Mechanism Of Nucleosome Targeting By Pioneer Transcription Factors

Meilin Mary Fernandez Garcia
University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/edissertations

Part of the Biochemistry Commons, and the Biophysics Commons

Recommended Citation
https://repository.upenn.edu/edissertations/3624

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/3624
For more information, please contact repository@pobox.upenn.edu.
Mechanism Of Nucleosome Targeting By Pioneer Transcription Factors

Abstract
Transcription factors (TFs) forage the genome to instruct cell plasticity, identity, and differentiation. These developmental processes are elicited through TF engagement with chromatin. Yet, how and which TFs can engage with chromatin and thus, nucleosomes, remains largely unexplored. Pioneer TFs are TF that display a high affinity for nucleosomes. Extensive genetic and biochemical studies on the pioneer TF FOXA, a driver of fibroblast to hepatocyte reprogramming, revealed its nucleosome binding ability and chromatin targeting lead to chromatin accessibility and subsequent cooperative binding of TFs. Similarly, a number of reprogramming TFs have been suggested to have pioneering activity due to their ability to target compact chromatin and increase accessibility and enhancer formation in vivo. But whether these factors directly interact with nucleosomes remains to be assessed. Here we test the nucleosome binding ability of the cell reprogramming TFs, Oct4, Sox2, Klf4 and cMyc, that are required for the generation of induced pluripotent stem cells. In addition, we also test neuronal and macrophage reprogramming TFs. Our study shows that reprogramming TFs bind nucleosomes with a range of nucleosome binding affinities, indicating that although specific cocktails of TFs are required for reprogramming, mechanistically these TFs show differential nucleosome interacting behaviors. These results allowed us to assess differential features between TFs nucleosome binding ability and to correlate their binding with reprogramming potential.

To determine how general is nucleosome binding we extended our analysis to screen 593 of the 2,000 predicted human TFs in the genome for potential nucleosome binding and validated their binding in solution. Based on 3D structural analysis, we proposed that strong nucleosome binders anchor DNA through short -helixes and have a flexible and adaptable DNA binding domain while weak nucleosome binders use -sheets or unstructured regions and have a higher rigidity within their DNA binding domain. Through the experiments presented in this dissertation we present the first study revealing the shared structural features contributing to nucleosome binding potential of pioneer TFs and thus allow for predication of novel pioneer TFs with cell reprogramming potential.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Biochemistry & Molecular Biophysics

First Advisor
Kenneth S. Zaret

Keywords
Pioneer Transcription Factor

Subject Categories
Biochemistry | Biophysics

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/3624
MECHANISM OF NUCLEOSOME TARGETING BY PIONEER TRANScription FACTORS

Meilín M. Fernández García

A DISSERTATION

in

Biochemistry and Molecular Biophysics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

______________________________
Kenneth S. Zaret, Ph.D., Joseph Leidy Professor of Cell and Developmental Biology

Graduate Group Chairperson

______________________________
Kim A. Sharp, Ph.D., Associate Professor of Biochemistry and Biophysics

Dissertation Committee

Ronen Marmorstein, Ph.D., George W. Raiziss Professor

Ben E. Black, Ph.D., Professor of Biochemistry and Biophysics

Doris Wagner, Ph.D., Robert I. Williams Term Professor of Biology

Marisa Bartolomei, Ph.D., Professor of Cell and Developmental Biology
In memory of my Dad.

Dedicated to my mom and “abuela” for molding me into the strong and persistent woman I am today.
ACKNOWLEDGMENTS

A massive shout out to my thesis advisor and greatest cheerleader, Ken Zaret. I would not have made it this far without your dedicated mentorship and unending optimism. You believed in my skills even when I did not, pushed me enough, and above all valued my wellbeing. I could not have picked a better mentor for my Ph.D. It was truly a privilege to be part of your lab. I leave the lab with lasting fun memories knowing that I’ll always have an exceptional mentor.

Thanks you to my thesis committee, Drs. Ronen Marmorstein, Ben Black and Doris Wagner, for your guidance and for keeping me focused. Special thanks to Doris for teaching me about the joys of teaching and collaboration.

I would like to thank the BGS program, in particularly the Biochemistry and Molecular Biophysics Program, the Epigenetics Program, the Department of Cell and Developmental Biology and the Office of Research and Diversity Training for creating a fun, supportive, diverse and collaborative training environment receptive to change. I am very grateful for the opportunity to participate and create many fun events as part of BGS. Special thanks to Kim Sharp, Marisa Bartolomei, Shelly Berger, Ben Garcia, Vera Moiseenkova-Bell, Erica Korb and Elizabeth Heller who in many ways advised me in project and career exploration.
To my dear lab members, I miss you all already. The everyday failures of science became more bearable because you were there with me. You became my family. Special thanks to Abdenour Soufi, who trusted in me and took me under his wing. Thanks to Naomi Takenaka, who was always there to share my excitement and downfalls with a smile, to my forever baymate Dario Nicetto for smacking me down to earth when needed and being next to me all throughout my Ph.D., to the grads of the lab Justin Becker, Katherine Palozola, Jessica Grindheim, Megan Frederick and Kelsey Kaeding for teaching me about life in particular to Katherine Palozola for closely sharing with me my last few months in the lab and being my personal editor. Finally, to the post-docs for sharing their wisdom. This thesis was a lab effort I could not have made it without your comments, discussion and advise.

To my friends in the Smilow 9th floor. Thanks for all the lunches and coffee shared, the dancing nights and the late work night in great company.

Endless thanks to all the wonderful friends I’ve made through gradschool. You made Philly feel like home. Thanks to Payal Jain, Mariel Coradin Aimee Juan, Erika Boutefeur, Melody Esmaeli, and Izmarie Poventud my partners in crime and being part of my girls squad. Gradschool would not have been the same without you. And to my almost human friend Charlie, who silently reminded to take constants breaks and unconditional enthusiasm.

My most sincere gratitude to my family for their continued support and unconditional love. I am grateful to my sister Emily Fernandez who befriend me
the moment we started gradschool and since then has gone along with all of my
crazy life projects. To my brother Miguel Fernandez Garcia who constantly
reminds me to take risks and to stay young in spirit. I also thank my mom for
prioritizing our education above all and giving me the freedom to leave my home
and family to pursue my career.

Finally, I would like to thank myself for making it till the end.
ABSTRACT

MECHANISM OF NUCLEOSOME TARGETING BY PIONEER TRANSCRIPTION FACTORS

Meilín M. Fernández García
Kenneth S. Zaret

Transcription factors (TFs) forage the genome to instruct cell plasticity, identity, and differentiation. These developmental processes are elicited through TF engagement with chromatin. Yet, how and which TFs can engage with chromatin and thus, nucleosomes, remains largely unexplored. Pioneer TFs are TF that display a high affinity for nucleosomes. Extensive genetic and biochemical studies on the pioneer TF FOXA, a driver of fibroblast to hepatocyte reprogramming, revealed its nucleosome binding ability and chromatin targeting lead to chromatin accessibility and subsequent cooperative binding of TFs. Similarly, a number of reprogramming TFs have been suggested to have pioneering activity due to their ability to target compact chromatin and increase accessibility and enhancer formation in vivo. But whether these factors directly interact with nucleosomes remains to be assessed. Here we test the nucleosome binding ability of the cell reprogramming TFs, Oct4, Sox2, Klf4 and cMyc, that are required for the generation of induced pluripotent stem cells. In addition, we also test neuronal and macrophage reprogramming TFs. Our study shows that reprogramming TFs bind nucleosomes with a range of nucleosome binding
affinities, indicating that although specific cocktails of TFs are required for reprogramming, mechanistically these TFs show differential nucleosome interacting behaviors. These results allowed us to assess differential features between TFs nucleosome binding ability and to correlate their binding with reprogramming potential.

To determine how general is nucleosome binding we extended our analysis to screen 593 of the 2,000 predicted human TFs in the genome for potential nucleosome binding and validated their binding in solution. Based on 3D structural analysis, we proposed that strong nucleosome binders anchor DNA through short α-helixes and have a flexible and adaptable DNA binding domain while weak nucleosome binders use β-sheets or unstructured regions and have a higher rigidity within their DNA binding domain. Through the experiments presented in this dissertation we present the first study revealing the shared structural features contributing to nucleosome binding potential of pioneer TFs and thus allow for predication of novel pioneer TFs with cell reprogramming potential.
# TABLE OF CONTENT

ACKNOWLEDGMENTS .................................................................................. III

ABSTRACT ................................................................................................. VI

LIST OF TABLES ..................................................................................... XI

LIST OF ILLUSTRATIONS .................................................................. XII

CHAPTER 1: INTRODUCTION ........................................................................ 1
  1.1 TRANSCRIPTION FACTOR MEDICATED CELL REPROGRAMMING: A LOW EFFICIENCY
      PROCESS .......................................................................................... 1
  1.2 CELL REPROGRAMMING TRANSCRIPTION FACTORS .................................. 3
  1.3 PIONEER TRANSCRIPTION FACTORS IN CELLULAR REPROGRAMMING AND
      DEVELOPMENT .................................................................................. 5
  1.4 CHROMATIN BARRIERS TO TRANSCRIPTION FACTOR DNA ACCESSIBILITY ........10
  1.5 TRANSCRIPTION FACTOR SCREENS: DNA AND CHROMATIN INTERACTING BEHAVIORS
      ........................................................................................................... 12
  1.6 CONCLUSIONS ................................................................................... 14
  1.7 FIGURES .............................................................................................. 16

CHAPTER 2: PIONEER TRANSCRIPTION FACTORS TARGET PARTIAL DNA MOTIFS ON
NUCLEOSOMES TO INITIATE REPROGRAMMING ......................................... 20
  2.1 PREFACE ............................................................................................ 20
  2.2 RESPECTIVE CONTRIBUTIONS ................................................................ 21
  2.3 ABSTRACT ........................................................................................... 22
  2.4 INTRODUCTION .................................................................................. 23
  2.5 RESULTS ............................................................................................ 26
  2.6 DISCUSSION ........................................................................................ 43
  2.7 MAIN FIGURES ................................................................................... 48
  2.8 SUPPLEMENTAL FIGURES .................................................................... 59
  2.9 SUPPLEMENTAL TABLES ..................................................................... 71
  2.10 MATERIALS AND METHODS ............................................................ 72

  Protein Expression and Purification ........................................................... 72
  Nucleosome Reconstitution ...................................................................... 73
  DNA Binding Reactions ........................................................................... 75
  DNase Footprinting .................................................................................. 77

  Western Blotting After EMSA (WEMSA) .................................................. 78
  Genomic Data Analysis ........................................................................... 80
  Motif Analysis ........................................................................................ 81
  Molecular Modeling .................................................................................. 82

  DNA Binding Sites ................................................................................... 84
  PCR Primers ........................................................................................... 84

CHAPTER 3: STRUCTURAL FEATURES OF TRANSCRIPTION FACTORS ASSOCIATING WITH
NUCLEOSOME BINDING ............................................................................. 86
  3.1 PREFACE ............................................................................................ 86
  3.2 RESPECTIVE CONTRIBUTIONS ............................................................ 87
  3.3 ABSTRACT ........................................................................................... 88
3.4 INTRODUCTION..............................................................................................89
3.5 RESULTS ....................................................................................................92
3.6 DISCUSSION .............................................................................................105
3.7 MAIN FIGURES ........................................................................................110
3.9 TABLES ....................................................................................................134
3.10 SUPPLEMENTAL TABLES .........................................................................135
3.11 KEY RESOURCES .....................................................................................136
3.11 MATERIALS AND METHODS .................................................................140
   Experimental model and subject details ....................................................140
   Genomic Data Analysis .............................................................................140
   Protein Expression and Purification ..........................................................140
   Cy5 DNA Labeling of Short Canonical-DNA Binding Sites .......................143
   Nucleosome Preparation .........................................................................144
   DNA and Nucleosome Binding Reactions ................................................145
   Western Blotting After EMSA (WEMSA) ..................................................146
   DNase Footprinting on Free DNA and Nucleosomes .................................147
   Protein Microarrays ..................................................................................148
   3D Structural Model Assemblies ................................................................149
   EMSA Processing ......................................................................................150
   Dissociation Constant Analysis ................................................................150
   Protein Microarray Analysis ......................................................................151
   Heatmap Representation of DNA and Nucleosome Binding by Transcription Factors ........................................................................................................152
   Data and Software Availability ................................................................152

CHAPTER 4: PERSPECTIVES AND FUTURE DIRECTIONS ................................153
  4.1 SUMMARY OF MAJOR CONCLUSIONS AND IMPLICATIONS ..................153
  4.2 FUTURE DIRECTIONS .............................................................................159
  4.3 FINAL REMARKS ...................................................................................165

REFERENCES ................................................................................................166

APPENDIX A: ENZYMATIC 3’ DNA LABELING WITH CY5 FLUOROPHORE ..............................187
   I. OBJECTIVE ..............................................................................................187
   II. REAGENTS AND KITS ...........................................................................187
      Reagents ..................................................................................................187
      DNA Sequences ......................................................................................188
      Buffers .....................................................................................................188
   III. PROTOCOL ...........................................................................................188

APPENDIX B: WIDOM 601 DNA SEQUENCE AS A TEMPLATE FOR TF-NUCLEOSOME INTERACTIONS .............................................................................................................192
   I. OBJECTIVE ..............................................................................................192
   II. REAGENTS ............................................................................................192
      DNA Sequences .......................................................................................193
      Purified Proteins .....................................................................................194
   III. ASSESSMENT OF FOXA1-SPECIFIC BINDING AT MULTIPLE ROTATIONAL POSITIONS ON 601 NUCLEOSOMES .................................................................194
# LIST OF TABLES

## CHAPTER 2: PIONEER TRANSCRIPTION FACTORS

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Recombinant O, S, K, and M Show a Range of Affinities to Nucleosomes</td>
<td>71</td>
</tr>
</tbody>
</table>

## CHAPTER 3: STRUCTURAL FEATURES OF TRANSCRIPTION FACTORS ASSOCIATING WITH NUCLEOSOME BINDING

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Recombinant TF Dissociation Constants</td>
<td>134</td>
</tr>
<tr>
<td>3.S1</td>
<td>Fortuitous Motif Occurrence in DNA Sequences by Motif Scan</td>
<td>135</td>
</tr>
<tr>
<td>3.S2</td>
<td>Protein Microarray Data</td>
<td>135</td>
</tr>
<tr>
<td>3.S3</td>
<td>Short DNA oligonucleotides containing TFs binding sites used as unlabeled competitors</td>
<td>135</td>
</tr>
</tbody>
</table>
**LIST OF ILLUSTRATIONS**

<table>
<thead>
<tr>
<th>Figure</th>
<th>CHAPTER 1: INTRODUCTION</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Transcription Factor Mediated Cell Reprogramming</td>
<td>16</td>
</tr>
<tr>
<td>1.2</td>
<td>Mechanisms of Chromatin Organization</td>
<td>17</td>
</tr>
<tr>
<td>1.3</td>
<td>Chromatin Binding by Pioneer Transcription Factors</td>
<td>18</td>
</tr>
<tr>
<td>1.4</td>
<td>Surface topology and charge of the nucleosome core particle</td>
<td>19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Figure</th>
<th>CHAPTER 2: PIONEER TRANSCRIPTION FACTORS TARGET PARTIAL DNA MOTIFS ON NUCLEOSOMES TO INITIATE REPROGRAMMING</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Oct4, Sox2, Klf4, and c-Myc Display Differential Affinity to Nucleosomes In Vitro</td>
<td>48</td>
</tr>
<tr>
<td>2.2</td>
<td>The Contribution of Non-Specific Binding to Nucleosome Targeting In Vitro</td>
<td>50</td>
</tr>
<tr>
<td>2.3</td>
<td>Oct4, Sox2, Klf4, and c-Myc Display a Range of Nucleosome Targeting In Vivo</td>
<td>52</td>
</tr>
<tr>
<td>2.4</td>
<td>Oct4, Sox2, and Klf4 Recognize Partial Motifs on Nucleosomes</td>
<td>54</td>
</tr>
<tr>
<td>2.5</td>
<td>c-Myc Recognition of Degenerate E-Box on Nucleosome Is Assisted by Binding with Co-Factors</td>
<td>56</td>
</tr>
<tr>
<td>2.6</td>
<td>The Folding Extent of bHLH Basic Helix-1 on DNA Anti-Correlates with Targeting Centrally Degenerate E-Box Motifs on Nucleosomes</td>
<td>57</td>
</tr>
<tr>
<td>2.S1</td>
<td>Recombinant Oct4, Sox2, Klf4, and c-Myc Show Specific DNA-Binding Activities In Vitro</td>
<td>59</td>
</tr>
<tr>
<td>2.S2</td>
<td>Recombinant Oct4, Sox2, Klf4, and c-Myc Show a Range of Specificity to Free DNA versus Nucleosomal DNA In Vitro</td>
<td>61</td>
</tr>
<tr>
<td>2.S3</td>
<td>Oct4, Sox2, Klf4, and c-Myc Show a Range of Affinity and</td>
<td>63</td>
</tr>
</tbody>
</table>
### Specificity to Nucleosomes In Vivo

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.S4</td>
<td>Oct4, Sox2, Klf4, and c-Myc Bind to Nucleosome Containing H2A and H3 Histones</td>
<td>65</td>
</tr>
<tr>
<td>2.S5</td>
<td>The Apparent Flexibility of Oct4, Sox2, Klf4, and c-Myc DBDs Correlate with Their Nucleosome Binding Compatibility</td>
<td>66</td>
</tr>
<tr>
<td>2.S6</td>
<td>The HLH Factors Compatibility with Nucleosomes Correlates with Central Degenerate E-Box Motif</td>
<td>68</td>
</tr>
<tr>
<td>2.S7</td>
<td>Nucleosome Binding Compatibility of Bipartite DBDs</td>
<td>70</td>
</tr>
</tbody>
</table>

**Figure**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Endogenous TF-Nucleosomal Targets Assemble into Stable Nucleosomes In Vitro</td>
<td>110</td>
</tr>
<tr>
<td>3.2</td>
<td>Reprogramming TFs bind Nucleosomes with Nanomolar Affinity</td>
<td>112</td>
</tr>
<tr>
<td>3.3</td>
<td>TFs Bind Nucleosomes with Specificity</td>
<td>114</td>
</tr>
<tr>
<td>3.4</td>
<td>ZnF TF Zelda Bind to Nucleosomes</td>
<td>116</td>
</tr>
<tr>
<td>3.5</td>
<td>Systematic Assessment of Nucleosome Binding of Human TFs with Protein Microarrays</td>
<td>118</td>
</tr>
<tr>
<td>3.6</td>
<td>Strong Nucleosome Binding TFs Recognize DNA with Short Recognition α-Helixes</td>
<td>120</td>
</tr>
<tr>
<td>3.S1</td>
<td>Reprogramming TFs Target MNase Enriched Target Sites Associated with cell identity Genes</td>
<td>122</td>
</tr>
<tr>
<td>3.S2</td>
<td>Recombinant Purified Full-Length TFs Bind Their Canonical Motifs on Short dsDNA</td>
<td>124</td>
</tr>
<tr>
<td>3.S3</td>
<td>Recombinant Purified Full-Length TFs Bind with Specificity to DNA and Nucleosomes</td>
<td>126</td>
</tr>
<tr>
<td>3.S4</td>
<td>Pipeline for Protein Microarray Data Processing</td>
<td>128</td>
</tr>
<tr>
<td>3.S5</td>
<td>3D Representation of DNA Sequences Showing Positioning of TFs Binding Motifs of Tested Factors</td>
<td>130</td>
</tr>
</tbody>
</table>
Figure APPENDIX B: WIDOM 601 DNA SEQUENCE AS A TEMPLATE FOR TF-NUCLEOSOME INTERACTIONS

B1 Design of 601-eG nucleosomes 197

B2 FOXA1 binds with specificity to 601–DNA containing the eG site at three different rotational positions 198

B3 Mutagenesis of 601 H3/4 TA steps does not enhance FOXA1 nucleosome binding 199

Figure APPENDIX C: FOXA1 MUTANTS BINDING TO NUCLEOSOMES

C1 FOXA1 WT and Mutants Bind Short dsDNA with Specificity 206

C2 Mutation of FOXA1 K270 or C-term αHelix Does not Inhibit Nucleosome Binding 210

C3 Recombinant Purified Full-Length FOXA1 WT and Mutants Bind Free DNA and Nucleosomes with Similar Nucleosome Binding Affinities 211

C4 FOXA1 Residue K270 Contributes to FOXA1 Histone Octamer Interaction 212
CHAPTER 1: INTRODUCTION

1.1 Transcription Factor Medicated Cell Reprogramming: A Low Efficiency Process

During development, cells commit to a resolved identity that dictates their specialized functions. A stable and specialized cellular identity is determined by the epigenetic and transcriptional states and the genome configuration that a cell acquires during differentiation stages. Differentiation stages are orchestrated by the hierarchical expression of transcription factors (TFs), key proteins that decode the information in our DNA to regulate expression and repression of genes. Pioneering work by John Gurdon first demonstrated that when Xenopus nuclei from a specialized cell are transferred into an enucleated oocyte it results in the full development of a frog. These studies showed that specialized cells still contain the developmental capacity to drive Xenopus oocytes, through embryonic development (Gurdon, 1962). Later work by Shinya Yamanaka demonstrated that the identity of a specialized cell can be reverted back to a pluripotency state by forced expression of specific TFs, a process known as cell reprogramming, which results in induced pluripotent stem cells (Figure 1.1) (Nakagawa et al., 2008; Takahashi and Yamanaka, 2006). These breakthrough experiments revolutionized the field of stem cell biology and defined TFs as the master regulators of cell identity conversion and plasticity.
Stem cells and *in vitro* reprogrammed cells have greatly advanced the field of regenerative medicine, as these cells may be used for cell replacement therapies in cases of tissue damage. For example, heart regeneration may be achieved by induced cardiomyocyte replacement, and liver damaged can be repaired by transplantation of induced hepatocytes (Huang et al., 2011; Ieda, 2013; Park et al., 2019; Song et al., 2012). In addition, reprogrammed cells provide the unique opportunity for disease-modeling with patient-specific cells in cases where the relevant cell types are otherwise unattainable, such as neurodegenerative disorders of the brain (Wang et al., 2019).

Over the last decade much work has been done to understand the pathways underlying cell reprogramming of somatic cells to pluripotency. Integration of population and single cell genome-wide technologies have allowed the description of a comprehensive road map of cells progression through fate change. These studies included the identity of surface markers crucial for iPSC isolation, single cell transcriptome analysis, chromatin landscape transitions, 3D chromatin organization, and TF dynamics during reprogramming (Beagan et al., 2016; Buganim et al., 2012a; Cacchiarelli et al., 2015; Chronis et al., 2017; Hansson et al., 2012; Hussein et al., 2014; Li et al., 2017; Paik et al., 2018; Polo et al., 2012; Soufi et al., 2012; White et al., 2016). Still, generation of mature reprogrammed cell populations remains a problematic, highly variable, and inefficient process with reported reprogramming efficiencies ranging from 0.00002 – 1 % (Malik and Rao, 2013; Omole and Fakoya, 2018). Therefore,
considerable barriers need to be overcome in order to facilitate clinical translation of cell reprogramming technologies. These challenges stress the need to better understand the unique characteristics that allow reprogramming TFs to induce such drastic changes in cell fate.

1.2 Cell Reprogramming Transcription Factors

MyoD, a skeletal muscle-specific TF required for myogenesis, was the first reported TF to induce direct cell reprogramming of fibroblasts to myoblasts, a process where cells undergo reprogramming without transitioning through a pluripotent state (Davis et al., 1987). Two decades later, Oct4, Sox2, Klf4 and cMyc (OSKM) were identified as the embryonic stem cell (ESCs)-specific TFs cocktail, from a 24 gene pool, required for the reprogramming of fibroblasts to induced pluripotent stem cells (Figure 1.1) (Takahashi and Yamanaka, 2006). These cells had ESC-like phenotype and were therefore named induced pluripotent stem cells (iPSCs).

These early examples of TF-mediated reprogramming triggered numerous screens to identify cell lineage-specific TFs that could induce direct cell reprogramming of somatic cells into distantly related cell types from all three germ layers (Figure 1.1) (Huang et al., 2011; Laiosa et al., 2006; Song et al., 2012; Vierbuchen et al., 2010). These screens used cDNA pools of cell type-specific TFs followed by single TF cDNA subtraction to narrow down genes required for direct cell reprogramming. Results from these experiments revealed
that only a subset of the tested TFs were capable of inducing direct cell reprogramming from fibroblasts, including: PU1, C/EBPα, and GATA1 for hematopoietic reprogramming (Feng et al., 2008; Laiosa et al., 2006); Foxa, Gata4, and Hnf1α for hepatocyte reprogramming (Figure 1.1) (Huang et al., 2011; Huang et al., 2014; Sekiya and Suzuki, 2011; Yu et al., 2013); and Ascl1, Myt1l, and Brn2 for neuronal cell reprogramming (Vierbuchen et al., 2010). These experiments revealed the existence of lineage-specific TF combinations capable of erasing mature cell identity and driving acquisition of a new identity. However, how these reprogramming TFs initially interact with repressed chromatin to reactivate a developmentally silenced program was not understood.

Pervasiveness of a select group of TFs required for cell reprogramming opened the question as to what makes these TFs unique and what features enable TFs with the ability to reprogram cells. It has been suggested that within this subset of factors there is a hierarchy of reprogramming dominance. Although multiple studies have identified alternative TF combinations to OSKM for iPSCs reprogramming, Oct4 and Sox2 remained constant within most TFs combinations (Feng et al., 2009; Nakagawa et al., 2008; Yu et al., 2007). Furthermore, expression of Oct4 alone is sufficient to generate iPSCs from mouse neuronal stem cells (Kim et al., 2009). In contrast, Klf4 can be replaced by Esrrb while cMyc can be omitted for reprogramming and is thus dispensable for this process (Feng et al., 2009; Nakagawa et al., 2008; Yu et al., 2007). However, in most circumstances S, K, and M are also needed for efficient reprogramming to
iPSCs (Takahashi and Yamanaka, 2006).

In neuronal cell reprogramming Ascl1 is dominant over Myt1l and Brn2 (Wapinski et al., 2013). Similar to Oct4, expression of Ascl1 alone in fibroblast is sufficient to drive neuronal cell reprogramming (Wapinski et al., 2013). Still, Brn2 and Myt1l were required to enhance reprogramming efficiency. Additionally, in hepatocyte reprogramming from fibroblasts, three different studies identified TFs that could promote hepatocyte conversion, but among them only FoxA was a common TF (Huang et al., 2014; Sekiya and Suzuki, 2011; Yu et al., 2013). Removal of FoxA2 or FOXA3 from hepatocyte reprogramming protocols inhibited the formation of epithelial colonies, but expression of Hnf4α was not required. These findings support the hypothesis that within a subset of reprogramming factors, there is a hierarchy in reprogramming ability. Early work on one of the reprogramming TFs FOXA, provided clues about the molecular features of strong reprogramming TFs and hierarchy of reprogramming ability.

1.3 Pioneer Transcription Factors in Cellular Reprogramming and Development

Dominant reprogramming TFs play a crucial role in cell development. FoxA, the dominant reprogramming factor required for fibroblast to hepatocyte conversion (Du et al., 2014; Huang et al., 2011; Huang et al., 2014; Yu et al., 2013) is also required for hepatic specification of the gut endoderm (Gualdi et al., 1996). During liver development, FoxA directly accesses the albumin enhancer
when it is packaged in nucleosomal DNA within transcriptionally silent, closed chromatin. This binding promotes accessibility of the enhancer and subsequent activation of *albumin*, a highly liver-specific gene (Chaya et al., 2001; Cirillo et al., 2002; Cirillo and Zaret, 1999a; Gualdi et al., 1996). *In vitro*, FoxA can remodel the compacted *albumin* enhancer into a local open state by displacing linker histone in the absence of ATP-chromatin remodelers (Cirillo et al., 2002). Unlike FoxA, ATP-dependent chromatin remodelers are sensitive to linker histone-bound nucleosomes (Hill and Imbalzano, 2000; Ramachandran et al., 2003). FoxA1-mediated DNA accessibility, as determined by an increase in DNAse hypersensitivity, leads to cooperative binding of subsequent TFs (Figure 2) (Cirillo et al., 2002; Cirillo et al., 1998; Cirillo and Zaret, 1999a; Iwafuchi-Doi et al., 2016). Therefore, unlike most TFs, FoxA binding to DNA is not inhibited by nucleosome structure. These distinguishing features of FoxA led to its identification as a “pioneer factor”– a novel class of developmental regulators with nucleosome binding activity and ATP-independent local chromatin remodeling activity (Figure 1.2).

As cells undergo mitosis, chromosomes condense and transcriptional levels decrease (Naumova et al., 2013; Palozola et al., 2017). Still, FoxA is retained in mitotic chromatin due to its high nucleosome binding affinity. Furthermore, genomic sites that retain FoxA binding through mitosis correlate with higher than predicted nucleosome occupancy (Caravaca et al., 2013; Kadauke et al., 2012; Sekiya et al., 2009). Therefore, FoxA1 was shown to play a
role in the maintenance of cellular memory through mitosis. This work highlighted the importance of TFs interactions with silenced chromatin for the initiation of chromatin accessibility and progression through cell differentiation and reprogramming.

Similar to FoxA1, dominant reprogramming TFs have been suggested to have pioneering activity through their engagement with target sites within compacted chromatin and the resulting chromatin changes that initiate cell reprogramming. Work from the Zaret lab used genome-wide nucleosome occupancy data to reveal that the iPSC TFs OSK, but not cMyc, co-occupy closed chromatin regions that contain key pluripotency genes only 48 hr into reprogramming (Soufi et al., 2012). Furthermore, cooperative binding of OSK is crucial for the initiation of proper enhancer selection and thus for fibroblast identify silencing. However, the addition of cMYC strongly biased OSKM to promoters poised for activation (Chronis et al., 2017).

Genome-wide mapping of Ascl1, Brn2, and Myt1L during neuronal reprogramming showed that Ascl1 preferentially localizes to closed chromatin and is targeted to neuronal-specific enhancers (Wapinski et al., 2013). In contrast, Brn2 and Myt1L cannot access closed chromatin, preferentially target open chromatin and thus have a less dominate role in at the initial stages of neuronal reprogramming (Wapinski et al., 2013). Expression of Ascl1 alone followed by single-cell transcriptome analysis in fibroblast showed that Ascl1 is sufficient to robustly induce reactivation of neuronal genes, downregulation of cell
cycle genes, and initiate neuronal reprogramming (Chanda et al., 2014; Treutlein et al., 2016). In contrast, Myt1L localizes to open and accessible promoters to execute transcriptional repression and downregulate non-neuronal genes (Mall et al., 2017). These findings show that neuronal reprogramming is dependent on the precise match between dominant TFs and chromatin context for the onset of neuronal identity acquisition followed by the repression of alternative identities.

Hematopoietic TFs PU1, CEBPα and CEBPβ have been demonstrated to reprogram fibroblast to macrophage like cells with PU1, CEBPα and CEBPβ suggested to act as pioneer factors in different contexts; PU1 in macrophage conversion, CEBPα in B Cell conversion and CEBPβ in adipose cells (Feng et al., 2008; Heinz et al., 2010; Pundhir et al., 2018; Siersbaek et al., 2011; Tagore et al., 2015; van Oevelen et al., 2015). PU.1 and CEBPα TFs suggested to have pioneering roles during B cell to macrophage reprogramming by distinct mechanisms of chromatin targeting. In vivo genomic and chromatin state analysis by MNase-seq of PU.1 and CEBPα during B cells to macrophages reprogramming revealed that PU1 target nucleosome-enriched chromatin and induces chromatin accessibility, while CEBPα binding to nucleosome-enriched regions is strongly dependent on PU.1 and EBF1 (Barozzi et al., 2014; Boller et al., 2016; Li et al., 2018; van Oevelen et al., 2015). EBF1 has also been shown to bind compacted chromatin and induce lineage-specific chromatin accessibility (Boller et al., 2016; Li et al., 2018). Moreover, knock down of PU.1 in B cells compromises CEBPα binding to nucleosome-enriched enhancers (Heinz et al.,
2010; Tagore et al., 2015). This leaves the open question of whether CEBPα truly targets nucleosomes and thus acts as a pioneer.

While significant research has been done to parse the epigenetic mechanism of TFs driving direct cell reprogramming toward neuronal and hematopoietic cell types (Vignoles et al., 2019), the epigenetic mechanism of direct cell reprogramming toward cardiac, pacemaker cells, keratinocytes suggesting novel pioneer TFs such as Isl1, Hand2, and TFAP2C/p63 respectively are just beginning to be explored (Fernandez-Perez et al., 2019; Gao et al., 2019; Li et al., 2019). Moreover, direct cellular reprogramming toward peripheral red bloods cells remains difficult and TFs cocktails have not yet been identified. Therefore, characterization of the unique features of pioneer transcription factors will allow the prediction of TFs with pioneering and reprogramming potential.

Studies correlating TF binding co-occurrence with chromatin accessibility in vivo suggest a hierarchical model of TF-mediate reprogramming. In this model, factors sufficient for reprogramming access compacted chromatin, leading to the recruitment of additional factors. Finally, supporting TFs target open chromatin and repress the host identity (Mall et al., 2017; Wapinski et al., 2013). The hierarchical model place TF nucleosome binding ability as a determinant of reprogramming potential. Nonetheless models based on genetic correlations alone are insufficient to distinguish between intrinsic nucleosome targeting ability and chromatin engagement through cooperative interactions. Moreover, whether
these factors have the inherent biochemical ability to bind nucleosomes remains to be assessed. Thus, the requirement for nucleosome binding by these factors as a seminal event in the initiation steps of reprogramming has not been determined.

1.4 Chromatin Barriers to Transcription Factor DNA Accessibility

TFs work in the context of chromatin where DNA is hierarchically packaged into a 3D genome conformation that controls the diverse regulatory mechanisms of gene activation and repression (Figure 1.3) (Luger, 2003; Luger et al., 1997). Regulation of the 3D genome organization and its dynamics across time and space results in a 4D nucleome network that controls cellular differentiation into specialized cells (Figure 1.1) (Dekker et al., 2017; Mirny et al., 2019; Zheng and Xie, 2019).

Nucleosomes are the principal packing elements of DNA and therefore, the fundamental determinants of DNA accessibility for essentially all DNA-templated processes. They are formed by the association of 146 bp of DNA wrapped around two histone dimers of H2A-H2B and a H3-H4 tetramer (Figure 1.4) (Luger, 2003; Luger et al., 1997). The histone octamer forms histone-DNA contacts at the interior minor groove of DNA, serving as the first level of DNA constraint modulating the binding of TFs to DNA and chromatin regulators (Luger, 2003; McGinty and Tan, 2015). The tight packing of DNA by histones impedes accessibility of TFs to DNA and inhibits TF transcriptional activity.
(Blomquist et al., 1996; Schild-Poulter et al., 1996; Taylor et al., 1991). Moreover, rotational phasing of the TF DNA binding sites within the nucleosome has been shown to limit TF binding (Liu and Kraus, 2017; Sekiya et al., 2009).

As described above, pioneer TFs can overcome such constraints. Nucleosomes can also allow binding of diverse proteins such as pioneer TFs, epigenetic modifiers and chromatin remodelers that alter the nucleosome structure, positioning and compaction (Hughes and Rando, 2014; McGinty and Tan, 2015). The nucleosome's acidic patch, a negatively charged region on the nucleosome surface, serves as the landing pad for nucleosome binding proteins including regulator of chromosome condensation RCC1, polycomb repressive complex PRC1, centromere protein, CENPC, among others (Figure 1.4) (Allu et al., 2019; Kato et al., 2013; Makde et al., 2010; McGinty et al., 2014; Wang et al., 2013). However, these nucleosome binding proteins interact with chromatin through non-specific interactions, which does not allow the targeting to cell type specific genes. On the other hand, TFs recognize unique sequences within DNA. Therefore, pioneer TFs play a crucial role in the gene specific targeting and recruitment of chromatin regulators and allow access of nucleosome-inhibited binding of TFs to target sites in closed chromatin to direct competence for a specific cell lineage.

TFs display sensitivity to the multiple states of chromatin compaction where nucleosome presence is inhibitory for the binding of most TFs. Our lab has identified heterochromatic domains which are refractory to OSKM binding during
iPSCs reprogramming (Soufi et al., 2012). Differentially bound regions (DBRs) were initially described as histone 3 lysine 9 trimethylation (H3K9me3)-enriched mega-base scale domains (Figure 1.3) (Soufi et al., 2012). These DBRs contain pluripotency genes with restricted expression to few cells in the late stages of iPSCs reprogramming stages (Buganim et al., 2012b). More recently our lab showed that H3K9me3 heterochromatin domains are biophysically distinct domains that are sonication resistant therefore revealing structural subtypes of heterochromatin (Becker et al., 2017). Additionally, liver specific genes located with these sonication resistant domain failed to activate during direct cell reprogramming of fibroblast to hepatocytes therefore, impeding efficient reprogramming (Becker et al., 2017). These studies, highlighted the role of TFs chromatin state sensitivity in restricting manipulation of cell fate. Continued research on the impediment of TF chromatin restriction will allow the design of better reprogramming protocols and facilitate cell reprogramming translation for therapeutic technologies.

1.5 Transcription Factor Screens: DNA and Chromatin Interacting Behaviors

Proteins that recognize the nucleosome as a substrate cannot be reliably predicted by genomic analysis and while there are over 2,000 predicted human TFs in the genome (Kummerfeld and Teichmann, 2006; Lander ES, 2001; Messina et al., 2004; Vaquerizas et al., 2009; Venter JC, 2001), only a minority of
sequence-specific DNA chromatin binding proteins have been assayed for nucleosome interaction (Blomquist et al., 1996; Haswell and O'Shea, 1999; Liu and Kraus, 2017; Perlmann, 1988; Schild-Poulter et al., 1996).

Multiple high throughput studies have described the profile of TFs sequence-specific DNA interactome and their sensitivity for DNA methylation by systematic evolution of ligands by exponential enrichment (SELEX) and ChIP-seq (Hu et al., 2009; Jolma et al., 2013; Yin et al., 2017). More recently the development of combinatorial technologies that allow parallel analysis of DNA accessibility and gene expression have allowed the discovery of associations between chromatin state, TF motif enrichment and gene expression (Cao et al., 2018; Clark et al., 2018; Cusanovich et al., 2015). But these technologies had been unable to incorporate TF-chromatin interaction profiles into the analysis due to limitations in the high number of cells required for these assays and the requisite of DNA-sequence specificity of TFs. These constraints limit the identification of TFs with nucleosome binding potential and make it challenging for computational tools to predict the general rules underlying nucleosome binding and pioneer activity.

To elucidate the molecular mechanisms that mediate TF access to nucleosomal sites and the nucleosome-interacting behaviors of such TFs, two recent publications developed novel approaches (Yan et al., 2018; Zhu et al., 2018). Yan et. al. systematically compared the properties of yeast nucleosome-displacing factors (NDFs). Similar to pioneer TFs, NDFs access embedded sites
within compacted DNA, but unlike pioneer TFs, NDFs induce depletion of nucleosomes (Yan et al., 2018). This study identified TFs with strong and weak nucleosome displacing activity. Furthermore, this study showed that strong NDFs antagonize nucleosome formation through DNA replication dependent mechanisms in yeast (Yan et al., 2018).

Research by Zhu et al from Taipale’s group developed nucleosome consecutive affinity purification –SELEX to test nucleosome binding of over 200 TFs. CAP-SELEX consists of pull down assays with recombinant nucleosomes assembled with 147 bp and 200 bp DNA libraries followed by analysis of TF sequence-specific enrichment (Zhu et al., 2018). In agreement with previous findings, nucleosomes inhibited the binding of most TFs but also revealed position-specific binding such as, periodic and dyad binding to nucleosomal DNA. In conclusion these publications revealed the interaction landscape between TFs and the nucleosome (Yan et al., 2018; Zhu et al., 2018). Nonetheless the molecular features of TFs that permit or preclude binding to nucleosomal DNA and the chromatin perturbations that immediately succeed binding, to promote cooperative interactions with other factors, remain to be understood.

1.6 Conclusions

Based on the shared occupancy at silent chromatin of FoxA and reprogramming TFs, I sought to investigate the nucleosome binding abilities of reprogramming TF as a required step for the initiation of reprogramming.
Although it has been proposed that TF-mediated chromatin opening is a generalized mechanism to access regulatory regions as an initial step in reprogramming and differentiation, these models have been based on genome-wide ChIP-seq data, which cannot differentiate between direct interaction with chromatin or close proximity mediated by intermediate proteins. Therefore, my thesis work tests the ability of reprogramming factors to directly engage with nucleosomes (chapter 2) and systematically assess the ability of the direct reprogramming TFs toward multiple cell types to bind reconstituted nucleosomes (chapter 3). Additionally, by TF screening methods I was able to identify TFs with previously unreported pioneering activity (Chapter 3). Indeed, revealing that nucleosome binding is a characteristic of strong reprogramming TFs and further predicting TFs for unreported reprogramming protocols such as erythrocyte direct cell reprogramming (Figure 1.1). My thesis work describes the unique structural features of reprogramming TFs and provides insights into the initial step in the mechanisms of cell reprogramming.
Transcription factors (TFs) that mediate reprogramming of fibroblast toward induced pluripotent stem cells or the direct cell reprogramming towards induced neurons, induced hepatocytes, induced macrophages and other hematopoietic lineages.

**Figure 1.1 | Transcription Factor Mediated Cell Reprogramming**

Transcription factors (TFs) that mediate reprogramming of fibroblast toward induced pluripotent stem cells or the direct cell reprogramming towards induced neurons, induced hepatocytes, induced macrophages and other hematopoietic lineages.
Figure 1.2 | Mechanisms of Chromatin Organization

DNA is packaged into nucleosomes by histone octamers composed of H2A, H2B, H3 and H4. These are positively-charged and adhere strongly to negatively-charged DNA. Nucleosomes then fold into heterogeneous groups of nucleosomes clutches (Ricci et al., 2015). These clutches are compressed and folded in interphase into loops by architectural proteins like CTCF and cohesion. DNA loops are then arranged into compartments by transcriptional state into heterochromatin and euchromatin (Mirny et al., 2019). DNA compartments and DNA loops are then studied through time to build spatial models of the dynamic spatiotemporal regulation of chromatin (Dekker et al., 2017). Figure adapted from (Dekker et al., 2017; Mirny et al., 2019).
Before gene activation DNA is highly compacted into low signal or repressed heterochromatin inhibiting binding of most TFs. Pioneer TFs scan low signal chromatin and access target sites within silent chromatin but this binding is inhibited at differentially bound regions (DBRs) enriched for H3K9me3. Upon binding pioneer TFs then induce local chromatin accessibility and through cooperative interactions allow subsequent binding of TFs and chromatin modifiers to induce cell type specific gene activation.

**Figure 1.3 | Chromatin Binding by Pioneer Transcription Factors**
Before gene activation DNA is highly compacted into low signal or repressed heterochromatin inhibiting binding of most TFs. Pioneer TFs scan low signal chromatin and access target sites within silent chromatin but this binding is inhibited at differentially bound regions (DBRs) enriched for H3K9me3. Upon binding pioneer TFs then induce local chromatin accessibility and through cooperative interactions allow subsequent binding of TFs and chromatin modifiers to induce cell type specific gene activation.
Figure 1.4 | Surface topology and charge of the nucleosome core particle.
(A) Surface of nucleosome core particle viewed down the DNA superhelical axis in space-filling representation.
(B) Surface electrostatic potential of nucleosome core particle contoured from $-5$ to $+5$ kT/e calculated with ABPS.164 Location of acidic patch is indicated (Figure from (McGinty and Tan, 2015))
CHAPTER 2: PIONEER TRANSCRIPTION FACTORS TARGET
PARTIAL DNA MOTIFS ON NUCLEOSOMES TO INITIATE
REPROGRAMMING

2.1 Preface

The manuscript presented in this chapter was originally published online
16 April 2015 in Cell, Volume 161 (Soufi et al., 2015b). It has been reformatted
here in accordance with University of Pennsylvania dissertation formatting
guidelines.

Authors: Abdenour Soufi¹,³, Meilin Fernandez Garcia¹, Artur Jaroszewicz²,
Nebiyu Osman¹, Matteo Pellegrini², Kenneth S.Zaret¹,³,*

Affiliations:
¹Institute for Regenerative Medicine and Epigenetics Program, Department of
Cell and Developmental Biology, University of Pennsylvania Perelman School of
Medicine, Smilow Center for Translational Research, Building 421, 3400 Civic
Center Boulevard, Philadelphia, PA 19104-5157, USA
²Department of Molecular, Cell and Developmental Biology, UCLA, Box 951606,
Los Angeles, CA 90095-1606, USA
³Lead Contact
*Correspondence: zaret@pennmedicine.upenn.edu
2.2 Respective Contributions

Dr. Abdenour Soufi and Dr. Kenneth S. Zaret. conceived the study. Under the guidance of Dr. Soufi and Dr. Zaret I carried out plasmid cloning, DNA purifications, histones purifications, nucleosome reconstitutions, EMSAs and competition experiments as well as dissociation constant analyses resulting in Figure 2.1D, Table 1, Figure 2.2A and supplemental Figures 2.S1, 2.S2 presented. Dr. Soufi, Arthur Jaroszewicz, and Dr. Mateo Pellegrini performed genome sequencing experiments and analysis including MNase-seq data analysis. With the guidance of Dr. Soufi, Nebiyu Osman carried out the recombinant protein purifications of transcription factors. Dr. Soufi performed DNAse footprinting experiments with samples that I prepared. Dr. Soufi also carried out the motif and 3D structure analysis. Dr. Soufi and Zaret contributed to supervision of personnel, data interpretation, and writing the manuscript. While I contributed to the manuscript revision and discussions that led to the final proposed model of transcription factors flexibility within the DNA binding domain.
2.3 Abstract

Pioneer transcription factors (TFs) access silent chromatin and initiate cell-fate changes, using diverse types of DNA binding domains (DBDs). FoxA, the paradigm pioneer TF, has a winged helix DBD that resembles linker histone and thereby binds its target sites on nucleosomes and in compacted chromatin. Herein, we compare the nucleosome and chromatin targeting activities of Oct4 (POU DBD), Sox2 (HMG box DBD), Klf4 (zinc finger DBD), and c-Myc (bHLH DBD), which together reprogram somatic cells to pluripotency. Purified Oct4, Sox2, and Klf4 proteins can bind nucleosomes in vitro, and in vivo they preferentially target silent sites enriched for nucleosomes. Pioneer activity relates simply to the ability of a given DBD to target partial motifs displayed on the nucleosome surface. Such partial motif recognition can occur by coordinate binding between factors. Our findings provide insight into how pioneer factors can target naive chromatin sites.
2.4 Introduction

Silent chromatin is packed with nucleosomes, acting as a barrier to targeting by most transcription factors (TFs) (Adams and Workman, 1995; Mirny, 2010). However, a select group of transcription factors (TFs) known as pioneer factors have the combined ability to access their target sites in silent chromatin and initiate cell-fate changes (Iwafuchi-Doi and Zaret, 2014; Zaret and Carroll, 2011). The winged-helix DNA binding domain (DBD) of the pioneer factor FoxA (Clark, 1993), which is similar to that of linker histone (Ramakrishnan et al., 1993), allows the protein to bind its DNA motif exposed on a nucleosome and access to silent chromatin (Cirillo et al., 2002; Cirillo et al., 1998; Cirillo and Zaret, 1999a). Such activity is necessary for liver induction (Lee et al., 2005). Other TFs involved in cell reprogramming can target their sites in silent chromatin (Montserrat et al., 2013; Soufi et al., 2012; Takahashi and Yamanaka, 2006; Wapinski et al., 2013), but they possess DBDs that differ from that of FoxA. Whether such reprogramming factors directly bind nucleosomes and how the structures of their respective DBDs relate to nucleosome binding, and hence pioneer activity, has not been assessed.

Transcription factors containing major structural classes of DBDs, including Pit-Oct-Unc (POU), Sry-related High Mobility Group (HMG), Zinc Fingers (ZF), and basic-helix-loop-helix (bHLH), represented by O, S, K, and M, respectively, have been used in the most dramatic example of cellular reprogramming: the conversion of differentiated cells into induced pluripotent
stem cells (Takahashi and Yamanaka, 2006). We previously compared genomic chromatin features of human fibroblasts, prior to the ectopic expression of OSKM, to where the factors first bind the genome during their initial expression (Soufi et al., 2012). This allowed us to assess how OSKM target pre-existing states in chromatin, as opposed to assessing chromatin states after the factors are bound. The data showed that Oct4, Sox2, and Klf4, but not c-Myc, could function as pioneers during reprogramming by virtue of their ability to mostly target “closed” chromatin sites that are DNase I resistant and “naive” by virtue of lacking evident active histone modifications (Soufi et al., 2012). Recently, single-molecule imaging analysis using fluorescently tagged proteins monitored in living cells proposed that Sox2 guides Oct4 to its target sites (Chen et al., 2014); the chromatin status of the sites was unknown. However, we previously found that the ectopic Oct4 and Sox2 bind most extensively to separate sites in chromatin (Soufi et al., 2012), leaving open how the bulk of chromatin targeting is achieved. While many of initial binding events were promiscuous and not retained in pluripotent cells, many others occurred at target genes that are required for conversion to pluripotency.

Ascl1, Pax7, and Pu.1 have emerged as pioneer transcription factors based on targeting closed chromatin and their ability to reprogram cells, though assessments of direct interaction with nucleosomes has been lacking (Barozzi et al., 2014; Budry et al., 2012; Wapinski et al., 2013). In light of the bHLH factor c-Myc being unable to bind closed chromatin on its own (Soufi et al., 2012), it was
surprising that Ascl1, another bHLH factor, can bind closed chromatin during reprogramming fibroblasts to neuron-like cells (Wapinski et al., 2013). Studies that have examined the correlation between co-existing TF binding and nucleosome occupancy, without characterizing the “pre-bound” chromatin state, could not address questions about initial chromatin access.

Generating induced pluripotent stem (iPS) cells, using the OSKM factors, has proved to be highly valuable for research, with great potential for regenerative medicine (Robinton and Daley, 2012). In an attempt to increase the efficiency of reprogramming, efforts have focused on explaining how somatic cells respond to the ectopic expression of OSKM (Buganim et al., 2012b; Papp and Plath, 2013; Soufi, 2014). To gain insights into the molecular mechanisms that impart OSKM access to closed chromatin, we measured the fundamental interaction between the factors and nucleosomes, in vivo and in vitro, by three mutually supportive approaches: biochemical assays, genomics, and structural analysis. We find that the inherent ability of DBDs to recognize one face of DNA on nucleosome, as seen by targeting a part of their canonical motif on nucleosome-enriched sequences in chromatin, is the primary determinant of pioneer factor activity. These findings can explain the pioneer activity of a diverse set of reprogramming factors containing different structural classes of DBDs as well as the synergistic behavior of pioneer and non-pioneer factors.
2.5 Results

**Oct4, Sox2, Klf4, and c-Myc Show a Range of Nucleosome Binding In Vitro**

The interaction of full-length O, S, K, and M, as used in reprogramming, with nucleosomes is not known. Therefore, we purified and refolded the full-length O, S, and K factors, along with c-Myc and its obligate heterodimerization partner Max from bacterial cells, representing post-translational unmodified proteins (Figure 2.1A; Figure 2.S1A). We also obtained the full-length O, S, K, and M expressed in human HEK293 cells and purified under native conditions, representing post-translational modified versions of the proteins (Figure 2.1A). To quantify the DNA binding activities of the proteins, the apparent equilibrium dissociation constants (KD) were determined using two different methods: from the decrement in the amount of free DNA (total KD) and from the appearance of the first DNA-bound complex (specific KD), in electrophoretic mobility shift assays (EMSA). As expected, the bacterial (bact.) and the mammalian (mamm.) expressed, recombinant O, S, K, and M proteins bound to DNA probes containing canonical motifs, as previously reported for the purified DBDs (Farina et al., 2004; Nakatake et al., 2006; Rodda et al., 2005) (Figure 2.S1B; Table 2.1), and bound with much lower affinity to non-specific DNA sequences of the same length (Figure 2.S1C). The bact. reconstituted Myc:Max heterodimers formed a complex that migrated more slowly than Max homodimers, and no protein-DNA complexes with similar mobility to Max homodimers were observed even at the highest concentrations, confirming that the c-Myc:Max preparation did not
contain Max homodimers (Figure 2.S1B). The mamm. c-Myc did not show any specific DNA binding activity in the absence of its partner Max, as seen previously (Wechsler et al., 1994). These data demonstrate that the recombinant full-length OSKM proteins were highly active in specific DNA binding.

To measure the direct interactions between OSKM and nucleosomes, we identified a nucleosome-enriched site in the fibroblast genome that is efficiently targeted by OSKM (Soufi et al., 2012), focusing on the LIN28B locus that is important for reprogramming and pluripotency (Shyh-Chang et al., 2013; Yu et al., 2007) RNA sequencing (RNA-seq) data showed that LIN28B is silent in human fibroblasts and remains silent after 48 hr OSKM induction, revealing that OSKM binding precedes LIN28B gene activation (data not shown). We selected a region downstream of the LIN28B poly(A) site that is strongly enriched for a nucleosome in pre-induced human fibroblasts, as measured by MNase sequencing (MNase-seq) (Kelly et al., 2012) and was targeted by all four factors at 48 hr post-induction (Figure 2.1B). We used PCR on human fibroblast DNA to generate a 162-bp, Cy5-labeled LIN28B-DNA, which was assembled into nucleosomes (LIN28B-nuc) by salt gradient dilution with purified recombinant human histones (Figure 2.S1D). The nucleosomes exhibited protection from low concentrations of DNase I except at the ends of the LIN28B fragment, compared to free DNA, indicating translational positioning around the center of the 162-bp LIN28B sequence (Figure 2.1C, top two boxes), similar to the observed position of the center of the MNase-seq peak (Figures 2.1B and 2.1C). Ten-fold higher
concentrations of DNase generated an approximately 10-bp DNase-cleavage repeat pattern on LIN28B-nuc, reflecting rotational positioning of nucleosomes within the population (Figure 2.1C, bottom).

It is generally accepted that nucleosomes act as a barrier to DNA binding by TFs (see Introduction), though exceptions have been noted (Perlmann, 1988). Interestingly, Oct4, Sox2, and Klf4, but not c-Myc:Max, showed binding to the LIN28B-nuc (Figure 2.1D). Remarkably, both mamm. and bact. Oct4 and Sox2 showed similar or lower apparent KD values for LIN28B-nuc compared to LIN28B-DNA, indicating similar or higher affinity to nucleosome than to free DNA (Figure 2.1D; Table 2.1). On the other hand, Klf4 was able to bind LIN28B-nuc with a higher apparent KD value compared to free DNA, indicating substantial nucleosome binding, but at a lower affinity than to free DNA (Figure 2.1D; Table 2.1). c-Myc:Max did not yield saturated binding to LIN28B-nuc, even at the highest concentrations of protein used, and thus the apparent KD must be in the μM range (Figure 2.1D; Table 2.1). In conclusion, both mammalian and bacterial expressed O, S, K, and M exhibit the same relative range of affinities to LIN28B-nuc, and O, S, and K have an independent nucleosome binding activity.

**Specific and Non-Specific DNA Interactions Contribute to Nucleosome Binding**

It is well recognized that TFs show both sequence-specific and non-specific interactions with their DNA targets (Biggin, 2011). To measure the
contribution of specificity on OSK binding to LIN28B nucleosomes, we carried out EMSA in the presence of increasing amounts of specific and non-specific DNA sequences that we had already characterized as competitors (Figures 2.S1B and 2.S1C; Table 2.1). EMSA competition experiments show that a 40-fold molar excess of non-labeled DNA probes containing specific binding sites, but not probes containing non-specific sequences, can displace LIN28B-DNA complexes with each of the OSKM proteins, indicating specific interaction with LIN28B-DNA (Figure 2.2A, left panel), similar to OSKM interaction with their canonical sites (Figure 2.S2A). As expected, bact. and mamm. O, S, or K in complexes with LIN28B-nuc were displaced in the presence of a 40× molar excess of unlabeled, specific competitors (Figure 2.2A, lanes 16, 19, and 22). A 40× or lower (range from 5× to 20×) molar excess of non-specific DNA failed to displace bact. and mamm. Oct4 from the LIN28B-nuc (Figures 2.2A, lane 17, and 2.S2B, lanes 14–16), demonstrating specific binding by Oct4 to the nucleosomes in vitro.

By contrast, a 40× excess of non-specific DNA competed almost all of Sox2 and Klf4 from binding to LIN28B-nuc (Figure 2.2A, lanes 20 and 23). Importantly, lower levels of non-specific competitor, from 5× to 20×, did not compete to the same extent as specific competitor with LIN28B-nuc for binding either Sox2 or Klf4 (Figures 2.S2C and 2.S2D, compare lanes 10 to 11–13 versus 14–16). Thus, both specific and non-specific interactions contribute to Sox2 and Klf4 binding to nucleosomes in vitro.
DNase footprinting showed that each of the O, S, K, and M factors protect sequences on LIN28B-free DNA that resemble their canonical motifs (Figures 2.2B and 2.2C, dash boxes). In addition, at the concentrations used for footprinting, Sox2, Klf4, and c-Myc also show non-specific protection of the LIN28B-free DNA (Figure 2.2B, peaks labeled by asterisks). DNase footprinting of LIN28B-nuc bound to Oct4 and Sox2 show that the factors protect part of their canonical motifs, agreeing with the specific binding to nucleosomes seen with EMSA competition experiments (Figures 2.2B and 2.2C). However, Sox2 and Klf4 protect both specific and non-specific nucleotides on LIN28B-nuc, supporting the non-specific contribution of Sox2 and Klf4 to nucleosomes as seen in EMSA competition experiments (Figure 2.2B). The Klf4 binding site is close to the predicted nucleosome dyad axis, where DNase cleavage is minimal, thus precluding an accurate assessment of specific footprinting. Expectedly, c-Myc showed minimal protection of LIN28B-nuc, confirming the weak affinity to nucleosomes. Altogether, the O, S, and K reprogramming factors employ specific and nonspecific nucleosome interactions to different extents.

Range of Nucleosome Binding In Vitro Is Observed in Genome Targeting In Vivo

We assessed whether OSKM, 48 hr post-induction, targeted sites with pre-existing nucleosome enrichment in fibroblast chromatin. Pooling seven replicates from the MNase-seq data set (GSM543311) allowed a high-resolution
map of nucleosomes with 6.6-fold genome coverage. First, we curated the sites where O, S, K, or M targeted alone, by identifying O, S, K, or M peaks that are 500 bp or more apart from each other. The sites were arranged in rank order by the number of chromatin immunoprecipitation sequencing (ChIP-seq) tags in the central 200 bp, from high- to low-affinity sites. This analysis confirms that each of the O, S, K, and M factors is highly enriched at the central 200 bp within a 2-kb region (Figure 2.3A, blue boxes). Interestingly, Sox2 bound most frequently alone (n = 41,107) compared to Oct4 (n = 22,495), Klf4 (n = 28,212), and c-Myc (n = 23,885). Subsequently, MNase tags across the respective 2-kb regions were counted, reflecting local nucleosome enrichment. Read-density heatmaps showed a range of nucleosome enrichment at the central 200-bp regions that were targeted by O, S, K, or M factors alone (Figure 2.3A, red boxes). Notably, Oct4 targets were the most highly enriched for nucleosomes, followed by Sox2, and then Klf4 throughout the respective TF rank-ordered binding profiles. By contrast, MNase tags in the c-Myc targeted sites were diminished. Also, we did not observe pre-phased arrays of nucleosomes at OSKM target sites, indicating that the initial association with nucleosomes proceeds repositioning, if any. Remarkably, the extent of nucleosome targeting of O, S, K, and M in vivo correlates with the relative abilities of the factors to bind nucleosomes in vitro (Figure 2.1D; Table 2.1).

To assess the contribution of non-specific binding in vivo, we counted the number of O, S, K, and M peaks at 48 hr post-induction as function of false
discovery rate (FDR) threshold. Remarkably, while O, K, and M peak numbers begin to stabilize above an FDR of 0.5% (used in our study) (slopes of 1.6, 1.5, and 1.3 respectively), the number of Sox2 peaks continues to increase (slope of 2.1) with higher FDR (Figure 2.S3A). Thus, it appears that Sox2 employs a measure of non-specific targeting in vivo, as we observed in vitro.

O, S, K, and/or M Synergistic Targeting of Nucleosomes In Vivo and In Vitro

It has been previously suggested that transcription factors can access nucleosomal DNA by cooperative binding in order to compete with histones (Polach and Widom, 1996). To investigate the contribution of synergy between O, S, K, and/or M to nucleosome targeting, we studied sites that were co-targeted by multiple factors within a range of 100 bp or less from each other, i.e., within one nucleosome. In general, we observed that all possible O, S, K, and/or M combinations targets were enriched for nucleosomes except for KM targets, and the co-bound sites, on average, were more enriched for nucleosomes than singly bound sites (Figures 2.3B and 2.S3B). Notably, there were more S, K, and/or M combinations that included Oct4 and showed higher nucleosome enrichment at initially targeted sites, compared to binding combinations lacking Oct4 (Figures 2.3B and 2.S3, compare C–I to J–M). For example, c-Myc showed the most nucleosome targeting when co-bound with Oct4, followed by with Sox2, while c-Myc showed weak targeting to nucleosomes with Klf4 (Figure 2.S3, compare E to K and M). Interestingly, the KM combination was the most frequent
at nucleosome-depleted promoters, similar to KM targeting DNase hypersensitive regions (Soufi et al., 2012) (Figure 2.S3M, red plot). Nevertheless, KM still targeted nucleosome-enriched sites at TSS-distal regions (Figure 2.S3M, blue plot).

To further investigate synergistic targeting with Oct4, we assessed binding by each of the bact. Sox2, Klf4, and c-Myc:Max (1 nM) to the reconstituted LIN28B-nuc (2 nM) in the presence of low amounts of Oct4 (0.3 nM). EMSA showed that all the three recombinant proteins are able to bind with Oct4 to nucleosomal DNA in vitro, forming higher order complexes (Figure 2.3C). Notably, c-Myc:Max binding to LIN28B-nuc was enabled in the presence of Oct4 (Figure 2.3C, right panel). To assess the presence of histones in the LIN28B-nuc in the complexes, we transferred the proteins from an EMSA gel to a polyvinylidene fluoride (PVDF) membrane and blotted for H3 and H2B (Figure 2.S4). Though the c-Myc antibody was the weakest, all LIN28B-nuc-bound complexes showed detectable amounts of H3, and to a lesser extent H2B, indicating the factors bind together to nucleosomes. In summary, Oct4, Sox2, and Klf4 enable c-Myc to target nucleosomal sites both in vivo and in vitro.

O, S, and K Separately Recognize Partial Motifs on Nucleosomes

To identify DNA motifs that are associated with O, S, and K alone targeting to nucleosomes in vivo, the respective targeted sites were rank ordered according to nucleosome enrichment in the central 200 bp. This allowed us to
separate nucleosome-enriched from nucleosome-depleted regions that were individually targeted by O, S, or K. By these criteria, 85%, 80%, and 65% of the genomic sites initially targeted by Oct4, Sox2, and Klf4, respectively, were enriched for nucleosomes (Figures 2.4A–2.4C, red boxes). We used de novo motif analysis, separately analyzing the targets that were enriched for nucleosomes (Figures 2.4A–2.4C, red boxes, upper portion) from those that were depleted of nucleosomes, i.e., free DNA targets (Figures 2.4A–2.4C, red boxes, lower portion). While O, S, and K primarily targeted sequences similar to their canonical motifs at nucleosome-depleted and nucleosome-enriched sites, motifs occurring at nucleosome-enriched sites showed distinctive features (Figures 2.4D–2.4F).

Strikingly, while Oct4 targeted its canonical octamer sequence at nucleosome-depleted sites (~49% of n = 3,375), Oct4 targeted hexameric motifs resembling one or another half of the octamer motif at nucleosome-enriched sites (42% and 28%, respectively, of n = 19,120) (Figure 2.4D). Sox2 targeted its canonical HMG box motif at nucleosome-depleted sites (64% of n = 8,221), while targeting a more degenerate motif lacking the sixth “G” nucleotide in the nucleosomal motif (~74% out of n = 32,886) (Figure 2.4E, arrowhead). Finally, Klf4 alone targeted its nonameric motif at nucleosome-depleted sites (94% of n = 9,874), whereas Klf4 targeted a hexameric motif that was missing the three terminal nucleotides at nucleosome-enriched sites (90% of n = 18,338) (Figure 2.4F, see dashed lines).
These findings agree with the above DNase footprinting of LIN28B-nuc bound to the factors (Figure 2.2B, right panels), with Oct4 and Sox2 protecting a part of their canonical motifs on one side of the LIN28B-nuc DNA (Figures 2.2B and 2.2C; right). On free DNA, Klf4 protected the first three nucleotides of its motif on the upper strand while protecting the remaining six nucleotides of its motif on the bottom strand (Figure 2.S5A). However, Klf4 did not protect the first three nucleotides on the upper strand of LIN28B-nuc, as they were not exposed to DNase I digestion, indicating that Klf4 may be interacting with part of its motif exposed on the other strand (Figures 2.2B and 2.2C).

These data show that the O, S, and K factors can independently target nucleosomes using partial or degenerate motifs, and that each of the factors targets their full canonical motif in the absence of nucleosomes at a target site. Targeting of partial motifs at nucleosomal sites by OS or OK together also reveals partial motifs for each of the factors (data not shown).

The Molecular Basis for O, S, and K Nucleosomal Targeting

In order to define the molecular basis that govern O, S, and K interactions with nucleosomal DNA, we interrogated the three dimensional structures of O, S, and K DBDs in complexes with their canonical motifs that were deposited in the RCSB Protein Data Bank. Oct4 contains a bipartite POU domain, composed of an N-terminal POU-specific (POUS) and a C-terminal POU-homeodomain (POUHD), separated by a linker region. The X-ray structure of Oct4-POU-DNA
complex confirms that the POUS and POUHD each bind one-half of the octameric motif on DNA (Esch et al., 2013) (Figure 2.4G, lower panels). The truncated POUS and POUHD can bind their respective half motif DNA probes in vitro, independently from each other (Verrijzer et al., 1992). Interestingly, the isolated DNA-bound state of either POUS or POUHD accommodates less than half of the DNA surface across the circumference of the double helix (DNA surface occupied 606 and 718 Å2, respectively), leaving the opposite DNA surface solvent-exposed and potentially free to interact with histones in a nucleosome conformation (Figure 2.4G, red dashed arrows in upper panels). However, once both POUS and POUHD are bound to the full motif (1,321 Å2), less than a quarter of the DNA circumference is solvent-exposed and hence would be incompatible with nucleosome binding, due to steric hindrance (Figure 2.4G, red dashed arrow in lower panel). Thus, the two POU domains do not target directly adjacent half sites on nucleosomes, as seen in free DNA, but the exposure of the separate half sites on nucleosomes is enough for Oct4 initial targeting.

Sox2 binds DNA through its HMG box, inducing a sharp bend and widening of the minor groove (Remenyi et al., 2003) (Figure 2.4H, lower-left panel). Our motif analysis showed that Sox2 targets a degenerate motif within nucleosomes, missing one “G” nucleotide at the sixth position (Figure 2.4E). This “G” nucleotide is positioned at the angle of the induced bend and makes direct contacts with the N46 residue at the N-terminal tail of Sox2-HMG (Remenyi et al.,
2003) (Figures 2.4E and 2.4H, arrowhead). Remarkably, mutation of this one amino acid (N46Q) within Sox2-HMG results in a significant decrease in DNA-bending ability without affecting DNA binding (Scaffidi and Bianchi, 2001). In transient transfection assays, the Sox2-N46Q mutant displays higher transactivation activity from the Fgf4 enhancer compared to Sox2 wild-type (Scaffidi and Bianchi, 2001). Furthermore, mutation of the “G” nucleotide in the sixth position of the motif has the unique ability, among all mutations tested, to abolish DNA bending by wild-type Sox2 (Scaffidi and Bianchi, 2001). Together these data indicate that Sox2 would not induce extensive DNA-distortion when targeting the nucleosomonal motif, since that motif lacks the “G” nucleotide. To further support these observations, we superimposed the 3D structure of DNA bound by wild-type Sox2 and Sox2-N46Q mutant on nucleosomal DNA and after 1,000 cycles refinement we calculated the root-mean-square deviation (RMSD) as a measure of the average distance between the phosphate backbone for the best fit. These analysis reveal that the less distorted DNA is more compatible with nucleosomal DNA (RMSD = 0.86 Å) compared to the extensively distorted DNA (RMSD = 6.83 Å) (Figure 2.4H, right panel). In conclusion, our data indicate that Sox2 engages nucleosomes by recognizing a degenerate motif that involves less DNA distortion, better filling the curvature and widened minor groove of DNA around the histone octamer.

Klf4 recognizes the nonameric DNA motif using all three C2H2-type ZFs (three nucleotides per ZF) located at the C terminus (Schuetz et al., 2011)
(Figure 2.4F). However, we identified a hexameric motif, lacking the last three nucleotides, enriched within nucleosomal targets (Figure 2.4F, 90%). Mutagenic studies have shown that the hexameric motif represents the minimal essential binding site for Klf4 (Shields and Yang, 1998). Recently, X-ray crystallography has revealed the structures of Klf4 bound to the hexameric and nonameric sites (Schuetz et al., 2011) (Figure 2.4I). Klf4 uses its two most C-terminal ZFs, out of the three, to recognize the hexameric motif, occupying one side of the DNA double helix (595 Å²) and leaving more than half of the opposite surface potentially free to interact with histones in a nucleosome (Figure 2.4I, red dashed arrow in upper-right panel). Klf4 bound to the nonameric motif, with all three ZFs, fills up more than half of the DNA surface (847 Å²) and would hinder binding to nucleosomes (Figure 2.4I, red dashed arrow in lower-right panel). This analysis suggests that Klf4 employs two of its three ZFs to engage nucleosomes. Interestingly, the observed adaptability of O, S, and K to recognize partial motifs correlates with the apparent flexibility of their respective DBDs that we modeled during their transition from the DNA-free to the DNA-bound states (Figures 2.5B–2.5G).

c-Myc Recognizes a Partial Motif Enriched on Nucleosomes through Co-Binding with Other Factors

Using the partitioning method in Figures 2.4A–2.4C, a subset of c-Myc targeted sites (33%, n = 5,494) were enriched for nucleosomal DNA, while the
majority of sites (77%, n = 18,391) did not exhibit enrichment (Figure 2.5A). Motif analysis revealed that c-Myc nucleosomal targets were enriched for an E-box motif that is missing the two central nucleotides (CANNTG) compared to the canonical E-box (CACGTG) (Figure 2.5B, double arrowheads in top panel). However, nucleosome-depleted targets were enriched for a less degenerate E-box motif that we and others have previously reported to be associated with c-Myc binding at enhancers (Lin et al., 2012; Nie et al., 2012; Soufi et al., 2012) (Figure 2.5B, single arrowhead in bottom panel). Interestingly, c-Myc-alone (i.e., without OSK) nucleosomal targets were additionally enriched for a homeobox (73%) motif that is highly similar to the POUHD motif, compared to nucleosome-depleted sites (48%) (Figure 2.5C). Likewise, the majority of c-Myc sites that co-targeted with Oct4 (76%, n = 2,219) that are enriched for nucleosomes contain centrally a degenerate E-box motif similar to that identified in nucleosomal c-Myc-alone targets (Figures 2.5D and 2.5E). The separate halves of the POU motif were also enriched at the OM targeted sites, indicating that Oct4 uses one or the other DBD while co-binding with c-Myc (Figure 2.5F). In conclusion, c-Myc targets nucleosomal sites either with O, S, K, or with endogenous homeodomain factors, recognizing a centrally degenerate E-box motif.

The basic region of bHLH domain, not bound to DNA, appears to be unfolded in solution (Sauve et al., 2004) (Figure 2.6A; Figure 2.S6A). Upon DNA binding, the basic region folds as an extension of helix-1 and will be referred to as basic-helix-1 (bH) (Nair and Burley, 2003) (Figures 2.6D and 2.S6B, blue
helices). Notably, the most conserved four nucleotides of the E-box (CANNTG) face toward the interaction interface between bHLH and DNA, while the degenerate central two nucleotides (CANNTG) face the exterior part of the DNA helix (Figure 2.6B, see cyan and magenta arrowheads). The transition between DNA free and DNA bound by molecular morphing indicates that the bH follows a gradual folding trajectory across the major groove of DNA (Figures 2.6A–2.6D and 2.6B). The interaction between a partially folded bHLH and the CANNTG drives the initial recognition of the E-box without making contacts with the central nucleotides (NN), resulting in the centrally degenerate E-box motif that we observed for c-Myc at the nucleosome-enriched sites (Figure 2.6B).

Importantly, the partially folded c-Myc only occupies one-half the DNA helix surface, leaving the other half solvent-exposed and potentially nucleosome compatible (Figure 2.6B, red dashed arrow). Apparently, the partially folded c-Myc-DNA complex requires further assistance from other factors such as Oct4 or other homeodomain-containing proteins to remain associated with DNA. The interaction between a partially folded bHLH and a centrally degenerate E-Box motif has been observed by X-ray crystallography for Mitf, which shares 86% sequence homology across the basic region with c-Myc (Figure 2.6C) (Pogenberg et al., 2012). Once fully folded, the c-Myc bHLH adopts a rigid structure, stabilizing DNA binding and resulting in less-degenerate E-box motif, which would be incompatible with nucleosomes (Figure 2.6D). We conclude that
partially unfolded c-Myc targets a centrally degenerate E-box motif, thereby adapting to a nucleosome template when assisted by other factors.

**Predicting Pioneer Activity among Different bHLH Factors in Reprogramming**

To gain insights on how bHLH proteins may differentially target nucleosomes in reprogramming, we examined the 3D structures of a range of bHLH-DNA complexes that have been used in reprogramming experiments (El Omari et al., 2013; Longo et al., 2008; Ma et al., 1994). Interestingly, the basic helix-1 from the different bHLH domains extends across the DNA helix to variable extents (Figures 2.6E–2.6I). Motif analysis was also carried out on genomic sites bound by these factors from available ChIP-seq data. Notably, in conjunction with our findings on c-Myc, the length of the bH α helix negatively correlates with the degeneracy of the central nucleotides (CANNTG) of the de novo motifs that we identified for each factor (Figures 2.6E–2.6I).

To further test this correlation, we examined the recent findings that the bHLH factor Ascl1 can act as a pioneer factor during reprogramming fibroblasts to neurons (Wapinski et al., 2013). We measured nucleosome enrichment in pre-induced mouse embryonic fibroblasts (MEF) within Ascl1 initial targets in MEFs after 48 hr induction (Teif et al., 2012; Wapinski et al., 2013). Unlike c-Myc, the majority of Ascl1 sites (73%, n = 3,019) were enriched for nucleosomes (Figure 2.S6D). Importantly, the basic helix-1 of Ascl1 is considerably shorter compared
to that of c-Myc, leaving more of the DNA surface solvent exposed (Figure 2.6E). Similar to c-Myc, Ascl1 target nucleosomes were enriched (99.3%) for an E-box motif with degenerate central two nucleotides (CANNTG) compared to the E-box seen in 98.7% of sites depleted from nucleosomes (Figure 2.S6E). Ascl1 nucleosomal targets contain an extra “G” nucleotide at the 3’-end of the E-box motif, which is missing in the nucleosome-depleted sites, resulting in more specific targeting of nucleosomes despite the centrally degenerate E-box (Figures 2.6E and 2.S6E).

Ascl1 and Olig2 exhibited the shortest bH regions, by molecular modeling, compared to X-ray crystals of NeuroD, MyoD, and Tal1, with longer bHs. To verify that the observed bH lengths were not due to the methodology, we examined the amino-acid composition of the basic regions in all bHLH factors (Figure 2.6J). The bH-DNA interaction is mainly driven by positively charged residues (and hence the name basic). Interestingly, the Ascl1 bH ends at the last (N-terminal end) basic residue (arginine), which is positioned further upstream (toward the C terminus) compared to the other factors (Figures 2.6J and 2.6R, residues in blue boxes). The last basic residue of Olig2-bH falls in between Ascl1 and the rest of the factors. In conclusion, the basic helix-1 of pioneer bHLH factors such as Ascl1 is intrinsically shorter, allowing the factors to bind nucleosomes more efficiently.
2.6 Discussion

The introduction of a defined set of TFs, such as OSKM, into differentiated cells can result in cell-fate conversion (Takahashi and Yamanaka, 2006), and yet it has been clear that the different factors have different contributions or “strengths” in cell-type conversion. This provided the basis for our effort to tackle the long-standing problem of how TFs initially target their sites in closed chromatin. The pioneer factor theory partly answers this question by suggesting that a select group of TFs, such as FoxA, access closed chromatin by a direct interaction with nucleosomal DNA through a DBD that resembles the structure of a linker histone (Iwafuchi-Doi and Zaret, 2014; Zaret and Carroll, 2011). We previously found that the diverse set of DBDs exhibited by O, S, K, and M, which are structurally different from a linker histone, have differential abilities to access closed chromatin (Soufi et al., 2012). Here, we revealed that the relative tendencies of O, S, K, and M to initially target nucleosomal sites in reprogramming reflect their inherent ability to bind nucleosomes in vitro and their ability to recognize partial motifs on nucleosomes in vivo. This is different from what was observed for FoxA1, which recognizes the same motif on free DNA and nucleosomes (Cirillo et al., 1998; Li et al., 2011). Factors that cannot bind nucleosomes on their own, such as c-Myc, associate with other factors to target degenerate E-boxes on nucleosomes. Our new approach is in contrast to the previous predictions of pioneer factors by fitting fully folded DBDs, in their naked DNA-bound state, on nucleosomes through a docking mechanism.
We found that the bipartite POU domain of Oct4 can target partial motifs exposed on nucleosomes using separate PouS or PouHD domains. The single motif targeted by each domain is longer than each half of the octamer motif, thus providing greater binding specificity than a half motif. In addition, mass spectroscopy analysis has identified histones as interacting partners of Oct4 in mouse ES cells (Pardo et al., 2010), indicating an additional affinity contribution by protein-histone interactions. The bipartite domain-Pax family of TFs can bind DNA using both domains and still occupy half of the DNA surface and would therefore be compatible with nucleosome binding (Garvie et al., 2001; Xu et al., 1999) (Figure 2.S7, right, compared to POU TFs). This agrees with the finding that Pax7 is a pioneer factor that uses full motif recognition during initial targeting (Budry et al., 2012). Thus, bipartite TFs have to either employ one DBD or position both DBDs on the same surface of DNA in order to interact with nucleosomes. Notably, the pioneer activity of a Zebrafish homolog of an Oct protein was observed during the maternal-to-zygote transition (Lee et al., 2013; Leichsenring et al., 2013), suggesting that targeting nucleosomal sites may be a general method for de novo programming of the genome.

The high affinity of Sox2 for nucleosomes may be due to the pre-bent conformation of DNA, which widens the DNA minor groove and favors initial minor groove sensing. While bending naked DNA by Sox2 requires minimal work (Privalov, 2009), the energy cost would impede Sox2 to further bend DNA on nucleosomes. We find that Sox2 would not further bend nucleosomal DNA
because it recognizes a partial motif that diminishes the extreme bending of the full motif. Sox family members share the recognition of the core motif but display diverse preferences outside the core in naked DNA (Badis et al., 2009). Our findings reveal greater flexibility with regard to Sox2 core motif preferences on nucleosomes than was previously recognized. In addition, we showed evidence for both specific and nonspecific binding by Sox2 in vitro and in vivo. The stable, motif-driven targeting by Sox2 on nucleosomes in the ChIP-seq data show much lower co-binding with Oct4 (Soufi et al., 2012) than seen in live imaging (Chen et al., 2014), leaving open whether the latter approach depicts nucleosomal or free DNA binding during genome scanning.

Klf4 showed higher affinity to free DNA compared to nucleosomes in vitro, and its initial targets in vivo were enriched for nucleosomes, though less so than compared to Oct4 and Sox2. Klf4 targets nucleosomes in vivo using two out of its three zinc fingers, recognizing a hexameric motif. This explains how the affinity of Klf4 to nucleosomes is lower than that to free DNA. The pioneer factor GATA4 binds nucleosomes modestly in vitro (Cirillo and Zaret, 1999a) and targets a hexameric motif in vivo (Zheng et al., 2013). Notably, GATA4 only contains two zinc fingers. The Gils zinc finger family 1 (Gli1) greatly enhances reprogramming when co-expressed with OSK (Maekawa et al., 2011). Interestingly, despite containing five ZFs, Glis1 only employs two ZFs (number four and five) to recognize its targets (Pavletich and Pabo, 1993). The repressor ZFP57/Kap1, which is known to be associated with closed chromatin, also recognizes a
hexameric motif despite containing an array of seven zinc fingers (Quenneville et al., 2011). This suggests that zinc finger proteins in general may use two zinc fingers to initially target hexameric motifs exposed on nucleosomes. Klf4 also showed non-specific interactions with nucleosomes, suggesting a similar genome searching mechanism as Sox2.

Various examples have been reported on the overexpression of bHLH factors in cancer, including c-Myc, Tal1, and Olig2 (Lin et al., 2012; Nie et al., 2012; Palii et al., 2011; Sanda et al., 2012; Suva et al., 2014). In all of these cases, the bHLH factors have been associated with degenerate E-box motifs and co-binding with other factors. We propose that the extent to which basic helix-1 lays on DNA and co-binds with pioneer factors is reflected in the recognized motif, predicting bHLH ability to bind nucleosomes and access closed chromatin. Interestingly, the mutation of two amino acids within the basic helix-1 that interacts with central E-box makes the non-myogenic bHLH factor E12 able to convert fibroblasts to muscle cells (Davis and Weintraub, 1992). The homeodomain factor PBX primes MyoD targets to induce myogenic potential (Maves et al., 2007). Furthermore, the hematopoietic TAL1-E45 heterodimer employs one of the two bHLH domains using LMO2 as an adaptor to interact with GATA1 (El Omari et al., 2013). Hence, in addition to their intrinsic structures, bHLH factors co-binding with DNA-binding and non-DNA binding proteins appear to be involved in stabilizing the interaction of the partially folded bHLH factors to
nucleosomes. These features are relevant to the multitude of bHLH factors functioning in development, cancer, and reprogramming experiments.

The differential ability of TFs to recognize their target sites on nucleosomes supports a hierarchical model where pioneer factors are the first to gain access to their targets in silent chromatin. We also observe that the initial targeting can occur for non-pioneer proteins when they bind in conjunction with pioneer factors that allow the former to recognize their DBDs to a reduced motif that is compatible with nucleosome binding. Further studies are needed to understand the secondary events that lead to subsequent changes in local chromatin structure and the formation of large complexes at gene regulatory sequences. By understanding the mechanistic basis by which certain transcription factors are especially capable of initiating cell-fate changes, we hope to modulate the process and ultimately control cell fates at will.
2.7 Main Figures

Figure 2.1 | Oct4, Sox2, Klf4, and c-Myc Display Differential Affinity to Nucleosomes In Vitro
(A) Recombinant purified mammalian and bacterial O, S, K, M, and bacterial Max (X) proteins analyzed by SDS-PAGE and Coomassie staining. The respective OSKMX bands run at the expected sizes when compared to the sizes of protein standards. The OSKM DNA binding activity and specificity are shown in Figures 2.S1A–2.S1C.
(B) O, S, K, and M ChIP-seq profiles (blue, red, orange, and green, respectively) 48 hr post-induction and MNase-seq profile (black) in fibroblasts across the LIN28B locus within the displayed genomic location.

(C) DNase I footprinting showing the protection of LIN28B-DNA before and after nucleosome reconstitution in vitro. Electropherograms of 5′-6FAM end-labeled LIN28B (top strand) oligonucleotides generated by digesting free DNA (blue) and nucleosomal DNA (red) with DNase I. The amount of DNase I used is indicated on top of each panel. Shaded boxes represent the DNase-I-protected regions within LIN28B-nuc in the expected ∼10-bp pattern. See Figure 2.S1D for details about nucleosome reconstitution.

(D) Representative EMSA showing the affinity of increasing amounts of recombinant O, S, K, and M proteins (bact. top panels and mamm. bottom panels) to Cy5-labeled LIN28B-DNA (left panels) and LIN28B-nucleosome (right panels). EMSA of O, S, K, and M to DNA probes containing specific and non-specific targets are shown in Figures 2.S1B and 2.S1C.
Figure 2.2 | The Contribution of Non-Specific Binding to Nucleosome Targeting In Vitro

(A) Representative EMSA showing the affinity of recombinant O, S, K, and M proteins (bact. top panels and mamm. bottom panels) to LIN28B-DNA (left panel).
panels) and LIN28B-nucleosome (right panels) in the presence of 40-fold molar excess of specific competitor (“s” lanes) or non-specific competitor (“n” lanes) or absence of competitor (“-” lanes). Competition assays showing the specificity of O, S, K, and M to their canonical DNA probes and to LIN28B DNA and nucleosome under lower titration of competitor are shown in Figure 2.S2.

(B) DNase I footprinting showing the protection of LIN28B-DNA (left panels) and LIN28B-nuc (right panels) in the absence (blue lines) or presence (red lines) of O, S, K, and M. Electropherograms of 5’-6FAM end-labeled LIN28B (top strand) oligonucleotides generated by DNase I digestion of DNA (0.006 U) and nucleosomal DNA (0.06 U). Dashed boxes and stars represent specific and non-specific sites protected by O, S, K, and M, respectively.

(C) A cartoon representation of the 162-bp LIN28B DNA (left) and nucleosome (right) highlighting the binding sites of O, S, K, and M in vitro in blue, red, orange, and green, respectively, as measured by DNase I footprinting. The protected DNA sequences are indicated.
Figure 2.3 | Oct4, Sox2, Klf4, and c-Myc Display a Range of Nucleosome Targeting In Vivo

(A) Read density heatmaps (in color scales) showing the intensity of O, S, K, and M ChIP-seq signal (blue) and MNase-seq (red) spanning ±1 kb from the center of the O, S, K, and M peaks where each factor binds alone within 500-bp threshold. The analyzed sequences were organized in rank order, from high to low number ChIP-seq reads within the central 200 bp (double arrows). The number of targeted sites is indicated.
(B) As in (A), but showing where the OS, OK, and OM factors peaks are within 100 bp or less apart from each other. The full possible OSKM combinations are shown in Figure 2.S3.

(C) The binding affinity of S, K, and M (1 nM) in the presence of Oct4 (0.3 nM) to LIN28B nucleosomal DNA (lanes 4, 6, and 8, respectively) or absence of Oct4 (lanes 3, 5, and 7). The binding of Oct4 on its own (lane 2) and free LIN28B nucleosomes (lane 1) are indicated. The histone content of the nucleosome bound complexes is shown in Figure 2.S4.
Figure 2.4 | Oct4, Sox2, and Klf4 Recognize Partial Motifs on Nucleosomes

(A–C) Same as in Figure 2.3A, but the sites were organized in a descending rank order according to the MNase-seq tags within the central 200 bp. The nucleosome-enriched sites were separated from the nucleosome-depleted sites (dashed line) for each factor.

(D–F) Logo representations of de novo motifs identified in the O, S, and K nucleosome-enriched targets (top) and nucleosome-depleted targets (bottom). The motifs were aligned to canonical motifs (middle). The number of targets analyzed and percentage of motif enrichments are indicated.

(G–I) Cartoon representations of the 3D structures of O (PDB-3L1P), S (PDB-1GT0), and K (PDBs-2WBS and 2WBU) DBDs in complexes with DNA.
containing canonical motifs. Side and top views are shown for O and K, and dashed curved arrows are shown to represent the extent of exposed DNA surface (G and I). The 3D structure of the less distorted DNA (top) and extensively distorted DNA (bottom) were superimposed on nucleosomal DNA (PDB-3LZ0, gray) to display the extent Sox2-nucleosome binding compatibility by measuring RMSD of the fit.
Figure 2.5 | c-Myc Recognition of Degenerate E-Box on Nucleosome Is Assisted by Binding with Co-Factors

(A–F) Same as shown in Figures 2.4A–2.4F, but for c-Myc alone and OM targets. (C) The enrichment of an associated motif (HD) is measured within c-Myc alone targets containing or depleted from nucleosomes. The data indicate that c-Myc is driven to a degenerate E-box on nucleosomes, in part, by homeodomain factors **p < 0.001.
Figure 2.6 | The Folding Extent of bHLH Basic Helix-1 on DNA Anti-Correlates with Targeting Centrally Degenerate E-Box Motifs on Nucleosomes

(A–D) The folding trajectory of basic helix-1 of c-Myc upon DNA binding showing the possible conformations of c-Myc:Max heterodimers (B and C) that are compatible with nucleosome binding. See Figure 2.S6A for c-Myc Morph. The initial DNA-free state (A) and the fully folded DNA-bound state (D), which is incompatible with nucleosome binding, are indicated. The associated motifs for each c-Myc:Max conformation are shown in the left. See Figure 2.S6B for Mitf structure in complexes with E-box with variable central nucleotides.

(E–I) Cartoon representations of various bHLH reprogramming factors in complexes with DNA containing their canonical motifs (right). The de novo motifs identified for each factor from ChIP-seq data are indicated (left). The cyan and pink arrows represent the position of the exposed nucleotides within the central E-box motif not making base-contacts with the relative bHLH conformation. The
central two nucleotides (CANNTG) are colored in purple in the DNA cartoon. The color scheme of the bHLH along with leucine zipper (LZ) is shown at the bottom. (J) Alignment of amino-acid sequences of the basic region of Ascl1, Olig2, NeuroD, MyoD, Tal1, and c-Myc. The last basic residue at the C-terminal end is highlighted in blue. See Figures 2.S6D and 2.S6E for MNase enrichment and motif analysis of Asl1.
2.8 Supplemental Figures


B O, S, K, and M:K show less affinity to DNA probes containing non-specific sites.

C Oligo conc. (1 nM)

FGF4 promoter

LEFTY1 promoter

CDKN2D promoter

D The reconstitution of LIN28B nucleosomes in vitro
Figure 2.S1 | Recombinant Oct4, Sox2, Klf4, and c-Myc Show Specific DNA-Binding Activities In Vitro, Related to Figure 2.1

(A) Schematic diagram showing the DNA-binding domains organization of the full-length Oct4, Sox2, Klf4, c-Myc and Max (O, S, K, M and X) proteins. The number of amino acids is indicated.

(B) Representative EMSAs showing the affinity of increasing amounts of recombinant bacterial (bact.) and mammalian (mamm.)–expressed O, S, K, and M proteins to Cy5-labeled DNA probes containing their respective canonical binding sites. The concentrations used for each protein (nM) are indicated above each lane. Lanes are numbered underneath each gel. Black arrows indicate the migration of Free DNA and DNA-protein complexes. DNA sequences of the Cy5-labeled probes are shown in the Extended Experimental Procedures.

(C) Same as in (B) but showing EMSAs with DNA probes containing non-specific sequences for each protein.

(D) SDS-PAGE and coomassie staining showing the homogeneity of the recombinant human histones (H2A, H2B, H3, and H4) purified from bacteria under denaturing conditions (left panel) and then refolded to H2A/H2B dimers and H3/H4 tetramers (middle-panel). EMSA (right panel) showing free Cy5-labeled LIN28B DNA and Cy5-labeled LIN28B DNA assembled to nucleosomes in vitro by salt gradient dilution with the refolded H2A/H2B dimers and H3/H4 tetramers. LIN28B DNA was generated by PCR using the primers shown in Extended Experimental Procedures. DNA was visualized using Ethidium-Bromide staining (Et-Br) and Cy5 fluorescence (Cy5) as indicated underneath each gel in the right panel. The sizes of protein standards in kDa and DNA standards in bp are shown.
Figure 2.S2 | Recombinant Oct4, Sox2, Klf4, and c-Myc Show a Range of Specificity to Free DNA versus Nucleosomal DNA In Vitro, Related to Figure 2.2

(A) Representative EMSA showing the affinity of recombinant bact. O, S, K, M-X and X proteins (1 nM) to Cy5-labeled probes (2 nM) containing canonical sites in
the presence of 40 fold molar excess of specific non-labeled competitor (s) or non-specific non-labeled competitor (n) or absence of competitor (−).

(B–E) Representative EMSAs showing the affinity of recombinant O, S, K, and M proteins (bact. top panels and mamm. bottom panels) (1 nM) to Cy5-labeled LIN28B free DNA (lanes 1–8) and Cy5-labeled LIN28B nucleosomal (nuc.) DNA (lanes 9–16) (2 nM) in the presence of 5, 10 and 20-fold molar excess of non-labeled specific competitor (s) (lanes 3–5 and 11–13) or non-labeled non-specific competitor (n) (lanes 6–8 and 14–16) or absence of competitor (−) (lanes 2 and 10). Concentrations of competitors in nM are indicated above each lane. Lanes are numbered underneath each gel. Full black arrow heads indicate free and TF-bound LIN28B-DNA, and white arrow heads indicate free and TF-bound LIN28B nucleosomal DNA. Brown boxes show Klf4-LIN28B-nucleosome complexes under prolonged exposure.
Figure 2. S3 | Oct4, Sox2, Klf4, and c-Myc Show a Range of Affinity and Specificity to Nucleosomes In Vivo, Related to Figure 2.3

(A) The O, S, K, and M ChIP-seq peaks at 48 hr post induction in human fibroblasts were called using different FDR thresholds to show the extent of non-specific DNA binding for each factor in vivo. The plots are color coded as indicated.

(B) Nucleosome enrichment as measured by MNase-seq in human fibroblasts within regions bound by O, S, K, and/or M combinations at 48 hr post induction in fibroblast.
The bottom and top of the box represent the 25th and 75th percentile and the middle band is the 50th percentile of the MNase-seq value; whisker ends represent the min and max values. Outlier values are eliminated.

(C–M) Read density heatmaps (top panels) in red color scale (0—20) showing the intensity of MNase-seq tags, spanning ± 1 kb from the center of the O, S, K, and/or M peaks where the factors bind within 100 bp or less from each other. The number of targeted sites is indicated above. Metaplots (bottom panels) showing the average nucleosome enrichment (MNase-seq tags) within the same O, S, K, and/or M sites shown above but separated into TSS-proximal (red) and TSS-distal groups (blue). Sites that were within 1 kb to the nearest TSS were considered proximal, while sites that were more than 1 kb away from the nearest TSS were considered distal.
O, S, K, and M:X bind to nucleosomes containing H3 and H2B histones (EMSA-Westerns; WEMSA)

Figure 2.4: Oct4, Sox2, Klf4, and c-Myc Bind to Nucleosome Containing H2A and H3 Histones, Related to Figure 2.3
Representative EMSAs showing the binding of Oct4 (1 nM) on its own and in combination with Sox2, Klf4, and c-Myc-Max (3 nM) (left panels). The proteins from EMSA were transferred onto a PVDF membrane (WEMSA) and blotted for H3, H2B, Oct4, and/or Sox2, Klf4 and Myc as indicated (the three panels on the right). Black arrow heads indicate the observed TF-nucleosome complexes.
Figure 2.5 | The Apparent Flexibility of Oct4, Sox2, Klf4, and c-Myc DBDs Correlate with Their Nucleosome Binding Compatibility, Related to Figure 2.4
(A) DNase-I footprinting showing the protection of LIN28B-DNA in the absence (blue lines) or presence (red lines) of Klf4. Electropherograms of 5'-6FAM end-labeled LIN28B oligonucleotides generated by DNase-I digestion of DNA (0.006 U) of the top-strand (top panel) or the bottom-strand (bottom panel). Dashed
boxes represent specific sites protected by Klf4 and the corresponding sequence is indicated underneath.

(B, C, and D) Cartoon representation showing the three dimensional structures of O, S, and K free of DNA as determined by NMR. All the NMR-determined 3D states are aligned and shown by transparent colors to indicate the measured flexibility of the free DBDs. The PDB ids of each structure are indicated.

(E, F, and G) The gradual transition of O, S, and K DBDs from DNA-free to DNA-bound was measured by morphing (Extended Experimental Procedures). Arrows and color transparency indicate the extent of the apparent flexibility of each DBD. The used color scheme is shown at the bottom.
The apparent rigidity of bHLH of c-Myc restricts its binding to nucleosomes.

Figure 2.5 | The HLH Factors Compatibility with Nucleosomes Correlates with Central Degenerate E-Box Motif, Related to Figures 2.5 and 2.6

(A) Same as Figure 2.5 (B, C, and D) for Myc-Max bHLH hetero dimer not bound to DNA.

(B) Same as Figure 2.5 (E, F, and G) for the transition of Myc-Max bHLH from DNA-free to DNA-bound states.

(C) Pair-wise sequence alignment (left panel) of the basic region of Mitf and c-Myc showing identical amino-acids (∗) and highly similar amino acids (;, .). Cartoon representations of Mitf bHLH in complexes with DNA containing the canonical E-box motif (middle panel) and centrally degenerate E-box (right panel). The motifs bound and PDB ids are indicated above. The cyan and pink
arrows represent the position of the exposed nucleotides within the central E-box motif not making base-contacts with the relative bHLH conformation. The central two nucleotides (CANNTG) are colored in purple in the DNA cartoon. The color scheme of the bHLH along with leucine zipper (LZ) is shown at the bottom.

(D) Read density heatmaps (in color scales) showing the intensity of Ascl1 ChIP-seq signal (blue) 48 hr post induction in MEFs and MNase-seq (red) in non-induced MEFs spanning ± 1 kb from the center of the Ascl1 peaks. The analyzed sequences were organized in a descending rank order according to the MNase-seq tags within the central 200 bp (double arrows). The number of targeted sites is indicated. The nucleosome enriched sites were separated from the nucleosome depleted sites (dashed line).

(E) Logo representations of de novo motifs identified in Ascl1 nucleosome-enriched targets (top) and nucleosome-depleted targets (bottom). The motifs were aligned to canonical motifs (middle). The number of targets analyzed and percentage of motif enrichments are indicated.
Variable bipartite DBDs lay on free DNA with different orientations, covering the DNA surface to different extents.

(A) Cartoon representations of the 3-D crystal structures of the Pou domains (blue) of Oct4 and Brn5 in complexes with DNA (red) containing canonical motifs. Side and top views are shown and dashed curved arrows are shown to represent the extent of exposed DNA surface. The PDB ids are indicated.

(B) Same as (A) for the Paired domains of Pax5 and Pax6.

(C) Cartoon representation (left panel) showing the three dimensional structures of the bipartite paired (PRD and HD) domain of Pax8 free of DNA as determined by NMR. All the NMR-determined 3D states are aligned and shown by transparent colors to indicate the measured flexibility of the free DBD. The gradual transition of the paired domain from DNA-free to DNA-bound was measured by morphing (Extended Experimental Procedures). Color transparency indicates the extent of the apparent flexibility. The motif recognized by the paired domains is shown above. The color scheme is shown at the bottom.

Figure 2.S7 | Nucleosome Binding Compatibility of Bipartite DBDs, Related to Figure 2.6

(A) Cartoon representations of the 3-D crystal structures of the Pou domains (blue) of Oct4 and Brn5 in complexes with DNA (red) containing canonical motifs. Side and top views are shown and dashed curved arrows are shown to represent the extent of exposed DNA surface. The PDB ids are indicated.

(B) Same as (A) for the Paired domains of Pax5 and Pax6.

(C) Cartoon representation (left panel) showing the three dimensional structures of the bipartite paired (PRD and HD) domain of Pax8 free of DNA as determined by NMR. All the NMR-determined 3D states are aligned and shown by transparent colors to indicate the measured flexibility of the free DBD. The gradual transition of the paired domain from DNA-free to DNA-bound was measured by morphing (Extended Experimental Procedures). Color transparency indicates the extent of the apparent flexibility. The motif recognized by the paired domains is shown above. The color scheme is shown at the bottom.
### 2.9 Supplemental Tables

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total canonical</td>
<td>0.61</td>
<td>0.64</td>
<td>0.37</td>
<td>0.98</td>
<td>2.49</td>
<td>1.46</td>
<td>1.88</td>
<td>ND</td>
</tr>
<tr>
<td>Specific canonical</td>
<td>0.76</td>
<td>1.04</td>
<td>0.45</td>
<td>1.50</td>
<td>3.18</td>
<td>1.95</td>
<td>0.77</td>
<td>ND</td>
</tr>
<tr>
<td>Total LIN28B DNA</td>
<td>0.75</td>
<td>0.93</td>
<td>0.38</td>
<td>1.46</td>
<td>1.25</td>
<td>0.41</td>
<td>8.28</td>
<td>ND</td>
</tr>
<tr>
<td>Specific LIN28B DNA</td>
<td>0.92</td>
<td>2.05</td>
<td>0.68</td>
<td>3.83</td>
<td>2.26</td>
<td>1.12</td>
<td>6.25</td>
<td>ND</td>
</tr>
<tr>
<td>Total LIN28B nuc.</td>
<td>1.09</td>
<td>1.34</td>
<td>0.34</td>
<td>1.06</td>
<td>5.96</td>
<td>3.45</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Specific LIN28B nuc.</td>
<td>1.17</td>
<td>1.84</td>
<td>0.39</td>
<td>1.43</td>
<td>7.21</td>
<td>13.97</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

**Table 2.1 | Recombinant O, S, K, and M Show a Range of Affinities to Nucleosomes**

Apparent dissociation constants (KD) were derived from EMSA to represent the relative affinities of bacterial (bact.) and mammalian (mamm.) O, S, K, and M to their canonical sites, LIN28B-free DNA, and LIN28B nucleosomes (nuc.). Apparent KD were derived from two separate binding curves representing two experimental replicates, fitted to the experimental data within R2 values of ~0.97, and expressed in nM units. Apparent KD were quantified from the fractional decrement of free DNA or nuc, designated as “total” binding, or from the first bound-DNA/nuc complexes, representing “specific” binding. ND, not determined.
2.10 Materials and Methods

Protein Expression and Purification

The bacterial expression plasmids pET-28B-huOct4, pET-28B-huSox2, pET-28B-huKlf4, and pET-28B-huMyc encode the human O, S, K, and M, respectively, fused to an N-terminal 6X histidine tag. The O, S, K, and M cDNA sequences were generated by PCR from the respective lentiviral constructs used for generating human iPS cells (Hockemeyer et al., 2008), introducing a NotI and EcoRI restriction sites for inserting into the pET-28B plasmid. The histidine-tagged O, S, K, and M proteins were expressed in E. Coli Rosetta (DE3) pLysS (Novagen # 70956-3). Transformed cells were grown at 37°C to a density of 0.5 at A600 nm and protein expression was induced by 0.5 mM IPTG for 4 hr for Oct4, 2 hr for Sox2, and overnight for Klf4 and c-Myc at 30°C. The proteins were purified over Hi-trap HP nickel-charged columns (GE healthcare # 17-5248-01) under denaturing conditions. The purified proteins Oct4 and Sox2 were refolded by initially dialyzing to 2 M Urea in 2 M increment gradients and then to 0 M Urea using a desalting column (GE healthcare # 17-1408-01). The purified denatured Klf4 was refolded by dialyzing to 2 M Urea in 2 M increment gradients and refolded by diluting directly to 1 μM concentration in DNA binding buffer (10 mM Tris-HCl pH 7.5, 1 mM MgCl2, 10 uM ZnCl2, 10 mM KCl, 1 mM DTT, 5% glycerol, 0.5 mg/ml BSA). The c-Myc:Max heterodimer complex was reconstituted as described previously (Farina et al., 2004). The recombinant
human histones were expressed and purified as described previously (Tanaka et al., 2004).

The mammalian expressed human O, S, K, and M recombinant proteins were obtained from OriGene (Oct4 # TP311998, Sox2 # TP300757, Klf4 # TP306691, c-Myc # TP301611). The DDK-tagged mammalian proteins were expressed in HEK293 cells (human embryonic kidney cells) and purified under native conditions using anti-DDK affinity column followed by conventional chromatography steps.

Due to the presence of contaminants (Figure 2.1A), the mammalian protein concentrations were calculated by quantifying the intensity of each of the O,S,K,M bands running at the expected sizes in SDS-PAGE (without including the contaminants) and comparing it with their respective bacterial counterparts; the latter having been quantitated by direct protein measurements using absorbance at 280 nM. To reduce error the band intensities were quantified under variable concentrations.

**Nucleosome Reconstitution**

The 162-bp LIN28B DNA fragment was created by PCR with end-labeled primers. The fluorescent-tagged DNA fragments were gel extracted and further purified using ion-exchange liquid chromatography by MonoQ (GE Healthcare). The nucleosomes were reconstituted by mixing purified human H2A/H2B dimers
and H3/H4 tetramers with LIN28B-DNA at 1:1 molar ratio of histone octamer:DNA using a salt-urea gradient.

The 162 bp LIN28B DNA fragment corresponds to the genomic location:

hg18-chr6:105,638,004-105,638,165
AGTGGTATTAACATACATCCTCAGTGGTGAGTATTAACATGGAACTTACTCCAACAATACA
GATGCTGAATAAATGTAATGCTCTAGTGAAGGAAAGGAAAGGTGGGAGCTGCCATCACT
CAGAATTGTCCAGCGAGGTGATTGTGAAGCTTGTGAATAAAGACA

The DNA sequence was created by PCR with end-labeled primers (see below for sequences). The 162 bp fluorescent-tagged DNA fragments were gel extracted and further purified using ion-exchange liquid chromatography on a Mono-Q column and 2 M salt step gradient. The nucleosomes were reconstituted as described previously (Tanaka et al., 2004). Briefly, 10 μg of Cy5 or FAM end labeled PCR fragment of LIN28B DNA was mixed with purified and refolded H2A/H2B dimers and H3/H4 tetramers at a 1:1 DNA:Histone-octamer molar ratio in 10 mM Tris-HCl pH 8, 5 M Urea, 2 M NaCl, 1 mM EDTA, 0.1 mg/ml BSA. The nucleosomes were assembled using salt-urea gradient by dialyzing against a solution containing 2, 1.5, 1, 0.8, and then 0.6 M NaCl and 10 mM Tris–HCl pH 8.0, 5 M Urea, 1 mM EDTA, 10 mM 2-mercaptoethanol for 4 hr in each buffer at 4°C. The nucleosomes were then dialyzed against a no Urea buffer containing 0.6 M NaCl and 10 mM Tris–HCl pH 8.0, 1 mM EDTA, 1 mM 2-mercaptoethanol and then the same buffer containing 0.1 M NaCl for 8 hr at 4°C. The reconstituted nucleosomes were heat shifted by incubating at 37°C for 6 hr.
DNA Binding Reactions

The end-labeled oligonucleotides containing specific or non-specific sites (see below for sequences), LIN28B-DNA, and LIN28B-nucleosomes were incubated with recombinant proteins in DNA-binding buffer (10 mM Tris-HCl (pH7.5), 1 mM MgCl2, 10 μM ZnCl2, 1 mM DTT, 10 mM KCl, 0.5 mg/ml BSA, 5% Glycerol) at room temperature for 60 min. Free and bound DNA were separated on 4% non-denaturing polyacrylamide gels run in 0.5X Tris–borate–EDTA and visualized using a PhosphorImager using Cy5 fluorescence setting (excitation at 633 nm and emission filter 670 BP 30) and high sensitivity setting. The apparent dissociation constant (Kd) was calculated in two ways:

1. Total Kd was calculated to quantify the total affinity of each protein to DNA accounting for both specific and non-specific binding by assuming that the amount of nonspecific binding is linearly proportional to the concentration of protein used. Total amount of DNA was quantified from Cy5 fluorescence of the free DNA band at 0 nM protein concentration. The amount of free DNA at each protein concentration was determined from the intensity of Cy5 fluorescence of the free DNA bands using Multi-Gauge software (Fujifilm Science lab). The fraction of bound DNA was calculated using the equation below:

\[
\text{fraction } DNA_{bond} = 1 - \left( \frac{DNA_{free}}{DNA_{total}} \right)
\]
Binding curves describing the fraction of bound DNA as a function of protein concentration [TF] from two separate experiments were fitted to the data using nonlinear regression in GraphPad Prism software (version 6.04 for windows). The goodness of the fit was assessed using an R2 greater than 0.97. The Kd for each protein was calculated using the equation below and fixing Bmax to a maximum of 1, where NS is the slope of nonspecific binding.

Specific Kd was calculated to quantify the specific affinity of each protein to DNA not accounting for non-specific binding. The amount of free DNA and bound DNA at each protein concentration was determined from the intensity of Cy5 fluorescence of the free DNA and the first DNA-TF complex bands using Multi-Gauge software (Fujifilm Science lab). The fraction of bound DNA was calculated using the equation below:

\[
\text{fraction } \frac{DNA_{\text{bound}}}{DNA_{\text{free}} + DNA_{\text{bound}}}
\]

Binding curves describing the fraction of bound DNA as a function of protein concentration [TF] from two separate experiments were fitted to the data using nonlinear regression in GraphPad Prism software (version 6.04 for windows). The goodness of the fit was assessed using an R2 greater than 0.97. The Kd for each protein, which is determined as the protein concentration at half-maximum DNA binding was calculated using the equation below and fixing Bmax to a maximum of 1.
For competition assays excessive amounts (from 5 to 40 fold) of non-labeled probes containing specific and non-specific sites were added to the binding reaction and incubated for 60 min at room temperature to reach equilibrium. The binding reactions were loaded on the 4% EMSA gels as described above. EMSA gels were run at 80 V at room temperature. As specific competitors, the following DNA probes were used: FGF4 promoter for Oct4 and Sox2, LEFTY1 promoter for Klf4, and CDKN2D promoter for c-Myc. As non-specific competitors the following DNA probes were used: NS for Oct4 and Sox2, NANO5 promoter for Klf4 and c-Myc. See below for the DNA probes sequence.

**DNase Footprinting**

DNase footprinting reactions were carried out by incubating the 6-FAM end labeled LIN28B free (50 ng) in the presence or absence of the purified TFs or histone octamers with 0.006 (DNA or DNA+TFs) or 0.06 (nuc. or nuc.+TFs) unit of DNase-I (Worthington) in 50 μl DNA-binding buffer (10 mM Tris-HCl (pH7.5), 1 mM MgCl2, 10 μM ZnCl2, 1 mM DTT, 10 mM KCl, 0.5 mg/ml BSA, 5% Glycerol) supplemented with additional 50 μl 10 mM MgCl2 and 5 mM CaCl2 at 25°C for 1 min. The reaction was stopped by adding 90 μl (200 mM NaCl, 30 mM EDTA, 1% SDS) and chilling on ice for 10 min. One tenth of reaction volume (~20 μl) of 3 M NaOAc (pH 5.2) was added to the reaction before the DNA fragments were extracted with Phenol–chloroform extraction. The DNA fragments were
further purified using MinElute PCR purification kit (QIAGEN) and eluted in 10 μl dH2O. The digested DNA fragments were separated by capillary electrophoresis as described previously (Zianni et al., 2006). Briefly, the digested DNA fragments (5 μl) were added to 4.9 μl HiDi formide (Applied Biosciences) and 0.1 μl GeneScan-500 LIZ size standards (Applied Biosciences). After denaturing at 95°C for 10 min, the samples were run on an ABI 96-capillary 3730XL Sequencer, using G5 dye setting, running a genotyping module with an increased injection time of 30 s and injection voltage of 3 kV. The generated electropherograms were analyzed using the peak scanner software (Applied Biosciences) and PeakStudio V 2.2.

**Western Blotting After EMSA (WEMSA)**

The EMSA was carried out as described above with 10-fold more protein and nucleosomes and run on a 1.5 mm thick mini-gel cassette (Life Technologies # NC2015) containing 5% polyacrylamide gel. To avoid Cy5 fluorescence saturation, 90% of the nucleosome used in binding reaction was not labeled. The gel was then visualized using Cy5 fluorescence as described above. To charge the proteins, the gel was incubated for 2 hr in denaturing buffer (1% SDS, 375 mM Tris-HCl pH 7.5) at 20°C. The proteins were transferred to a 0.22 μm Sequi-Blot PVDF membrane (Bio-Rad) using NuPAGE transfer buffer (Life Technologies...
supplemented with 0.1% SDS and 20% methanol for 1 hr at 100 Volts at 4°C.

The proteins were fixed to the membrane by incubating in 10% Glacial Acetic Acid for 15 min at room temperature. The membranes were blocked with PBS-0.1% Tween containing 10% non-fat dry milk overnight at 4°C. The primary antibody incubations with anti-human Oct4 antibody (0.5 μg/ml; Abcam # ab19857), human Sox2 antibody (1 μg/ml; R&D systems # AF2018), human KLF4 antibody (0.5 μg/ml; R&D systems # AF3640), human c-Myc antibody (1 μg/ml; R&D systems # AF3696), anti-human H3 (0.5 μg/ml; abcam # ab1791), and anti-human H2B (0.8 μg/ml; abcam # ab1790) were performed for 2 hr at room temperature. The secondary antibody incubations with goat anti-rabbit IgG-HRP (1:5000 dilution; Santa Cruz # sc-2004) and donkey anti- goat IgG-HRP (1:2000; Santa Cruz # sc-2020) were performed for 1 hr at room temperature. Blots were visualized by using SuperSignal West Pico chemiluminescent substrate (Thermo-Scientific # 34080) in Fujifilm LAS-4000 imaging system. The membranes were stripped by incubating with Restore Western-Blot Plus Stripping Buffer (Thermo-Scientific # 46430) for 30 min at RT and re-blocked after blotting with each antibody. The same membrane was serially blotted and stripped with all antibodies shown.
Genomic Data Analysis

The O, S, K, and M ChIP-seq aligned data along with the called peaks (FDR-controlled at 0.005) were obtained from the GEO database (GSE36570) (Soufi et al., 2012). The MNase-seq data (GSM543311) (Kelly et al., 2012) were aligned to build version NCBI36/HG18 of the human genome and seven replicates were pooled together generating 145,546,004.00 unique reads. The MNase-seq reads were extended to 150 bp to cover one nucleosome and thus resulting in 6.6 fold genome coverage.

To identify regions bound by single factors, we separated peaks if their centers were at least 500 bp apart from each other. Sites bound by all possible OSKM combinations were merged if their peak centers were within 100 bp or less from each other. Regions spanning 1 kb upstream and downstream from the center of the curated peaks were divided into 10 bp bins (n = 200). Tag counts from O, S, K, and M ChIP-seq and MNase-seq were assigned to each corresponding bin and used as a measure for enrichment. The curated genomic locations were organized in ascending rank-order according to the tag counts from the central 20 bins (200 bp) as described in the text. Sites were considered to be nucleosome-depleted if their central 200 bp tag counts were smaller than that of the average 200 bp flanking regions (ratio < 1).

bHLH factors ChIP-seq data were obtained from GEO with the accession code GSE43916 for Ascl1 (Wapinski et al., 2013), GSM1167583 for Tal1,
GSM1167584 for Mitf (Calero-Nieto et al., 2014), GSM1306365 and GSM1306367 for Olig2 (Suva et al., 2014), GSM751036 for NeuroD (Tennant et al., 2013), and GSE50415 for MyoD (MacQuarrie et al., 2013). MNase-seq data for MEFs were obtained from GSM1004654 (Teif et al., 2012). The ChIP-seq and MNase-seq data for the above factors were processed as described for OSKM in human fibroblasts. The mouse sequencing data were aligned to the Mouse genome built mm9, accordingly.

Motif Analysis

For de novo motif discovery, we used Discriminative DNA Motif Discovery algorithm (DREME) (Bailey, 2011). We focused on motifs occurring at the central 200 bp of O, S, or K peaks, using central motif enrichment analysis (CentriMo) (Bailey and Machanick, 2012). We quantified the occurrences of the first hits that returned with the most statistical significance within the O, S, K, and M sites using Find Individual Motif Occurrences (FIMO) (Grant et al., 2011). Motifs that showed most central enrichment were considered. Moreover, the newly discovered motifs were compared to the JASPAR and UniPROBE motif databases using the Motif comparison tool (TOMTOM) (Gupta et al., 2007; Mathelier et al., 2014; Newburger and Bulyk, 2009). The above tools are part MEME-ChIP suit v.4.9.1 (Machanick and Bailey, 2011), available at http://meme.nbcr.net.
Molecular Modeling

We have modeled the macromolecular motions that take place during the initial recognition of O, S, K, and M DBDs to their binding sites using the MORPH server as described previously (Krebs and Gerstein, 2000). Briefly, we used the DNA-free structures of O, S, K, and Myc:Max DBDs as the initial state and the DNA-DBDs complexes structures as the final state (see below for PDB ids used). Based on adiabatic mapping, the possible states accommodating the conformational space between free and bound states were calculated within the energy barriers constraints. By defining a set of hinges, the protein motion describing the rigid-body rotation of a small part “core” in relation to a larger part was directly linked protein flexibility. DNA flexibility was not accounted for in our molecular dynamics. The DNA-free states of Oct4-POUS, Oct4-POUHD, Klf4-3ZFs, and c-Myc-bHLH were built based on their sequence homology (92%, 85%, 93%, and 89%) to the experimental NMR structures of Oct1-POUS, Oct1-POUHD, Klf5-3ZFs and Max-bHLH, respectively (PDBs: 1POU, 1POG, 2EBT, and 1R05) (Assa-Munt et al., 1993; Sauve et al., 2004) using Modeler program (Sali and Blundell, 1993). The DNA-free structure of Sox2-HMG has been solved using NMR and submitted to the protein databank under the PDB id 2LE4. The structures of DNA in complex with Oct4-POU, Sox2-HMG, Klf4-2ZFs, Klf4-3ZFs and Myc:Max-bHLH (PDBs: 1GT0, 2WBS, 2WBU, and 1NKP) were solved using
X-ray crystallography (Esch et al., 2013; Nair and Burley, 2003; Remenyi et al., 2003; Schuetz et al., 2011). The Sox2-HMG N46Q mutant was modeled based on its sequence homology (94%) to hSRY-HMG mutant (PDB: 1J47) (Murphy et al., 2001) as described above. The DNA bound to Sox2 wt or Sox2 N46Q mt was superimposed on the nucleosomal DNA obtained from PDB-3LZ0 (Vasudevan et al., 2010) using the super command from Pymol (Version 1.5.0.1 Schrödinger, LLC) and the RMSD was calculated using the rms_curr command between the phosphate backbone carbon atoms. The DNA accessible surface area (ASA) exposed to solvent was calculated from free-DNA or bound to Oct4-POUS, Oct4 POUHD, Oct4-POUS-HD, Klf4-2ZF5s, and Klf4-3ZF5s from the corresponding crystal structures using areaimol from the CCP4 package (Lee and Richards, 1971).

The X-ray crystal structures of Mitf (4ATK, 4ATI), NeuroD (2QL2), MyoD (1MDY), and Tal1 (2YPB) were obtained from the RCSB protein data bank (El Omari et al., 2013; Longo et al., 2008; Ma et al., 1994; Pogenberg et al., 2012). The Ascl1 and Olig2 structures were obtained from the SWISS-Model server based on their sequence homology to NeuroD (2QL2) (Kiefer et al., 2009; Kopp and Schwede, 2006). The images used in the figures were ray-traced and created using the PyMOL molecular graphics system (Version 1.5.0.1 Schrödinger, LLC).
DNA Binding Sites

The DNA oligonucleotides used as binding sites (top and lower strands) are shown below. The Cy5 5’-end-labeled oligonucleotides were obtained from IDT (Integrated DNA Technologies). The double stranded probes were generated by annealing the single strands using the following reaction: 1 nano-moles of each strand (10 μl of 100 μM) were mixed in 50 μl final volume annealing buffer (20 mM Tris-HCl pH7.6, 50 mM NaCl, 0.1 mM DTT, 1 mM EDTA). The reaction was incubated at 70°C for 10 min, and slowly cooled at room temperature overnight.

<table>
<thead>
<tr>
<th>name_TF</th>
<th>Upper Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGF4</td>
<td>TTTAAGTATCCCATTTAGCATCCAAACAAAGAGTTTC</td>
</tr>
<tr>
<td>NANOG</td>
<td>CTTACAGCTTCTTTTGCAAAATGTCCATGGTGGGA</td>
</tr>
<tr>
<td>NS</td>
<td>CTGCAGGTGGATATTAACTGTGAATTCA</td>
</tr>
<tr>
<td>lEFTY</td>
<td>GAGCTCCCAGGAGGTCCCGAGGGGTGACCTCTCT</td>
</tr>
<tr>
<td>CDKN2D</td>
<td>AGGAGCCTGAGCTGCCACGTTGGGAAGGCTGAGAGCATAGT</td>
</tr>
</tbody>
</table>

PCR Primers

The DNA oligonucleotides used as primers for PCR to generate the LIN28B sequence (162 bp) from human genomic DNA are shown. The chemical modifications at the end of each oligonucleotide are also shown. All the DNA oligonucleotides were obtained from IDT (Integrated DNA Technologies).
<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>Bases</th>
<th>Sequence</th>
<th>Modifications and Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>lin28B-FWD</td>
<td>27</td>
<td>AGT GGT ATT AAC ATA TCC TCA GTG GTG</td>
<td>Standard Desalting</td>
</tr>
<tr>
<td>Cy5-lin28B-FWD</td>
<td>27</td>
<td>/5Cy5/AGT GGT ATT AAC ATA TCC TCA GTG GTG</td>
<td>5’ Cy5 HPLC Purification</td>
</tr>
<tr>
<td>6-FAM-lin28B-FWD</td>
<td>27</td>
<td>/56-FAM/AGT GGT ATT AAC ATA TCC TCA GTG GTG</td>
<td>5’ 6-FAM Standard Desalting</td>
</tr>
<tr>
<td>lin28B-RVS</td>
<td>25</td>
<td>TGT CTT TAT TCA CAA GCT TGC ACA A</td>
<td>Standard Desalting</td>
</tr>
<tr>
<td>Cy5-lin28B-RVS</td>
<td>25</td>
<td>/5Cy5/TGT CTT TAT TCA CAA GCT TGC ACA A</td>
<td>5’ Cy5 HPLC Purification</td>
</tr>
<tr>
<td>6-FAM-lin28B-RVS</td>
<td>25</td>
<td>/56-FAM/TGT CTT TAT TCA CAA GCT TGC ACA A</td>
<td>5’ 6-FAM Standard Desalting</td>
</tr>
</tbody>
</table>
CHAPTER 3: STRUCTURAL FEATURES OF TRANSCRIPTION

FACTORS ASSOCIATING WITH NUCLEOSOME BINDING

3.1 Preface

The manuscript presented in this chapter was originally published online July 11th, 2019 in Molecular Cell, Volume 75 (Fernandez Garcia et al., 2019). It has been reformatted here in accordance with University of Pennsylvania dissertation formatting guidelines. Supplemental tables S1 through S3 are publically-available with this dissertation at:

http://repository.upenn.edu/edissertations/.

Authors: Meilin Fernandez Garcia,¹ Cedric D Moore,² Katharine N. Schulz,³ Oscar Alberto,⁴ Greg Donague,⁴ Melissa M. Harrison,³ Heng Zhu,²,* and Kenneth S. Zaret ⁴,*

Affiliations:
¹Biochemistry and Molecular Biophysics Graduate Group, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104-5157, USA
²Department of Pharmacology and Molecular Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, 21205, USA
³Department of Biomolecular Chemistry, School of Medicine and Public Health, University of Wisconsin Madison, Madison, Wisconsin 53706, USA
⁴Institute for Regenerative Medicine, Department of Cell and Developmental Biology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104-5157, USA

*Lead contact: zaret@pennmedicine.upenn.edu
3.2 Respective Contributions

The majority of the experiments and analyses presented in this chapter were designed and performed by myself, under the guidance of Dr. Kenneth S. Zaret. I carried out genomic data analysis with the guidance of Greg Donahue from Dr. Zaret’s lab, recombinant protein purifications, nucleosome reconstitutions, EMSAs, DNase footprinting experiment, mid-stage and final data processing of protein microarray, and 3D protein structure modeling. Oscar Alberto purified CEBPβ DBD and mutants under my supervision. Dr. Katharine N. Schulz purified ZELDA and consulted on ZLD-binding experiments. Cedric D. Moore carried out protein microarray experiments and initial data processing collected. Dr. Kenneth S. Zaret supervised personnel, data interpretation, and writing the manuscript, while Dr. Melissa M. Harrison, Dr. Cedric Moore, Dr. Heng Zhu and Dr. Dario Nicetto provided comments to the manuscripts. The manuscript was written by me, with the assistance of Dr. Zaret.
3.3 Abstract

Fate-changing transcription factors (TFs) scan chromatin to initiate new genetic programs during cell differentiation and reprogramming. Yet the protein structure domains that allow TFs to target nucleosomal DNA remain unexplored. We screened diverse TFs for binding to nucleosomes containing motif-enriched sequences targeted by pioneer factors in vivo. FOXA1, OCT4, ASCL1/E12α, PU1, CEBPα, and ZELDA display a range of nucleosome binding affinities that correlate with their cell reprogramming potential. We further screened 593 full-length human TFs on protein microarrays against different nucleosome sequences, followed by confirmation in solution, to distinguish among factors that bound nucleosomes, such as the neuronal AP-2α/β/γ, versus factors that only bound free DNA. Structural comparisons of DNA binding domains revealed that efficient nucleosome binders use short anchoring α-helices to bind DNA, whereas weak nucleosome binders use unstructured regions and/or β-sheets. Thus, specific modes of DNA interaction allow nucleosome scanning that confers pioneer activity to transcription factors.
3.4 Introduction

Diverse genomic studies have established that a subset of fate-changing transcription factors (TFs) can target nucleosomal sequences in chromatin, and hence act as pioneer TFs (Heinz et al., 2010; Iwafuchi-Doi and Zaret, 2014; Li et al., 2018; Wapinski et al., 2013). However, the principles that predict whether a given TF may target sites on nucleosomal DNA are not clear. DNA sequences containing clusters of TF binding sites at active enhancers and promoters that were considered to be “nucleosome-free regions” have been shown to harbor histones variants H3.3 and H2A.Z (de Dieuleveult et al., 2016; Jin et al., 2009), histone modifications such as H3K4me1/2, H3K27Ac (Calo and Wysocka, 2013), and exposed or "fragile" nucleosomes (Iwafuchi-Doi et al., 2016; Mieczkowski et al., 2016). Still, the presence of a nucleosome can be refractory to TF binding and transcriptional activity (Blomquist et al., 1996; Taylor et al., 1991). In this study, we investigate the intrinsic nucleosome-binding properties of diverse TFs to identify features that enable nucleosome binding.

We focused on TFs that drive cell differentiation and reprogramming. Analysis of the pioneer factor FOXA, a winged helix factor required for liver development (Lee et al., 2005), revealed that FOXA1 can bind its target sequence on nucleosomes and induce local nucleosomal accessibility in vitro (Cirillo et al., 2002; Cirillo and Zaret, 1999a). In vivo, FOXA1 can displace linker histone to promote local chromatin opening and allow cooperative binding of other TFs (Iwafuchi-Doi et al., 2016). Similar to FOXA, the hematopoietic factors
PU1 and CEBPα target compacted chromatin and induce chromatin accessibility, while CEBPα binding to nucleosome-enriched regions is more dependent on PU.1 and EBF1 (Heinz et al., 2010; van Oevelen et al., 2015). EBF1 has also been shown to bind compacted chromatin and induce lineage-specific chromatin accessibility (Boller et al., 2016; Li et al., 2018).

Genomic assessment of the ectopic OCT4, SOX2, KLF4 (OSK) targeted sites and nucleosome occupancy by MNase-seq on pre-existing chromatin showed that OSK target partial motifs on nucleosomal DNA that mirrors their relative nucleosome affinity in vitro. By contrast, c-MYC bind nucleosomes poorly and localizes at its nucleosomal targets via cooperative interaction with other factors (Soufi et al., 2015b). Thus, TFs show a differential preference for nucleosomes and a hierarchy by which they enable cell fate changes. Yet the structural basis for nucleosome motif targeting, apart from the inherent adaptability of the DNA binding domains (DBDs) (Soufi et al., 2015b), remains unclear.

While there are over 2,000 predicted human TFs (Messina et al., 2004; Vaquerizas et al., 2009), only a fraction has been tested for nucleosome binding (Haswell and O'Shea, 1999; Hayes, 1992; Liu and Kraus, 2017; Perlmann, 1988). Recently, an assessment of TF DBDs interacting with nucleosomes, based on SELEX enrichment of bound populations in vitro, revealed diverse binding behaviors as revealed by the positions of motifs at different targeted positions across the nucleosome (Zhu et al., 2018). However, the study did not
provide a direct quantitative assessment of free DNA vs. nucleosome binding and left open the question of how protein structure relates to nucleosome binding. Transcription factors can exhibit slow on-rates to nucleosomes, allowing time for thermal motions to elicit motif exposure, which can be compensated by slow off-rates, resulting in nucleosome binding in the nanomolar range of dissociation constant (kD) as seen for free DNA (Cirillo and Zaret, 1999b; Donovan et al., 2019)

Protein microarrays are a powerful tool for high-throughput assessment of intermolecular interactions (Chen et al., 2008; Hu et al., 2009). However, interactions determined by solid-state methods may not truly reflect interactions under physiological conditions (Sun et al., 2013) and sequence-specific binding needs to be confirmed in solution. To gain a broader view of how TFs target nucleosomes, we used a high-throughput protein microarray that contained 593 unique full-length human TFs, after filtering for technical accessibility, followed by validation in-solution of the most significant interactors. Using in vitro reconstituted nucleosomes with three different endogenous sequences enriched in TF motifs, we find that TFs bind with a range of nucleosome binding affinities that positively correlate with TF reprogramming potential. Notably, we find that strong nucleosome binding is associated with an anchoring α-helix that interacts with no more than half of the DNA’s circumference, leaving the opposite side of DNA free to engage in histone interactions. These findings improve our ability to
predict transcription factors that can directly scan and bind nucleosomes and thus act as pioneer factors.

3.5 Results

Recombinant Nucleosomes from Endogenous Nucleosomal Sequences Targeted by Pioneer TFs

To establish conditions for a large scale study of TFs that can directly interact with nucleosomes, we selected seven TFs (Table 1) suggested to have pioneering activity on the basis of nucleosomal or closed chromatin targeting in vivo, during fibroblast reprogramming to hepatocytes, neurons, or macrophages (Heinz et al., 2010; van Oevelen et al., 2015; Wapinski et al., 2013). Pioneer TF FOXA1 can drive the conversion of fibroblasts to hepatocytes and, during liver development, directly interacts with nucleosome particle N1 within the ALB enhancer to stimulate expression (Chaya et al., 2001; McPherson et al., 1993; Sekiya and Suzuki, 2011). Therefore, FOXA1 was used as a positive control for nucleosome binding on ALBN1 in vitro, assembled into 160 bp nucleosomes containing motifs for FOXA1 and other liver TFs (Figure 3.1B, 3.1C).

ASCL1 and BRN2 drive reprogramming from fibroblasts to neurons (Vierbuchen et al., 2010), with ASCL1 targeting closed fibroblast chromatin (Wapinski et al., 2013). By integration of genomic datasets (Figure 3.1A) we identified an ASCL1 and BRN2 nucleosomal target sequence at the neuron-glia-CAM-related cell adhesion molecule (NRCAM) locus (Figure 3.1D). The NRCAM
gene plays a role in neurite outgrowth and schizophrenia development (Honer et al., 1997; Vawter et al., 1999). The NRCAM site was selected by consideration of its high nucleosomal MNase-seq signal enrichment in fibroblasts before Ascl1 and Brn2 expression (Figure 3.1D, 3.S1A), ASCL1 and BRN2 targeting to the site by ChIP-seq, the presence of respective DNA binding motifs (Figure 3.1D, 3.1E), its enrichment for enhancer associated marks H3K4me1, H3K27Ac, DNase-hypersensitivity (Figure 3.1D), and gene expression associating with neuronal function by gene ontology (Figure 3.S1A, 3.B).

Genomic and chromatin state analysis by MNase-seq of PU1, CEBPα, and CEBPβ targeting during B cell to macrophage reprogramming suggests that hematopoietic TFs PU1 and CEBPα target nucleosome-enriched chromatin (Barozzi et al., 2014; Heinz et al., 2010; van Oevelen et al., 2015), yet the pre-bound chromatin state was not evaluated. Comparison of fibroblast nucleosome occupancy by MNase-seq signals and TFs ChIP-seq in macrophages (Figure 3.S1C-D) revealed CX3 chemokine receptor 1 (CX3CR1) locus as a candidate nucleosomal targeted site (Figure 3.1F). CX3CR1-DNA shows enrichment for H3K4me1, H3K27Ac and DNase-hypersensitivity in macrophages (Figure 3.1F) and it contains the respective DNA binding motifs (Figure 3.1G). DNA fragments of 160 bp of ALBN1-DNA, and 162 bp of NRCAM-DNA and CX3CR1-DNA were PCR amplified, Cy5-end labeled (Figure 3.S1E), and prepared as free DNA and nucleosomal templates by urea-salt step dialysis (Figure 3.S1F).
At a low concentration of DNaseI, ALBN1, NRCAM, and CX3CR1 nucleosomes exhibited nearly complete resistance to cleavage, compared to free DNA (Figure 3.1H-J). Using 20 fold more DNase-I on nucleosomes elicited a markedly different cleavage pattern than seen on free DNA. Each of the nucleosomes revealed several interspersed ~10 bp cleavage repeat patterns, indicative of different rotational frames that expose diverse DNA binding motifs within 120 bp central core of the nucleosome populations (Figure 3.1H-J, bottom panels). Our method of curating endogenous nucleosomal sites enriched for pioneer factor binding in vivo appears robust for discovering DNA sequences that make stable nucleosomes in vitro.

**ASCL1 Heterodimerization with E12α Enhances DNA Binding and Specificity**

To establish conditions for a large-scale TF-nucleosome screen, we purified full-length FOXA1, GATA4, HNF1α, ASCL1, BRN2, PU1, and CEBPα from *E. coli*, validated their identity by predicted size and immunoblotting (Figure 3.2A, 3.S2A), and measured free DNA binding activity. Apparent dissociation constants were determined by electromobility shift assays (EMSAs) with short Cy5-DNA probes (Figure S2B-D) and quantified by two methods; total dissociation constant (total $K_D$), determined by decrement of free probe, and specific dissociation constant (specific $K_D$), determined by the appearance of specific TF-complexes (Soufi et al., 2015b). As expected, all TFs but ASCL1
displayed nanomolar dissociation constants for probes containing canonical DNA binding motifs (Figure 3.S2C, Table 1), but not for probes lacking binding motifs (Figure 3.S2D). We concluded that most of the purified full-length TFs were highly active and bound DNA specifically.

ASCL1 homodimer DNA binding was not detectable within the nM range, agreeing with previous studies estimating a high dissociation constant \( (K_D) \) 140 \( \mu \text{M} \) (Figure 3.S2E-F) (Meierhan et al., 1995). Basic helix-loop-helix (bHLH) TFs heterodimerize with other bHLHs (Longo et al., 2008; Powell and Jarman, 2008) and interaction network analysis revealed the ubiquitously expressed E12\( \alpha \) (Tcf3) as an ASCL1 interacting partner (Figure 3.S2G) (Henke et al., 2009). Therefore, we co-purified ASCL1/E12\( \alpha \) heterodimers (Figure 3.2A), which formed a DNA-bound complex that migrated faster than E12\( \alpha \) homodimers (Figure S2E, lanes 6, 11) and with a markedly increased DNA binding affinity of \( \sim3.1 \) nM (Figure 3.S2E,F, Table 3.1). We considered ASCL1/E12\( \alpha \) heterodimers suitable for further analysis.

Reprogramming TFs Interact with Nucleosomal Substrates with a Nanomolar Range of Affinities

TFs bound to their respective 160 bp free DNAs of \textit{ALBN1-DNA}, \textit{NRCAM-DNA}, \textit{CX3CR1} within the low nM range (Figure 3.2B-G, lanes 1-6). As expected, the TFs show a higher affinity for the longer 160 bp DNA, compared to shorter DNA probes (Table 3.1) due to increased non-specific DNA interactions with
longer DNA templates. Thus, at low concentrations of TFs (0.1-1 nM), each of the factors bound to the 160 bp DNA fragments in a single shifted band, while at higher concentrations we detected additional shifted complexes with distinct dissociation constants (Figure 3.2B-G, 3.S3A-B). The use of an EMSA assay allowed us to make these distinctions, which could correspond to redundant “periodic” TF binding to nucleosomes detected by SELEX (Zhu et al., 2018).

We then assessed binding of the TFs to the same sequences on nucleosomes. FOXA1, GATA4, ASCL1/E12α, BRN2, HNF1α, and PU1 showed high affinity for nucleosomes, with apparent $K_D$ of 3.0 nM, 3.8 nM, 2.6 nM, 2.0 nM, 4.7 nM, 5.6 nM, respectively (Figure 3.2B-G lanes 7-12, Table 3.1). Comparison of TFs by their $K_D^{DNA}$ and $K_D^{Nuc}$, including prior data on iPS reprogramming factors OCT4, SOX2, KLF4, and c-MYC (Soufi et al., 2015b), showed that nucleosome binding factors segregate among each other by a slight difference in DNA affinity, with strong reprogrammers showing a slightly higher DNA affinity (lower $K_D$) compared to BRN2, KLF4, and HNF1α which are suggested to have supporting roles in reprogramming (Figure 3.2H) (Chanda et al., 2014; Raposo et al., 2015; Wapinski et al., 2013). CEBPα displayed the lowest nucleosome affinity of the nucleosome-binding TFs, with a 5-fold higher $K_D$ of 18 nM compared to the other TFs (Figure 3.2G lanes 7-12, Table 3.1). In general, TF affinity for DNA was slightly higher than that for nucleosomes, demonstrating that the presence of histone octamer can attenuate, but not inhibit, DNA binding of the reprogramming pioneer TFs.

96
The isoforms CEBPα and CEBPβ target the same DNA motif and share a high degree of structural homology (Jakobsen et al., 2013; Nerlov, 2007). Posttranslational modifications within the DBD of CEBPs can negatively regulate DNA binding, with the few exceptions of positive DNA binding regulation by sequential phosphorylation of CEBPβ at residues T167 and S163, corresponding to S184 and T188 in CEBPβ LAP* isoforms (Li et al., 2007; Piwien-Pilipuk et al., 2002; Tang et al., 2005). We generated phospho-mimic variants of CEBPβ (T167D) and (S163D, T167D) of the LAP isoform (Figure 3.S3E). CEBPβ T167D showed a weak increase of free DNA binding (Figure 3.2I, lanes 3,5) and no discernable effect on nucleosome binding (Figure 3.2I, lanes 8,10). CEBPβ S163D, T167D showed no DNA binding or nucleosome binding enhancement (Figure 3.2I, lanes 3,12,8,14). Our results show that engagement of histones with DNA limit CEBPα/β binding to DNA and that post-translational modifications of CEBPβ at T167 or S163, T167 may not enhance nucleosome binding.

**Nucleosomes Enhance Reprogramming TF Binding Specificity**

TFs engage DNA through specific and non-specific DNA interactions (Biggin, 2011). To measure the contribution of sequence specificity on TF binding to DNA vs. nucleosomes, we carried out EMSA in the presence of increasing amounts of short specific and non-specific double stranded DNA (dsDNA) validated as competitors (Figure 3.S2C, D). An 80X molar excess of non-labeled short dsDNA containing specific binding sites minimally displaced free DNA-TF
complexes for FOXA1, ASCL1/E12α, PU1, and C/EBPα (Figure 3.S3F, lanes 2,5), similar to competition with non-specific probes (Figure 3.S3F, lane 2,8); while BRN2 exhibited specific competition. Strikingly, competition experiments on nucleosomes, with a low 20X molar excess of specific competitor, completely displaced all TF-nucleosome complexes (Figure 3.3A, B lanes 2,3). By contrast, an 80X excess of non-specific competitor failed to displace them (Figure 3.3A, B lanes 2,8). Moreover, specific ASCL1/E12α binding to nucleosomes occurs without histone octamer dissociation, as determined by EMSA followed by Western blotting with antibodies for core histones H2B and H3 (Figure 3.3C). It appears that TFs generally engage in specific interactions with nucleosomes due to decreased DNA availability by presence of histone octamers, while free DNA is more accessible, allowing the TFs to engage in more non-specific interactions.

DNase footprinting reveals that ASCL1/E12α blocks cleavages on two of its motifs on each of the NRCAM and CX3CR1 nucleosomes (Figure 3.3D, E, middle panels, filled circles). One of the ASCL1 motifs on the NRCAM nucleosome is in the nucleosome center and was DNase-resistant, precluding the ability to assess further protection by ASCL1/E12α. By contrast, the ALBN1 nucleosome lacks an ASCL motif and ASCL1/E12α gave no evidence of protections (Figure 3.3F, middle panel). PU1 exhibited protections on one of its two motifs on the NRCAM nucleosomes and protections on the single motifs of each of the CX3CR1 and ALBN1 nucleosomes (Figure 3.3D-F, lower panels, filled circles). ASCL1/E12α and PU1 also show protections and enhanced
DNase cleavages (Figure 3.3D-F, closed and open circles) at additional sites, compared to the nucleosomes alone, suggesting that the TFs stabilize a preferred rotational frame that exposes the target motif (Shim et al., 1998) or induce local conformational changes in the nucleosomal DNA. The data indicate that ASCL1/E12α and PU1 bind their motifs in multiple nucleosome sequences contexts.

**Zygotic Genome Activator Zelda Binds Nucleosomes Like Human Pioneer Factors**

We compared the ability of the *Drosophila* TF Zelda (ZLD) to bind nucleosomes like the human TFs. ZLD reprograms the zygotic genome for transcriptional activation during the maternal-to-zygotic transition (Foo et al., 2014; Liang et al., 2008; Schulz et al., 2015; Sun et al., 2015) and binds in vitro assembled nucleosomes from the bottleneck locus (McDaniel, 2019). Recombinant full-length ZLD binds a short dsDNA containing a canonical ZLD-binding motif (CAGGTAG), but not to dsDNA with mutated motif, in agreement with previous studies (Figure 3.4A) (Hamm et al., 2015).

We identified fortuitous ZLD binding motifs (CAGGCAG) in the NRCAM-DNA sequence (Figure 3.4B). ZLD robustly bound both NRCAM-DNA and NRCAM-NUC (Figure 3.4C), without histone octamer displacement (Figure 3.4E) with a $K_{D}^{\text{Nuc}}$ of 1.8 nM (Figure 3.2H), comparable to strong reprogramming TFs. Similar to the human reprogramming TFs we tested, EMSA competition
experiments identified non-specific binding of ZLD to free DNA (Figure 3.4D left panel). However, sequence-independent features also contributed to the binding of ZLD to nucleosomes, as demonstrated by the requirement for 160X unlabeled specific competitor to compete the ZLD-NUC complex as compared to 20X competitor for the previously tested human TF-nucleosome complexes (Figure 3.4D, right panel). Thus, ZLD can directly engage nucleosomes, consistent with chromatin targeting seen in vivo, and supports a comparable role of nucleosome binding for reprogramming factors in multiple organisms.

**Protein Array Screen for Nucleosome Binding TFs Identified TFAP-2 and HMG Transcription Factors as Strong Nucleosome Binders**

Having detailed the nucleosome binding of a subset of human-reprogramming factors and a well-studied insect TF, we were interested in more generally identifying TFs with the fundamental feature of nucleosome binding. To date only a small subset of the estimated 2,000 TFs in the human genome has been tested for direct nucleosome binding (Kummerfeld and Teichmann, 2006; Messina et al., 2004; Vaquerizas et al., 2009). We therefore screened human full-length TFs, purified from yeast and printed on microarrays in duplicates, for binding to Cy5-labeled *ALBN1, NRCAM, CX3CR1* as free DNA and nucleosomes (Figure 3.5A) (Hu et al., 2013). The three different templates have a fortuitous occurrence of 116, 121, and 112 TF binding motifs, respectively, as determined by TRANSFAC database, enabling us to assess many TFs binding to these
natural nucleosome sequences (Table 3.S1) (Matys et al., 2006). Moreover, the occurrence of multiple rotational positions on the nucleosomes provide a myriad of accessible motifs (Figure 3.1H-J). Cy5 fluorescence of probed protein microarrays, incubated with Cy5-labeled ALBN1, NRCAM, CX3CR1 DNAs or nucleosomes, was measured and positive protein-nucleic acid signals were processed (Figure 3.S4A) by protein auto-fluorescence subtraction (Figure 3.S4B), minimal protein amount per spot cutoff (Figure 3.S4C,D), background normalization (Figure 3.S4E), and determination of positive versus negative signal thresholds (Figure 3.S4F). Protein microarrays contained 1,755 nuclear proteins spotted, which after the cutoffs resulted in 1,592 proteins. Finally, out of 1,755 proteins, 593 TFs showed a fluorescent intensity after probing for free DNA and nucleosomes that passed the positive/negative signal threshold; we considered these proteins to be experimentally accessible (Table 3.S2).

The protein microarray screen provisionally identified 326 TFs that interact with both DNA and nucleosomes, 104 that interact with DNA-only, and 163 that interact with nucleosomes-only. Protein microarray screening with Cy5 NRCAM free DNA and nucleosomes identified TFAP2α as a possible nucleosome binding TF (Figure 3.5B). AP2 TFs are crucial for neuronal crest development and regulate the fragile X mental retardation-1 gene (Schorle et al., 1996; Zhang et al., 1996). Validation of TFAP2α-nucleosome interaction with an in-solution EMSA showed that TFAP2α binds DNA and nucleosomes comparable to ASCL1/E12α (Figure 3.5C).
To further validate interactions detected from the screen, we purified 81 more of the proteins. Surprisingly, 57 of the proteins exhibited no free DNA or nucleosome binding by EMSA, including 9 transcription factors that scored as "nucleosome only binders" on the arrays. Twenty-four full-length human TFs resulted in detectable DNA binding by EMSA on one of the ALBN1, NRCAM, or CX3CR1 free DNA and nucleosome sequences that contained motifs for each respective TF located within the central domain of the nucleosomes (Figure 3.S5). TF-nucleosome binding was tested at 3X, 6X, or 9X nM higher concentration of TF compared to TF-DNA concentrations (Figure 3.S6A). Quantification of the free DNA and nucleosome bound fraction by EMSA revealed that 11 of the 24 TFs bound to DNA and nucleosomes, 13 to DNA only, and none to nucleosomes only (Figure 3.5D, 3.S6B-E). Thus, the nucleosome-only interactors by protein microarrays could not be confirmed in solution. Furthermore, the DNA only TFs, IRF3 and CREM1, lacked nucleosome binding in multiple nucleosome sequences contexts harboring their target motifs (Figure 3.S6F-G). The IRF3 DBD is a modified version of the helix-turn-helix (HTH) motif that includes a four-stranded antiparallel β-sheet (De Ioannes et al., 2011). Our results showed that lack of nucleosomes binding is not sequence context dependent.

Clustering of TFs by bound fraction, quantified from EMSAs, identified TFs as strong DNA and nucleosome binders (cluster 1), low DNA binders (cluster 2), and high DNA binders with low nucleosome affinity (cluster 3) (Figure 3.5D).
Cluster 1 TFs with high nucleosome binding affinity also display high DNA affinity (Figure 3.5D). This included TFAP-2 isoforms TFAP-2β and TFAP-2γ which also showed strong nucleosome binding similar to reprogramming TFs, as quantified by EMSA (Figure 3.5D). In addition, we also observed that in solution, members of the high mobility group superclass HMGA1, HMGN1, HMGN5, and SOX5 show high affinity to nucleosomes (Figure 3.S6D, lane 10, 3.S6E, lanes 8, 12, and 16). Thus, we were able to identify the TFAP-2 and HMG families of TFs as nucleosome binding factors.

In contrast, sorting TFs by the nucleosome bound fraction revealed a more heterogeneous DNA binding preference, with IRF3, T-Box factors TBX20 and Brachyury (T), and the TALE-MEINOX TF PKNOX displaying high DNA binding but undetectable nucleosome binding, regardless of a higher 9X fold TF concentration on nucleosomes compared to DNA (Figure 3.5E, 3.S6B) and the presence of DNA binding motifs (Figure 3.S5). We conclude that high DNA affinity does not necessarily translate to nucleosome binding. Thus, nucleosome binding TFs are also strong DNA binders and yet nucleosome binding is not an obligate feature of TFs with high DNA affinity.

**Common Structural Features Associated with Pioneer TFs programming TFs**

To identify structural commonalities among strong relative to weak nucleosome binders, we compared the 3D structures of the above tested
transcription factors’ DBDs. TFs were classified into strong nucleosome binders (group I) and weak nucleosome binders (group II). Group I TFs, validated in solution, exclusively include structural DBD superclasses characterized by α-helical folds (Wingender et al., 2015), including basic helix-loop-helix (bHLH), helix-turn-helix (HTH), homeodomains (HD), and zinc fingers (ZnF) (Figure 3.6A). ZLD is a C2H2-ZnF TF that could fit within group I, even though its structure is not known. Two main types of DNA anchoring modes are observed; a scissor-like binding mode by dimers of bHLH TFs, and the HTH module characteristic of HMG TFs such as SOXs, homeodomain TFs OCT4, BRN2, and ETS TFs PU1. Although structurally different, we observed that strong nucleosome binders anchor DNA through a short recognition α-helix (Figure 3.6A). TALE-PBC class TFs such as PBX adopt a DNA binding arrangement with a kink in the recognition helix, resulting in a truncated recognition helix similar to group I short recognition α-helix (Figure 3.6D). This is in agreement with in vivo data suggesting PBX1 has pioneer activity (Berkes et al., 2004). Therefore, structural comparison of PBX1 with group I strong nucleosome binders predicts strong nucleosome binding potential.

Group II TFs include weak nucleosome binders, which we further subdivided into groups IIA and IIB. Similar to group I TFs, group IIA TFs MYOG, CREM, and previously tested weak nucleosome binder USF1 (Adams and Workman, 1995) display a scissor-like arrangement, but contain an extended recognition α-helix that protrudes past the diameter of the DNA helix (Figure
This suggests that a scissor-like recognition mode with an extended anchoring α-helix results in weak nucleosome binding (Soufi et al., 2015b) and would include bZIP TFs such as cMYC, MYOG, and CEBP.

Group IIB weak nucleosome binders TFs Brachyury and TBX20, homologous to the TBX1 DBD, lack a recognition α-helix and instead use short helical twist or unstructured regions (Figure 3.6C). These factors are members of the immunoglobulin superclass characterized by a β-sheet core structure, hypervariable loops, and DNA recognition through unstructured regions (Bork et al., 1994; Wingender et al., 2015). Interestingly, group IIB TFs have been previously shown to engage in a multitude of cooperative interactions with homeodomain TFs binding partners resembling group I strong nucleosome binders, such as PITX2 and NKX2 (Naiche and Kelly, 2005), and thus appear more dependent on direct nucleosome binders for their genetic activity.

3.6 Discussion

It has been well established that the presence of nucleosomes can impair DNA binding of many transcription factors and DNA repair enzymes (Schild-Poulter et al., 1996; Taylor, 1991). We have addressed how lineage-specific TFs engage nucleosomes to initiate regulatory events in silent chromatin during development and cell programming. We previously showed that the pioneer TFs FOXA and OSK are able to target nucleosomes in vivo and comparably interact with nucleosomes in vitro (Cirillo et al., 1998; Iwafuchi-Doi et al., 2016; Soufi et
Here we show that TFs that drive reprogramming toward hepatic, neuronal, macrophage lineages and the maternal-to-zygotic transition can directly interact with nucleosomes. Additionally, we identify TFs with previously unreported nucleosome binding ability. Given that nucleosome binding is the defining characteristic of pioneer factors, our study defines the DNA binding domain characteristics that endow pioneer activity.

Using nucleosomes assembled with endogenous sequences, we show that ASCL1 heterodimerization with E12α markedly increases ASCL1 DNA and nucleosome binding affinity (Figure 3.S2E,F). E12α is ubiquitously expressed in fibroblasts, suggesting that, during neuronal reprogramming, ASCL1 heterodimerizes with ubiquitously expressed fibroblast TFs to target neuron specific regulatory elements in nucleosome enriched chromatin. This agrees with previous findings, where tethering of ASCL1 to E12α results in increased transcriptional activity (Henke et al., 2009). Additionally, we find that BRN2 also directly interacts with nucleosomes (Figure 3.2F). This is in agreement with in vivo chromatin targeting of these factors during neuronal reprogramming where ASCL1 and BRN2 engage chromatin (Wapinski et al., 2013).

PU1 is a strong nucleosome binder, while CEBPα shows a weaker nucleosome affinity (Figure 3.2H). During B cell reprogramming, knockdown of PU1 results in a decreased CEBPα binding to pre-existing and de novo enhancers, suggesting that in the absence of PU1, CEBPα targeting to closed chromatin might be driven by cooperative interactions (van Oevelen et al., 2015).
Comparison among reprogramming TFs show a range of DNA and nucleosome binding affinities that correlate with TF reprogramming potential and chromatin targeting strength, suggesting that the intrinsic nucleosome binding preferences of these TFs contributes to the hierarchy driving cellular reprogramming.

By utilizing protein microarrays as a discovery tool, we identified the TFAP-2 family as a family of pioneer TFs. By contrast, TBX20 and Brachyury of the T-BOX family lack detectable nucleosome binding, even while possessing high DNA affinity to sequences containing specific motifs. Nucleosome-only binders, suggested by protein microarrays, could not be validated in solution, probably due to aberrant protein folding or unnatural domain exposure in a solid-state environment. We note that our nucleosome binding results with full-length proteins agree partially but not entirely with Zhu et. al., who mostly studied DBDs in a SELEX assay. Regardless, both studies find that high DNA affinity does not necessarily result in nucleosome binding, which presents a different model than where TFs would engage nucleosomal sites simply by displacing the octamer.

Interestingly, sequence alignment of HMGN proteins and SOX2, SOX5, and SOX9 revealed a high degree of conservation between the SOX group B homology domain, with unknown function (Weina, 2014), and the C-term of the HMGN nucleosome binding domain (NBD) (Figure 3.S6H). C-term truncation of the HMGN1 NBD greatly impairs chromatin binding (Ueda et al., 2008), suggesting that SOX homology group B, outside of the SOX DBD, is required for the nucleosome binding and chromatin targeting ability of SOX TFs.
Pioneer TFs contain diverse DBD structures (Figure 3.6A), suggesting diverse modes of DNA targeting in chromatin. For example, FOXA has been suggested to compete with linker histone, due to its resemblance with H1 DBD (Clark, 1993; Iwafuchi-Doi et al., 2016), while OCT4 has been suggested to initially interact with chromatin using either of its POU or homeodomain DNA binding modules (Soufi et al., 2015b). We find that DNA recognition via a short scissor-like module or a HTH module is compatible with nucleosome binding (Figure 3.6A). Recent studies in agreement with our findings showed that Reb1, a yeast HTH factor, displays slower dissociation rates from nucleosomes, compared to free DNA, without histone octamer eviction (Donovan et al., 2019). More recently additional HTH TFs, such as Isl1, have been proposed play a pioneering role in cardiomyocyte reprogramming (Gao et al., 2019).

In contrast, we predict that TFs of the immunoglobulin superclass, lacking an α-helix recognition helix, such as T-Box factors, would need cooperative interactions for nucleosome binding (Figure 3.6C). In agreement with our findings, the TBOX factor Tpit binding to closed target sites in vivo is restricted and mainly driven by the pioneer TF PAX7 (Mayran et al., 2018). Furthermore, TP53, an immunoglobulin fold TF, was shown to bind nucleosomes at their edges, at histone-free DNA (Yu and Buck, 2019). Similarly, BZLF1, a bZIP TF similar to group IIB, showed minimal nucleosome binding at high TF concentrations (Schaeffner et al., 2019).
Our study reveals that nucleosome binding positively correlates with reprogramming potential and suggests a hierarchical model where pioneer TFs work as trailblazers for epigenome regulators and modifiers during reprogramming. Furthermore, we reveal commonalities and structural length limitations among the DNA binding modes of pioneer TFs, compared to non-pioneer TFs. Our study provides insight into how diverse groups of TFs engage in nucleosomal interactions via common structural features. Further studies of the domains of pioneer factors that modulate nucleosomal interactions will unveil how other factors are enabled to bind closed, silent chromatin and initiate cell fate changes.
3.7 Main Figures
Figure 3.1 | Endogenous TF-Nucleosomal Targets Assemble into Stable Nucleosomes In Vitro

(A) Schematic diagram showing genomic data processing for the identification of TFs nucleosomal targets.

(B-G) ChIP-seq profile for reprogramming TFs at identified nucleosomal targets and 3D representation of the DNA sequences used for nucleosome assembly containing TFs canonical motifs indicated (yellow). (B) FOXA1, GATA4 and HNF1α ChIP-seq (red) in liver and MNase-seq profile (green) in fibroblasts across the ALBN1 enhancer within the displayed genomic location. (C) 3D representation of the 160 bp- ALBN1-DNA. (D) ASCL1 and BRN2 ChIP-seq (red) at 48 hr induction in fibroblast and MNase-seq profile in fibroblasts near the NRCAM gene. (E) 162 bp NRCAM-DNA. (F) PU1, CEBPα, and CEBPβ ChIP-seq in macrophages and MNase-seq profile in fibroblasts near the CX3CR1 gene. (G) 162 bp- CX3CR1-DNA.

(H-J) DNase-I footprinting showing the protection of (H) ALBN1-DNA, (I) NRCAM-DNA, (J) CX3CR1-DNA before and after nucleosome reconstitution in vitro. Electropherograms generated by digesting 5'-6 FAM end-labeled free DNA (top panel) and nucleosomes with low DNaseI (middle panel) and high DNaseI (bottom panel). Concentrations of DNase-I indicated. Dashed lines indicate central histone octamer protection within nucleosomes.
null
**Figure 3.2 | Reprogramming TFs bind Nucleosomes with Nanomolar Affinity**

(A) Recombinant purified full-length TFs analyzed by SDS-PAGE and Coomassie staining. The factors are grouped by reprogramming to iHEP (induced hepatocytes), iN (induced neurons), and iMAC (induced macrophages). Recombinant single purification of ASCL1, E12α, and co-purification of ASCL1/E12α (right panel).

(B-G) Representative EMSA showing the affinity of increasing amounts of TFs (B) FOXA1, (C) PU1, (D) ASCL1/E12α, (E) GATA4, (F) BRN2 and (G) CEBPα to Cy5-labelled DNA (lanes 1-6) and nucleosome (lanes 7-12). Black arrowheads indicate TF-DNA complexes. White arrowheads indicate TF-nucleosome complexes.

(H) 2D plot of TFs dissociation constants for DNA (x-axis) and nucleosomes (y-axis).

(I) Representative EMSA showing the affinity of CEBPβ WT, mutants T167D and S163D, T167D to CX3CR1-DNA and nucleosomes.
Figure 3.3 | TFs Bind Nucleosomes with Specificity

(A-B) Representative EMSA of competition assays showing the affinity of recombinant (A) FOXA1, ASCL1/E12α and PU1 to ALBN1-NUC, NRCAM-NUC, and CX3CR1-NUC in the presence of 20-, 40- and 80-fold molar excess of specific competitor ("s" lanes) or non-specific competitor ("ns" lanes) or absence of competitor ("-" lanes). (B) Same as (A) for BRN2 and CEBPα.

(C) Representative WEMSA showing the binding of ASCL1/E12α to NRCAM-NUC. ASCL1/E12α :Nuc complex from EMSA were transferred onto a PVDF membrane (WEMSA) and blotted for H3, H2B, ASCL1 as indicated (the three panels on the right). White arrow heads indicate the observed TF-nucleosome complexes.

(D-F) DNase-I footprinting electropherograms of 5′-6 FAM-labeled (D) NRCAM-NUC, (E) CX3CR1-NUC, and (F) ALBN1-NUC in absence (top) or presence of ASCL1/E12α (middle) or PU1 (bottom) end-labeled free DNA (top strand). 3D DNA representation (red) with each TF motif (yellow). Filled circles, protections; open circles, enhancements. "deg. motif" = degenerate motif for PU1.
Figure 3.4 | ZnF TF Zelda Bind to Nucleosomes

(A) Representative EMSA of ZLD showing affinity to short dsDNA probes a containing canonical ZLD-binding motif ("s" lanes) or mutated motif ("ns" lanes).

(B) Graphical representation of ZLD motifs (yellow) identified on NRCAM-DNA sequence.

(C) Representative EMSA showing the affinity of increasing amounts of recombinant ZLD to Cy5- NRCAM-DNA (lanes 1-5) and nucleosome (lanes 6-10).

(D) Quantification competed fraction of ZLD:DNA (left panel) or ZLD:nucleosome (right panel) complexes by addition of molar excess of specific competitor ("s" lanes), non-specific competitor ("ns" lanes) or absence of competitor ("-" lanes). Molar excess listed.
(E) Representative WEMSA of ZLD :Nuc complex from EMSA transferred onto a PVDF membrane (WEMSA) and blotted for H3, H2B as indicated (the two panels on the right). White arrow heads indicate the observed TF-nucleosome complexes.
Figure 3.5 | Systematic Assessment of Nucleosome Binding of Human TFs with Protein Microarrays

(A) Graphical scheme used to identify nucleosome interacting human TF using DNA and nucleosome probes binding to protein microarrays.
(B) Cy5 fluorescence of TAP2-α on protein microarray printed spot in duplicate in absence (“-“) or hybridized with NRCAM-DNA or NRCAM-NUC.
(C) Representative EMSA comparing the affinity of ASCL1/E12α with TFAP-2α to NRCAM-DNA or NRCAM-NUC. Black arrow heads indicate TF-DNA complexes. White arrow heads indicate the TF-nucleosome complexes.
(D-E). Heatmap representations of TFs bound fractions to DNA (left) and nucleosome (right) by quantification from EMSA. (D) Clustered heatmap showing strong DNA and nucleosome binders (red-cluster 1), low DNA binders and (yellow-cluster 2), high DNA binders with low nucleosome affinity (blue-cluster 3) (E) Heatmap sorted on TF nucleosome bound fraction. TFs concentrations used on Figure 3.S6A.
A  Group I- Pioneer Factors - Nucleosome binders
recognition helix (+)
one side DNA binding or short scissor-like binding

FOXA3  HTH
OCT4  POU_HD
PU1  ETS
ASCL1  bHLH
KLF4  ZnF
GATA3  ZnF

B  Group IIA- Free DNA Binding
recognition helix (+)
extended scissor-like DNA binding

MYOG  cMYC/ MAX
CREM  CEBPα
USF1

C  Group IIB- Cooperative Binding
recognition helix (-)
Immunoglobulin fold

TBX1
Brachyury
NF-κB p50
GAL4

D  PBX1  EXD
HOX9A
UBX
Figure 3.6 | Strong Nucleosome Binding TFs Recognize DNA with Short Recognition α-Helices

(A) Group I pioneer TFs DBDs crystal structures of FOXA3 (pdb 1VTN) (Clark, 1993), OCT4 (pdb 3L1P) (Esch et al., 2013), PU1 (pdb 1PUE) (Kodandapani et al., 1996), KLF4 (pdb 2WBS) (Schuetz et al., 2011) and, GATA3 (pdb 4HC9) (Chen et al., 2012). ASCL1 (SMR P50553).

(B) Group IIA TFs with scissor-like DBDs crystal structures and extended recognition α-helices of cMYC/MAX (pdb 1NKP) (Nair and Burley, 2003), CEBPα (pdb 1NWQ) (Miller et al., 2003), USF (pdb 1AN4) (Ferre-D'Amare et al., 1994), MYOG (SMR P15173), and CREM (SMR Q03060).

(C) Group IIB TFs with immunoglobulin-like fold DBDs crystal structures of TBX1 (pdb 4A04) (El Omari et al., 2012), Brachyury (pdb 1XBR) (Muller and Herrmann, 1997), NF-kB p50 subunit (pdb 1SVC) (Muller et al., 1995) and, GAL4 (pdb 3COQ) (Hong et al., 2008).

(D) TALE-PBC PBX1 (pdb 1PUF) (LaRonde-LeBlanc and Wolberger, 2003) and UBX (pdb 4UUS) (Foos et al., 2015) crystal structures showing scissor-like binding in dimer form with HOX TFs showing kink in recognition α-helices (blue arrow).
3.8 Supplemental Figures

A. ASCL1 and BRN2 co-targeted peaks
- 414 peaks
- 365 inter peaks
- 388 nearest genes
- 314 RNA-seq Hits
- 113 Down
- 108 +Nuc
- 7 +Nuc
- 201 Up
- 182 +Nuc
- 18 +Nuc

B. Brain
- Cerebellum: 30
- Nervous system development: 14
- Negative regulation of Notch signaling: 4
- Mammary gland alveolus: 4
- Righting reflex: 3
- Neprhon development: 3
- Membrane: 21
- Cytoplasm: 18
- Integral component of plasma membrane: 12
- Cell junction: 11
- Synapse: 9
- Postsynaptic density: 9
- Postsynaptic membrane: 9

C. PU1, CEBPα, CEBPβ co-targeted peaks
- 7554 peaks
- 7418 inter peaks
- 7418 nearest genes
- 4554 RNA-seq Hits
- 2029 Down
- 1780 +Nuc
- 249 +Nuc
- 2525 Up
- 2245 +Nuc
- 280 +Nuc

D. Biological Function
- Macrophage
- Embryonic stem cell
- Activated glia
- Kidney tissue
- Bone marrow
- Biological Function
- Cell Compartment
- Toll-like receptor 4 activity
- Lipopolysaccharide activity
- Positive regulation of B-cell activation
- Positive regulation of T-cell activation
- Positive regulation of TNF-alpha secretion
- Neutrophil chemotaxis
- Chemokine activity
- Cellular response to inflammatory stimulus
- Positive regulation of chemokine activity
- Ruffle membrane
- Late endosome
- Endosome membrane
- Early endosome
- External surface of plasma membrane
- Membrane ruffle
- Recycling endosome
- Focal adhesion
- Lysosome
- Focal adhesion
- Endosome

E. dsDNA with 5' extensions
1. dsDNA with 5' extensions
2. Filling
   - CyS-dCTP
   - dGTP
3. Purification

F. NRCAM, NRCAM, ALB1, CX3CR1
- bp
- M DNA nuc
- DNA nuc DNA nuc
- DNA nuc DNA nuc
- EtBr
- Cy5
Figure 3.1 | Reprogramming TFs Target MNase Enriched Target Sites Associated with cell identity Genes. Related to Figure 3.1.

(A-B) Genomic data integration for ASCL1 and BRN2 nucleosomal target site identification. (A) Venn diagram of TFs ChIP-seq peaks association with upregulated/downregulated nearest genes and their nucleosome occupancy as determine by MNase-seq tag enrichment or depletion at 48 hr ASCL1, BRN2 and MYT1L ectopic expression in fibroblast. (B) Gene ontology analysis of nearest genes associated with TFs ChIP-seq peaks in 48 hr ASCL1, BRN2 and MYT1L ectopic expression in fibroblast.

(C-D) Venn diagram of PU1, CEBPα and, CEBPβ ChIP-seq peaks association with upregulated/downregulated nearest genes in macrophages and nucleosome occupancy of fibroblast determine by MNase-seq tag enrichment at expected target sites according to macrophage PU1, CEBPα and, CEBPβ ChIP-seq targeting. (D) Gene ontology analysis of nearest genes associated with for PU1, CEBPα and, CEBPβ ChIP-seq peaks in macrophages.

(E) Schematic diagram showing 3’ end enzymatic Cy5 labeling of DNA with Klenow<sup>exo</sup>

(F) Nucleosome reconstitution of <i>ALBN1-NUC</i>, <i>CX3CR1-NUC</i> and <i>NRCAM-NUC</i> (white arrows) compared to free DNA (black arrows).
A Western blot analysis of purified full-length TFs

B Cy5 signals from short specific binding site probes (probe size shown)

C EMSA of binding to specific DNA

D EMSA of binding to nonspecific DNA

E EMSA of binding to non-specific DNA

F Fraction Bound

G Neuro2

H Neuro1
Figure 3.S2 | Recombinant Purified Full-Length TFs Bind Their Canonical Motifs on Short dsDNA. Related to Figure 3.2.

(A) TFs identity verification by Western blot analysis.
(B) Cy5 detection of labeled short double stranded DNA containing canonical binding motifs for reprogramming TFs. Base pair length of each DNA shown at the bottom.
(C-D) TFs binding curves showing the affinity of E. coli. expressed full-length TFs to Cy5 labelled DNA probes containing or lacking TFs respective canonical binding sites, (C) specific DNA and (D) non-specific respectively. TFs were titrated at 0, 0.1, 0.3, 1, 3 and 9 nM concentrations with 1 nM DNA. DNA sequences of the Cy5-labelled probes are in the star methods and equations for total and specific dissociation constants (K_d^Total, K_d^Specific are described in (Soufi et al., 2015b).

(E) Representative EMSA showing the affinity of increasing amounts of purified ASCL1, ASCL1/E12α and, E12α homodimers proteins to Cy5- MCK-DNA containing an ASCL1 E-box binding motif.
(F) Binding curves or ASCL1, ASCL1/E12α and, E12α to Cy5-MCK DNA based on (E).
(G) ASCL1 STRING functional association network of protein-protein interactions.
Figure 3.S3 | Recombinant Purified Full-Length TFs Bind with Specificity to DNA and Nucleosomes. Related to Figures 3.2 and 3.3.
(A-D) TFs binding curves related to EMSAs fraction bound quantification of Figures 3.2B-G showing the affinity of TFs to Cy5-160bp-DNA (left panels) or Cy5-160bp-NUC (right panels). TFs were titrated on 1 nM DNA or NUC.
(E) Recombinant purified full-length CEBPβ WT and, CEBPβ-T163, CEBPβ-S163D, T167D phosphomimetic mutants analyzed by SDS-PAGE and
coomassie staining. The TFs bands run at the expected sizes when compared to the sizes of protein standards. All proteins in the same gel.

(F) FOXA1, ASCL1/E12α, PU1, BRN2, CEBPα binding to ALBN1-DNA, NRCAM-DNA and CX3CR1-DNA in the presence of 20-, 40- and 80-fold molar excess of specific competitor ("s" lanes) or non-specific competitor ("ns" lanes) or absence of competitor ("-" lanes).
Figure 3.S4 | Pipeline for Protein Microarray Data Processing. Related to Figure 3.5.

(A) Schematic diagram showing protein microarray data processing for the identification novel nucleosome binding TFs.

(B) Representative protein microarray chip showing printed proteins auto-fluorescence on chip edge.
(C) Representative protein microarray zoom showing relative printed protein amounts per spots differences by fluorescence detection with primary anti-GST-Tag antibody and Cy5-conjugated secondary antibody.
(D) Removal of proteins with low concentration per spot to yield positive signal. Scatter plot showing protein amounts of all protein spots in all experiments (left panel), and scatter plot showing proteins removed from the analysis due to low estimated concentration (right panel).
(E) Representative protein microarrays probed with Cy5 labeled ALBN1, NRCAM and CX3CR1 DNA or NUC showing the difference in background intensity.
(F) Examples of spots duplicates showing the range of fluorescence intensities (shown above each spot) detected when probed with Cy5 labeled ALBN1, NRCAM and CX3CR1 DNA or NUC. Threshold for a positive hit (binding) and negative hit (not binding) was determined to be F-B of 80 (black arrow).
<table>
<thead>
<tr>
<th>TF</th>
<th>Sequence</th>
<th>Canonical motif</th>
<th>Motif of sequence</th>
<th>Strand</th>
<th>3D DNA-motif representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BTF3</td>
<td>ALBN1</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>GATA1</td>
<td>ALBN1</td>
<td>TGATA</td>
<td>CCAGGGAAATG</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>RBPJ</td>
<td>ALBN1</td>
<td>TGTGGT</td>
<td>ATCTC</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>FOXA1</td>
<td>ALBN1</td>
<td>TGGTTG; TGGTTGC</td>
<td>TTAAT</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>HNF1A</td>
<td>ALBN1</td>
<td>minor groove AT-rich</td>
<td>AAAAATTTAAAAAATTC</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>HMG1A</td>
<td>ALBN1</td>
<td>NKGNNMKCACGCT</td>
<td>TGGGTTTGGCA</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>NHLH2</td>
<td>NRCAM</td>
<td>GGGCMN</td>
<td>TCCACGACGCTGCTGCCT</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>ASCL1</td>
<td>NRCAM</td>
<td>CAGCTG; CAGCTG</td>
<td>CAGCTG; CAGCTG</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>E12-alpha</td>
<td>NRCAM</td>
<td>CANNTG</td>
<td>CCCCTG; CCCCTG</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>MYOG</td>
<td>NRCAM</td>
<td>ACGAGCTG; GACAGCTG</td>
<td>ACGAGCTG; GACAGCTG</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>TFAP2A/B/C</td>
<td>NRCAM</td>
<td>GGCACAGGCA</td>
<td>GGCACAGGCA</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ZNF250</td>
<td>NRCAM</td>
<td>NTAGGCGCTAN</td>
<td>TGTGCCACAC</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>ELK1</td>
<td>NRCAM</td>
<td>CAGGAAAG</td>
<td>CAGGAAAG</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Bm2 (POU3F2)</td>
<td>NRCAM</td>
<td>ATCAAAAGT</td>
<td>ATCGA, ATACAAAAAGT</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PKNOX</td>
<td>NRCAM</td>
<td>TGGTTTGGCA</td>
<td>TGGTTTGGCA</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>TBX20</td>
<td>NRCAM</td>
<td>NAGGTGTGAAN</td>
<td>ACATCACATCTCTGTG</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CREM</td>
<td>NRCAM</td>
<td>NRTGAYGTGTCAYNATCACATCC</td>
<td>NRTGAYGTGTCAYNATCACATCC</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>SPDEF</td>
<td>NRCAM</td>
<td>AACAGGATG</td>
<td>AACAGGATG</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Bm5 (POU6F1)</td>
<td>NRCAM</td>
<td>ATGCAAAAGT</td>
<td>ATACAAAAAGT</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>HMGN1</td>
<td>CX3CR1</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>HMGN5</td>
<td>CX3CR1</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>ESRRG</td>
<td>CX3CR1</td>
<td>ATGGGAATCAGCAG</td>
<td>CTGAAGATCAGCAG</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CEBPα</td>
<td>CX3CR1</td>
<td>GTGGCT; GTGGCA; AGCAAT; TAGCAAT</td>
<td>AGCAAT; TAGCAAT</td>
<td>+,+,:+</td>
<td></td>
</tr>
<tr>
<td>CRX</td>
<td>CX3CR1</td>
<td>ATGTTTGG</td>
<td>ATGTTTGG</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>NFATC1</td>
<td>CX3CR1</td>
<td>NTTCCCGGGAAN</td>
<td>GGGGAATGGGAAGATATCCAT</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>IRF3</td>
<td>CX3CR1</td>
<td>ATCGGGAAAGTA</td>
<td>ATCGGGAAAGTA</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ELF2</td>
<td>CX3CR1</td>
<td>NACCGAGGAAGTNAATCGAGAAGTATTC</td>
<td>NACCGAGGAAGTNAATCGAGAAGTATTC</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PU1</td>
<td>CX3CR1</td>
<td>GGAAGT, GGAATC</td>
<td>GGAAGT, GGAATC</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>CX3CR1</td>
<td>TCACAAAAATAGGTCAG</td>
<td>TCACAAAAATAGGTCAG</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.S5 | 3D Representation of DNA Sequences Showing Positioning of TFs Binding Motifs of Tested Factors. Related to Figure 3.5.

For each TF tested by EMSA (see Figure 3.5D) the DNA and nucleosome sequence which the TF was tested for bindings is shown, consensus motif for the factor, motif sequence within the tested DNA sequence, strand location of the motif, and 3D representation of tested sequences showing TF motif (yellow) localization within the sequence are shown.
Figure 3.S6 | HMG TFs Interact with Nucleosomes. Related to Figures 3.5 and 3.6.

(A) Heatmap representation of TFs DNA bound fraction showing tested TFs (nM) concentration on DNA and nucleosomes. Concentrations were determined experimentally.
(B-E) Representative EMSA of identified microarray hits for nucleosome interaction with (B) NRCAM-DNA, (C,E) CX3CR1-DNA and, (C) ALBN1-DNA (black arrows) or nucleosome (white arrows). (E) Representative EMSA showing DNA and nucleosome binding of HMG TFs. (F-G) EMSA of DNA only binders (F) IRF3 (G) CREM1 to ALBN1 and CX3CR1 DNA and nucleosomes containing TFs binding motifs. (H) Sequence alignment of SOX2, SOX5, SOX9 with HMGN1, HMGN2, HMGN5 TFs showing high conservation between SOX group B homology domain and HMGN C-term nucleosome binding domain.
### 3.9 Tables

<table>
<thead>
<tr>
<th>Apparent (nM)</th>
<th>$K_D$</th>
<th>FOXA1</th>
<th>GATA4</th>
<th>HNF1A</th>
<th>ASCL1/E12α</th>
<th>E12α</th>
<th>BRN2</th>
<th>PU1</th>
<th>CEBPα</th>
<th>ZLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total canonical</td>
<td>1.3</td>
<td>2.4</td>
<td>2.7</td>
<td>3.1</td>
<td>ND</td>
<td>1.2</td>
<td>4.5</td>
<td>ND</td>
<td>ND</td>
<td>-</td>
</tr>
<tr>
<td>Specific canonical</td>
<td>0.9</td>
<td>2.1</td>
<td>2.5</td>
<td>3.0</td>
<td>31.4</td>
<td>4.9</td>
<td>3.0</td>
<td>7.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total DNA</td>
<td>0.3</td>
<td>0.5</td>
<td>1.1</td>
<td>0.7</td>
<td>-</td>
<td>0.7</td>
<td>0.6</td>
<td>4.6</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Specific DNA</td>
<td>0.4</td>
<td>0.5</td>
<td>1.7</td>
<td>0.9</td>
<td>-</td>
<td>2.3</td>
<td>0.7</td>
<td>5.3</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>Total nuc</td>
<td>2.5</td>
<td>1.6</td>
<td>3.5</td>
<td>1.6</td>
<td>-</td>
<td>1.4</td>
<td>4.9</td>
<td>13</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>Specific nuc</td>
<td>3.0</td>
<td>3.8</td>
<td>4.7</td>
<td>2.6</td>
<td>-</td>
<td>2.0</td>
<td>5.6</td>
<td>18</td>
<td>1.8</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.1 | Recombinant TF Dissociation Constants**

Apparent dissociation constants ($K_D$) were derived from quantified EMSA bands to represent the relative affinities of TFs to their canonical sites or respective free DNA and nucleosomes (nuc). Apparent $K_D$s were derived from two separate binding curves, each representing two experimental replicates, and fitted to the experimental data within $R^2$ values of 0.8-0.99, expressed in nM units. Apparent $K_D$s were quantified from the fractional decrement of free DNA or nuc, designated as “total” binding, or from the first bound-DNA/nuc complexes, representing “specific” binding. ND, not determined. “-” not measured (Figure 3.S2C, 3.S3A-D).
3.10 Supplemental Tables

Table 3.S1 | Fortuitous Motif Occurrence in DNA Sequences by Motif Scan. Related to Figure 3.5. (see accompanying online spreadsheet)
Each of the ALBN1-DNA, NRCAM-DNA and CX3CR1-DNA sequences were scanned with PROMO (Messeguer et al., 2002) for identification of putative TFs binding sites (TFBS) defined by TRANSFAC databased and used as weight matrices for TFBS prediction.

Table 3.S2 | Protein Microarray Data. Related to Figure 3.5. (see accompanying online spreadsheet)
Unprocessed and processed protein microarray data containing the 1,755 proteins name, identifier lists and their respective (F-B)^Cy5 signals for in each experiment, 1,592 proteins after data processing and 593 positive hits TFs in at least one experiment.

Table 3.S3 | Short DNA oligonucleotides containing TFs binding sites used as unlabeled competitors. Related to Figure 3.3 (see accompanying online spreadsheet)
Oligonucleotide DNA sequences used as dsDNA containing TFs binding sites in EMSA competition experiments as specific or non-specific competitors.
### 3.11 Key Resources

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibodies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anti-mouse ASCL1</td>
<td>Abcam</td>
<td>ab74065</td>
</tr>
<tr>
<td>anti-human H2B</td>
<td>Abcam</td>
<td>ab1790</td>
</tr>
<tr>
<td>anti-human H3</td>
<td>Abcam</td>
<td>ab1791</td>
</tr>
<tr>
<td>Cy5-dCTP</td>
<td>GE Healthcare Life Sciences</td>
<td>PA55021</td>
</tr>
<tr>
<td>Klenow Fragment (3ÅL→5ÅL exo-)</td>
<td>New England Biolabs</td>
<td>M0212S</td>
</tr>
<tr>
<td>Bacterial and Virus Strains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6HIS_pRSF_TEV</td>
<td>Gift from Ronen Marmorstein lab</td>
<td>6HIS_pRSF_TEV</td>
</tr>
<tr>
<td>FUW-TetO-Gata4</td>
<td>Buganim et al., 2012</td>
<td>Addgene 41084</td>
</tr>
<tr>
<td>pCMW-SPORT6-Hnf1</td>
<td>Dharmaco mammalian gene collection</td>
<td>MMM1013-202761012</td>
</tr>
<tr>
<td>Tet-O-FUW-Ascl1</td>
<td>Vierbuchen et al., 2010</td>
<td>Addgene 27150</td>
</tr>
<tr>
<td>pBABE-E12-cTAP</td>
<td>Yang et al., 2009</td>
<td>Addgene 20916</td>
</tr>
<tr>
<td>pET-28b-FoxA1</td>
<td>Zaret Lab</td>
<td>pET-28b-FoxA1</td>
</tr>
<tr>
<td>pET-28b-Gata4</td>
<td>Zaret Lab</td>
<td>pET-28b-Gata4</td>
</tr>
<tr>
<td>pET-28b-Hnf1a</td>
<td>Zaret Lab</td>
<td>pET-28b-Hnf1a</td>
</tr>
<tr>
<td>pET-28b-Ascl1</td>
<td>Zaret Lab</td>
<td>pET-28b-Ascl1</td>
</tr>
<tr>
<td>pET-28b-Tcf3</td>
<td>Zaret Lab</td>
<td>pET-28b-Tcf3</td>
</tr>
<tr>
<td>pRSFDUET1-Ascl1+Tcf3</td>
<td>Zaret Lab</td>
<td>pRSFDUET1-Ascl1+Tcf3</td>
</tr>
<tr>
<td>pET-28d-Brn2</td>
<td>Gift from Marius Wernig</td>
<td>pET-28d-Brn2</td>
</tr>
<tr>
<td>pET-28b-Spi1</td>
<td>Zaret Lab</td>
<td>pET-28b-Spi1</td>
</tr>
<tr>
<td>pET-28b-Cebpa</td>
<td>Zaret Lab</td>
<td>pET-28b-Cebpa</td>
</tr>
<tr>
<td>pET-28b-Cepbb</td>
<td>Zaret Lab</td>
<td>pET-28b-Cepbb</td>
</tr>
<tr>
<td>pET-28b-Cepbb-T167D</td>
<td>Zaret Lab</td>
<td>pET-28b-Cepbb-T167D</td>
</tr>
<tr>
<td>pET-28b-Cepbb-S163D,T167D</td>
<td>Zaret Lab</td>
<td>pET-28b-Cepbb-S163D,T167D</td>
</tr>
<tr>
<td>Chemicals, Peptides, and Recombinant Proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>His6x-FOXA1</td>
<td>Zaret Lab</td>
<td>His6x-FOXA1</td>
</tr>
<tr>
<td>His6x-GATA4</td>
<td>Zaret Lab</td>
<td>His6x-GATA4</td>
</tr>
<tr>
<td>His6x-HNF1α</td>
<td>Zaret Lab</td>
<td>His6x-HNF1α</td>
</tr>
<tr>
<td>His6x-ASCL1</td>
<td>Zaret Lab</td>
<td>His6x-ASCL1</td>
</tr>
<tr>
<td>His6x-E12α</td>
<td>Zaret Lab</td>
<td>His6x-E12α</td>
</tr>
<tr>
<td>His6x-ASCL1/E12α</td>
<td>Zaret Lab</td>
<td>His6x-ASCL1/E12α</td>
</tr>
<tr>
<td>GST-His6x-BRN2</td>
<td>Zaret Lab</td>
<td>GST-His6x-BRN2</td>
</tr>
<tr>
<td>His6x-PU1</td>
<td>Zaret Lab</td>
<td>His6x-PU1</td>
</tr>
<tr>
<td>His6x-CEBPα</td>
<td>Zaret Lab</td>
<td>His6x-CEBPα</td>
</tr>
<tr>
<td>His6x-CEBPβ WT</td>
<td>Zaret Lab</td>
<td>His6x-CEBPβ WT</td>
</tr>
<tr>
<td>His6x-CEBPβ T167D</td>
<td>Zaret Lab</td>
<td>His6x-CEBPβ T167D</td>
</tr>
<tr>
<td>His6x-CEBPβ S163D,T167D</td>
<td>Zaret Lab</td>
<td>His6x-CEBPβ S163D,T167D</td>
</tr>
<tr>
<td>GST-MYOG</td>
<td>Hu et al., 2013</td>
<td>GST-MYOG</td>
</tr>
<tr>
<td>GST-TBX20</td>
<td>Hu et al., 2013</td>
<td>GST-TBX20</td>
</tr>
<tr>
<td>GST-NHLH2</td>
<td>Hu et al., 2013</td>
<td>GST-NHLH2</td>
</tr>
<tr>
<td>GST-CREM1</td>
<td>Hu et al., 2013</td>
<td>GST-CREM1</td>
</tr>
<tr>
<td>GST-IRF3</td>
<td>Hu et al., 2013</td>
<td>GST-IRF3</td>
</tr>
<tr>
<td>GST-RBPJ</td>
<td>Hu et al., 2013</td>
<td>GST-RBPJ</td>
</tr>
<tr>
<td>GST-HMGN1</td>
<td>Hu et al., 2013</td>
<td>GST-HMGN1</td>
</tr>
<tr>
<td>GST-HMGN5</td>
<td>Hu et al., 2013</td>
<td>GST-HMGN5</td>
</tr>
<tr>
<td>GST-HMGA1</td>
<td>Hu et al., 2013</td>
<td>GST-HMGA1</td>
</tr>
<tr>
<td>GST-SOX5</td>
<td>Hu et al., 2013</td>
<td>GST-SOX5</td>
</tr>
</tbody>
</table>

Critical Commercial Assays

- **MinElute PCR purification kit**
  - QIAGEN
  - 28004

- **Quick change II XL Site-Directed mutagenesis Kit**
  - Agilent
  - 200521

Deposited Data

- **ASCL1 and BRN2 ChIP-seq**
  - Vierbuchen et al., 2013
  - GEO GSE43916

- **PU1, CEBPα and CEBPβ ChIP-seq**
  - Heinz et al., 2010
  - GSE21512

- **FOXA1 and HNF1α ChIP-seq**
  - Faure et al., 2012
  - E-MTAB-941

- **GATA4 ChIP-seq**
  - Zheng et al., 2013
  - GSE49132

- **Fibroblast MNase-seq**
  - Teif et al., 2012
  - GSE40910

- **FOXA3-DBD:DNA co-crystal structure**
  - Clark et al., 1993
  - RCSB PDB 1VTN

- **OCT4-DBD:DNA co-crystal structure**
  - Esch et al., 2013
  - RCSB PDB 3L1P

- **PU1-DBD:DNA co-crystal structure**
  - Kodandapani et al., 1996
  - RCSB PDB 1PUE

- **KLF4-DBD:DNA co-crystal structure**
  - Schuetz et al., 2011
  - RCSB PDB 2WBS

- **GATA3-DBD:DNA co-crystal structure**
  - Chen et al., 2012
  - RCSB PDB 4HC9

- **ASCL1**
  - SWISS-MODEL
  - SMR P50553

- **cMYC/MAX-DBD:DNA co-crystal structure**
  - Nair and Burley, 2003
  - RCSB PDB 1NKP

- **CEBPα-DBD:DNA co-crystal structure**
  - Miller et al., 2003
  - RCSB PDB 1NWQ

- **USF-DBD:DNA co-crystal structure**
  - Ferre-D`Amare et al., 1994
  - RCSB PDB 1AN4

- **MYOG**
  - SWISS-MODEL
  - SMR P15173

- **CREM**
  - SWISS-MODEL
  - SMR Q03060

- **TBX1-DBD:DNA co-crystal structure**
  - EI Omari et al., 2012
  - RCSB PDB 4A04

- **Brachyury-DBD:DNA co-crystal structure**
  - Muller and Herrmann, 1997
  - RCSB PDB 1XBR

- **NF-kB p50-DBD:DNA co-crystal structure**
  - Muller et al., 1995
  - RCSB PDB 1SVC

- **GAL4-DBD:DNA co-crystal structure**
  - Hong et al., 2008
  - RCSB PDB 3COQ

- **PBX1-DBD:DNA co-crystal structure**
  - LaRonde-LeBlanc and Wolberger, 2003
  - RCSB PDB 1PUF

- **UBX-DBD:DNA co-crystal structure**
  - Foos et al., 2015
  - RCSB PDB 4UUS

- **Raw gel images**
  - Mendeley
  - 10.17632/5dnznn9kwt.1

Experimental Models: Cell Lines

- **Rosetta (DE3)pLysS Competent Cells**
  - Novagen
  - Novagen # 70956-3
<table>
<thead>
<tr>
<th>Experimental Models: Organisms/Strains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Oligonucleotides**

Table S3

Recombinant DNA

<table>
<thead>
<tr>
<th>Oligonucleotide Sequence</th>
<th>Origin</th>
<th>Length</th>
<th>DNA Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATCCTGTCTGCTGCTGACAGGGGAATGGTTTTGTCTAAATAACCATTCCAGGGACGTTGGTGGCTTTGCCATGCAGAAGAGTTTTGACTGATCAATACAGTCTGTCTGCTTAAATACCATCATGTCGTTTGCC</td>
<td>Zaret Lab</td>
<td>160 bp</td>
<td>ALBN1-DNA</td>
</tr>
<tr>
<td>GATCCATTATTCTGAAACAGATGACTCCCAAGAGCAGGCTGCCTGGCCCACAGGGGCTTGCTGCCCTGCATGACAGCTGCACATCACATCCTGTTGGCCTGACTACTTCAGCCAGATACAAAAGTGGGTGGGGAACATAGGCCAGATACAAAAGTGGGTGGGGAACATAG</td>
<td>Zaret Lab</td>
<td>162 bp</td>
<td>NRCAM-DNA</td>
</tr>
<tr>
<td>GATCCGCAGGGCCTCTCGTCTGATCTTGAGCTGGCTGAGAGTTGCAGCATTGCTGAGTCTTAGCAATGGATACTTCCCGATTCCCCTCACAAAAATAGGTCAGTCTGTCTGGCTAGTTCTGTACTTGCTGACACAGGGCATGTGGGGTTCGGA</td>
<td>Zaret Lab</td>
<td>162 bp</td>
<td>CX3CR1-DNA</td>
</tr>
</tbody>
</table>

**Software and Algorithms**

<table>
<thead>
<tr>
<th>Software/Algorithm</th>
<th>Source</th>
<th>Cat no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Mapper v4.1</td>
<td>Thermo Fisher Scientific Cloud</td>
<td>4440915</td>
</tr>
<tr>
<td>Peak Scanner module</td>
<td>Thermo Fisher Scientific Cloud</td>
<td></td>
</tr>
<tr>
<td>Pymol</td>
<td>Schrodinger</td>
<td></td>
</tr>
<tr>
<td>PRISM 7</td>
<td>GraphPad</td>
<td></td>
</tr>
<tr>
<td>PROMO</td>
<td>Messeguer et al., 2002</td>
<td></td>
</tr>
<tr>
<td>GenePix Pro 7</td>
<td>Molecular Devices, LLC.</td>
<td></td>
</tr>
<tr>
<td>R studio</td>
<td><a href="https://www.rstudio.com/">https://www.rstudio.com/</a></td>
<td></td>
</tr>
<tr>
<td>Fiji is Just ImageJ (Fiji)</td>
<td><a href="https://imagej.net/">https://imagej.net/</a></td>
<td></td>
</tr>
<tr>
<td>SWISS-MODEL</td>
<td>Bienert et al., 2017</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>Human TF Protein microarrays</td>
<td>Hu et al., 2013</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------------------------</td>
<td>-----------------</td>
</tr>
</tbody>
</table>
3.11 Materials and Methods

Experimental model and subject details

*Escherichia coli* bacterial strain Rosetta (DE3)pLysS Competent Cells were used for recombinant protein expression.

Genomic Data Analysis

The ASCL1 and BRN2 ChIP-seq were obtained from GEO GSE43916 (Vierbuchen et al., 2013); PU1, CEBPα and CEBPβ from GSE21512 (Heinz et al., 2010); FOXA1 and HNF1α from E-MTAB-941 (Faure et al., 2012); and GATA4 from GSE49132 (Zheng et al., 2013). ChIP-seq and MNase-seq data GSE40910 (Teif et al., 2012) were aligned to build version NCBI37/mm9 of the mouse genome. The MNase-seq reads were extended to 150 bp to cover one nucleosome.

Protein Expression and Purification

The bacterial expression plasmids: pET-28b-FoxA1, pET-28b-Gata4, pET-28b-Hnf1a, pET-28b-Hnf4a, pET-28b-Ascl1, pET-28b-Tcf3, pRSFDUET1-
Ascl1+Tcf3, pET-28d-Brn2, pET-28b-Spi1, pET-28b-Cebpa, pET-28b-Cebpb encode the mouse FOXA1, GATA4, HNF1α, HNF4α, ASCL1, E12α, ASCL1+E12α, BRN2, PU1, CEBPα, and CEBPβ proteins respectively, fused to an N-terminal 6X histidine tag. BRN2 was expressed fused to a N-terminal 6X histidine tag followed by a GST-tag. pRSFDUET1 contained 6X his tagged Ascl1 in MCS1 between restriction sites BamH1 and HindIII and untagged E12α (tcf3 gene) in MCS2 between restriction sites Ndel and Xhol. Plasmid containing TFs cDNA were obtained as follow; FUW-TetO-Gata4 (Addgene plasmid #41084) (Buganim et al., 2012b), pCMW-SPORT6-Hnf1a from Dhharmacon mammalian gene collection (MMM1013-202761012), pGCDNsam-Hnf4a (Addgene plasmid #33002) (Sekiya and Suzuki, 2011), Tet-O-FUW-Ascl1 (Addgene plasmid #27150) (Vierbuchen et al., 2010), pBABE-E12-cTAP (Addgene plasmid #20916)(Yang et al., 2009), pET28d-mBrn2 was a gift from Marius Wernig, LZRS PU.1 WT (Addgene plasmid #34835) (Anderson et al., 2002), CEBPalpha NGFR (Addgene plasmid #44627) (Del Real and Rothenberg, 2013), pcDNA 3.1 (-)mouse C/EBP beta (LAP) was a gift from Peter Johnson (Addgene plasmid #12557). TF cDNA sequences were generated by PCR from the respective constructs, introducing restriction sites; Ndel and HindII for FoxA1, Hnf1a, Hnf4a; Ndel and Xhol for Gata4, Spi1, Cebpa, Cebpb, Ascl1; XbaI and Xhol for Brn2 for insertion into their respective plasmid. pET-28b-Cebpb-T167D and pET-28b-Cebpb-S163D,T167D plasmids were generated from pET-28b-Cebpb plasmid with Quick change II XL Site-Directed mutagenesis Kit (Agilent).
The histidine-tagged proteins were expressed in *E. Coli* Rosetta (DE3) pLysS (Novagen # 70956-3). Transformed cells were grown at 37 °C to a density of 0.5-0.7 at an absorbance of 600 nm and protein expression was induced with; 1 mM IPTG at 37 °C for 4 hr with a 30 min delay addition of 20 mg/mL rifampicin for FOXA1; 1 mM IPTG at 16 °C for 16 hr for ASCL1, BRN2 and PU1; 1 mM IPTG at 37 °C for 4 hr for HNF1α, HNF4α, CEBPβ; 2 mM IPTG at 37 °C for 4 hr GATA4 and CEBPα. ZLD was purified as previously described (Harrison et al., 2010; McDaniel, 2019)

The proteins were purified over Hi-trap HP nickel-charged columns (GE healthcare #17-5248-01) or with Ni-NTA resin under denaturing conditions (20 mM Tris-HCl pH 8.0, 0.5M NaCl, 6M Urea) with 5 mM imidazole and 20-300 mM single step imidazole changes follow by 4M and 2M urea step dialysis. The recombinant human full length histones H2A, H2B, H3, and H4 were expressed and purified as described previously (Tanaka et al., 2004). Histone expression plasmids were a gift from Shelley Berger. Protein concentrations were calculated by quantifying the intensity of each of the protein bands running at the expected sizes in SDS-PAGE fitted to a BSA standard curve. To reduce error, the band intensities were quantified at various concentrations.
Cy5 DNA Labeling of Short Canonical-DNA Binding Sites

The DNA oligonucleotides used as binding sites are shown in Table S3. The expected binding sites are highlighted in red. The short oligonucleotide fragments were labeled with Cy5-dCTP (GE Healthcare Life Sciences) by end-repair with Klenow fragment DNA polymerase (3’→5’ exo-) (NEB) as follows. Cy5 5’-end-labelled double stranded probes were generated by annealing complementary single strand DNA probes to give rise to a two nucleotide 3’ overhang (with the last annealed nucleotide being a G) using the following reaction: 1 nanomoles of each ssDNA strand (10 μl of 100 μM) were mixed in 50 μl final volume annealing buffer (20 mM Tris-HCl pH7.5, 50 mM NaCl, 0.1 mM DTT, 1 mM EDTA). The reaction was incubated at 70 ºC for 10 min, and slowly cooled at room temperature overnight. Cy5-DNA labeling reactions were carried out at final concentrations of 1.26 μM dsDNA, 4 μM Cy5-dCTP and 0.5 U/uL Klenow fragment in the presence of excess 4 mM dATP, 4 mM dTTP, and 4 mM dGTP in NEB buffer 2 (50 mM NaCl, 10 mM Tris-HCl, 10 mM MgCl₂, 1 mM DTT). The reaction was incubated at 37 ºC for 1 hr. After Cy5 labeling, the probe was purified using Illustra MicroSpin G-25 columns (GE Healthcare Life Sciences).
Nucleosome Preparation

The 160 bp ALBN1-DNA fragment containing BamH1 flanking sequences corresponds to the genomic location: mm9 chr5:90879148-90879300

GATCCTGTCTCCTGCTCTGCGACGAGCAGCTACTTGCTGATCCAGGGAATGTGTTTTG
TTCTTAATACCATCATCTCGGGACGTGTGTTGCTTGGCCAGTTTTCCATGTACATGCAG
AAAGAAGCTTTGGACTGATCAATACAGTCCTCTGCCTGCTGGATC

The 162 bp NRCAM-DNA fragment containing BamH1 flanking sequences corresponds to the genomic location: mm9 chr12:45445284-45445450

GATCCATTACTTCTGAAACAGATGACTCCCAGCAGCTGCTGCCTGTGGCCCACAGGGCT
TCCTGGCCTGCGATGACGGGCCATCATCAATGCTGTCATGATGACCATGCAGCCTGC
TACGAGCCGATACAAAGTGGGTTGGGAAACATAGAGGAAGGATC

The 162 bp CX3CR1-DNA containing BamH1 flanking sequences fragment corresponds to the genomic location: mm9 chr9:119946611-119946762

GATCCGCGCCGCTCCGGCTGCTGATTTCTCGGAGAAGTTGCTGCACTTGACGATTGCC
TGATCTTACGCAATGGATACGACCTTCCCCTCACAAAAATAGGCTAGTCTGTGTGCTG
CTAGTTCTGTAATCAGACACAGGGCATGTGGGGTTGGGTCGATC

The DNA sequences were created by PCR of genomic mouse DNA with EcoRI-BamHI and BamH1-XbaI restriction sites into pUC19 plasmid to generate pUC19-EB-NRCAM-BX, pUC19-EB-ALBN1-BX, and pUC19-EB-CX3CR1-BX. Plasmid were amplified in *E. coli*, purified, and digested with BamH1 to release fragments. DNA sequences were purified by agarose gel electroelution, phenol:chloroform extraction, and ethanol precipitation followed by Cy5 enzymatic labeling as described above and purified with QIAquick PCR purification kit (Qiagen 28106)
The nucleosomes were reconstituted by dialysis as described previously (Tanaka et al., 2004). Briefly, 10 μg of Cy5 end labelled DNA sequences were mixed with purified and refolded H2A/H2B dimers and H3/H4 tetramers at a 1:1 DNA:histone-octamer molar ratio in 10 mM Tris-HCl pH 8.0, 4 M Urea, 2 M NaCl, 0.1 mg/ml BSA. The nucleosomes were assembled by salt-urea step dialysis against buffers containing 1.5, 1, 0.8, and then 0.6 M NaCl with 10 mM Tris–HCl pH 8.0, 5 M Urea, 1 mM EDTA, 10 mM 2-β mercaptoethanol for 2-4 hr at 4 °C. The nucleosomes were then dialyzed against a no urea buffer containing 0.6 M NaCl and 10 mM Tris–HCl pH 8.0, 1 mM EDTA, 1 mM 2-β mercaptoethanol and then the same buffer containing 0.1 M NaCl for 6 hrs at 4 °C. The reconstituted nucleosomes were then heat shifted by incubating at 42 °C for 2 hr, and further purified with a 10-30% glycerol gradient in 50 mM Tris pH 7.5, 1 mM EDTA and 0.03 mg/mL BSA at 35,000 rpm for 18 hr at 4 °C. Gradients were fractionated and nucleosome-containing factions were pooled and concentrated.

DNA and Nucleosome Binding Reactions

The end-labelled oligonucleotides containing specific or non-specific sites (see above for sequences), free DNA sequence and nucleosomes were incubated with recombinant proteins in DNA-binding buffer (10 mM Tris-HCl pH 7.5, 1 mM MgCl2, 10 μM ZnCl2, 1 mM DTT, 50 mM KCl, 3 mg/ml BSA, 5% Glycerol) at room temperature for 30 min. Free and bound DNA were separated
on 4% non-denaturing polyacrylamide gels run in 0.5X Tris–borate–EDTA. Gels were visualized using with an Amersham Typhoon RGB Biomolecular Imager using Cy5 fluorescence setting (excitation at 633 nm and emission filter 670 BP 30) and a high sensitivity setting. The apparent dissociation constant (Kd) were calculated as previously described (Soufi et al., 2015b). For competition assays excessive amounts (from 20 to 80 fold) of non-labelled probes containing specific and non-specific sites were added to the binding reaction and incubated for an additional 30 min at room temperature to reach equilibrium. The binding reactions were loaded on the 4% EMSA gels as described above. EMSA gels were run at 90 volts at room temperature.

**Western Blotting After EMSA (WEMSA)**

EMSAs were carried out as described above with 10-fold more protein and nucleosomes run on a 1.5 mm thick mini-gel cassette (life technologies # NC2015) containing 5% non-denaturing polyacrylamide gels run in 0.5X Tris–borate–EDTA. To avoid Cy5 fluorescence saturation, 90% of the nucleosomes used in binding reactions was not labelled. The gel was then visualized using Cy5 fluorescence as described above. Western blot was done as described previously (Soufi et al., 2015b). Primary antibody incubations with anti-mouse ASCL1 antibody (1:1000; Abcam # ab74065), anti-human H3 (0.5 μg/ml; abcam # ab1791), and anti-human H2B (0.8 μg/ml; abcam # ab1790) were performed
for 2 hr at room temperature. The secondary antibody incubations with goat anti-rabbit IgG-HRP (1:5000 dilution; Santa Cruz # sc-2004) and donkey anti-goat IgG-HRP (1:2000; Santa Cruz # sc-2020) were performed for 1 hr at room temperature. Blots were visualized by using SuperSignal West Pico chemiluminescent substrate (Thermo-Scientific # 34080) in Fujifilm LAS-4000 imaging system. The membranes were stripped by incubating with Restore Western-Blot Plus Stripping Buffer (Thermo-Scientific # 46430) for 30 min at RT and re-blocked after blotting with each antibody. The same membrane was serially blotted and stripped with all antibodies shown.

**DNase Footprinting on Free DNA and Nucleosomes**

DNase footprinting reactions were carried out as previously described (Soufi et al., 2015b). In brief, 6-FAM end labeled free DNA or nucleosomes (50 ng) were incubated in the presence or absence or the purified ASCL1/E12α (60 nM) and PU1 (100 nM) for 45 min at room temperature in a total volume of 50 μl in DNA-binding buffer (10 mM Tris-HCl pH 7.5, 1 mM MgCl2, 10 μM ZnCl2, 1 mM DTT, 50 mM KCl, 3 mg/ml BSA, 5% glycerol). The binding reactions were then supplemented with an additional 50 μl of 10 mM MgCl2 and 5 mM CaCl2 at room temperature for 1 min (total 100 μl). Binding reactions were treated with DNase-I (Worthington) by addition of 3 μl of DNase-I yielding a final concentration of $9 \times 10^{-5}$ U/μl for free DNA or $1.8 \times 10^{-3}$ U/μl for nucleosomes for 1 min at room
temperature. The reactions were stopped by adding 90 μl of 200 mM NaCl, 30 mM EDTA, 1% SDS buffer immediately followed by phenol extraction. One tenth of reaction volume (20 μl) of 3 M NaOAc (pH 5.2) was added to the reaction before the DNA fragments were extracted with saturated phenol. The DNA fragments were further purified using MinElute PCR purification kit (Qiagen) and eluted in 10 μl of EB buffer. The digested DNA fragments were separated by capillary electrophoresis. Briefly, the DNA fragments (5 μl) were added to 4.9 ml HiDi formamide (Applied Biosciences) and 0.1 ml GeneScan-500 LIZ size standards (Applied Biosciences). After denaturing at 95 °C for 10 min, the samples were run on an ABI 96-capillary 3730XL Sequencer, using G5 dye setting, running a genotyping module with an increased injection time of 30 s and injection voltage of 3 kV. The generated electropherograms were analyzed using the Gene mapper V4.1 (Applied Biosciences) and Peak Scanner module (Thermo Fisher Scientific Cloud).

Protein Microarrays

Full length human proteins were purified as GST fusion proteins from yeast using a high-throughput protein purification protocol as described previously (Zhu et al., 2001). Purified human proteins predicted to bind DNA were arrayed in a 384-well format and printed on FAST slides (Whatman, Germany) in duplicate, as described previously (Hu et al., 2009). To estimate
protein amount printed per spot, prepared chips were incubated with anti-GST antibody conjugated with Cy5 binding targeting the printed proteins GST-tags. The protein microarrays were probed without (- control) or with Cy5-labeled ALBN1, NRCAM, CX3CR1 DNA or NUC using a similar protocol described previously (Hu et al., 2009). In brief, a protein chip was blocked for 3 hr at 4 °C with 3% BSA in 25 mM HEPES pH 8.0, 50 mM KGl, 8 mM MgCl2, 3 mM DTT, 10% glycerol, 0.1% triton x-100 buffer and then incubated with Cy5-labeled ALBN1, NRCAM, CX3CR1 DNA or NUC at a final concentration of 20 nM in 80 μL of binding buffer (10 mM Tris-HCl pH 8.0, 1 mM MgCl2, 10 μM ZnCl2, 1 mM DTT, 50 mM KCl, 3 mg/ml BSA, 5% Glycerol) overnight at 4 °C. The chip was washed once with TBST (0.1% triton x-100) for 5 min at 4 °C, rinsed with water and spun dry. The slides were finally scanned with a GenePix 4000B scanner (Molecular Devices, Sunnyvale, CA) and the binding signals were acquired by aligning an array list over the chip, foreground and background signals for each protein spot were then extracted by the using the GenePix Pro 7 software, and fluorescence minus background (F-B) signals were obtained for each spots.

3D Structural Model Assemblies

MYOG, CREM, and ASCL1/E12α structural models (SMR P15173, SMR Q03060, SMR P50553) were constructed based on SWISS-MODEL sequence alignment (Bienert et al., 2017) with MYOD (pdb 1MDY) showing 74.2% DBD
sequence identity, CREB1 (pdb 1DH3) showing 76.3% DBD sequence identity with CREM1, and NEUROD1 of the NEUROD1/E12α (pdb 2QL2) showing 42.86% DBD sequence identity with ASCL1. Structural models assemblies were aligned to published crystal structures of MYOD, CREB1, and NEUROD1/E12α using Pymol.

**EMSA Processing**

For image processing, we used ImageJ version 2.0.0-rc/1.51f. Before displaying, the images were corrected for brightness and contrast; a linear range was maintained.

**Dissociation Constant Analysis**

Apparent Kₐₛ were derived from two separate binding curves, each representing two experimental replicates as described by (Soufi et al., 2015b). Experimental data was fitted to non-linear regressions “One site – Specific binding”, “One site – Total” or Specific binding with Hill slope” with a Bmax less than 1 constrain within R² values of 0.8-0.99, expressed in nM units using GraphPad Prism 7 software.
To quantify the signal intensity for each spot, the signal intensities were calculated as the fluorescent median intensity minus its local background median intensity (F-B) followed by averaging of printed duplicate spots F-B (Table S2)(Hu et al., 2009). F-B signals were further processed to minimize artifacts (Figure 3.S4A). Auto-fluorescent edging (Figure 3.S4B) was corrected by subtraction of the F-B signal of microarrays incubated with binding buffer (F-B)\textsuperscript{ctrl} from the F-B signal of microarrays incubated with Cy5-labeled samples (F-B)\textsuperscript{Cy5}. Heterogeneity in printed protein amounts per spot (Figure 3.S4C) was considered by estimating relative protein levels per spot amount using chips incubated with anti-GST antibody conjugated to Cy5 (F-B)\textsuperscript{GST}, a \( (F-B)^{GST} < 170 \) signal cutoff was assigned (Figure 3.S4D). The cutoff removed proteins spots with below threshold levels amounts to yield significant binding events from our analysis. To eliminate artifacts resulting from uneven washing and drying of the chips after incubation with Cy5 probes (Figure 3.S4E), we performed a within-chip normalization by the median normalization method. Known chromatin modifiers were removed from our analysis. Finally, to identify proteins that bind to DNA or nucleosomes, an intensity cutoff value of (F-B)\textsuperscript{Cy5} equal to 80 was assigned (Figure 3.S4F), where spots producing a signal greater than the cutoff were identified as “positive hits".
Heatmap Representation of DNA and Nucleosome Binding by Transcription Factors

Heatmaps in Figure 3.5D, E were generated by first quantifying the TFs total bound fraction from EMSAs as described in the methods section “DNA and Nucleosome Binding Reactions” and in (Soufi et al., 2015a). Total bound fraction for each TF was measured for their respective free DNA and nucleosome template shown in Figure 3.S5, at the concentrations shown in Figure 3.S6A. K-means clustering of the TF total bound fraction for free DNA and nucleosomes was done with R studio software and the pheatmap package. We defined a K means clustering function with an optimal number of 6 cluster centers (k); no Z score normalization was applied to the data. Cluster number was further simplified by manually regrouping of highly similar clusters resulting in three clusters.

Data and Software Availability

All the original unprocessed gel images in this manuscript have been deposited with Mendeley and can be accessed with https://doi.org/10.17632/5dnznn9kwt.1.
CHAPTER 4: PERSPECTIVES AND FUTURE DIRECTIONS

4.1 Summary of Major Conclusions and Implications

Early work on FOXA1 interactions with chromatin in vivo and in vitro defined pioneer TFs by their ability to directly associate with nucleosomes and induce local changes in chromatin accessibility, to allow competence for gene expression at the initial stages of cell differentiation (Cirillo et al., 1998; Iwafuchi-Doi and Zaret, 2014; McPherson et al., 1993; Zaret and Carroll, 2011). The subsequent identification of FOXA1 as a direct reprogramming TF (Huang et al., 2011; Huang et al., 2014) then raised questions about the potential pioneering activity of other reprogramming TFs and the role of TF nucleosome binding during cell reprograming.

Our lab previously demonstrated that OCT4, SOX2, and KLF4 (OSK) can access closed chromatin during the initial days of fibroblast to iPSCs reprogramming (Soufi et al., 2012). Furthermore, OSK was found to directly interact with nucleosomes and target nucleosomes-enriched sites in vivo, while cMYC target sites are mainly nucleosome-depleted (Figure 2.4,5). We therefore showed that in vivo OSKM chromatin targeting tendencies are mirrored by the TFs ability to directly interact with nucleosomes in vitro (Figure 2.1). Thus, we established the identity of OSK but not M as pioneer TFs (Chapter 2). Based on these findings, we propose that the inherent ability of these factors to target chromatin is determined by their intrinsic ability to recognize partial motifs on
nucleosomes and by the structural adaptability of the TFs DBD (Soufi et al., 2015b).

Studies correlating TF binding co-occurrence with chromatin accessibility in vivo suggest a hierarchical model of TF-mediated genome re-activation during reprogramming where early expressed developmental TF are also sufficient for reprogramming, such as OSK. These early developmental and now reprogramming TFs preferentially access compacted chromatin, induce chromatin opening and subsequently recruit factors with weak nucleosome binding activity (Mall et al., 2017; Soufi et al., 2012; Soufi et al., 2015b; Wapinski et al., 2013). The reprogramming hierarchical model places TF nucleosome binding ability and thus pioneering activity as a determinant of a TF reprogramming potential. Nonetheless, what dictates the preference of OSK to closed chromatin remains unknown. The studies presented in Chapter 2 and 3 were based on mono-nucleosomes substrates. Therefore, we suggest that a more complex chromatin structure allows a higher level of nucleosome interactions and provides multiple binding sites for TFs that could increase the nucleosome affinity of these factors in chromatin. Therefore, a more complex substrate such as chromatin could explain the preferential targeting of these TFs to closed chromatin.

Following our findings of OSKM pioneer activity in vivo and in vitro during iPSC reprogramming, I extended our analysis to include TFs driving the direct cell reprogramming of fibroblast to neurons, hepatocytes, macrophages and B
cells (Chapter 3). We purified 9 cell reprogramming TFs under denatured conditions and measured their apparent dissociation constant on short dsDNA containing the respective TFs motif to validate specific DNA binding of the TFs compared to non-specific DNA (Figure 3.S2C-D). We showed that each TF is able to specifically interact with their respective motif by EMSA, competition experiments and DNase I footprinting (Figure 3.3). Additionally, by estimation of the percent of fraction bound by the TF to free DNA at 1 nM we estimate that 90% of the purified FOXA1, PU1 and ASCL1/E12α are properly folded since a 1:1 molar ratio of TF:DNA shows almost complete saturation of binding to the free DNA probe (Figure 3.2B-D). We estimate GATA4 and BRN2 to be 75% folded by this method (Figure 3.2E-F). Still, more quantitative experiments such as isothermal titration calorimetry can be done to determine the stoichiometric ratio of TF:DNA binding in order to accurately quantify the percentage of folded TF. Assessment of the nucleosome binding of these cell reprogramming TFs revealed these factors display a range of nucleosome binding affinities that correlate with the reprogramming strength of TFs (Fernandez Garcia et al., 2019). Moreover, by DNase I footprinting experiments, we observe that upon TF binding, nucleosomes become DNase I protected at TF binding sites but also show enhancement of DNase I sensitivity at other regions (Figure 3.3). Therefore, suggesting that TF binding to nucleosomes can induce allosteric conformational changes in the histone octamer that lead to changes in nucleosomes rotational positioning.
Transcription factors cannot be functionally understood without knowledge of the DNA sequences that they bind. Therefore, many efforts have been made to understand the DNA binding preferences of TFs resulting in the creation of TF motif databases such as JASPAR, TRANSFAC, UNiPROBE, and CIS-BP that use TFs ontology and structural databases (Jolma et al., 2013; Mathelier et al., 2014; Matys et al., 2006; Yin et al., 2017). Still, nucleosome-TFs studies have been mainly restricted to low-throughput analyses due to the requirement of specific DNA sequences for TFs binding. Moreover, the complexity of TFs post-translational modifications, TF dimerization, and TF binding partners, all of which contribute to TF stability and DNA binding, have made it challenging to screen for sequence-specific nucleosome binders and to determine the underlying features that permit or inhibit nucleosome binding. This study precisely addressed this gap in our knowledge.

Using protein microarray technology, we screened 593 TFs for nucleosome binding on sequence specific DNA and identified factors with previously unreported pioneer activity, including the AP2 TF family (Figure 3.5C). In solution validation, of sequence specific nucleosome-TF interactions identified from the protein microarray screen, showed that while nucleosome binding TFs typically have a high DNA binding affinity this does not equate to nucleosome binding activity (Figure 3.5D). Based on these findings and earlier work on pioneer FOXA1 (Cirillo et al., 2002) we suggest that DNA curvature induced by
association with the histone octamer or with histones themselves plays a role facilitating TFs-nucleosome binding interactions.

TFs are classified by their DBDs into 10 structural superclasses that represent the topology of the TFs DBD. The 3 largest superclasses, zinc-coordinating, helix-turn-helix, and basic domain, account for 90% of all human TFs. Following in size are the superclasses containing β-sheet structures including the immunoglobulin fold DBDs, and the β-hairpin exposed superclasses (Lambert et al., 2018; Wingender et al., 2015). My comparison between strong and weak nucleosome binders revealed that strong nucleosome binding factors engage DNA through a short anchoring α-helix while weak TFs engage DNA through an extended anchoring α-helix (Figure 3.6A,B). A long anchoring α-helix that extends past the diameter of the DNA can be incompatible with nucleosome structure and cause steric hindrance to the association of the histone octamer with DNA. It has now become evident that pioneer TFs fall within superclasses with α-helical structures (Figure 3.6).

We also identified a second group of weak nucleosome binders that contains DBDs mainly composed of β-sheets and that interact with DNA through short helical twists or unstructured loops. This group of non-pioneer TFs include TFs of the immunoglobulin-like superclass family. Therefore, this group of non-pioneer TFs are part of the β-sheet containing TF superclasses (Figure 3.6C). Not surprisingly, these TFs which include p63 and TBX5, have been shown to be insufficient for direct cell reprogramming protocols versus. For example, p63 is
insufficient for keratinocyte reprogramming and requires pioneer TFs AP2 or KLF4 (Li et al., 2019; Lin-Shiao et al., 2019). Additionally, reprogramming into cardiomyocytes with MEF2A and TBX5 is also insufficient for reprogramming and requires pioneer TF GATA4 (Ieda, 2013; Ieda et al., 2010). MEF2A is classified as a MADS-box TF which contains a DBD structure mainly composed of α-helical structures exposed by β-sheet scaffold (Santelli and Richmond, 2000; Wu et al., 2010). Comparison of these factors with non-pioneer TBX20 (Figure 3.5D), suggests that β-sheet-containing DBDs are insufficient for nucleosome binding and thus cell reprogramming.

Structural comparison of weak versus strong nucleosome binders revealed key structural differences between pioneer and non-pioneer factors. Identification of such features revealed DNA binding modes associated with nucleosome binding strength and reprogramming dominance (Fernandez Garcia et al., 2019). The findings from my thesis work provide strong evidence to suggest that the intrinsic nucleosome binding preferences of pioneer TFs contribute to their hierarchy of cellular reprogramming (Fernandez Garcia et al., 2019). Furthermore, the work presented here gives insight into how such a diverse group of TFs from multiple structural superclasses engage in nucleosomal interactions via common structural features.

My thesis works provides a knowledge base that allows prediction of the TFs with pioneer and reprogramming potential based on their TFs DBD structures. Such knowledge may be used to improve the generation of induced
reprogrammed hematopoietic stem and progenitor cells (iHSPCs), since efficient generation of Induced HSPCs from somatic cells has yet to be developed (Ebina and Rossi, 2015). Runx3 has been proposed to initiate chromatin accessibility during memory cytotoxic T lymphocyte differentiation (Wang et al., 2018), yet RUNX is insufficient for reprogramming protocol toward hematopoietic lineages and reprogramming (Ebina and Rossi, 2015). Current RUNX-based strategies for iHSPCs reprogramming using RUNX, from closely related lymphoid progenitors, require the use of at least 7 TFs for efficient reprogramming (Wang et al., 2018). RUNX DBD adopts an immunoglobulin fold conformation (Tahirov et al., 2001) and can therefore be predicted as a non-pioneer TF. We speculate that RUNX inability to efficiently reprogram cells is due to structural similarities with non-pioneer TFs. DBD-based prediction of pioneer TFs will instruct the informed selection of better reprogramming TFs cocktails for the reprogramming of cells yet to be produced.

4.2 Future directions

Advancement in structural biology techniques, growth in the number of nucleosome structures, and proteomic based mass-spectrometry has allowed the analysis of nucleosome-protein complexes with higher resolution and has provided new insight the landscape of the nucleosome interactome (Kale et al., 2019). These advancements have revealed the principles of nucleosome recognition and the effect of histone and DNA post translational modifications on
nucleosome recognition by epigenetic factors (Bartke et al., 2010; Dann et al., 2017; Makowski et al., 2018; Nguyen et al., 2014). However, these studies mainly provided understanding of non-specific nucleosome binding proteins by relaying on nucleosomes assembled with DNA sequences shown to restrict TFs nucleosome binding, due to high rigidity (Appendix B) (Takizawa et al., 2018). Currently, there is no published structure of full-length sequence-specific TFs bound to nucleosomes, which therefore limits our understanding on the nucleosome recognition mode of full-length, sequence-specific factors.

The development of nucleosome libraries assembled with diverse DNA sequences bypassed the limitations of TF DNA sequence specificity (Zhu et al., 2018). The study revealed distinct nucleosome interacting behaviors of TF DBDs based on sequence-specific DNA motif analysis. Still, further experiments need to be done for the validation of such nucleosome interacting behaviors. Structural manipulation of the DNA anchoring α-helix of pioneer TFs DBDs followed by nucleosome affinity assays will determine the role of the proposed secondary structures in dictating nucleosome interacting behaviors (Chapter 3). Conversely, it would be interesting to determine whether non-pioneer TFs can acquire pioneering activity and reprogramming activity by substitution of DNA-anchoring loops with an anchoring α-helix.

Different mechanisms have been suggested for the action of pioneer TFs to induce local chromatin changes in DNA accessibility. For example, comparison of FOXA with histone 1 (H1) showed that the FOXA DBD contains a
winged-helix domain resembling H1 (Clark, 1993). This extends to other FOX family TFs such as FOXE1 and FOXO1, which also interact with nucleosome through a winged-helix domain (Hatta and Cirillo, 2007). Interestingly, H1 and FOX TFs execute opposing functions – H1 compacts nucleosomes while FOXA induces DNA chromatin accessibility (Cirillo et al., 2002; Osipova et al., 1980). These differences are potentially dictated by accessory domains outside of the DBD since both H1 and FOXA1 interact near the nucleosome dyad (Chaya et al., 2001; McPherson et al., 1993; Sekiya et al., 2009). For instance, H1 is known to interact with linker DNA while FOXA is not (Allan et al., 1986; Thomas, 1999). Nonetheless, structural similarities between FOXA and H1 suggest FOXA-displacement of H1 may play a role in the pioneer function of FOXA1 (Iwafuchi-Doi et al., 2016).

Our lab has also shown that FOXA1 contain a highly conserved C-terminal \(\alpha\)-helix that establishes direct histone contacts with the histone octamer (Cirillo et al., 2002). Deletion of FOXA \(\alpha\)-helix in the C-term (FOXA1\(\Delta\alpha\)-helix) inhibits FOXA1 chromatin accessibility activity in nucleosomes arrays in vitro (Cirillo et al., 2002). Therefore, the presence of FOXA C-term-\(\alpha\)-helix suggests the existence of external DBD domains that mediate nucleosome binding and the chromatin opening activity of pioneer TFs. Interestingly, preliminary experiments measuring dissociation constants of FOXA1\(\Delta\alpha\)-helix showed it retain nucleosome affinity similar to FOXA1 WT (Appendix C, Figure C3). Additional experiments on FOXA1-DBD fragment showed that the FOXA DBD alone is sufficient for
nucleosome binding, albeit with a lower affinity (Cirillo et al., 2002). Preliminary
work from our lab indicates that a FOXAΔα-helix mutants triggers an increase in
chromatin compaction at FOXA1 target sites in undifferentiated mouse embryonic endoderm (manuscript submitted). These results indicate that
domains outside of FOXA1-DBD are needed for chromatin opening (Appendix
C). Similarly, studies of the pioneer TF EBF1 have also shown EBF1 promotes
chromatin accessibility of B-cell specific genes (Boller et al., 2016; Li et al.,
2018). Detailed analysis of EBF1 ΔC-term mutants showed a decrease in
chromatin accessibility and limited the co-occupancy of secondary TFs at these
genomic regions (Boller et al., 2016). Therefore, EBF1 C-term confers chromatin
opening ability. These results show that nucleosome binding ability can be
uncoupled from the chromatin opening activity of pioneer TFs. Based on these
findings we suggest that accessory domains external to TFs DBDs mediate the
establishment of TF-histone interaction and disruption of nucleosome-
nucleosome histone contacts to induce local chromatin accessibility.

As discussed in Chapter 3, through my thesis work I identified previously
unreported pioneer TFs and confirmed the pioneering activity of proposed
reprogramming TFs that direct reprogramming toward three different cell types
(Fernandez Garcia et al., 2019). Identification of pioneer TFs form diverse
structural families now place us in position to determine whether these factors
also interact with histones and which domains are involved in these interactions.
Studies comparing pioneer TFs specific nucleosome binding with non-specific
nucleosome binding proteins will determine whether TFs also establish similar histone-nucleosome interactions through association of the histone acidic patch (Figure 1.4) or through distinct interactions (McGinty and Tan, 2016). Detailed structural studies in TF-histone interactions will aid in the elucidation of the similar or differential mechanisms by which pioneer TFs interact with nucleosomes and will determine whether nucleosome binding by TFs stabilize or destabilize nucleosome structures to induce chromatin accessibility. Future research on pioneer TF mechanisms of nucleosome binding will elucidate the initial steps silent chromatin targeting by TFs and will unveil the mechanisms by which cells reactivate developmentally silenced gene networks.

At a higher level of DNA compaction our lab has shown that OSK cannot access highly compacted chromatin that is enriched for H3K9me3, thus revealing the sensitivity of these pioneer TFs to highly compacted chromatin subtype (Becker et al., 2017; Soufi et al., 2012). As the chromatin sensitivity of these pioneer factors limits their ability to induce activation of late reprogramming genes resulting in incomplete activation of new cell fate regulatory networks (Becker et al., 2017; Soufi et al., 2012). Knowledge of the histone marks, DNA modifications, and chromatin-specific interacting proteins that influence pioneer TFs chromatin sensitivity may be used to enhance cellular reprogramming. These gaps in knowledge can be addressed by the use of DNA-barcoded mono-nucleosome libraries of chemically-modified histones, histone variants and methylated DNA (Dann et al., 2017; Nguyen et al., 2014). Such libraries would be

163
generated with native DNA sequences specific to TF of interest (Figure 3.1) in order to profile the biochemical susceptibility of a pioneer TF to diverse types of sequence-specific nucleosomes.

In addition, we have shown that although some reprogramming TFs can associate with closed chromatin, not all TFs have nucleosome binding activity (Chapter 2). c-MYC was shown to have weak nucleosome binding and a minority of cMYC target sites within nucleosome-enriched chromatin in vivo are also enriched for Oct4 and Sox2 binding sites (Soufi et al., 2015b). Therefore, we suggest that cMYC associates with chromatin through cooperative interactions with pioneer TFs. Similarly, CEBPA has been associated with the induction of chromatin accessibility during B-cell reprogramming, although we showed that CEBPA displays the weakest nucleosome binding affinity of the TFs tested in Chapter 3. Furthermore, a majority of CEBP target sites have shown to be dependent on PU1 (van Oevelen et al., 2015). Therefore, future studies should also address the role of cooperative binding of non-pioneering TFs at closed chromatin in inducing chromatin accessibility during reprogramming.

The question of which reprogramming TFs can directly induce chromatin opening in an ATP-independent mechanism, and whether chromatin opening activity is limited by nucleosome affinity can be directly tested in vitro by high precision mapping techniques such as DNAse footprinting using compacted chromatin arrays. Our lab has shown that FOXA1 and GATA4 can bind nucleosomes. GATA4 showed a lower nucleosome affinity and decreased
chromatin opening compared to FOXA1 on H1 compacted arrays (Cirillo and Zaret, 1999a). DNAse footprinting experiments of in vitro reconstituted chromatin can therefore elucidate the dependence of chromatin opening activity on nucleosome binding affinities of the newly identified pioneer TFs. Furthermore, coupling of these experiments with newly developed DNA-sequencing technologies, in vitro, will provide further details and mapping of the ATP-independent mechanism by which pioneer TFs induce chromatin opening.

4.3 Final Remarks

Just about over 10 years ago the explosion and refinement of high-throughput genome-wide technologies have highlighted the importance of chromatin compaction and 3D genome folding in the establishment of cell identity. Discoveries of genome wide changes in chromatin accessibility during cell differentiation shifted the perception of transcription factors as merely static DNA binding proteins into proteins that target nucleosomes and lead to local structural perturbations. Now, thanks to the extensive identification and validation of pioneer TFs in addition to FOXA, after much skepticism pioneer TFs have now become widely accepted as mediators of chromatin accessibility. Therefore, pioneer TFs nucleosome and chromatin binding preference should be incorporated into the functional analysis of TFs in order to understand how TFs regulate gene expression in the nucleus, a highly compacted and tightly regulated environment.
REFERENCES


the SCL:E47bHLH heterodimer to hematopoietic-specific transcriptional targets. Cell Rep 4, 135-147.


Feng, B., Ng, J.H., Heng, J.C., and Ng, H.H. (2009). Molecules that promote or enhance reprogramming of somatic cells to induced pluripotent stem cells. Cell Stem Cell 4, 301-312.


specific regulatory codes for homeostatic and cell cycle gene batteries. Genome Res 23, 592-603.


APPENDIX A: ENZYMATIC 3’ DNA LABELING WITH CY5 FLUOROPHORE

This appendix provides a detailed overview of the enzymatic 3’ DNA labeling with Cy5 fluorophore protocol used in Chapters 2 and 3. The protocol has been robustly used by other labs, including Doris Wagner and Shelly Berger’s labs. The protocol was developed and optimized by myself with guidance from Dr. Zaret.

I. Objective

Label DNA at the 3’ end with Cy5 fluorophore for EMSA experiments and for Cy5 fluorescence detection.

II. Reagents and Kits

Reagents

Cy5-dCTP [1mM]
Unlabeled dATP [100 mM]
Unlabeled dTTP [100 mM]
Unlabeled dGTP [100 mM]
Klenow fragment (-exo)
Purification column illustra MicroSpin G-25 columns (GE healthcare)
Note: If using DNA longer than ~150 bp DNA, cleanup can be done with Qiagen PCR Purification Kit.

**DNA Sequences**

Double stranded DNA (dsDNA) should be longer than 32 bps otherwise the purification (removal of unincorporated dNTPs) by size exclusion will not be efficient as the size difference between dNTP and DNA will be minimal. Designed oligonucleotides should contain sticky ends of at least 2 bp on each side, the last of which will be labeled with a Cy5 labeled C. This protocol works for both annealed ssDNA (top strand) and ssDNA (bottom strand) or dsDNA fragments.

**Buffers**

10X Annealing Buffer: 200 mM Tris, pH 7.6; 500 mM NaCl; 1 mM DTT; 10 mM EDTA

**III. Protocol**

Note: Skip to Step 2 if starting from dsDNA (Cy5 labeling reaction).

1. Anneal complementary strands of ssDNA:

   a. Prepare annealing reaction:

      10 uL ssDNA (top) 100 µM stock
      10 uL ssDNA (bottom) 100 µM stock
      5 uL 10X annealing buffer
      25 uL ddH2O
b. Heat the reaction at 70°C for 10 min. on a heat block. Turn the heat block off and let the sample gradually cool down in the heat block overnight.

c. Determine the DNA concentration on a Nanodrop.

2. Labeling reaction:

   a. Prepare 5X dNTP mix:

<table>
<thead>
<tr>
<th>Stock</th>
<th>Final Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 mM dATP</td>
<td>20 mM</td>
</tr>
<tr>
<td>100 mM dTTP</td>
<td>20 mM</td>
</tr>
<tr>
<td>100 mM dGTP</td>
<td>20 mM</td>
</tr>
<tr>
<td>1 mM Cy5-dCTP</td>
<td>20 μM</td>
</tr>
</tbody>
</table>

   b. Prepare reaction:

<table>
<thead>
<tr>
<th>Stock</th>
<th>Final Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 dsDNA (50 ug)</td>
<td>[1.260 μM ] final concentration</td>
</tr>
<tr>
<td>10X NEB Buffer #2</td>
<td>1X</td>
</tr>
<tr>
<td>5X dNTP Mix</td>
<td>1X (4mM dATP, dTTP, dGTP, 0.004 mM Cy5-dCTP)</td>
</tr>
<tr>
<td>Klenow Fragment (-exo) 5 U/µL</td>
<td>0.5 U/µL</td>
</tr>
<tr>
<td>ddH2O</td>
<td>Subtract everything from total volume</td>
</tr>
<tr>
<td>Total volume</td>
<td>Adjust to DNA final conc.</td>
</tr>
</tbody>
</table>

   Calculate how much DNA you want to label: 50 ug 160 bp DNA

   Then calculate the total volume needed to have 50 ug DNA with a final concentration of 1.26 μM. Do not exceed 10 μL if DNA < 150 bp. Do not exceed the maximum loading volume if using G-25 spin columns since this will decrease the purification efficiency. Additionally, if more DNA is
desired make sure to keep the ratio of DNA:Cy5-dCTP the same
otherwise efficiency of labeling can be compromised.

3 Calculate how much 10X buffer, 5X dNTP mix and Klenow you need to
add to have 1X, 1X and 0.5 U/µL final concentrations.

4 Subtract the volume of DNA, buffer, dNTPs, and Klenow from the total
to get the volume of H₂O to add.

c. Incubate at 37˚C for 1 hr in a water bath.

3. DNA purification

a. DNA < 150 bp: proceed as per the illustra MicroSpin G-25 column
manufacturer protocol, with a few modifications:

i. Load the sample into the column and add an additional 10 µL
ddH₂O for a final volume of 20 µL.

ii. Spin at 2,800 RPM for 1.5 min.

iii. To verify purity of the sample run on a 10 % Acrylamide, 0.5X TBE
gel (non-denaturant). Make a ∼1:60 dilution of your sample and
load 10 µL of the dilution + 2 µL 30% sucrose. Additionally, run one
lane with 2 µL 6X DNA loading dye just to have an estimate of
how far off your sample has run. Run the gel in 0.5X TBE at 90 V
for ∼1.5 hr. Scan the gel in a fluorimeter scanner with Cy5
detection. Note: be sure to use sucrose when loading your sample,
as DNA loading dye will obscure the Cy5 signal.

b. DNA ≥ 150 bp: purify using Qiagen PCR Purification Kit.
i. Note: Do not load more than 10 μg DNA per column otherwise purification yields will be lower.

ii. Elute 10 μg in 50 μL EB buffer.
APPENDIX B: WIDOM 601 DNA SEQUENCE AS A TEMPLATE FOR TF-NUCLEOSOME INTERACTIONS

This appendix contains unpublished analyses which test FOXA1 nucleosome binding affinity for Widom 601 asymmetric nucleosomes containing or lacking FOXA1 binding sites at multiple nucleosome rotational phasing positions. Experimental design, reagents, experiments, and analysis was done by me.

I. Objective

The goal of the experiments presented here is to determine the viability of using a synthetic DNA sequence, such as the Widom 601 asymmetric DNA, as a universal and modifiable template for nucleosome assembly. Such a template will be useful for testing the binding affinities of multiple transcription factors on nucleosomes containing binding sites at various rotational positions. The Widom 601 sequence has been the standard for in vitro nucleosome biochemical assays due to its exceptional histone octamer affinity and favorable DNA bending and distortion, both of which allows stable nucleosome assembly in vitro (Vasudevan, J. Mol. Bio, 2010; Chua et al., NAR, 2012; Lowary & Widom, J. Mol. Bio, 1998). The 601 sequence is extensively used for nucleosome structural analyses (Makde et al., 2010) nucleosome interaction assessment (Lone, et al, PLOS Genetics, 2013), and nucleosome mechanics (Chua et al., NAR, 2012) and
disassembly studies (Gansen et al. PNAS 2009). Therefore, we did preliminary work testing FOXA1 binding to 601 nucleosomes constraining or lacking a FOXA binding site at multiple rotational positions.

II. Reagents

DNA Sequences

601 asymmetric-162 bp

5’-ATCGATGTATATATCTGACACGTGCTGGAGACTAGGGAGTAATCCCTTGGCGGTTAACACCGGGACACCGGCTAGTGCGCTTATAGCGGTGTAGCTGCTAGCAGCTACGAGCGGCTCTGGCACCAGGGATTCTGAT-3’

601-eG (D)

5’-ATCGATGTATATATCTGACACGTGCTGGAGACTAGGGAGTAATCCCTTGGCGGTTAAAGAATTGGTTAGTTAGTACGTGCGTTTAAGCGGTGCTAGAGCTGCTCTACGACAAATTGAGCGGCTCCTGGCACCAGGGATTCTGAT-3’

601-eG (D-1)

5’-ATCGATGTATATATCTGACACGTGCTGGAGACTAGGGAGTAATCCCTTGGCGGTTAAAGAATTGGTTAGTTAGTACGTGCGTTTAAGCGGTGCTAGAGCTGCTCTACGACAAATTGAGCGGCTCCTGGCACCAGGGATTCTGAT-3’

601-eG (D-5): 5 bp shift causes a shift of 34° degrees.

5’-ATCGATGTATATATCTGACACGTGCTGGAGACTAGGGAGTAATCCCTTGGCGGTTAAAGAATTGGTTAGTTAGTACGTGCGTTTAAGCGGTGCTAGAGCTGCTCTACGACAAATTGAGCGGCTCCTGGCACCAGGGATTCTGAT-3’

601-TA mut at superhelical location -0.5, 1.5

5’-ATCGATGTATATATCTGACACGTGCTGGAGACTAGGGAGTAATCCCTTGGCGGTTGAAAGACGGGGACACCGGCCTCCTGGCACCAGGGATTCTGAT-3’

601-eG (D)-TA mut at superhelical location -0.5, 1.5

5’-ATCGATGTATATATCTGACACGTGCTGGAGACTAGGGAGTAATCCCTTGGCGGTTGAAAGACGGGGACACCGGCCTCCTGGCACCAGGGATTCTGAT-3’
Purified Proteins

Mouse FOXA1 full-length

III. Assessment of FOXA1-specific binding at multiple rotational positions on 601 nucleosomes

Goal

To compare FOXA1 binding to Widom 601 free DNA and Widom 601 nucleosomes with or without a single FOXA1 binding motif and determine whether different rotational phasing inhibits or allows FOXA1 binding to nucleosomes.

Approach

pUC19-601 plasmid (gift from Karolyn Luger) was used as a template for mutagenesis of the FOXA eG binding site at positions near the 601 nucleosome dyad (D), 1 bp away from the dyad towards the DNA super helix (SH) 1.5, or 5 bp away from dyad toward SH 1.5 (Figure B1B). The FOXA1 eG site in the dyad is considered assessible. From the dyad, a 1 bp shift of the eG site in the nucleosome causes a 34° rotation and a 5 bp shift causes it to be inaccessible to FOXA1, as shown by molecular modeling (Figure B1C). The eG site orientation was placed so that wing 2 of the FOXA1-DBD structure is closest to the dyad,
similar to the orientation that FOXA1 adopts on the alb1- N1 nucleosomes (Figure B1D).

PCR with Cy5 labeled primers was performed to generate 162-bp of 601, 601-eG (D), 601-eG (D-1), and 601-eG(D-5) DNAs. The sequences were then assembled into nucleosomes by urea-salt gradient dilution with purified recombinant, full-length human histones. FOXA1 free DNA and nucleosome binding was determined by Electromobility Shift Assay (EMSA).

Results

FOXA1 binds with higher affinity to 601 DNA containing the eG site at positions D, D-1 and D-5 (Figure B2A). In contrast, FOXA1 shows minimal binding to all 601 nucleosomes containing or lacking the eG site (Figure B2B).

IV. Assessing FOXA1 binding to less rigid 601 nucleosomes

Goal

To make flexible 601 nucleosomes that are less rigid and more closely resemble natural nucleosome structures to allow FOXA1 binding.

Approach

Widom 601 nucleosomes have been characterized as very rigid nucleosomes compared to nucleosomes assembled with the 5S DNA sequence or with the albN1 enhancer. Therefore, to generate a more natural and flexible
nucleosome I mutated the H3/H4 TA steps super helical locations -0.5 and 1.5 closest to the dyad (Figure B3A,B). H3/4 TA steps serve as histone-DNA interaction points. These two TA steps were mutated to GC in order to increase nucleosome flexibility near dyad.

Results

Mutagenesis of 601 H3/4 TA steps near the dyad and eG site did not affect FOXA1 specific binding to 601 free DNA. EMSA of FOXA1 to 601-eG (D) TA mut nucleosomes showed minimal FOXA1 nucleosome binding enhancement compared to 601-eG(D) nucleosomes (Figures B2B-C) and nucleosome saturation could only be observed at 90 nM FOXA1 concentrations.

V. Conclusions

Here I have shown that the Widom 601 asymmetric DNA can be used to test TF-specific binding with the insertion of TF target site and that mutagenesis of 601 does not disrupt nucleosome assembly. Nonetheless, 601 nucleosomes showed to be too rigid and inhibited FOXA1 nucleosome binding ability, regardless of the DNA binding motif orientation within the nucleosome structure. These conclusions are in agreement with a recent publication comparing the albN1 nucleosomes to 601 nucleosomes by cryo-EM (Takizawa et al., 2018).
**VI. Figures**

**Figure B1 | Design of 601-eG nucleosomes.**

(A) Schematic diagram showing plasmid generation to DNA purification workflow of Cy5-601 DNA sequences and FOXA1 eG binding site insertion.

(B) 3D representation of the 601 DNA sequences used for nucleosome assembly containing FOXA1 eG binding site (yellow) and DNA nucleotide sequence of FOXA1 with highlighted FOXA1 specific contacts (green and red) and DNase protected nucleotides shown by McPherson et al 1993, Cell.

(C) Modeled crystal structure of 601 nucleosomes (pdb 3LZ0) with docked FOXA3 DBD (pdb IVNT) by alignment of FOXA3 eG containing DNA from crystal structure with 601 DNA at eG site insertion positions shown in (B).

(D) AlbN1-DNA sequence with MNase and DNAase I cleavages in liver nuclei. Arrows indicate sites of enhanced MNase cleavage in liver nuclei; large arrows represent the strongest cleavages that define the real nucleosome boundaries. Sequences contained within the N1site (red box) which were inserted into 601 (B) show the orientation of the FOXA3 secondary structural folds with respect to the dyad. Bars indicate areas protected from cleavage by DNAase I in liver nuclei, and dots indicate positions of enhanced liver-specific cleavage by DNAase I. (Liu et al., 1991, McPherson et al 1993, Cell)
Figure B2 | The Widom 601 nucleosomes fail to Recapitulate FOXA1 nucleosome binding regardless of TF binding motif rotational positining.

(A) Representative EMSA showing the affinity of increasing amounts of FOXA1 WT to 160 bp 601 free DNAs. Black arrowheads indicate free DNA and DNA:FOXA1 complexes.

(B) Representative EMSA showing the affinity of increasing amounts of FOXA1 WT to 160 bp 601 free DNAs and nucleosomes. Black arrowheads indicate free DNA and DNA:FOXA1 complexes; white arrow indicate nucleosomes and nuc:FOXA1 complexes.
Figure B3 | Mutagenesis of 601 H3/4 TA steps does not enhance FOXA1 nucleosome binding.

(A) 601 nucleotide sequence showing the dyad position (Φ) and highlighted H3/4 TA steps (orange).

(B) Crystal structure of 601 nucleosomes (pdb 3LZ0) showing H3/4 TA steps (orange base pairs) and mutated super helical locations (SHL) -0.5 and 1.5.
APPENDIX C: FOXA1 CHROMATIN OPENING-DEFICIENT MUTANTS BINDING TO NUCLEOSOMES

This appendix contains unpublished analyses which measure FOXA1 WT and FOXA1 C-terminal mutant affinity to \textit{albN1} DNA and nucleosome. While FOXA1 double mutant protein expression plasmids and FOXA1 purifications were done by Naomi Takenaka, I generated the single FOXA1 mutant plasmids for protein expression, EMSA experiments, dissociation curve analysis, and histone octamer pulldown assays.

I. Objective

Unpublished work from our lab have shown that FOXA1 contains two highly conserved lysines, K270 and K414, that interact with the nucleosome core histone octamer by FOXA1–core histone octamer crosslinking followed by mass spectrometry (in revision; Iwafuchi-Doi et al, Nature Genetics). mFOXA1 residues K270 and K414 reside within the C-terminus of FOXA1. K270 is located in close proximity to the FOXA1 DBD and within highly conserved EKQ (269-271) residues while residue K414 resides within the FOXA1 C-terminal \( \alpha \)-helix structure (In revision; Iwafuchi-Doi et al, Mol Cell, 2019). Deletion of the FOXA1 C-term \( \alpha \)-helix was shown to decrease DNA accessibility at FOXA1 bound sites \textit{in vivo} observed by ATAq-seq experiments, while in \textit{in vitro} experiments it
causes chromatin opening inhibition and loss of FOXA1 histone interactions (In revision; Iwafuchi-Doi et al, Nature Genetics). Therefore, to determine the role of the K270 and K414 residues to FOXA1 nucleosome binding I measured free and DNA affinity of FOXA1 C-term mutants to 160 bp *albN1* DNA and nucleosomes.

**II. Reagents**

**DNA Sequences**

See section 3.11 Key Resource Table -160 bp *ALBN1-DNA*

**Purified Proteins**

- FOXA1 WT
- FOXA1 DBD aa168-268,
- FOXA1 eDBD aa144-294
- FOXA1 eDBD aa144-294: EKQ (269-271) AAA
- FOXA1 (A415P, A419P)
- FOXA1 ΔF413-Q421
- FOXA1 EQK (269-271) AAA
- FOXA1 EQK (269-271) AAA + (A415P, A419P)
- FOXA1 EQK (269-271) AAA + Δ413-421

**III. FOXA1 chromatin opening-deficient mutants binding to short dsDNA**

**Goal**

To compare the ability of WT and chromatin opening deficient FOXA1 mutants to specifically bind to short dsDNA.
Approach

FOXA1 WT and mutant plasmids were generated and recombinant proteins were purified (Figures C1A-C) from *E. coli* under denaturing conditions as described above. Purified recombinant proteins were tested and validated for specific and non-specific DNA binding to short ds-DNA containing or lacking the FOXA1 eG site by EMSA (Figure C1D).

Results

FOXA1 chromatin opening deficient mutants bind short dsDNA DNA with specificity and similar affinity compared to WT.

IV. FOXA1 chromatin opening-deficient mutants bind to 160 bp *albN1* free DNA and nucleosomes

Goal

Determine whether FOXA1 K270 or K414, are required for FOXA1 nucleosome binding ability.

Approach

Electromobility shift assays (EMSA) were used to determine the apparent dissociation constants of WT and mutant FOXA1 to 160 bp *albN1* free DNA and
nucleosomes (Figure C2A-F). Dissociation constants were quantified by two methods: total dissociation constant (total $K_D$), determined by decrement of free probe, and specific dissociation constant (specific $K_D$), determined by the appearance of specific TF-complexes (Soufi et al., 2015b) (Figures C3A-C).

Results

FOXA1 mutants bind 160 bp $albN1$ DNA and nucleosomes with similar affinities within the low nanomolar range of 1-2.5 nM for nucleosomes compared to WT, albeit with less affinity for nucleosomes compared to DNA, as was also seen for FOXA1 WT.

V. FOXA1 EKQ (269-271) residues contribute to FOXA1 interactions with the histone octamer

Goal

Determine FOXA1 EKQ (269-271) amino acid contribution to FOXA1-histone octamer interaction.

Approach

Constructs of FOXA1 truncations containing either the DBD lacking residues EKQ 269-271, an extended DBD (eDBD), or eDBD with EKQ (269-271) AAA mutation were generated (Figure C4A). These plasmids were then used for
protein expression and purification from *E. coli* under denatured conditions (Figure C4B). Purified recombinant proteins where initially tested and validated for specific and non-specific DNA binding to short dsDNA containing or lacking the FOXA1 eG site (Figure C4C). Histone octamer pulldown assays are described in Cirillo et al., 2002, Mol Cell (Figure C4C). Initially we compared histone interaction of FOXA1 full-length WT with FOXA1 full-length EKQ (269-271) AAA mutant followed by comparison of FOXA1 DBD lacking residues EKQ 269-271 and extended DBD (eDBD).

**Results**

FOXA1 eDBD, short DBD, and eDBD-AAA$^{mut}$ show similar affinities for short dsDNA containing a FOXA eG binding site (Figure C4C). We show that full-length FOXA1 is able to interact with histone octamers, in agreement with previous published experiments (Cirillo et al., 2002), but AAA substitution of EKQ residues in full-length FOXA1 causes a partial decrease in FOXA1-histone octamer interaction (Figure C4E). Further truncation of FOXA1 eDBD fragment which removes adjacent amino acids, including the EQK residues, greatly inhibits FOXA1 DBD fragment interactions with histone octamers (Figure C4F).

**VI. Conclusions**

Here we show that purified FOXA1 WT and chromatin opening-deficient mutants bind nucleosomes with similar affinities. Nonetheless mutation of residues EQK, adjacent to FOXA1 DBD, limit histone octamer interaction of
FOXA1. Previous work from our lab has shown that FOXA1 DBD fragment is sufficient for nucleosome binding but show decrease affinity compared to full-length FOXA1 (Cirillo et al., 2002). Moreover the DBD alone is insufficient for FOXA1 chromatin opening ability (Cirillo et al., 2002). Therefore, comparison of histone octamer interaction by pull down assay between FOXA1 eDBD (includes EQK) and short DBD fragments (exclude EQK), show that truncation of adjacent amino acids such as EQK result in inhibition of FOXA1-DBD interaction with histone octamers in vitro. These experiments provide evidence that FOXA1 nucleosome binding and chromatin opening functions can be uncoupled and that domains outside of the DBD might play a role mediating FOXA1 histone interactions.
VII. Figures

A. HPFSINNLMSSEQQKLDKLYEQALQYSPYGATLPASLPLGSASVATRSPIEPSALEPAYYQGVYSRPVLNTS

B. WT, DBD (168-268), eDBD (144-294), P/P, ΔαH, mAAA, P/P + mAAA, ΔαH + AAA

C. Gel images for WT, P/P, ΔαH, mAAA, P/P + mAAA, ΔαH + mAAA, WT, P/P, ΔαH, mAAA, P/P + mAAA, ΔαH + mAAA

D. WT, P/P, ΔαH, mAAA, P/P + mAAA, ΔαH + mAAA

E. WT, P/P, ΔαH, AAA\textsuperscript{mut}, P/P + AAA\textsuperscript{mut}, ΔαH + AAA\textsuperscript{mut}
Figure C1 | FOXA1 WT and mutants bind short dsDNA with specificity

(A) Amino acid sequence of FOXA1, highlighting the predicted α-helix and conserved region III (CR III) in red.
(B) Representation of FOXA1 WT, truncations, mutants and deletions displaying FOXA1 DNA binding domain (DBD), K270 and K414.
(C) Purified recombinant, full-length FOXA1 and mutants were analyzed by SDS-PAGE and Coomassie staining. The TFs bands run at the expected sizes when compared to the protein standards. All proteins were run on the same gel.
(D) Representative EMSA showing the affinity of increasing amounts of FOXA1 WT and mutants to specific, short dsDNA Cy5-eG and non-specific short, dsDNA Cy5-lefty. Black arrowheads indicate free DNA and DNA:FOXA1 complexes.
Figure C2 | Mutation of FOXA1 K270 or C-term αHelix does not inhibit nucleosome binding.

(A-F) Representative EMSAs showing the affinity of increasing amounts of FOXA1 WT and mutants to (A-C) Cy5- albN1 DNA and (D-F) nucleosomes. Black arrowheads indicate free DNA and DNA:FOXA1 complexes, white arrowheads indicate nucleosomes and nuc:FOXA1 complexes.
Figure C3 | Purified recombinant, full-Length FOXA1 WT and mutants bind free DNA and nucleosomes with similar nucleosome binding affinities.

(A) FOXA1 binding curves related to EMSAs fraction bound quantification of Figures B2-A,F showing the dissociation constants of FOXA1 mutants to Cy5-160bp-\textit{albN1} DNA (blue) or to Cy5-160bp-\textit{albN1-NUC} (orange). Proteins were titrated on 1 nM DNA or NUC.

(B) FOXA1 WT and mutant binding curves of total fraction bound to free DNA (left) and nuc (right) in a single graph for comparison.

(C) Same as B, but binding curves of specific fraction bound to free DNA (left) and nuc (right).
Figure C4 | FOXA1 residue K270 contributes to FOXA1 histone octamer interaction.

(A) Representation of FOXA1 DNA binding domain (DBD) truncations aa 168-268, extended DBD 144-294 (eDBD), which includes K270 and mutants eDBD, and 144-294 with EKQ (269-271) AAA triple alanine mutant.

(B) Purified recombinant FOXA1 DBD, eDBD, and eDBD EKQ (269-271) AAA truncations analyzed by SDS-PAGE and Coomassie staining. The TFs bands run at the expected sizes when compared to the protein standards. All proteins were run on the same gel.

(C) Representation of Sepharose-conjugated FOXA1 pull down with purified recombinant histone octamers.

(D) Histone octamer pulldown assays with FoxA1 full-length (FL) protein and FoxA1 FL-AAA mutant bound to sepharose beads to assess binding to histone octamers.

(E) Representative EMSA showing the affinity of increasing amounts of FOXA1 DBD, eDBD, and eDBD-AAA to specific “s” short dsDNA Cy5-eG and non-specific “ns” short dsDNA Cy5- lefty. Black arrowheads indicate free DNA.

(F) Same as E but with eDBD and DBD.