 acidic TAD is specific for reprogramming and cannot be replaced by TADs from Sp1 and VP16. Since an acidic TAD within p53 was shown to function by interacting with clathrin heavy chain [19-21], it will be interesting to test whether the Klf4 TAD can be replaced by this sequence.

Materials and Methods

Retrovirus Production

Retroviruses carrying mutant Klf4 constructs were produced according to the protocol of Takahashi and Yamanaka [22] with minor modifications. Klf4 variants were FLAG-tagged and cloned into pMXs using the In Fusion PCR Cloning System (Clontech). The Sp1 (NCBI Accession: NP_038700) transactivation domain fusion construct contains residues 145-494. The

VP16 (NCBI Accession: NP_044650) transactivation domain fusion construct contains residues 413-455. For each virus, a 10 cm plate of Plat-E cells at ~40% confluence was transfected with 12.5 ug of plasmid using PEI overnight. The following morning, the transfection mixture was removed and replaced with 8 ml of mES media containing 15% FBS. 24 h later, viral supernatent was collected and stored at 4°C. An additional 8 ml of media was added to the cells and collected the following day. Viral supernatents were pooled, aliquoted, frozen in liquid nitrogen, and stored at -80°C.

Reprogramming

MEFs, harvested from E14.5 embryos, were seeded onto 6-well plates in MEF media and allowed to expand to ~50% confluence. For each reprogramming experiment, media was removed and replaced with 1 ml of infection mixture overnight. This mixture contained 250 μl of each viral supernatent (Oct4, Sox2, and Klf4 variant), 250 μl of mES media containing 15% FBS, and 1 μg/ml polybrene. This mixture was replaced the following morning with mES media containing 15% FBS. After 2 days, reprogramming cultures were split 1:5 onto 22x22 mm glass coverslips (Fisher Scientific) and into separate wells to monitor factor expression by Western blotting and immunofluorescence. 5 days after initial viral infection, media was changed to mES media containing 15% KSR. Media was changed every 3 days until the experiment was stopped 12 days post-infection.

Western Blotting

For each reprogramming experiment, a single well of a 6-well plate was harvested 5 days post-infection for analysis by Western blotting to monitor factor expression. Cells pellets were disrupted by sonication in 250 μl lysis buffer containing 1% SDS in 1xPBS with 0.5 mM DTT and cOmplete protease inhibitor (Roche). Lysate was centrifuged and mixed with 4x LDS

sample buffer and 10x sample reducing agent and separated on a 4-12% Bis-Tris polyacrylamide gel (Invitrogen). Protein was transferred to a nitrocellulose membrane (Whatman) and Western blotting was performed using the LI-COR Odyssey system and reagents. Wash steps used 1xPBS + 0.1% Tween-20. The following antibodies and dilutions were used: α-FLAG (Sigma, F1804) 1:1,000; α-GAPDH 1:10,000 (Fitzgerald, 10R-G109a), IRDye 800 donkey anti-mouse IgG 1:20,000 (LI-COR).

Immunofluoresence

At 4 days post-infection, cells split onto 12 mm circle glass coverslips (Fisher Scientific) were analyzed by immunofluorescence to monitor infection efficiency, factor expression, and subcellular localization. Cells were washed in 1xPBS, fixed with 4% paraformaldehyde in 1xPBS, and permeablized with 0.5% Triton X-100 in 1xPBS. Coverslips were blocked with 0.2% fish skin gelatin, 0.2% Tween-20, and 5% goat serum in 1xPBS. Antibodies were diluted in blocking buffer and wash steps were carried out with 1xPBS + 0.2% Tween-20. Coverslips were mounted onto glass slides using Aqua-Poly/Mount (Polysciences). The following antibodies and dilutions were used: α-FLAG (Sigma, F1804) 1:200; Alexa Fluor 546 goat anti-mouse IgG 1:1,000 (Invitrogen, A-11003).

Reprogramming coverslips fixed 12 days post-infection were immunostained for the presence of Nanog using the procedure listed above. The following antibodies and dilutions were used: α-Nanog (Abcam, ab80892) 1:200; Alexa Fluor 488 goat anti-rabbit IgG 1:1,000 (Invitrogen, A-11008). Nanog+ colonies were counted using an upright fluorescence microscope (Zeiss Axio Imager). 7 non-overlapping strips representing the width of a 20x field and the length of the coverslip were counted for each coverslip. Cell clusters containing at least 5 Nanog+ cells were deemed to be iPS colonies.

Dual Luciferase Assay

Putative transactivation domains were cloned into pBXG1 to generate GAL4 fusion proteins using the In Fusion PCR Cloning System (Clontech). 2.5 x 10^4 V6.5 mES or primary E14.5 MEF cells were added in 750 µl media to a transfection mixture containing 50 µl OPTI-MEM, 7.5 µl 1 mg/ml PEI pH=7.2, 200 ng pBXG1 expression vector, 40 ng pGL4.75, and 800 ng G5E4T luc reporter vector in a 24-well plate. Plates for V6.5 mES cells had previously been coated with gelatin. Experiments were carried out in triplicate for each construct tested. Cells were harvested 36 h post-transfection by trypsinization and transferred into a 96-well plate. Luciferase readings were made according to the protocol of the Dual-Luciferase Reporter Assay System (Promega, E1910).

Production of GST Fusion Proteins

Klf4 wild-type and mutant transactivation domains were cloned into pGEX-4T-1 (GE Healthcare) using the In Fusion PCR Cloning System (Clontech). Plasmids were transformed into BL21-CodonPlus(DE3)-RIL E.coli (Stratagene) and a single colony was used to inoculate an overnight culture in LB ampicillin. The following morning, the overnight culture was diluted 1:100 and grown at 25°C to OD_{600}~0.8. IPTG was added to a final concentration of 1 mM and the culture was grown overnight at 14°C. The culture was harvested by centrifugation at 500 x g for 10 mins. The resultant pellet was resuspended in lysis buffer containing 1xPBS, 5% glycerol, 1 mM DTT, and cOmplete protease inhibitor (Roche) and disrupted with sonication pulses. After centrifugation, Triton X-100 was added to the supernatent to a final concentration of 0.1%. This lysate was bound to glutathione sepharose beads (GE Healthcare) for 1 h at 4°C via end-over-end rotation. Beads were washed 3 x 5 minutes with wash buffer containing 1xPBS, 5% glycerol, and 1 mM DTT. Purified protein was eluted in wash buffer with 10 mM reduced

glutathione adjusted to pH=8.0. Purification was monitored by Coomassie staining of fractions separated on an SDS-PAGE gel. Peak fractions were mixed 1:1 with storage buffer (1xPBS, 35% glycerol, 1 mM DTT) and aliquots were frozen in liquid nitrogen and stored at -80°C.

Affinity Purification

GST fusion proteins were dialyzed against 20 mM HEPES pH=7.9, 150 mM NaCl, 5% glycerol, 0.5 mM DTT in a 7,000 MWCO Slide-A-Lyzer Dialysis Cassette (Thermo Scientific, 66370) to remove residual glutathione. 1.4 mg of each protein was coupled to 150 µl Affi-Gel 15 resin (Bio-Rad) overnight at 4°C via end-over-end rotation. Resin was incubated with 100 mM Tris, 8 M Urea at 4°C via end-over-end rotation to remove non-covalently bound protein and block unreacted sites. Coupling was monitored by Coomassie staining of fractions separated on an SDS-PAGE gel and A_{280} spectroscopy. Each resin was incubated with 2.2 mg of E14 mES nuclear extract prepared by the method of Dignam et al. [24] at 4°C via end-over-end rotation. Wash steps were performed in 20 mM HEPES pH=7.9, 150 mM KCl, 0.2 mM EDTA, 20% glycerol, 0.5 mM DTT, cOmplete EDTA-free protease inhibitor (Roche). Bound proteins were eluted 100 mM Tris, 8 M Urea, 0.5 mM DTT at room temperature via end-over-end rotation.

Mass Spectrometry

Affinity purification eluates were digested with trypsin and Lys-C. Peptides from each sample were loaded onto a pulled silica capillary packed with strong cation exchange (Partisphere) and reversed phase (C-18) resins. Salt steps of increasing concentration of ammonium acetate driven by an HPLC pump (Agilent) were used to move subpopulations of peptides from the strong cation exchange resin to the reversed phase resin for further separation. As these peptides eluted from the reversed phase resin, they were subjected to electrospray ionization and analyzed using an LTQ-Orbitrap mass spectrometer (Thermoelectron). Tandem

mass spectra were then analyzed using SEQUEST to find the proteins that were the sources of the identified peptides.

Figure 3-1 Klf4 deletion analysis reveals several regions required for reprogramming.

A) Overview of reprogramming protocol. A Klf4 variant was delivered into MEFs along with Oct4 and Sox2 using retroviruses on day 0. Cells were transitioned to KSR mESC media on day 5. Cells were fixed on day 12 and immunostained for Nanog, which marks fully reprogrammed iPSCs. Example images of MEFs (left) and iPSCs (right) (Nanog - green, DAPI - blue). **B)** Klf4 deletion mutants reprogramming colony counts expressed as a percentage of wild-type control (1-483). The DNA binding domain (3xZF, green) and characterized TAD (AD, red) are indicated. Red lines highlight regions found to be important for reprogramming. Numbers indicate Klf4 residues. Graph contains data collected from separate reprogramming experiments. Each sample is normalized to the wild-type control performed within its respective experiment. Error bars represent standard deviation. Error bars for 1-483 are from the experiment with the largest standard deviation.

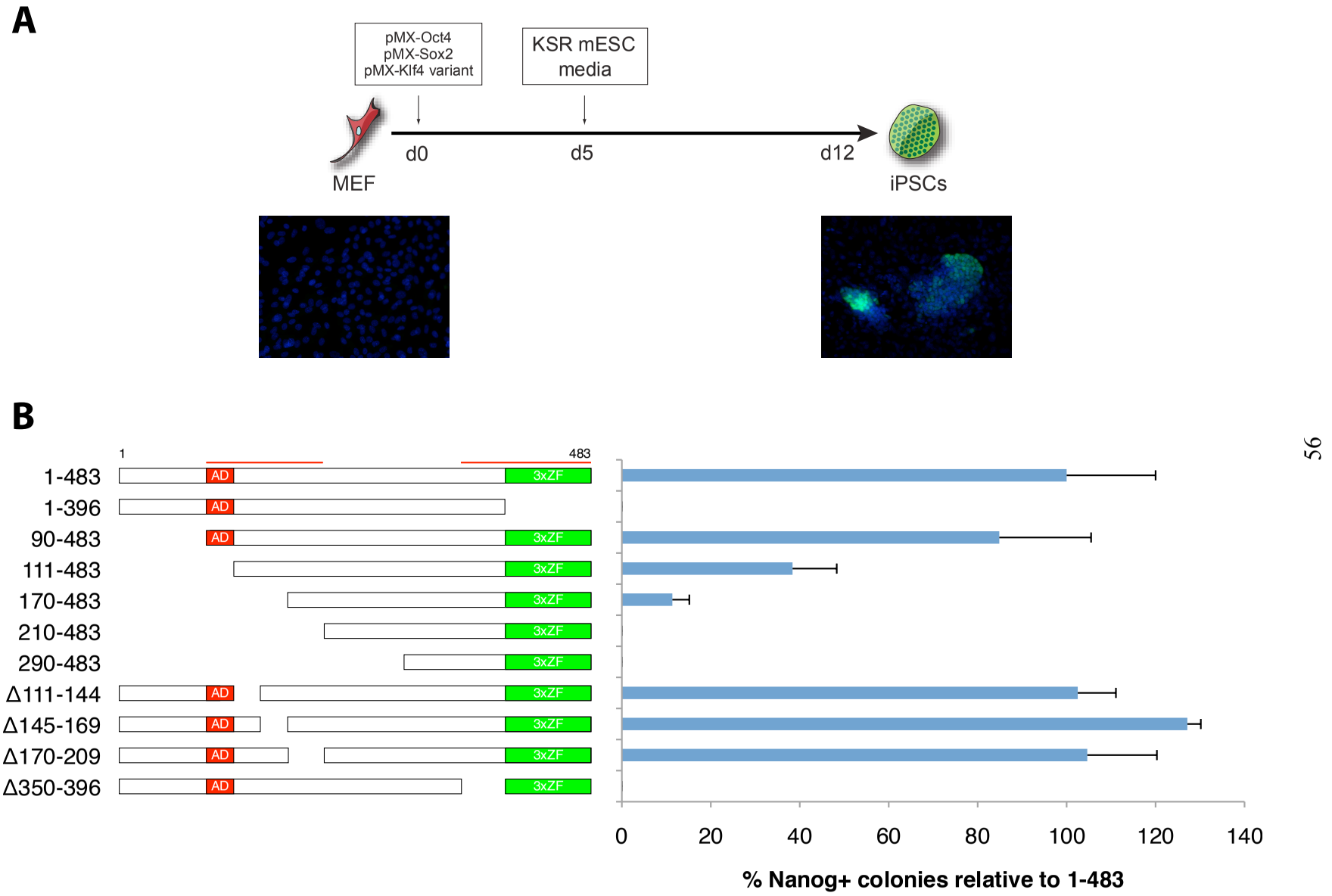
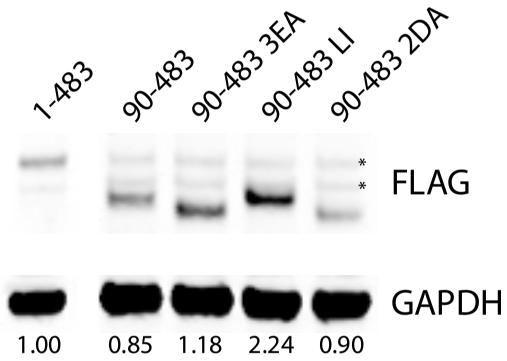

A
pMX-Oct4
pMX-Sox2
pMX-Klf4 variant
KSR mESC
media
d0
d5
d12
MEF
iPSCs
B
1
483
1-483
1-396
90-483
111-483
170-483
210-483
290-483
Δ111-144
Δ145-169
Δ170-209
Δ350-396
AD
3xZF
0
20
40
60
80
100
120
140
% Nanog+ colonies relative to 1-483

Figure 3-2 Representative analysis of Klf4 construct expression and subcellular localization patterns during reprogramming.

Data shown corresponds to reprogramming experiment from Figure 3-6a. **A)** Western blotting for FLAG, which recognizes N-terminal epitope tag on each construct, and GAPDH. Stars mark non-specific bands. Numbers indicate normalized band intensity of FLAG relative to GAPDH. Total protein lysates were isolated from day 5 of the reprogramming culture. **B)** Immunostaining for FLAG. Klf4 mutants all localize to the nucleus similar to the full-length construct (1-483). A similar level of infection efficiency was observed in each condition. Cells were fixed at reprogramming day 4.

A


1-483
90-483
90-483 3EA
90-483 LI
90-483 2DA
FLAG
GAPDH
1.00
0.85
1.18
2.24
0.90

B

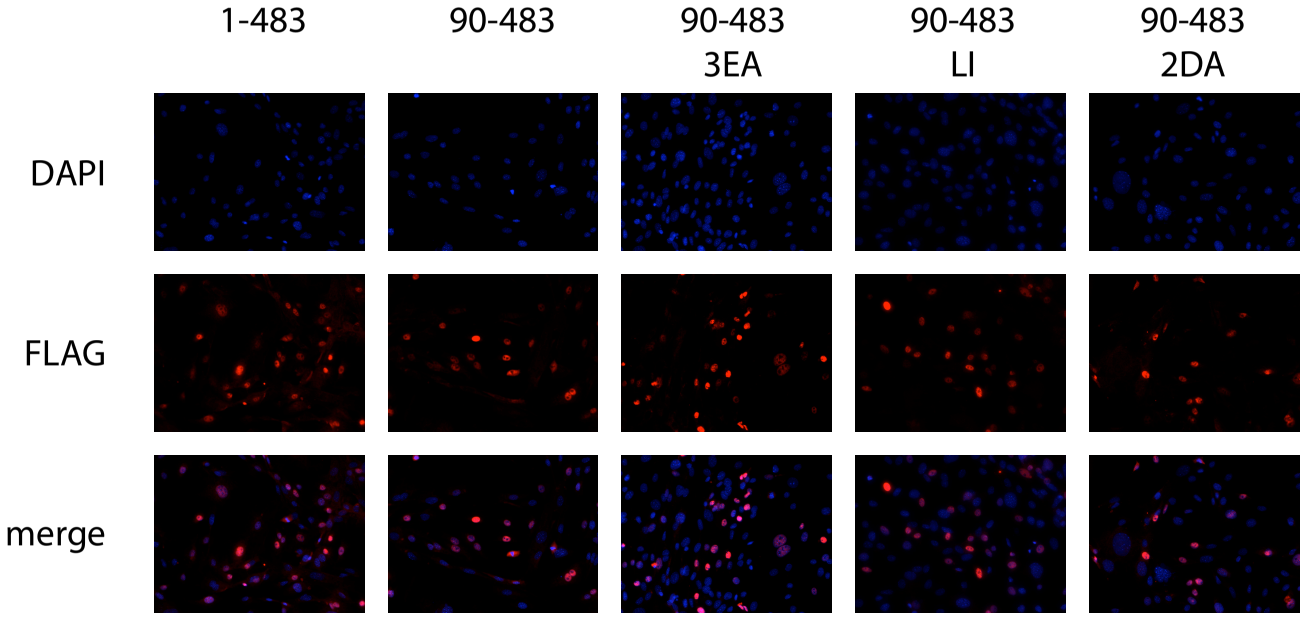
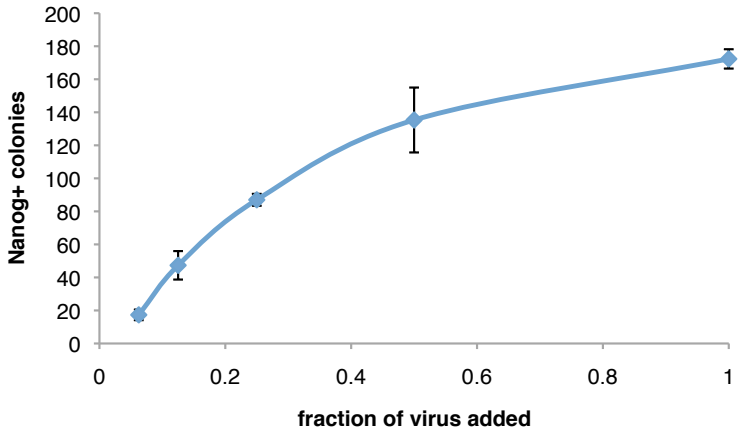

1-483
90-483
90-483 3EA
90-483 LI
90-483 2DA
DAPI
FLAG
merge

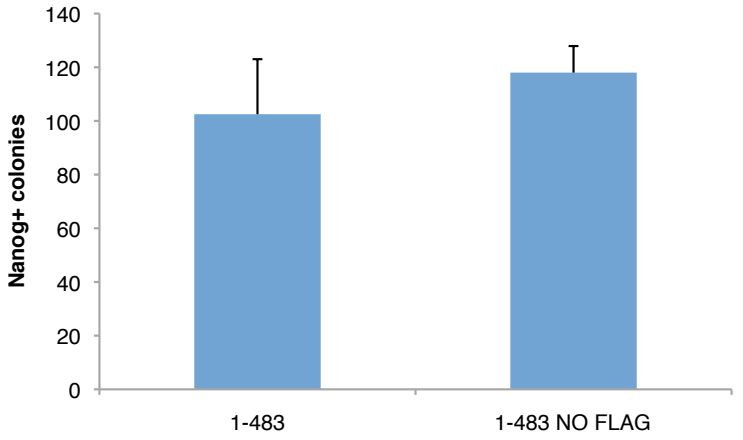
Figure 3-3 Analysis of the effect of viral titer and FLAG epitope tag on reprogramming efficiency.

A) Titration of Klf4 1-483 virus. 1 represents the quantity of Klf4 viral supernatent normally added to reprogramming experiments. Remainder of fraction was supplemented with media collected from mock transfected Plat-E cells. Error bars represent standard deviation. **B)** Addition of FLAG epitope to the Klf4 N-terminus does not alter the reprogramming efficiency of the full-length protein. Error bars represent standard deviation. **C)** Acidic FLAG residues are not responsible for high level of reprogramming activity seen in the 90-483 3EA construct or the residual level of reprogramming activity remaining in the 111-483 construct. Error bars represent standard deviation.

A


Nanog+ colonies
200
180
160
140
120
100
80
60
40
20
0
0
0.2
0.4
0.6
0.8
1
fraction of virus added

B


Nanog+ colonies
140
120
100
80
60
40
20
0
1-483
1-483 NO FLAG

C

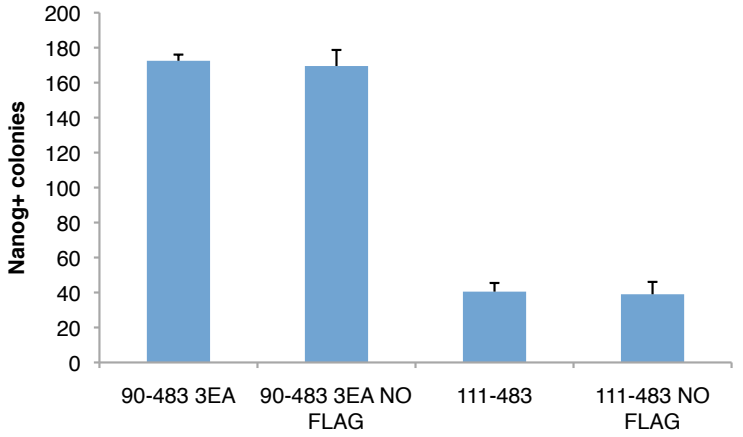

Nanog+ colonies
200
180
160
140
120
100
80
60
40
20
0
90-483 3EA
90-483 3EA NO FLAG
111-483
111-483 NO FLAG

Figure 3-4 Mutation of K275 SUMOylation site does not affect reprogramming.

K275R mutant has similar reprogramming activity to wild-type. Error bars represent standard deviation.

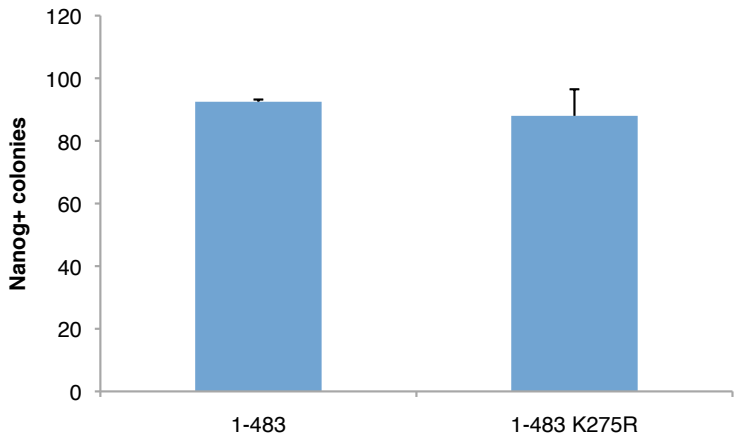
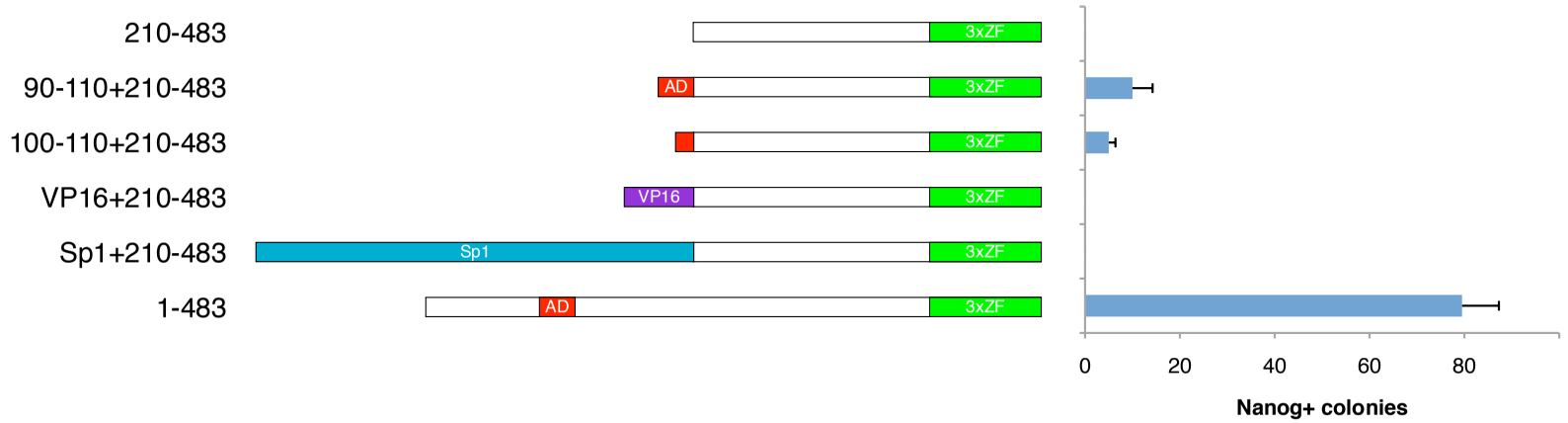

120
100
80
60
40
20
0
Nanog+ colonies
1-483
1-483 K275R

Figure 3-5 The Klf4 N-terminal TAD exhibits reprogramming-specific activity.

A) TADs from Klf4 (red), VP16 (purple), and Sp1 (blue) were fused to N-terminus of the reprogramming-deficient Klf4 210-483 construct. Only TAD sequence from Klf4 was able to partially rescue reprogramming function. **B)** Klf4 170-483 exhibits substantially reduced reprogramming activity relative to the full-length protein. Klf4 90-110 (red, AD), but not VP16 (purple), restores full reprogramming activity. Error bars represent standard deviation.

A


210-483
90-110+210-483
100-110+210-483
VP16+210-483
Sp1+210-483
1-483
3xZF
AD
VP16
Sp1
0
20
40
60
80
Nanog+ colonies

B

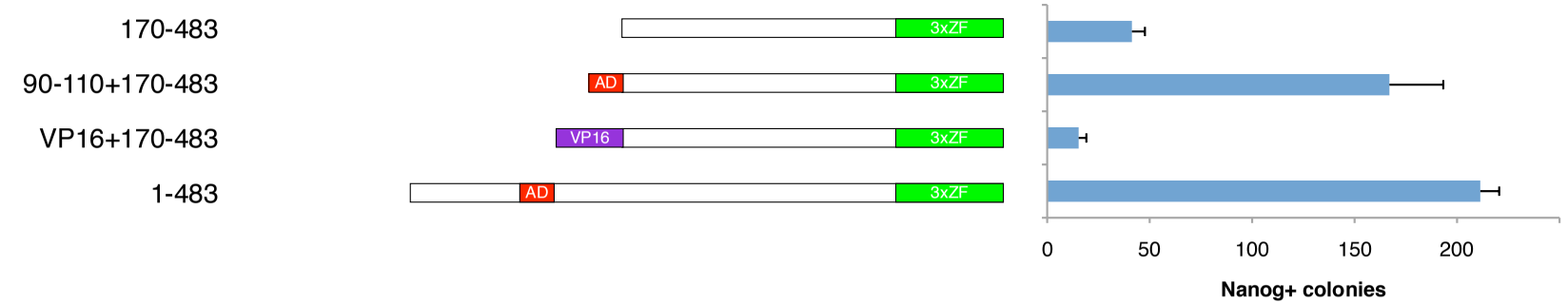
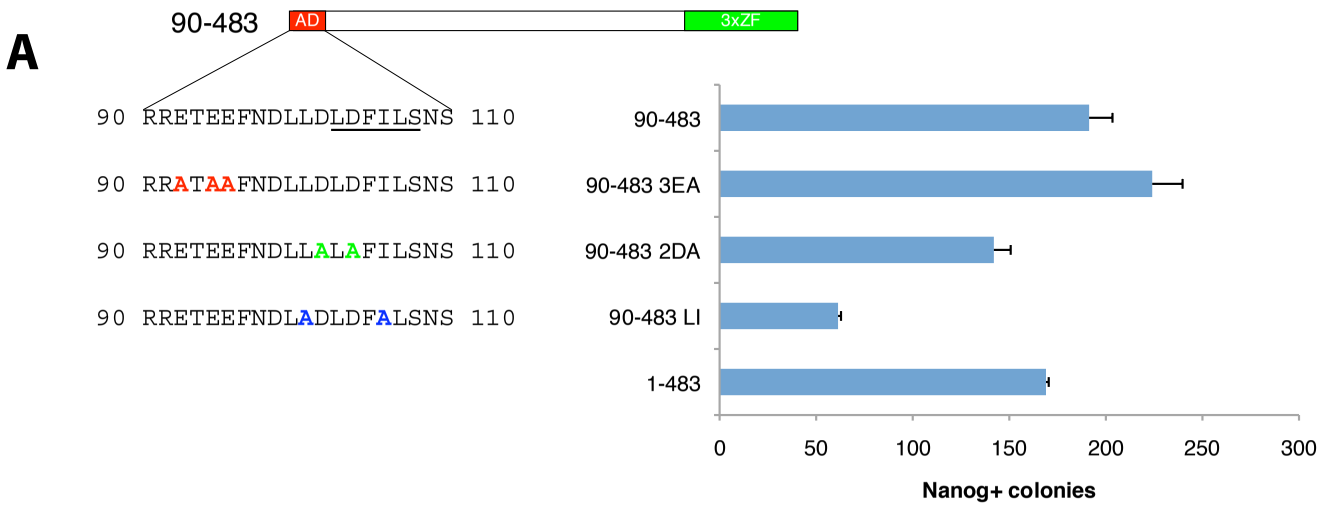
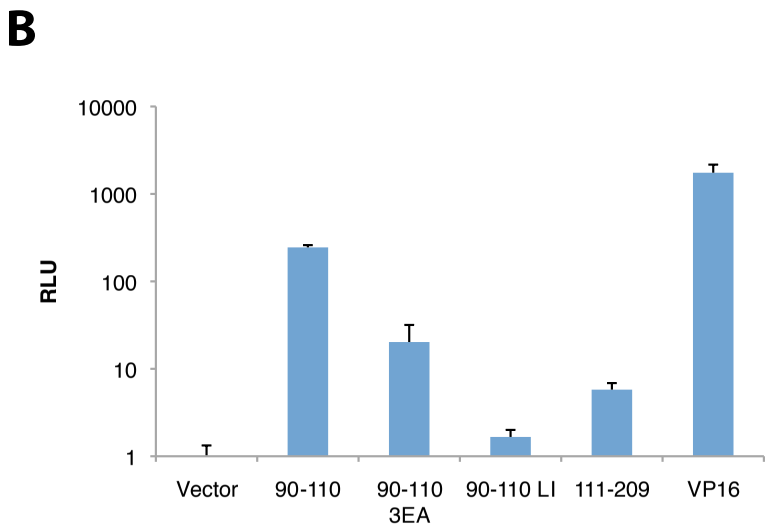

170-483
90-110+170-483
VP16+170-483
1-483
3xZF
AD
VP16
0
50
100
150
200
Nanog+ colonies

Figure 3-6 Hydrophobic residues are critical for reprogramming-specific transactivation. **A)** Reprogramming experiment analyzing the effect of point mutations within the 90-110 TAD. 3EA (red), 2DA (green), and LI (blue) mutations are highlighted. Underlined sequence indicates absolute conservation with Klf2. Error bars represent standard deviation. Dual luciferase assay performed in MEFs (**B**) and ESCs (**C**). Labels indicate sequences fused to GAL4 DNA binding domain. Error bars represent standard deviation.


A
90-483
AD
3xZF
90 RRETEEFNDLLDLDFILSNS 110
90 RRATAAFNDLLDLDFILSNS 110
90 RRETEEFNDLLALAFILSNS 110
90 RRETEEFNDLADLDFALSNS 110
90-483
90-483 3EA
90-483 2DA
90-483 LI
1-483
0
50
100
150
200
250
300
Nanog+ colonies


B
10000
1000
100
10
1
RLU
Vector
90-110
90-110 3EA
90-110 LI
111-209
VP16

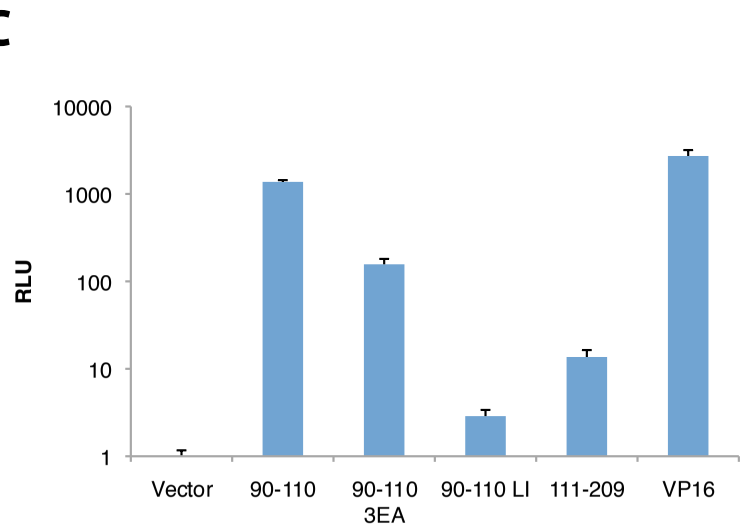
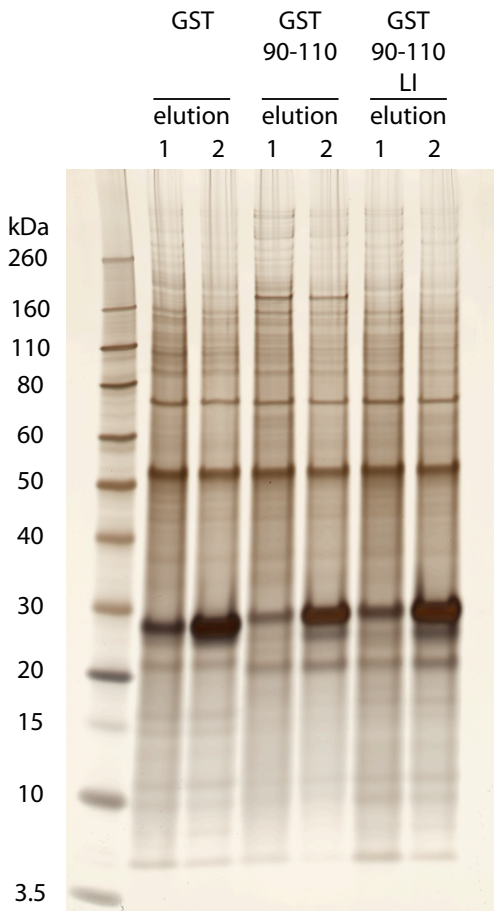

C
10000
1000
100
10
1
RLU
Vector
90-110
90-110 3EA
90-110 LI
111-209
VP16

Figure 3-7 Clathrin heavy chain binds to the 90-110 TAD through a consensus motif.

A) Silver-stained gel showing elution fractions purified from ESC extract using Klf4 90-110 variants fused to GST. Marker molecular weights are indicated. Note band between 160 and 260 kDa present only in the 90-110 wild-type purification. **B)** Identification of clathrin heavy chain by mass spectrometry. **C)** Klf4 90-110 contains a consensus clathrin binding motif (boxed). The LI mutation (blue) disrupts this sequence and leads to a loss of the interaction with clathrin heavy chain.

A



B

UniprotID	Name	MW (kDa)
Q68FD5	Clathrin heavy chain 1	191.56

	GST	GST 90-110	GST 90-110 LI
NSAF	0	356.09	0
Total Spectra	0	102	0
Unique Spectra	0	18	0

C

90 RRETEEFNDLLDLDFILSNS 110

90 RRETEEFNDLADLDFALSNS 110

References

1. **Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochiduki Y, Takizawa N, Yamanaka S: Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol* 2008, 26:101-106.**
2. **Wernig M, Meissner A, Cassady JP, Jaenisch R: c-Myc is dispensable for direct reprogramming of mouse fibroblasts. *Cell Stem Cell* 2008, 2:10-12.**
3. **Sridharan R, Tchieu J, Mason MJ, Yachechko R, Kuoy E, Horvath S, Zhou Q, Plath K: Role of the murine reprogramming factors in the induction of pluripotency. *Cell* 2008, 136:364-377.**
4. **Jiang J, Chan Y-S, Loh Y-H, Cai J, Tong G-Q, Lim C-A, Robson P, Zhong S, Ng H-H: A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol* 2008, 10:353-360.**
5. **Kaczynski J, Cook T, Urrutia R: Sp1- and Krüppel-like transcription factors. *Genome Biol* 2003, 4:206.**
6. **Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al: Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 2008, 133:1106-1117.**
7. **Shields JM, Yang VW: Identification of the DNA sequence that interacts with the gut-enriched Krüppel-like factor. *Nucleic Acids Research* 1998, 26:796-802.**
8. **Geiman DE, Ton-That H, Johnson JM, Yang VW: Transactivation and growth suppression by the gut-enriched Krüppel-like factor (Krüppel-like factor 4) are dependent on acidic amino acid residues and protein-protein interaction. *Nucleic Acids Research* 2000, 28:1106-1113.**
9. **Shields JM, Yang VW: Two potent nuclear localization signals in the gut-enriched Krüppel-like factor define a subfamily of closely related Krüppel proteins. *J Biol Chem* 1997, 272:18504-18507.**
10. **Courey AJ, Tjian R: Analysis of Sp1 in vivo reveals multiple transcriptional domains, including a novel glutamine-rich activation motif. *Cell* 1988, 55:887-898.**
11. **Du JX, McConnell BB, Yang VW: A small ubiquitin-related modifier-interacting motif functions as the transcriptional activation domain of Krüppel-like factor 4. *J Biol Chem* 2010, 285:28298-28308.**
12. **Kim MO, Kim S-H, Cho Y-Y, Nadas J, Jeong C-H, Yao K, Kim DJ, Yu D-H, Keum Y-S, Lee K-Y, et al: ERK1 and ERK2 regulate embryonic stem cell self-renewal through phosphorylation of Klf4. *Nature Structural & Molecular Biology* 2012:1-9.**
13. **Hirai H, Tani T, Kikyo N: Structure and functions of powerful transactivators: VP16, MyoD and FoxA. *Int J Dev Biol* 2010, 54:1589-1596.**
14. **Wang Y, Chen J, Hu J-L, Wei X-X, Qin D, Gao J, Zhang L, Jiang J, Li J-S, Liu J, et al: Reprogramming of mouse and human somatic cells by high-performance engineered factors. *EMBO Rep* 2011, 12:373-378.**
15. **Hirai H, Tani T, Katoku-Kikyo N, Kellner S, Karian P, Firpo M, Kikyo N: Radical Acceleration of Nuclear Reprogramming by Chromatin Remodeling with the Transactivation Domain of MyoD. *Stem Cells* 2011, 29:1349-1361.**

16. **Yet SF, McA'Nulty MM, Folta SC, Yen HW, Yoshizumi M, Hsieh CM, Layne MD, Chin MT, Wang H, Perrella MA, et al: Human EZF, a Kruppel-like zinc finger protein, is expressed in vascular endothelial cells and contains transcriptional activation and repression domains.** ***J Biol Chem*** **1998, 273:1026-1031.**
17. **Mas C, Lussier-Price M, Soni S, Morse T, Arseneault G, Di Lello P, Lafrance-Vanasse J, Bieker JJ, Omichinski JG: Structural and functional characterization of an atypical activation domain in erythroid Kruppel-like factor (EKLF).** ***Proc Natl Acad Sci USA*** **2011, 108:10484-10489.**
18. **Lemmon SK, Traub LM: Getting in Touch with the Clathrin Terminal Domain.** ***Traffic (Copenhagen, Denmark)*** **2012.**
19. **Enari M, Ohmori K, Kitabayashi I, Taya Y: Requirement of clathrin heavy chain for p53-mediated transcription.** ***Genes & Development*** **2006, 20:1087-1099.**
20. **Ohmori K, Endo Y, Yoshida Y, Ohata H, Taya Y, Enari M: Monomeric but not trimeric clathrin heavy chain regulates p53-mediated transcription.** ***Oncogene*** **2008, 27:2215-2227.**
21. **Ohata H, Ota N, Shirouzu M, Yokoyama S, Yokota J, Taya Y, Enari M: Identification of a function-specific mutation of clathrin heavy chain (CHC) required for p53 transactivation.** ***J Mol Biol*** **2009, 394:460-471.**
22. **Takahashi K, Yamanaka S: Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** ***Cell*** **2006, 126:663-676.**
23. **Dell'Angelica EC: Clathrin-binding proteins: got a motif? Join the network!** ***Trends Cell Biol*** **2001, 11:315-318.**
24. **Dignam JD, Lebovitz RM, Roeder RG: Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei.** ***Nucleic Acids Research*** **1983, 11:1475-1489.**

CHAPTER 4

FUNCTIONAL DIVERGENCE WITHIN THE KLF/SP FAMILY DETERMINES DNA BINDING DOMAIN REPROGRAMMING ACTIVITY

Introduction

Krüppel-like factor/Sp (Klf/Sp) proteins are transcription factors that play diverse roles at multiple stages of vertebrate development [1]. The protein family is characterized by their highly conserved DNA binding domains, which bear homology to the *Drosophila* protein, Krüppel [1]. These regions consist of 3 tandem C2H2 zinc finger motifs that make sequence-specific contacts with GC-rich binding sites [1].

A member of the Klf/Sp family, Klf4, was found to reprogram somatic cells, such as mouse embryonic fibroblasts (MEFs), to a pluripotent state when expressed along with Oct4 and Sox2 [2-5]. These induced pluripotent stem cells (iPSCs), which can be generated from an individual patient, harbor immense therapeutic potential as part of cell replacement strategies. The role of Klf4 in the reprogramming process remains under intense investigation. Some clues as to its specific functional requirements are provided by the observation that it can be replaced by several close family members [2]. This result suggests that these proteins have functional overlap in both their transactivation and DNA binding elements. However, the extent to which other zinc finger DNA binding domains can function in reprogramming is unclear.

In this study, we assayed zinc finger domains from Glis and Klf/Sp proteins for their ability to mediate reprogramming by replacing the DNA binding domain within Klf4. We identified a set of reprogramming-competent DNA binding domains within an evolutionarily distinct branch of the Klf/Sp family. The reprogramming activity within this subfamily can be attributed to functional divergence within the second and third zinc fingers. These evolutionary differences lead to multiple effects, including altered DNA binding specificity. A lysine residue within the third zinc finger of reprogramming-incompetent Klf/Sp DNA binding domains creates an additional two base pair sequence preference likely through simultaneous hydrogen bonding

with the stacked bases. The presence of this residue may lead to a more restricted DNA binding pattern, which prevents the recognition of crucial reprogramming target genes.

Results

DNA Binding Domains from a Distinct Klf Subfamily are Functionally Redundant in Reprogramming

The reprogramming activity of Klf4 is absolutely dependent upon its C-terminal DNA binding domain (data not shown). This region (3xZF, Figure 4-1a) likely functions in reprogramming by directing transactivation elements within the N-terminal portion of the protein (residues 1-396, Figure 4-1a) to the regulatory sequences of important target genes. Additionally, it is possible that the DNA binding domain itself may recruit important cofactors that help to regulate reprogramming-specific transcription. In order to gain insight into the function of this region in reprogramming, we sought to assess the ability of each DNA binding domain in the Klf/Sp family to mediate reprogramming within the context of a protein chimera (Figure 4-1a). This approach isolated the effects due to the presence of the individual DNA binding domains by fusing each of them to the N-terminal region of Klf4 (Figure 4-1a). These chimeric proteins were expressed in MEFs, along with Oct4 and Sox2, using retroviruses and reprogramming activity was quantified by counting Nanog+ iPSC colonies after 12 days (Figure 4-1b). In addition to the Klf/Sp family, we tested portions of the zinc finger DNA binding domains from less related factors, Glis1 and Glis2. Glis1 was identified in a screen to replace Klf4 in reprogramming [6], suggesting that these proteins may bind to a critical set of overlapping target genes.

Analysis of the protein chimera reprogramming experiments revealed that only DNA binding domains from a specific Klf subfamily containing Klf4 were able to function in

reprogramming (Figure 4-1c). The alignment tree illustrates the evolutionary relationship between the zinc finger domains based on primary sequence (Figure 4-1c). Within the reprogramming-competent Klf subfamily, we observed significant variation in reprogramming efficiency (Figure 4-1c). Most notably, the Klf6ZF construct exhibited very slight reprogramming activity. Two protein chimeras (Klf14ZF and Klf16ZF) did not exhibit stable expression in infected MEFs, and therefore, were not able to be analyzed (Figure 4-1c, data not shown). The expression levels, infection efficiencies, and subcellular localization of all other chimeras were found to be similar by Western blotting and immunofluorescence (data not shown).

Members of the Klf/Sp family whose DNA binding domains did not possess reprogramming activity contain a C-terminal extension following their zinc fingers (Klf10, Figure 4-1a). This loss of this region from Sp1 had no effect on its ability to bind DNA, but reduced its transactivation activity [7, 8]. To assess the effect of this region on reprogramming, we extended the Klf10ZF chimera to include its C-terminal sequence (Klf10ZF-C, Figure 4-2). Addition of the C-terminal extension to the Klf10ZF chimera did not lead to iPSC generation (Figure 4-2).

Reprogramming Activity Differences are Attributable to Zinc Fingers 2 and 3

To identify functionally divergent regions within the reprogramming-competent and -incompetent DNA binding domains that explain their respective reprogramming activities, we selected Klf4 and Klf10 as model proteins from each group and made finer-scale chimeras. The DNA binding domain can be separated into 5 parts - 3 zinc fingers (ZFs) and 2 linker sequences (Figure 4-3). Protein chimeras were generated containing either Klf4 (red) or Klf10 (yellow) sequence in each of these positions (Figure 4-3). Only chimeras containing both the second and

third ZFs from Klf4 were able to efficiently generate iPSCs (Figure 4-3). Each of these regions was separately found to require Klf4 sequence for reprogramming activity, since replacement of either ZF2 or ZF3 with Klf10 sequence in an otherwise Klf4 background leads to loss of function (Figure 4-3). Very slight reprogramming activity remained in some of the constructs that had Klf10 sequence for either ZF2 or ZF3, indicating that loss of function was not always absolute as was seen when both of these regions derive from Klf10 (Figure 4-3).

The functional divergence within the second and third ZFs mirrors the distinct evolutionary separation between the reprogramming-competent and -incompetent subfamilies observed by sequence alignment of these regions (Figure 4-4b,c). In contrast, ZF1 does not observe a similar evolutionary division (Figure 4-4a). Overall, ZF1 appears to be under less selective pressure, especially in its N-terminal half (Figure 4-4a).

The importance of ZF2 and ZF3 for Klf4 DNA binding is demonstrated by a crystal structure of its zinc finger domain in complex with DNA [9]. Binding energy obtained from interactions with the DNA bases comes primarily through arginine:guanine contacts from ZF2 and ZF3 [9]. However, these arginines are conserved throughout the entire Klf/Sp family. Thus, we further examined the ZF2 and ZF3 regions to elucidate the basis of their functional divergence in reprogramming.

Residues within the ZF2 β-sheet Determine Reprogramming Activity

Each zinc finger consists of a zinc atom sandwiched between a small, antiparallel β-sheet and an α-helix [10]. The zinc atom is coordinated by side chains projecting from both secondary structural motifs [10]. When in contact with DNA, residues at the -1, +3, and +6 positions of the α-helix extend into the major groove and have the potential to make specific contacts with the DNA bases [10].

The ZF2 regions of Klf4 and Klf10 are divided by a central stretch of conserved residues (orange, Figure 4-5a). We made protein chimeras by varying the sequence flanking either side of this block of conservation to examine the functional divergence between the α-helical and β-sheet regions within the zinc finger (Figure 4-5a). Reprogramming efficiency was reduced as a result of inserting Klf10 sequence into either of these positions (Figure 4-5a). However, the presence of Klf10 sequence in the β-sheet region of ZF2 had a much more dramatic effect, leading to an almost complete loss of reprogramming activity (Figure 4-5a). The partial loss of function due to the Klf10 α-helix can be explained by the mutation of a lysine in Klf4 that makes a contact with the DNA backbone (Figure 4-5a) [9]. Point mutation of this residue alone to its corresponding residue in Klf10, which would likely result in reduced DNA binding affinity, leads to a similar reduction in reprogramming activity to the ZF2 - 4/10 chimera (Figure 4-5a). The DNA binding domain chimera containing ZF2 from Klf10 exhibited a similar *in vitro* DNA binding preference to Klf4 (Figure 4-5b), confirming that base-specific contacts are unaltered by this mutation. In summary, the functional divergence between ZF2 mainly arises from the β-sheet region and does not involve a change in DNA binding specificity.

Eight residues differ between the ZF2 regions of Klf4 and Klf10 (Figure 4-5c). We analyzed these residues by mutating each of them individually to their Klf10 counterparts and measuring reprogramming activity (Figure 4-5c). Of the five residues that lie within the β-sheet, only mutation of D435 led to reduced iPSC formation (Figure 4-5c,d). Surprisingly, the W439R mutation resulted in a large gain of reprogramming function (Figure 4-5c). None of these mutations by themselves explains the dramatic reduction in reprogramming that occurred when the entire β-sheet was mutated to Klf10. Thus, the nature of this difference is likely due to the combined action of several of these residues.

ZF3 Functional Divergence is Primarily Due to Changes in DNA Binding Specificity

Upon sequence analysis of the third ZF within the Klf/Sp family, we noticed a leucine to lysine difference that correlates with reprogramming activity in our protein chimeras (Figure 4-6a, Figure 4-4c). This residue lies in the +6 position of the recognition helix and may contribute to DNA binding preference by interfacing with the DNA bases (Figure 4-6a). Mutation of L477 within Klf4 to lysine substantially reduces, but does not completely eliminate, the appearance of Nanog+ colonies (Figure 4-6a). Conversely, mutation of the equivalent lysine to leucine in a construct containing ZF3 from Klf10 produces iPSC colonies at levels slightly lower than wild-type Klf4 (Figure 4-6a). These functional changes in reprogramming activity corresponded nicely with changes in DNA binding specificity measured *in vitro* (Figure 4-6b). However, given that loss and gain of reprogramming function due to these point mutations was not complete, other residues within ZF3 may play a minor role in the functional divergence observed in the Klf/Sp family.

Lysine at the ZF3 +6 Position Alters DNA Binding Preference at Two Positions through Specific Hydrogen Bonding

DNA binding by the Klf/Sp family is anchored by the R-E-R recognition residues in ZF2, which prefer GCG [9, 11]. T is also somewhat tolerated in place of C in the central position (Figure 4-6b) [12]. Arginine and histidine residues in the -1 and +3 positions, respectively, of ZF3 prefer GG, thereby extending this core sequence to 5'-GGG[C/T]G-3' (Figure 4-6b). The residue in the ZF3 +6 position would be predicted to be oriented towards the proximal base upstream of this core motif [10]. Thus, it was surprising to find that the presence of lysine in the +6 position created a strong binding preference for G at two consecutive bases (Figure 4-6b).

In order to further interrogate the effect of the leucine to lysine switch at the ZF3 +6 position on DNA binding specificity, we generated scatterplots from protein binding microarray data displaying the enrichment scores (E-scores) for all 8-mers (Figure 4-6c). In agreement with what was observed in the logo plots (Figure 4-6b), Klf4 or Klf10 DNA binding domains with lysine in the ZF3 +6 position strongly preferred GG in the positions immediately upstream of the core motif (yellow, Figure 4-6c). 8-mers matching the core motif, but containing other bases besides G in these positions, were disfavored by proteins with the ZF3 +6 lysine residue (red, Figure 4-6c). Sequences with G at one of the two positions displayed an intermediate preference (blue, Figure 4-6c). These results demonstrate that changing leucine to lysine at the ZF3 +6 positon creates a strong DNA binding preference for guanine at two base pairs.

Assessment of the crystal structure of Klf4 bound to DNA indicates that L477 does not contact the DNA bases (Figure 4-7a) [9], thereby explaining its inability to generate a DNA binding preference. In an attempt to understand the molecular basis for the altered binding preference detected in the Klf4 L477K mutant, we created a structural model based on the wild-type Klf4 structure (Figure 4-7b) [9]. This model reveals that the amino group on the lysine side chain can be positioned to be simultaneously within hydrogen bonding distance (3.2 and 2.6 Å) of the N7 atoms of the two guanine bases adjacent to the core motif (Figure 4-7b). Thus, the DNA binding preference established by the ZF3 +6 lysine can be explained by these contacts.

Reprogramming-competent Klf DNA Binding Domains Bind a Wider Range of Sequences In vitro

We wondered whether the alteration in DNA binding specificity due to the presence of the ZF3 +6 lysine could be generalized across the entire Klf/Sp family and if this difference may help to explain the reprogramming capacity of each of their DNA binding domains. We focused

our analysis on 8-mers that are recognized to an extent *in vitro* by recombinant zinc finger domains where they are likely to be bound in an *in vivo* setting by the cognate DNA binding protein (E-score > 0.40) [13]. The datasets for the zinc finger domains cluster according to similar DNA binding preferences (Figure 4-8). The Klf/Sp family is divided into two groups, correlating perfectly with the reprogramming activity of the individual DNA binding domains (Figure 4-8). Glis1 and Glis2 occupy a separate group that has largely non-overlapping DNA binding specificity (Figure 4-8).

Individual 8-mers were partitioned into one of ten k-means clusters, and a composite motif was generated from each cluster (Figure 4-8). Clusters with similar composite motifs are grouped together and a common composite motif is displayed for the group as a whole (Figure 4-8). The Klf/Sp family is separated into three groups (red, orange, yellow), distinguished by their preference for guanine at the positions adjacent to the core motif contacted by ZF3 +6 lysine (Figure 4-8). The reprogramming-competent Klf/Sp family members are capable of binding all three of these groups, while the reprogramming-incompetent family members are restricted to interacting with 8-mers within the orange and yellow groups (Figure 4-8). However, the zinc finger domains with lysine in the ZF3 +6 position score more highly than their leucine-containing counterparts, likely due to increased binding energy derived from hydrogen bonding with the guanine bases (yellow, Figure 4-8). These data suggest that reprogramming-incompetent zinc finger domains within the Klf/Sp family are restricted in their DNA binding specificity due to the additional interaction with the DNA bases by their lysine side chain.

Two groups (tan and black) contain 8-mers strongly bound by the more distantly related Glis factors (Figure 4-8). The binding specificity we observed using the final three zinc fingers of Glis1 was similar to what was previously found using the entire 5 zinc finger domain from

Glis2 (Figure 4-8) [14]. This result is consistent with the finding that the majority of base-specific contacts occur through ZFs four and five [15]. The composite motifs derived from the tan and black groups correspond to overlapping portions of the Glis3 consensus binding sequence previously identified by *in vitro* selection - 5'-[G/C]TGGGGGGT[A/C]-3' (Figure 4-8) [16]. GGGGGGT is recognized with low affinity by the reprogramming-incompetent Klf/Sp subfamily, likely due to its similarity to the 5' portion of its binding site (tan, Figure 4-8). In contrast, GGGGTC elicits binding specifically by the Glis factors (black, Figure 4-8).

Discussion

In this study, we used reprogramming as an assay to assess the functional divergence within the zinc finger domains the Klf/Sp family. These results provide insight into the function of these DNA binding domains in general as well as into the nature of their reprogramming activity. We observed that only a subset of Klf/Sp DNA binding domains are suitable to mediate reprogramming. Our findings corroborate and expand on previous work showing that full-length Klf1, Klf2, and Klf5 can replace Klf4 in reprogramming [2]. It is difficult to properly examine the reprogramming activity of all of the proteins within the Klf family, since many of them express poorly in MEFs (data not shown). Thus, it remains an open question whether other full-length Klf factors with reprogramming-competent zinc finger domains could replace Klf4 in reprogramming if properly expressed.

We determined that the difference in reprogramming activity between the Klf4 and Klf10 DNA binding domains is due to differences within the second and third zinc fingers. Surprisingly, the difference within ZF2 was contained primarily within its β-strands, which face away from the double helix. The functional divergence within the ZF2 β-sheet region is unlikely to be the result of an altered interaction with an important cofactor, since no single point

mutation of the divergent residues in Klf4 led to a dramatic loss of function. This effect is also unlikely to be due to changes in DNA binding preference, since *in vitro* binding specificity was unaltered. However, we cannot rule out a reduction in binding affinity as an explanation for this phenomenon. In contrast, Klf10 sequence in the ZF2 recognition helix likely reduces reprogramming activity by disrupting an electrostatic interaction with the DNA backbone [9]. We predict that a point mutation restoring the DNA contact within this construct would fully rescue reprogramming activity.

ZF3 contains an amino acid difference at the +6 position within its recognition helix that splits the Klf/Sp family DNA binding domains along the lines of their reprogramming activity. The identity of this residue dictates DNA binding preference at the 5' end of their binding site. Furthermore, a shift in DNA binding specificity is sufficient to disrupt or activate reprogramming function.

DNA bases in the Klf/Sp binding site are contacted by relatively few amino acid side chains as compared to binding sites recognized by other zinc finger DNA binding domains [9, 10, 15]. However, strong planar contacts between arginine side chains and the electronegative surface of guanines, which face the major groove, generate substantial binding energy [9]. The histidine in the ZF3 +3 position can contact the hydrogen bond acceptor N7 atom of either purine base [9]. However, a stronger interaction is likely to occur with guanine due to its increased electronegativity. We found that guanine in this position was associated with higher 8-mer enrichment scores (data not shown). Interestingly, we show that lysine in the ZF3 +6 position creates a sequence preference at the two adjacent 5' bases within the binding site. This is somewhat surprising since residues in the recognition positions of zinc finger proteins generally contact only one base at a time. Our structural model demonstrates that a lysine side chain at the

ZF3 +6 position can simultaneously hydrogen bond with stacked guanine bases. The base preference and potential hydrogen bond with the position outside of the traditional zinc finger footprint has not been previously appreciated. The preference created at these positions within the binding site is not limited exclusively to guanine. Adenine and thymine were also observed to a lesser extent in these positions, consistent with their presentation of hydrogen bond acceptor atoms on the major groove surface. As seen at the position contacted by the ZF3 +3 histidine, guanine is also preferred most likely as a result of its increased electronegativity. Alternatively, the guanine-rich binding site may be required to form G-quadruplex structures that can be bound by the zinc finger domain [17].

The DNA binding motif generated *in vitro* using protein binding microarrays for Klf4 largely mirrors its *in vivo* DNA binding preference [18]. However, binding affinity and specificity may be altered by cofactor interactions and/or posttranslational modifications.

We observed a striking correlation between the DNA binding preferences and reprogramming activities of each zinc finger domain within the Klf/Sp family. Reprogramming-competent DNA binding domains recognize a much broader array of sequences due to the lack of specificity dictated by the ZF3 +6 position. Thus, we propose that the induction of pluripotency by Klf4 requires binding to gene regulatory elements through sites that lack guanine in 5' bases adjacent to the Klf/Sp core motif. At sites containing guanine in these positions, we suggest that Klf/Sp zinc fingers with lysine in the ZF3 +6 position may be able to bind with increased affinity relative to their leucine-containing counterparts due to the additional hydrogen bonding interactions. However, it was surprising to note the correlation between the presence of guanine at the 5' positions and higher enrichment scores even in the Klf/Sp subfamily whose ZF3

+6 residue was leucine. Thus, the presence of guanine may indirectly lead to enhanced binding through some yet undefined means.

Klf4 and Glis1 have been reported to be functionally interchangeable in reprogramming [6]. Although these factors may carry out this process by recognizing an overlapping set of important target genes, it is unlikely that they would do so through a common DNA binding site given the differences in their in vitro DNA binding preferences. Also, elements outside of the Glis1 DNA binding domain are likely to be specifically required for its reprogramming activity, since these regions could not be functionally replaced by the N-terminal portion of Klf4.

Materials and Methods

Retrovirus Production

Retroviruses carrying protein chimera constructs were produced according to the protocol of Takahashi and Yamanaka [3] with minor modifications. Each construct was FLAG-tagged and cloned into pMXs using the In Fusion PCR Cloning System (Clontech). For each virus, a 10 cm plate of Plat-E cells at ~40% confluence was transfected with 12.5 ug of plasmid using PEI overnight. The following morning, the transfection mixture was removed and replaced with 8 ml of mES media containing 15% FBS. 24 h later, viral supernatent was collected and stored at 4°C. An additional 8 ml of media was added to the cells and collected the following day. Viral supernatents were pooled, aliquoted, frozen in liquid nitrogen, and stored at -80°C.

Reprogramming

MEFs, harvested from E14.5 embryos, were seeded onto 6-well plates in MEF media and allowed to expand to ~50% confluence. For each reprogramming experiment, media was removed and replaced with 1 ml of infection mixture overnight. This mixture contained 250 μl of each viral supernatent (Oct4, Sox2, and Klf4 variant), 250 μl of mES media containing 15%

FBS, and 1 μg/ml polybrene. This mixture was replaced the following morning with mES media containing 15% FBS. After 2 days, reprogramming cultures were split 1:5 onto 22x22 mm glass coverslips (Fisher Scientific) and into separate wells to monitor factor expression by Western blotting and immunofluorescence. 5 days after initial viral infection, media was changed to mES media containing 15% KSR. Media was changed every 3 days until the experiment was stopped 12 days post-infection.

Western Blotting

For each reprogramming experiment, a single well of a 6-well plate was harvested 5 days post-infection for analysis by Western blotting to monitor factor expression. Cells pellets were disrupted by sonication in 250 μl lysis buffer containing 1% SDS in 1xPBS with 0.5 mM DTT and cOmplete protease inhibitor (Roche). Lysate was centrifuged and mixed with 4x LDS sample buffer and 10x sample reducing agent and separated on a 4-12% Bis-Tris polyacrylamide gel (Invitrogen). Protein was transferred to a nitrocellulose membrane (Whatman) and Western blotting was performed using the LI-COR Odyssey system and reagents. Wash steps used 1xPBS + 0.1% Tween-20. The following antibodies and dilutions were used: α-FLAG (Sigma, F1804) 1:1,000; α-GAPDH 1:10,000 (Fitzgerald, 10R-G109a), IRDye 800 donkey anti-mouse IgG 1:20,000 (LI-COR).

Immunofluoresence

At 4 days post-infection, cells split onto 12 mm circle glass coverslips (Fisher Scientific) were analyzed by immunofluorescence to monitor infection efficiency, factor expression, and subcellular localization. Cells were washed in 1xPBS, fixed with 4% paraformaldehyde in 1xPBS, and permeablized with 0.5% Triton X-100 in 1xPBS. Coverslips were blocked with 0.2% fish skin gelatin, 0.2% Tween-20, and 5% goat serum in 1xPBS. Antibodies were diluted

in blocking buffer and wash steps were carried out with 1xPBS + 0.2% Tween-20. Coverslips were mounted onto glass slides using Aqua-Poly/Mount (Polysciences). The following antibodies and dilutions were used: α-FLAG (Sigma, F1804) 1:200; Alexa Fluor 546 goat anti-mouse IgG 1:1,000 (Invitrogen, A-11003).

Reprogramming coverslips fixed 12 days post-infection were immunostained for the presence of Nanog using the procedure listed above. The following antibodies and dilutions were used: α-Nanog (Abcam, ab80892) 1:200; Alexa Fluor 488 goat anti-rabbit IgG 1:1,000 (Invitrogen, A-11008). Nanog+ colonies were counted using an upright fluorescence microscope (Zeiss Axio Imager). 7 non-overlapping strips representing the width of a 20x field and the length of the coverslip were counted for each coverslip. Cell clusters containing at least 5 Nanog+ cells were deemed to be iPS colonies.

Production of GST Fusion Proteins

Sequences encoding DNA binding domains were cloned into pGEX-4T-1 (GE Healthcare) using the In Fusion PCR Cloning System (Clontech). Plasmids were transformed into BL21-CodonPlus(DE3)-RIL E.coli (Stratagene) and a single colony was used to inoculate an overnight culture in LB ampicillin. Zinc acetate was added at a final concentration of 50 μM to all growth media and purification buffers for subsequent steps. The following morning, the overnight culture was diluted 1:100 and grown at 25°C to OD_{600}~0.8. IPTG was added to a final concentration of 1 mM and the culture was grown overnight at 14°C. The culture was harvested by centrifugation at 500 x g for 10 mins. The resultant pellet was resuspended in lysis buffer containing 1xPBS, 5% glycerol, 1 mM DTT, and cOmplete protease inhibitor (Roche) and disrupted with sonication pulses. After centrifugation, Triton X-100 was added to the supernatent to a final concentration of 0.1%. This lysate was bound to glutathione sepharose

beads (GE Healthcare) for 1 h at 4°C via end-over-end rotation. Beads were washed 3 x 5 minutes with wash buffer containing 1xPBS, 5% glycerol, and 1 mM DTT. Purified protein was eluted in wash buffer with 10 mM reduced glutathione adjusted to pH=8.0. Purification was monitored by Coomassie staining of fractions separated on an SDS-PAGE gel. Peak fractions were mixed 1:1 with storage buffer (1xPBS, 35% glycerol, 1 mM DTT) and aliquots were frozen in liquid nitrogen and stored at -80°C.

Protein Binding Microarray

Protein binding microarray experiments were performed following the protocol of Berger and Bulyk [19]. GST-tagged proteins were used at a concentration of 200 nM.

Sequence Analysis and Motif Construction

Data analysis and plot generation were performed using R. To generate consensus motifs, 8-mers were aligned using CLUSTALW [20] without gaps. Information from each position was extracted to form an alignment matrix. LOGO plots were generated using enoLOGOS [21].

Structure Modeling

DNA from 2WBU.pdb [9] minus the terminal base pairs was extended on either end *in silico* to represent the sequence of the Klf4 binding site within the *Nanog* enhancer sequence [22]. Mean structural parameters for individual base pair steps from the protein-DNA crystal structure library of Olson et al. (mean twist = 34.2˚) were added to the structural parameters file of 2WBU, which was converted to a DNA PDB file using the 3DNA rebuild software [23]. The Klf4 3 Zn finger peptide from 2WBU was docked onto the extended DNA by superimposition with 2WBU, Leu477 was substituted with lysine, and the rotamers of Lys477 and Arg481

adjusted in The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC. and Coot [24].

References

1. **Kaczynski J, Cook T, Urrutia R: Sp1- and Krüppel-like transcription factors. *Genome Biol* 2003, 4:206.**
2. **Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, Aoi T, Okita K, Mochiduki Y, Takizawa N, Yamanaka S: Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. *Nat Biotechnol* 2008, 26:101-106.**
3. **Takahashi K, Yamanaka S: Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006, 126:663-676.**
4. **Wernig M, Meissner A, Cassady JP, Jaenisch R: c-Myc is dispensable for direct reprogramming of mouse fibroblasts. *Cell Stem Cell* 2008, 2:10-12.**
5. **Wernig M, Meissner A, Foreman R, Brambrink T, Ku M, Hochedlinger K, Bernstein BE, Jaenisch R: In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* 2007, 448:318-324.**
6. **Maekawa M, Yamaguchi K, Nakamura T, Shibukawa R, Kodanaka I, Ichisaka T, Kawamura Y, Mochizuki H, Goshima N, Yamanaka S: Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1. *Nature* 2011, 474:225-229.**
7. **Courey AJ, Tjian R: Analysis of Sp1 in vivo reveals multiple transcriptional domains, including a novel glutamine-rich activation motif. *Cell* 1988, 55:887-898.**
8. **Kriwacki RW, Schultz SC, Steitz TA, Caradonna JP: Sequence-specific recognition of DNA by zinc-finger peptides derived from the transcription factor Sp1. *Proc Natl Acad Sci USA* 1992, 89:9759-9763.**
9. **Schuetz A, Nana D, Rose C, Zocher G, Milanovic M, Koenigsmann J, Blasig R, Heinemann U, Carstanjen D: The structure of the Klf4 DNA-binding domain links to self-renewal and macrophage differentiation. *Cell Mol Life Sci* 2011, 68:3121-3131.**
10. **Pavletich NP, Pabo CO: Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* 1991, 252:809-817.**
11. **Desjarlais JR, Berg JM: Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc Natl Acad Sci USA* 1992, 89:7345-7349.**
12. **Letovsky J, Dynan WS: Measurement of the binding of transcription factor Sp1 to a single GC box recognition sequence. *Nucleic Acids Research* 1989, 17:2639-2653.**
13. **Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML: Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 2006, 24:1429-1435.**
14. **Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al: Diversity and Complexity in DNA Recognition by Transcription Factors. *Science* 2009, 324:1720-1723.**
15. **Pavletich NP, Pabo CO: Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science* 1993, 261:1701-1707.**
16. **Beak JY, Kang HS, Kim Y-S, Jetten AM: Functional analysis of the zinc finger and activation domains of Glis3 and mutant Glis3(NDH1). *Nucleic Acids Research* 2008, 36:1690-1702.**

17. **Raiber E-A, Kranaster R, Lam E, Nikan M, Balasubramanian S: A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro.** ***Nucleic Acids Research*** **2012, 40:1499-1508.**
18. **Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al: Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.** ***Cell*** **2008, 133:1106-1117.**
19. **Berger MF, Bulyk ML: Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors.** ***Nat Protoc*** **2009, 4:393-411.**
20. **Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** ***Nucleic Acids Research*** **1994, 22:4673-4680.**
21. **Workman CT, Yin Y, Corcoran DL, Ideker T, Stormo GD, Benos PV: enoLOGOS: a versatile web tool for energy normalized sequence logos.** ***Nucleic Acids Research*** **2005, 33:W389-392.**
22. **Jiang J, Chan Y-S, Loh Y-H, Cai J, Tong G-Q, Lim C-A, Robson P, Zhong S, Ng H-H: A core Klf circuitry regulates self-renewal of embryonic stem cells.** ***Nat Cell Biol*** **2008, 10:353-360.**
23. **Lu X-J, Olson WK: 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.** ***Nucleic Acids Research*** **2003, 31:5108-5121.**
24. **Emsley P, Cowtan K: Coot: model-building tools for molecular graphics.** ***Acta Crystallogr D Biol Crystallogr*** **2004, 60:2126-2132.**

Figure 4-1 Evolutionary divergence within the Klf/Sp family DNA binding domains determines reprogramming activity.

A) Schematic of chimeric protein generation. Klf4 DNA binding domain (3xZF, green) and characterized transactivation domain (AD, red) are indicated. Protein chimeras were generated by replacing the DNA binding domain of Klf4 with homologous sequence from a related protein. In this case, the Klf10 DNA binding domain (purple) was fused to the N-terminal region of Klf4 (residues 1-396). **B)** Reprogramming procedure overview. MEFs were infected with retroviruses carrying Oct4, Sox2, and a Klf4-ZF chimera on day 0. The reprogramming culture was transitioned to KSR mESC media on day 5. Cells were fixed on day 12 and immunostained for the pluripotency marker, Nanog. Reprogramming was quantified by counting Nanog+ colonies. **C)** Nanog+ colony counts for Klf4-ZF protein chimeras expressed as a percentage of wild-type Klf4 colonies. Klf4-ZF protein chimeras are ordered by Clustal multiple sequence alignment of their DNA binding domain. Reprogramming-competent (red) and -incompetent (yellow) DNA binding domains within the Klf/Sp family are colored on the alignment tree. N.D. indicates that Klf4-ZF protein chimera was not expressed. Graph contains data collected from separate reprogramming experiments. Each sample is normalized to the wild-type Klf4 control performed within its respective experiment. Error bars represent standard deviation. Error bars for Klf4ZF are from the experiment with the largest standard deviation.

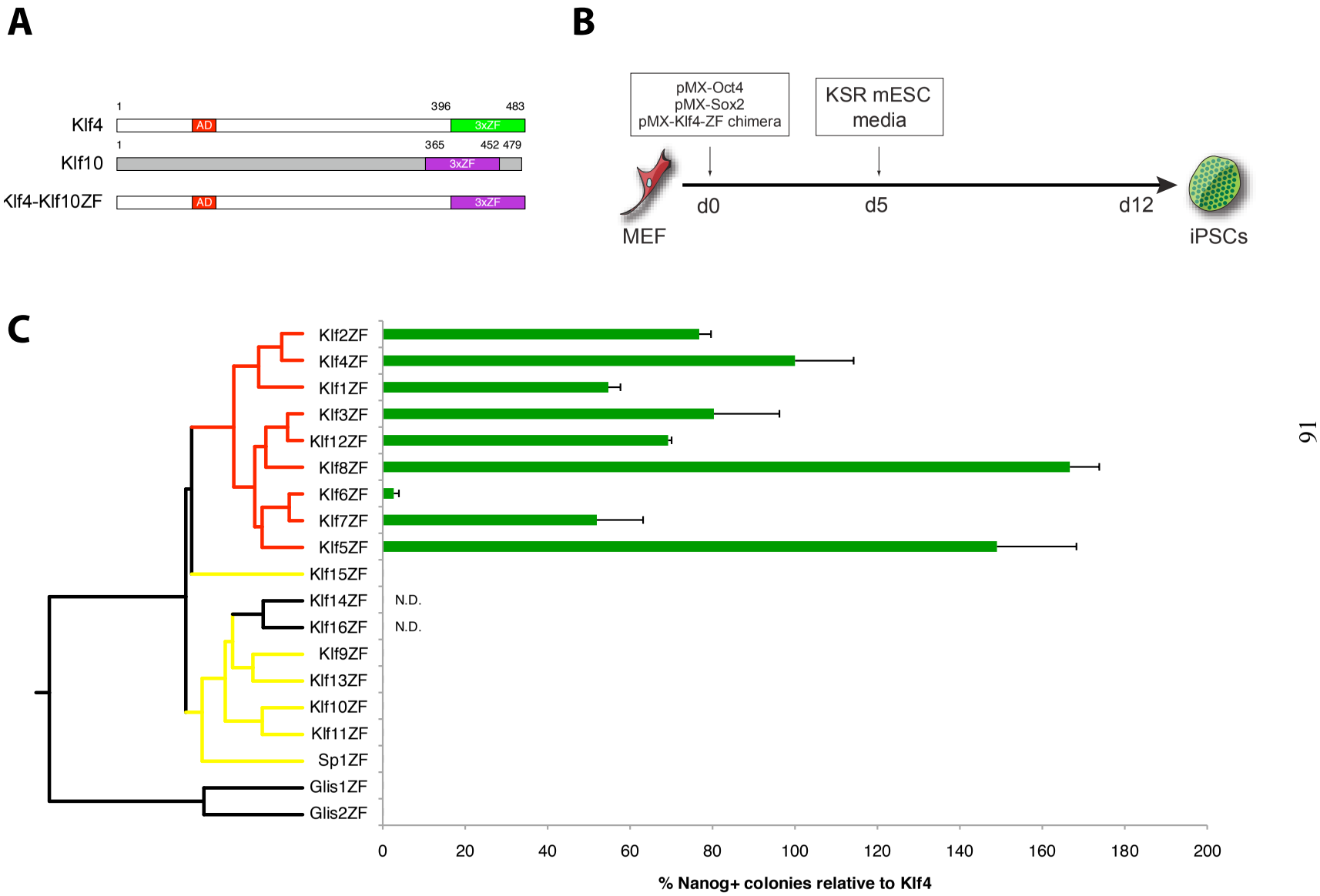

A
1
396
483
Klf4
AD
3xZF
1
365
452 479
Klf10
3xZF
Klf4-Klf10ZF
AD
3xZF
B
pMX-Oct4
pMX-Sox2
pMX-Klf4-ZF chimera
KSR mESC
media
MEF
d0
d5
d12
iPSCs
C
Klf2ZF
Klf4ZF
Klf1ZF
Klf3ZF
Klf12ZF
Klf8ZF
Klf6ZF
Klf7ZF
Klf5ZF
Klf15ZF
Klf14ZF
Klf16ZF
Klf9ZF
Klf13ZF
Klf10ZF
Klf11ZF
Sp1ZF
Glis1ZF
Glis2ZF
N.D.
N.D.
0
20
40
60
80
100
120
140
160
180
200
% Nanog+ colonies relative to Klf4

Figure 4-2 C-terminal sequence does not affect protein chimera reprogramming activity. Inclusion of C-terminal region in Klf4-Klf10ZF chimeric protein (Klf4-Klf10C) does not alter its reprogramming activity. Error bars represent standard deviation.

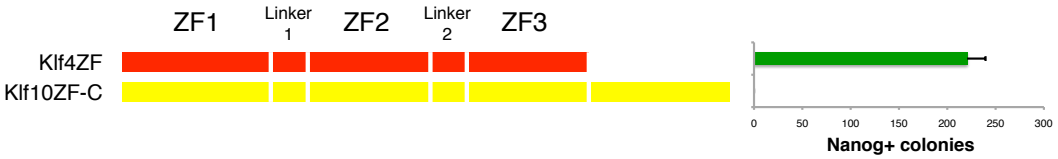

ZF1
Linker 1
ZF2
Linker 2
ZF3
Klf4ZF
Klf10ZF-C
0
50
100
150
200
250
300
Nanog+ colonies

Figure 4-3 Differences in the second and third zinc finger regions determine reprogramming activity.

The 3xZF DNA binding domain can be separated into 5 regions consisting of 3 zinc fingers (ZFs) and 2 linker sequences. Klf4 and Klf10 DNA binding domains were chosen as representatives of the reprogramming-competent and -incompetent subgroups, respectively, within the Klf/Sp family. Fine-scale protein chimeras were generated within the DNA binding domain containing a combination of Klf4 (red) and Klf10 (yellow) sequences. Efficient reprogramming required ZF2 and ZF3 to contain Klf4 sequence. Klf10 sequence in either of these regions disrupted reprogramming function. Graph contains data collected from separate reprogramming experiments. Each sample is normalized to the wild-type Klf4 control performed within its respective experiment. Error bars represent standard deviation. Error bars for Klf4ZF are from the experiment with the largest standard deviation.

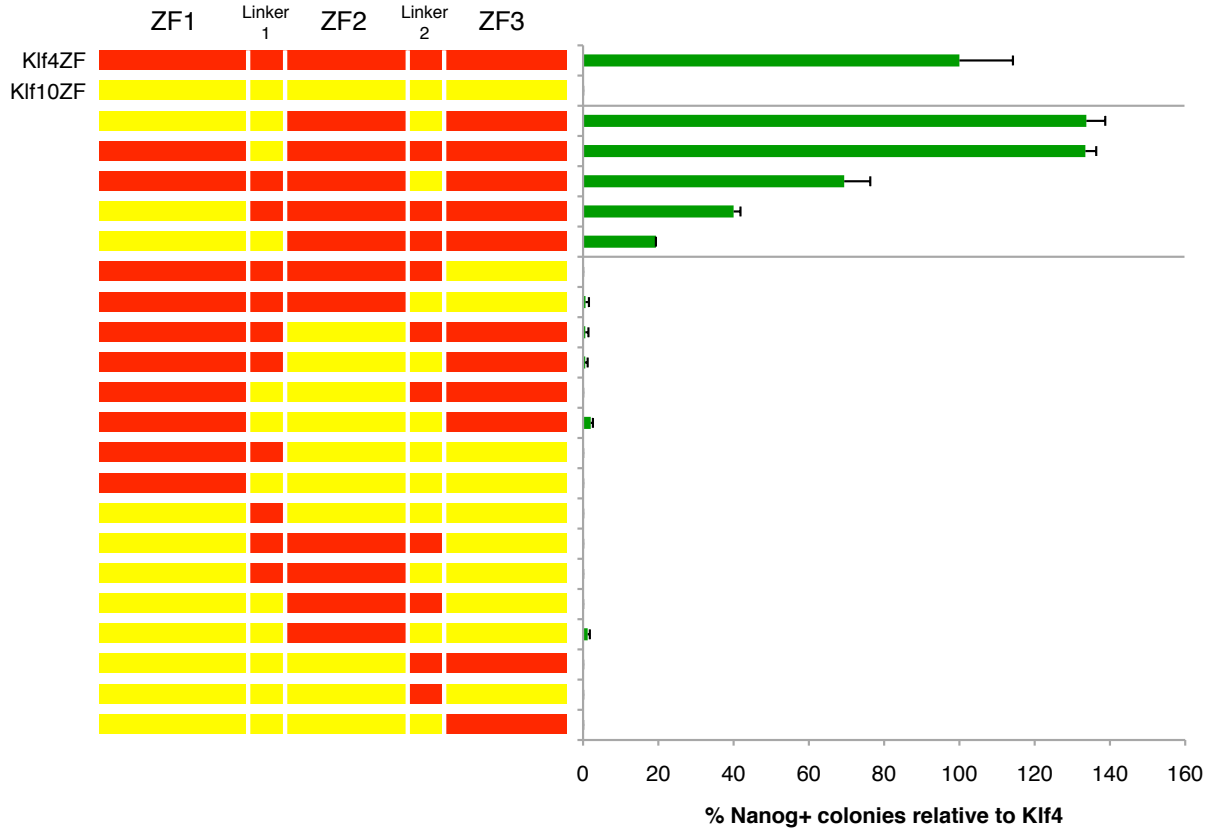

ZF1
Linker 1
ZF2
Linker 2
ZF3
Klf4ZF
Klf10ZF
0
20
40
60
80
100
120
140
160
% Nanog+ colonies relative to Klf4

Figure 4-4 Multiple sequence alignments for each zinc finger region.

ZF1 (**A**), ZF2 (**B**), and ZF3 (**C**) regions are shown. Absolutely conserved residues are highlighted in yellow.

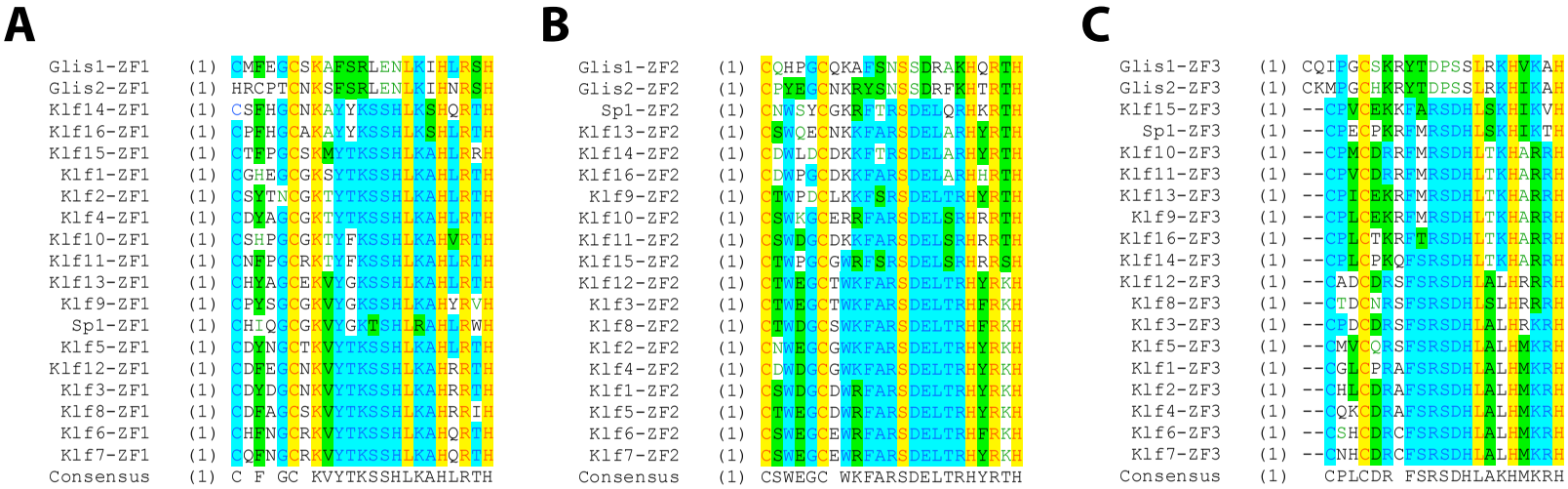
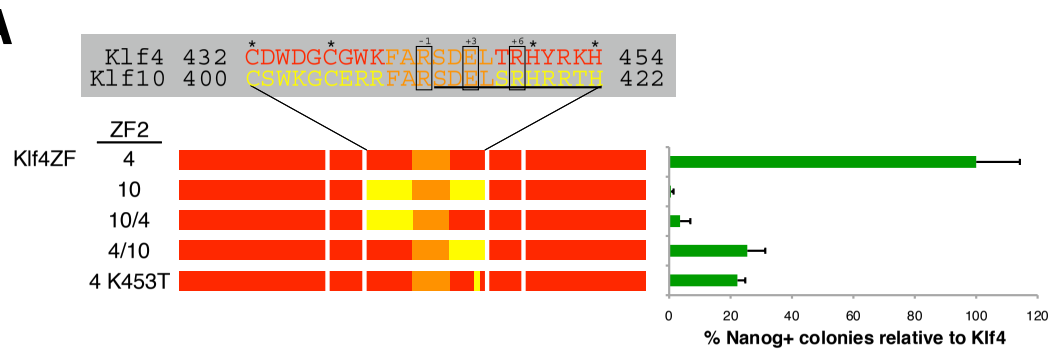

A
Glis1-ZF1 (1) CMFEGCSKAFSRLENLKIHLRSH
Glis2-ZF1 (1) HRCPTCNKSFSRLENLKIHNRSH
Klf14-ZF1 (1) CSFHGCNKAYYKSSHLKSHQRTH
Klf16-ZF1 (1) CPFHGCAKAYYKSSHLKSHLRTH
Klf15-ZF1 (1) CTFPGCSKMYTKSSHLKAHLRRH
Klf1-ZF1 (1) CGHEGCGKSYTKSSHLKAHLRTH
Klf2-ZF1 (1) CSYTNCGKTYTKSSHLKAHLRTH
Klf4-ZF1 (1) CDYAGCGKTYTKSSHLKAHLRTH
Klf10-ZF1 (1) CSHPGCGKTYFKSSHLKAHVRTH
Klf11-ZF1 (1) CNFPGCRKTYFKSSHLKAHLRTH
Klf13-ZF1 (1) CHYAGCEKVYGKSSHLKAHLRTH
Klf9-ZF1 (1) CPYSGCGKVYGKSSHLKAHYRVH
Sp1-ZF1 (1) CHIQGCGKVYGKTSHLRAHLRWH
Klf5-ZF1 (1) CDYNGCTKVYTKSSHLKAHLRTH
Klf12-ZF1 (1) CDFEGCNKVYTKSSHLKAHRRTH
Klf3-ZF1 (1) CDYDGCNKVYTKSSHLKAHRRTH
Klf8-ZF1 (1) CDFAGCSKVYTKSSHLKAHRRIH
Klf6-ZF1 (1) CHFNGCRKVYTKSSHLKAHQRTH
Klf7-ZF1 (1) CQFNGCRKVYTKSSHLKAHQRTH
Consensus (1) C F GC KVYTKSSHLKAHLRTH
B
Glis1-ZF2 (1) CQHPGCQKAFSNSSDRAKHQRTH
Glis2-ZF2 (1) CPYEGCNKRYSNSSDRFKHTRTH
Sp1-ZF2 (1) CNWSYCGKRFTRSDELQRHKRTH
Klf13-ZF2 (1) CSWQECNKKFARSDELARHYRTH
Klf14-ZF2 (1) CDWLDCDKKFTRSDELARHYRTH
Klf16-ZF2 (1) CDWPGCDKKFARSDELARHHRTH
Klf9-ZF2 (1) CTWPDCLKKFSRSDELTRHYRTH
Klf10-ZF2 (1) CSWKGCERRFARSDELSRHRRTH
Klf11-ZF2 (1) CSWDGCDKKFARSDELSRHRRTH
Klf15-ZF2 (1) CTWPGCGWRFSRSDELSRHRRSH
Klf12-ZF2 (1) CTWEGCTWKFARSDELTRHYRKH
Klf3-ZF2 (1) CTWEGCTWKFARSDELTRHFRKH
Klf8-ZF2 (1) CTWDGCSWKFARSDELTRHFRKH
Klf2-ZF2 (1) CNWEGCGWKFARSDELTRHYRKH
Klf4-ZF2 (1) CDWDGCGWKFARSDELTRHYRKH
Klf1-ZF2 (1) CSWDGCDWRFARSDELTRHYRKH
Klf5-ZF2 (1) CTWEGCDWRFARSDELTRHYRKH
Klf6-ZF2 (1) CSWEGCEWRFARSDELTRHFRKH
Klf7-ZF2 (1) CSWEGCEWRFARSDELTRHYRKH
Consensus (1) CSWEGC WKFARSDELTRHYRTH
C
Glis1-ZF3 (1) CQIPGCSKRYTDPSSLRKHVKAH
Glis2-ZF3 (1) CKMHGCHKRYTDPSSLRKHIKAH
Klf15-ZF3 (1) --CPVCEKKFARSDHLSKHIKVH
Sp1-ZF3 (1) --CPECPKRFMRSDHLSKHIKTH
Klf10-ZF3 (1) --CPMCDRRFMRSDHLTKHARRH
Klf11-ZF3 (1) --CPVCDRRFMRSDHLTKHARRH
Klf13-ZF3 (1) --CPICEKRFMRSDHLTKHARRH
Klf9-ZF3 (1) --CPLCEKRFMRSDHLTKHARRH
Klf16-ZF3 (1) --CPLCTKRFTRSDHLTKHARRH
Klf14-ZF3 (1) --CPLCPKQFSRSDHLTKHARRH
Klf12-ZF3 (1) --GADCDRSFSRSDHLALHRRRH
Klf8-ZF3 (1) --CTDCNRSFSRSDHLSLHRRRH
Klf3-ZF3 (1) --CPDCDRSFSRSDHLALHRKRH
Klf5-ZF3 (1) --CMVCQRSFSRSDHLALHMKRH
Klf1-ZF3 (1) --CGLCPRAFSRSDHLALHMKRH
Klf2-ZF3 (1) --CHLCDRAFSRSDHLALHMKRH
Klf4-ZF3 (1) --CQKCDRAFSRSDHLALHMKRH
Klf6-ZF3 (1) --CSHCDRCFSRSDHLALHMKRH
Klf7-ZF3 (1) --CNHCDRCFSRSDHLALHMKRH
Consensus (1) CPLCDR FSRSDHLAKHMKRH

Figure 4-5 ZF2 functional divergence lies within β-sheet and does not result in altered DNA binding specificity.

A) ZF2 sequences are displayed for Klf4 (red) and Klf10 (yellow). A central region of absolute conservation is highlighted in orange. Zinc-coordinating residues (stars), recognition helix (underline), and potential base contacting residues (boxes; -1, +3, +6 positions in α-helix) are indicated. Graph contains data collected from separate reprogramming experiments. Each sample is normalized to the wild-type Klf4 control performed within its respective experiment. Error bars represent standard deviation. Error bars for Klf4ZF are from the experiment with the largest standard deviation. **B)** DNA binding motifs determined *in vitro* using protein binding microarrays (PBMs) for DNA binding domains containing ZF2 derived from either Klf4 or Klf10. **C)** Table shows residues that differ between Klf4 and Klf10 within ZF2. Orange line indicates the position of the central conserved region. Point mutations were made at each position listed in the table and tested in reprogramming. Graph contains data collected from separate reprogramming experiments. Each sample is normalized to the wild-type Klf4 control performed within its respective experiment. Error bars represent standard deviation. Error bars for Klf4ZF are from the experiment with the largest standard deviation. **D)** Crystal structure of Klf4 DNA binding domain bound to DNA (PDB: 2WBU) [9] highlighting residues within the β-sheet that differ between Klf4 and Klf10 (purple).

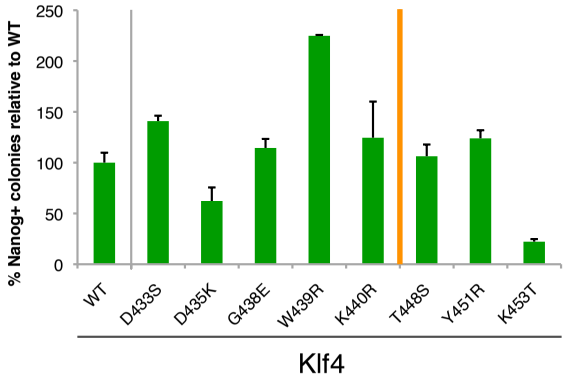
A



B

C

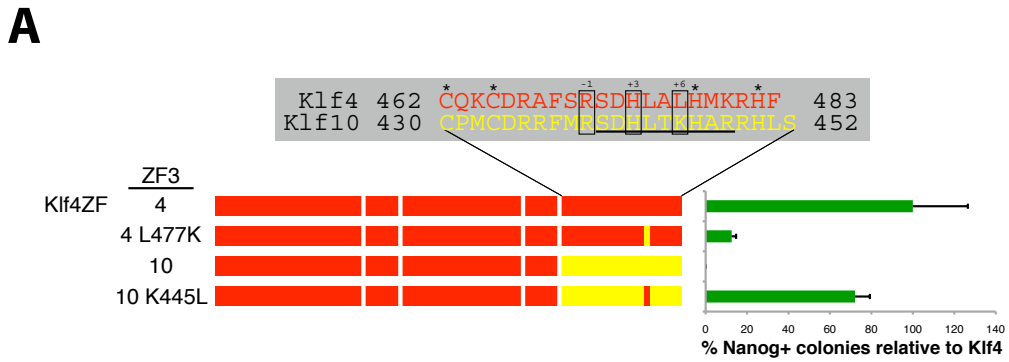
Klf4 Residue #	Klf4 Residue	Klf10 Residue	Klf10 Residue #
433	D	S	401
435	D	K	403
438	G	E	406
439	W	R	407
440	K	R	408
448	T	S	416
451	Y	R	419
453	K	T	421

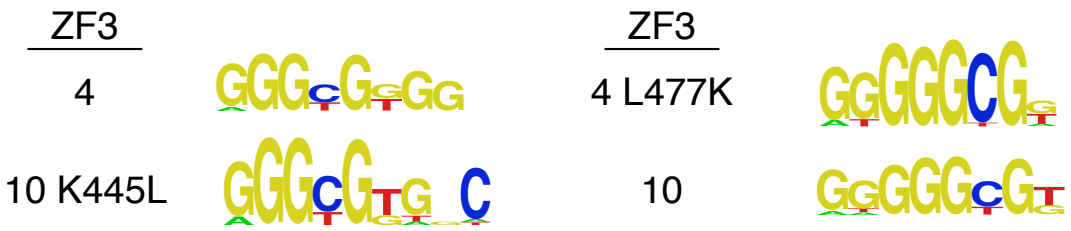


D

Figure 4-6 ZF3 functional divergence is due to altered DNA binding specificity.

A) ZF3 sequences are displayed for Klf4 (red) and Klf10 (yellow). Zinc-coordinating residues (stars), recognition helix (underline), and potential base contacting residues (boxes; -1, +3, +6 positions in α-helix) are indicated. Note difference at +6 position within the recognition helix (L in Klf4, K in Klf10). Graph contains data collected from separate reprogramming experiments. Each sample is normalized to the wild-type Klf4 control performed within its respective experiment. Error bars represent standard deviation. Error bars for Klf4ZF are from the experiment with the largest standard deviation. **B)** DNA binding motifs determined *in vitro* using protein binding microarrays (PBMs) for DNA binding domains containing ZF3 derived from either Klf4 or Klf10 along with point mutants. Reprogramming activity correlates with DNA binding preference. **C)** Scatterplots of 8-mers shows altered DNA binding specificity due to the residue in the +6 position. Klf4 and Klf10 are plotted against their respective point mutants. 8-mers containing indicated sequences are highlighted in yellow, blue, or red.


A
Klf4 462 CQKCDRAFSRSDHLALHMKRHF 483
Klf10 430 CPMCDRRFMRSDHLTKHARRHLS 452
ZF3
Klf4ZF
4
4 L477K
10
10 K445L
0 20 40 60 80 100 120 140
% Nanog+ colonies relative to Klf4


B
ZF3
4
ZF3
4 L477K
10 K445L
10

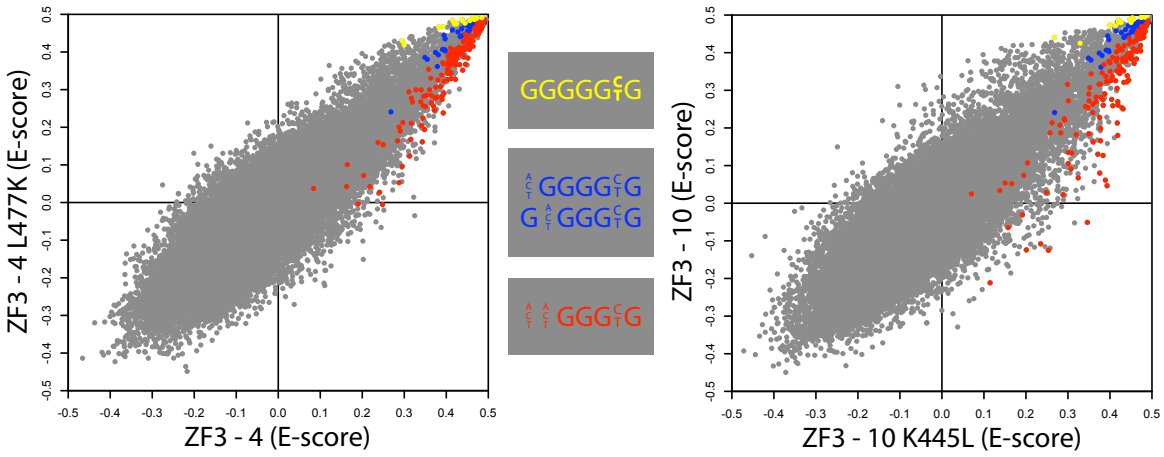
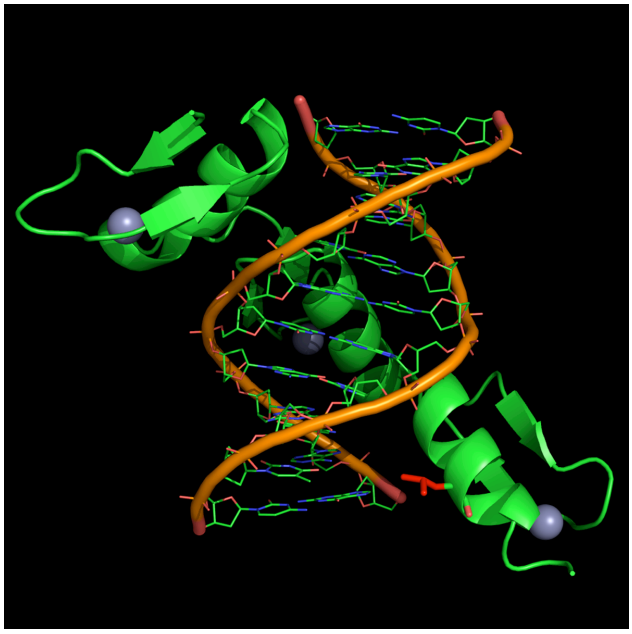

C
ZF3 - 4 L477K (E-score)
ZF3 - 4 (E-score)
GGGGGTG
ZF3 - 10 (E-score)
ZF3 - 10 K445L (E-score)

Figure 4-7 Structural model of ZF3 +6 lysine contacting guanine bases.

A) Crystal structure of Klf4 DNA binding domain bound to DNA (PDB: 2WBU) [9] highlighting L477 (red). **B)** Structural model of Klf4 L477K bound to DNA highlights potential molecular contacts driving altered DNA binding specificity. The lysine amino group is positioned within hydrogen-bonding distance of both N7 atoms of stacked guanine bases.

A



B

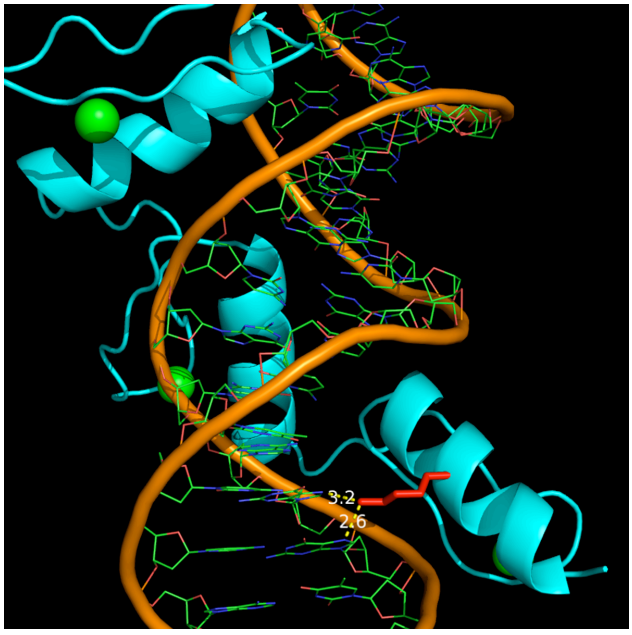
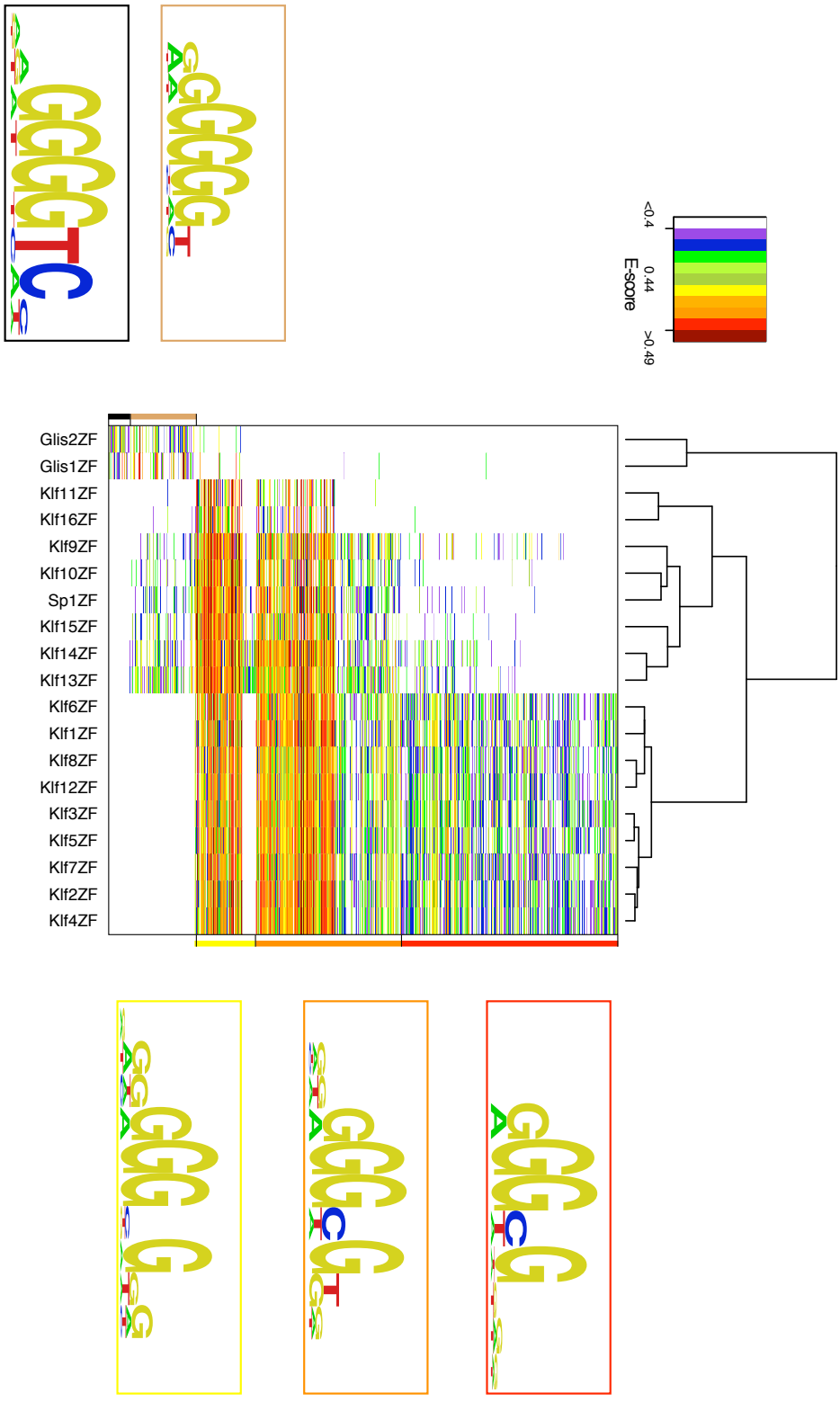


Figure 4-8 DNA binding preferences within the Klf/Sp family are split along evolutionary lines.

Heatmap contains 8-mers where the E-score for at least one protein exceeds 0.40. Dendrogram results from hierarchical clustering of datasets. Note the correlation between dendrogram, phylogenetic tree, and reprogramming activity in Klf4-ZF chimeras. 8-mers (rows) are clustered by k-means clustering (k=10). Clusters with similar DNA binding motifs were arranged into groups (red, orange, yellow, tan, black). One motif derived from alignment of all constituent 8-mers is displayed for each group.


Glis2ZF
Glis1ZF
Klf11ZF
Klf16ZF
Klf9ZF
Klf10ZF
Sp1ZF
Klf15ZF
Klf14ZF
Klf13ZF
Klf6ZF
Klf1ZF
Klf8ZF
Klf12ZF
Klf3ZF
Klf5ZF
Klf7ZF
Klf2ZF
Klf4ZF
E-score
<0.4
0.44
>0.49

CHAPTER 5

CONCLUSIONS

The work presented in this dissertation investigates the molecular mechanism of somatic cell reprogramming, focusing on the role of Klf4. Previous studies of the reprogramming factors mainly examined their genome-wide binding patterns and the associated expression changes of nearby genes. We sought to look within a reprogramming factor molecule to learn how it reaches these target sites and how it regulates gene expression once it has arrived. We took advantage of the essential nature of Klf4 in the Yamanaka reprogramming cocktail [1] to compare the function of mutants to the wild-type protein. Our experiments map, for the first time, the regions of this protein that function in the induction of pluripotency. Additionally, we carry out a fine-scale analysis of the transactivation and DNA binding domains of Klf4 to identify critical residues and demonstrate how they may contribute to reprogramming activity.

Understanding Reprogramming Factor Function

In chapter 2, we overviewed the changes in chromatin and gene expression states that occur during somatic cell reprogramming of MEFs. We then discussed the known functions of Oct4, Sox2, and Klf4 in this process. We outlined a framework for learning about the reprogramming mechanism through the identification and study of modifier factors. Enhancer factors increase the efficiency of reprogramming, and may do so in a cell proliferation-dependent or -independent manner. The mechanism by which cell cycling enhances the reprogramming process is not entirely clear and is an important area of future investigation. Cell proliferation-independent enhancer factors may act early or late within the reprogramming process. The timing of their action can be established by assessing their ability to promote the generation of pre-iPS cells or convert them to the fully reprogrammed state. Factors that contribute to induced pluripotency in place of Oct4, Sox2, or Klf4 can be classified by their similarity to the factor that they replace. Paralogs, which can also induce pluripotency, likely act on a common set of

critical genes using homologous functional domains. Thus, study of these proteins may help us to focus on the regions within the reprogramming factor protein that determine its activity. Dissimilar replacement factors may regulate common genes or pathways, but likely do so through distinct molecular mechanisms.

Analysis of gene expression data obtained over the course of the reprogramming process showed large-scale upregulation of genes encoding nuclear proteins, especially those involved in chromatin modification. Many of these proteins function as part of multisubunit complexes. Examination of the expression changes within individual complexes revealed outstanding expression patterns of complex substituents that may play important roles in dictating pluripotency. These proteins should be tested functionally in reprogramming assays to determine their role, and potentially the role of their associated histone modification, in the induction of pluripotency. The chromatin marks that these complexes deposit are just now being mapped during reprogramming. These data will contribute greatly to our understanding of how Oct4, Sox2, and Klf4 are able to initially bind to chromatin and how these proteins then reprogram the epigenome.

Finally, study of the reprogramming factors themselves is necessary to understand induced pluripotency. What are their important domains? Which specific functions do they contribute? How does the coordinated action of the reprogramming factors lead to the iPS cell state?

Identification and Characterization of Functional Domains in Klf4

In chapter 3, we used mutagenesis to identify regions of Klf4 that are required for its function in reprogramming. We found that Klf4 requires its C-terminal DNA binding domain, an adjacent domain of unknown function, and N-terminal transactivation domains (TADs). The

requirement for its DNA binding domain indicates that sequence-specific targeting of Klf4 to regulatory elements is an important component of its reprogramming function. While the importance of the DNA binding domain was not unexpected, we were surprised to discover that an adjacent region consisting of residues 350-396 was also essential for reprogramming activity. This region was previously shown to contain a nuclear localization sequence (NLS) [2]; however, nuclear localization was unaffected in our deletion mutant construct likely due to the presence of a second NLS within the DNA binding domain [2]. Our work is the first suggestion that this region is important for the activity of the Klf4 protein. Thus, follow-up experiments are necessary to determine whether the loss of residues 350-396 alters its ability to bind DNA or transactivate transcription. When we deleted the previously characterized acidic TAD, we observed that this region is important for the induction of pluripotency. However, the mutant lacking this domain still maintained some residual reprogramming activity. Further experiments indicated that latent transactivation activity, which functions in the absence of the acidic TAD, is possessed by an adjacent region (residues 145-209). This activity has not been observed previously in other systems, and thus represents a function that is specific to the reprogramming context.

After identifying important regions of Klf4 in reprogramming, we sought to determine the molecular mechanisms through which they act. We wondered if the acidic TAD (residues 90-110) within Klf4 simply contained a general transactivation function that could be replaced by well-studied TADs derived from Sp1 and VP16. We fused these TADs to reprogramming-deficient Klf4 mutants and observed that only the acidic TAD from Klf4 was able to rescue reprogramming activity. This result indicates that this domain possesses a unique activity in the context of somatic cell reprogramming that extends beyond its ability to transactivate a reporter

gene. We reasoned that this difference may be due to an interaction with a distinct cofactor, and we sought to narrow in on the residues that interact with this potential cofactor by creating point mutations within the 90-110 region. Using this approach, we identified hydrophobic residues that are critical for the function of this TAD in reprogramming. Interestingly, acidic residues that had been previously shown to be critical for the transactivation activity of Klf4 in other cell types were not important for its reprogramming activity [3, 4]. We then used an open-ended biochemical approach to isolate potential cofactor proteins that bind to the wild-type TAD but not to the TAD with mutated hydrophobic residues. Through this method, we demonstrated that clathrin heavy chain binds to the Klf4 TAD through a consensus binding motif that is disrupted by mutation of the hydrophobic residues. Clathrin heavy chain has previously been shown to function as a coactivator through interaction with an acidic TAD within p53 [5]. This interaction was also dependent on hydrophobic residues within the p53 TAD [6, 7]. However, p53 uses a distinct motif that recognizes a different region of the clathrin heavy chain molecule [6, 7]. Future experiments are necessary to validate the functional significance of the Klf4-clathrin heavy chain interaction. Also, given the shared interaction of the p53 and Klf4 TADs with clathrin heavy chain, it will be interesting to determine whether the p53 TAD can replace its Klf4 counterpart in reprogramming.

In chapter 4, we performed a functional comparison of the DNA binding domains within the Klf family in order to understand the characteristics of the Klf4 DNA binding domain that govern its reprogramming activity. We show that only a subset of zinc finger domains from the Klf family are able to replace the Klf4 DNA binding domain in reprogramming. This difference in reprogramming activity arises through variation in the second and third zinc fingers within the DNA binding domain. Each zinc finger can be separated into an antiparallel β-sheet and an α-

helix. Amino acid side chains extend from the α-helix and can make base-specific contacts with the DNA. We found that differences in reprogramming activity in the Klf family attributable to the second zinc finger lie mainly in its β-sheet. The function of this region of the DNA binding domain remains unclear. We speculated that this β-sheet might serve as a platform for the binding of an important protein cofactor. However, point mutagenesis across this region did not identify any one single residue as critical for reprogramming activity, arguing against this notion. Further experiments are necessary to determine if variation in the second zinc finger alters the conformation of the DNA binding domain in a manner that reduces binding affinity to target sites.

Reprogramming-competent and -incompetent DNA binding domains in the Klf family are distinguished by the identity of one of the base-contacting amino acids in the third zinc finger. Our work demonstrated that this change dictates *in vitro* DNA binding specificity and largely explains observed differences in reprogramming activity due to this zinc finger. We found that reprogramming-incompetent zinc fingers are restricted in their binding as a result of their preference for additional guanines within their recognition sites, and we presented a structural model to explain this phenomenon. We speculate that the subset of sites that are bound *in vitro* by reprogramming-competent Klfs but not their reprogramming-incompetent family members are likely to be contained in the regulatory elements of critical reprogramming targets. In the future, we hope to identify these genes by mapping the genome-wide binding patterns of protein chimeras containing various Klf DNA binding domains during the early stages of reprogramming.

Our finding that multiple DNA binding domains within the Klf family bind to the same sequence *in vitro* and are functionally redundant in reprogramming raises the question of why so

many derivatives of the ancestral Klf gene appear in mammalian genomes. One mechanism that functionally differentiates these proteins and could explain this phenomenon is the composition of the poorly conserved regions outside of the DNA binding domain. Klf proteins possess several distinct domains that have been implicated in gene activation or repression [8]. Targeting of Klf proteins to a gene regulatory element by similar DNA binding domains may lead to contrasting effects depending on the abilities of these domains to recruit different effector proteins. Additonally, the results from our experiments with the Klf4 acidic transactivation domain indicate that not all domains from a given class are functionally interchangeable. Thus, some genes may require specific subsets of coactivators or corepressors in order to turn them on or off, unlike artificial reporter constructs. Klf proteins may also be differentially regulated by cofactors or posttranslational modifications that alter their activity or stability. For example, binding of the E3 ubiquitin ligase, SIAH1, specifically to Klf10 leads to its ubiquitylation and degradation [9]. Finally, the presence of multiple Klf genes allows for individualized regulation of expression patterns since each is under the control of distinct regulatory elements. This allows their actions to be separated in time and space during development and enables them to be placed downstream of different stimuli. In gut epithelium, for instance, Klf5 and Klf4 exert opposing effects on cell proliferation through their mutually exclusive expression patterns. The proliferating cells of the intestinal crypt express Klf5, while their post-mitotic progeny in the adjacent villus express Klf4 [10].

Concluding Remarks

Somatic cell reprogramming is an incredible feat of cell fate engineering with great therapeutic potential. The discovery of this process was built on knowledge obtained through the study of transcription factors and their functions during development. This class of proteins will

play an important role in the differentiation of iPS cells into desired cell types as well as the development of future reprogramming protocols that do not transition through the pluripotent state. While much is known about model transcription factors that have been well-studied *in vitro*, we still lack the ability to predict where in the genome these proteins will bind and what the effect on gene expression will be once they get to their target sites. Additionally, we lack the ability to anticipate the combinatorial effects that occur when transcription factors are coexpressed in cell. Reprogramming of MEFs using Oct4, Sox2, and Klf4 represents a useful model system to begin to address these issues. The work presented in this dissertation advances the understanding of the molecular mechanism of Klf4 function during reprogramming and lays out an experimental approach that can be applied to the other reprogramming factors.

References

1. **Takahashi K, Yamanaka S: Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006, 126:663-676.**
2. **Shields JM, Yang VW: Two potent nuclear localization signals in the gut-enriched Krüppel-like factor define a subfamily of closely related Krüppel proteins. *J Biol Chem* 1997, 272:18504-18507.**
3. **Du JX, McConnell BB, Yang VW: A small ubiquitin-related modifier-interacting motif functions as the transcriptional activation domain of Krüppel-like factor 4. *J Biol Chem* 2010, 285:28298-28308.**
4. **Geiman DE, Ton-That H, Johnson JM, Yang VW: Transactivation and growth suppression by the gut-enriched Krüppel-like factor (Krüppel-like factor 4) are dependent on acidic amino acid residues and protein-protein interaction. *Nucleic Acids Research* 2000, 28:1106-1113.**
5. **Enari M, Ohmori K, Kitabayashi I, Taya Y: Requirement of clathrin heavy chain for p53-mediated transcription. *Genes & Development* 2006, 20:1087-1099.**
6. **Ohata H, Ota N, Shirouzu M, Yokoyama S, Yokota J, Taya Y, Enari M: Identification of a function-specific mutation of clathrin heavy chain (CHC) required for p53 transactivation. *J Mol Biol* 2009, 394:460-471.**
7. **Ohmori K, Endo Y, Yoshida Y, Ohata H, Taya Y, Enari M: Monomeric but not trimeric clathrin heavy chain regulates p53-mediated transcription. *Oncogene* 2008, 27:2215-2227.**
8. **Kaczynski J, Cook T, Urrutia R: Sp1- and Krüppel-like transcription factors. *Genome Biol* 2003, 4:206.**
9. **Subramaniam M, Hawse JR, Rajamannan NM, Ingle JN, Spelsberg TC: Functional role of KLF10 in multiple disease processes. *BioFactors* 2010, 36:8-18.**
10. **Nandan MO, Yang VW: The role of Krüppel-like factors in the reprogramming of somatic cells to induced pluripotent stem cells. *Histol Histopathol* 2009, 24:1343-1355.**