Title
Examining the Functions of Master Regulators in Maintaining Pluripotency and Inducing Reprogramming

Author
Lo, Hung-Hao

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Examining the Functions of Master Regulators in Maintaining Pluripotency and Inducing Reprogramming

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Molecular Biology

by

Hung-Hao Lo

2019

ABSTRACT OF THE DISSERTATION

Examining the Functions of Master Regulators in Maintaining Pluripotency and Inducing Reprogramming

by

Hung-Hao Lo

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2019

Professor Stephen Smale, Chair

Pluripotency factors Oct4, Sox2, and Nanog orchestrate an elaborate hierarchy of gene regulation governing embryonic stem cell (ESC) identity. The capability of differentiating into cell from any of the three germ layers offers a potential solution and path toward the discovery of novel therapies for devastating diseases. The differentiation of ESCs requires three fundamentally distinct transitions in the transcriptional state: 1) the activation of silent genes in ESCs that are transcribed in differentiated states, 2) the silencing of ESC-transcribed genes that are entirely inactive in the differentiated cells, 3) the modulation of transcription for genes that are expressed in both populations. Deciphering the mechanisms regulating the transitions of these transcriptional states will illuminate how the pluripotent state is established and maintained by the regulation of pluripotency factors. Because Oct4 and Sox2 are the core regulators of the transcriptional

regulatory network for pluripotency, we focused our studies on mechanisms regulated by Oct4 and Sox2 in ESCs.

Oct4 and Sox2 bind to thousands of enhancer composite sites at pluripotency genes and differentiation-promoting genes. These Oct4/Sox2 target genes exhibit distinct transitions of transcriptional states for the establishment of pluripotency. As distinct mechanisms are likely to regulate different transcriptional states, a critical step toward a mechanistic understanding of pluripotency is to delineate the genes with well-defined transcription characteristic. Because most studies have relied on low stringency criteria to define differential expression to infer regulatory mechanisms, they did not rigorously evaluate the selective functions of Oct4/Sox2 composite binding critical for establishing pluripotency. To scrutinize how Oct4 and Sox2 account for possible differences in the regulatory mechanisms necessary for distinct transcriptional states, we refined the gene classification and interrogated the role of Oct4/Sox2 binding at enhancer composite motifs at distinct gene classes. By combining RNA-seq, ChIP-seq, and ATAC-seq with functional validation employed by CRISPR/Cas9 mutagenesis, we discovered that Oct4/Sox2 function differently across gene groups with various transcriptional states. In addition to a role in transcriptional activation at ESC-specific and Dynamic genes, my data suggest that Oct4/Sox2 motifs at silent genes may mediate the transcriptional repression in ESCs. In Chapter 3, we also employed a gene-centric approach to quantitatively compare transcription factor co-binding, histone modifications, chromatin accessibility, enhancer properties, and the genomic context of transcriptional regulation by Oct4 and Sox2 in ESCs. Together, we extended our understanding of the critical role played by Oct4 and Sox2 in the establishment of pluripotency.

The dissertation of Hung-Hao Lo is approved.

Alexander Hoffmann

Jingyi Li

Kathrin Plath

Sriram Kosuri

Stephen Smale, Committee Chair

University of California, Los Angeles

2019

Dedication

In dedication to my parents, Chiu-Shen Lo and Su-Yun Fang, and my wife, Yi-Ting Chen;

for their inspiration, support, and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 2

LIST OF TABLES

CHAPTER 2

CHAPTER 3

ACKNOWLEDGEMENTS

I am fortunate to work closely with many talented people who keep encouraging and supporting me throughout my graduate studies. This adventurous journey would never be so rewarding without any of your companions. I would like to thank my thesis advisor, Stephen Smale, for being a model of scientist, critical thinker, patient mentor, and for supporting me getting over every challenge in the past years. I thank my doctoral committee members, Alexander Hoffmann, Jingyi Jessica Li, Sriram Kosuri, and Kathrin Plath for the insightful advice and helpful discussion to move this project forward. I greatly appreciate all past and current members of the Smale lab to train my technical skills, inspire our brainstorming, and identify significant correlation between food, beer, and science : Ann-Jay Tong, Brandon Thomas, Justin Langerman, Miguel Edwards, Prabhat Purbey, Philip Scumpia, Peter Kim, George Yeh, An-Chieh Feng, Allison Daly, Amber Ruccia, and Vasileios Ragkousis.

I thank my previous mentor Hsiao-Sheng Liu of my undergraduate research and Hsei-Wei Wang of my master thesis for their inspiration and mentorship. I thank Ren Sun and Genhong Cheng for guidance throughout my rotations at UCLA. I thank Justin Langerman and Constantinos Chronis for discussing and optimizing the stem cell experiments. I thank my funding support from Whitcome pre-doctoral training program.

Lastly, I would like to thank my wife, Yi-Ting, for her understanding and support while I was always working late night shift in the lab. Thank you for being such a

trustworthy partner so I can share every happiness, sadness, anxious, but not furious moment with you. My great appreciation goes to my families in Taiwan, especially my parents, who have always been supportive and giving me strengths.

VITA

2011	Bachelor of Science
	Medical Laboratory Science and Biotechnology
	National Cheng Kung University, Taiwan
2013	Master of Science, Microbiology and Immunology
	National Yang Ming University, Taiwan
2014 – present	Molecular Biology Institute
	University of California, Los Angeles
2016 – 2017	Teaching Assistant – Life Science 3
	University of California, Los Angeles
2017 – 2018	Whitcome Pre-doctoral Fellowship
	University of California, Los Angeles

PUBLICATIONS

Smale ST, Lo HH (2018). A Quantitative Perspective of Transcription in Pluripotent and Differentiated Cells. *Invited Speaker Presentations /Experimental Hematology 64: S23–S39*

Wang HW, Sun HJ, Chang TY, Lo HH, Cheng WC, Tseng GC, Lin CT, Chang SJ, Pal N, Chung IF (2015). Discovering monotonic stemness marker genes from time-series stem cell microarray data. *BMC Genomics,* 16(Suppl 2):S2

Remenyi R, Qi H, Su SY, Chen Z, Wu NC, Arumugaswami V, Truong S, Chu V, Stokelman T, Lo HH, Olson CA, Wu TT, Chen SH, Lin CY, Sun R (2014). A

Comprehensive Functional Map of the Hepatitis C Virus Genome Provides a Resource for Probing Viral Proteins. *MBio,* 30;5(5):e01469-14.

Wang HW*, Lo HH*, Chiu YL, Chang SJ, Huang PH, Liao KH, Tasi CF, Wu CH, Tsai TN, Cheng CC, Cheng SM (2014). Dysregulated miR-361-5p/VEGF axis in the Plasma and Endothelial Progenitor Cells of Patients with Coronary Artery Disease. *PLOS ONE*, 27;9(5):e98070 * equal contribution

Wang HW*, Huang TS*, Lo HH*, Huang PH, Lin CC, Chang SJ, Liao KH, Tsai CH, Chan CH, Tsai CF, Cheng YC, Chiu YL, Tsai TN, Cheng CC, Cheng SM (2014). Deficiency of the miR31-miR720 Pathway in the Plasma and Endothelial Progenitor Cells from Patients with Coronary Artery Disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 34(4):857-69. * equal contribution

PRESENTATIONS

JCCC Gene regulation monthly meeting. University of California, Los Angeles, Dec. 4th, 2018. Oral presentation.

Cold Spring Harbor Laboratory Meetings: Epigenetics and Chromatin, Cold Spring Harbor, New York, Sept. 11th-15th, 2018. Poster presentation.

MBI Program Retreat, University of California, Los Angeles. 2018. Poster presentation.

UCLA Biomedical and Life Science Innovation Day, University of California, Los Angeles. 2018. Poster presentation.

UCLA I3T Retreat, University of California, Los Angeles. 2018. Poster presentation.

MBI Program Retreat, University of California, Los Angeles. 2017. Poster presentation.

CHAPTER 1

Introduction

Master Regulators and Transcriptional Regulation in Pluripotency and Reprogramming

A. Pluripotency and Reprogramming

Embryonic stem cells (ESC), derived from the inner cell mass (ICM) of the preimplantation blastocyst, was firstly isolated and cultured by Martin Evans and Gail Martin back in 1981 (Evans and Kaufman, 1981; Martin, 1981). The capabilities of indefinite proliferation (self-renewal) and potential differentiation (pluripotency) are two fundamental signatures of ESCs (Ito and Suda, 2014). These pluripotent stem cells can further differentiate into a multitude of distinct cell types, which are crucial for embryonic development. Both human and mouse demonstrated in vitro differentiation and germline transmission to generate cells from all three germ layers. With proper growth supplements and signaling induction, ES cell will initiate the differentiation processes. These include the activation of differentiation-promoting genes and the development of tissue-specific functions. The homeostasis between pluripotency and differentiation are considered to be regulated by a well-governed gene regulatory program (Kim et al., 2008a).

The discovery of ESCs has reshaped the fields of transplantation and regenerative medicine. The pluripotent stem cells offer a potential solution for graft-versus-host disease (GVHD) and a path toward the discovery of novel therapies for degenerative diseases, including spinal cord injury, Huntington disease, Parkinson's disease, or heart disease (Trounson and DeWitt, 2016). The fusion of ESCs with somatic cells convert the somatic cells to a pluripotent state (Cowan et al., 2007; Hochedlinger et al., 2004; Tada et al., 2001). Furthermore, ectopic expression of Oct4, Sox2, Klf4, and c-Myc (OSKM) in mouse fibroblast induces pluripotency, and these cells give rise to tissue development both in vitro and in vivo (Maherali et al., 2007; Takahashi and Yamanaka, 2006;

Yamanaka and Blau, 2010). These induced pluripotent stem cells (iPSCs) display the same morphology and growth properties as ESCs. iPSCs can also differentiate into muscle and all three germ layers, including cartilage, neural tissues, and epithelium, providing the evidence of pluripotency. It is impressive how a limited number of transcription factors can maintain and induce pluripotency. Defining the precise of OSKM targets and delineating how the transcription factors function concretely will undoubtedly facilitate our understanding of the pluripotent state and reprogramming.

B. Transcriptome Profiling of ESCs and Somatic Cells

While all cells in the complex multicellular organisms contain identical DNA context, the distinct functions of hundreds of different cell types depend on the differences in gene expression. The transcriptome determines the proteins to be translated, which collectively contribute to the functions of a given cell. Thus, it is vital to define and quantify the transcriptome in the context of the level of gene expression and the differential expression between different cell types. The transcriptome profiling will also help us study a variety of fundamental biological phenomena from development to disease, and uncover the mechanisms that govern the cell identity. To profile the transcript level genome-wide, RNA sequencing (RNA-seq) provides us with an unparalleled opportunity to identify and accurately quantify transcripts (Mortazavi et al., 2008). Although RNA-seq provides a high-throughput platform to measure the transcript level genome-wide, a new era of the challenge has come to the approach of bioinformatics analysis. One of the challenges to analyze transcriptome quantitatively is that most of the RNA-seq datasets are not analyzed with sufficient depth to perform a precise quantitative analysis of the dynamic

range of gene expression. The measurement of differential expression between cell types often relies on statistical analysis with low stringency criteria. An accurate and quantifiable assessment of transcriptomes in different cell types and developmental stages will be indispensable in our efforts to further our understanding of gene regulation.

B1. Tissue Specificity

Many efforts have attempted to identify and understand tissue-specific genes and their regulation during embryonic development (Carninci et al., 2006; Efroni et al., 2008; Ernst et al., 2011). However, the previous studies often used limited fold differences to define tissue specificity, which grouped thousands of tissue-specific genes with various degrees of the dynamic range of expression (Carninci et al., 2006; Efroni et al., 2008; Meister et al., 2010; Song et al., 2013). Subsequent analysis of transcriptional regulation, such as transcription factor binding or chromatin landscape, focused on these gene sets could barely reveal the fundamental regulatory mechanisms necessary for such dynamic regulation. Using deep nascent transcript RNA-seq (~ 500 million mapped reads), the nascent transcript profiles allowed us to accurately measure the transcripts for the genes expressed at a very low level and prevented the influences of post-transcriptional RNA stability. This approach provided the power to calculate the dynamic range of expression profiles of ESC in comparison with three other primary differentiated cell types (cortical neurons, double-positive thymocytes, and bone marrow-derived macrophages). Thus, we can distinguish genes exhibiting a largest dynamic range of expression among all cell types from those showing the modest dynamic range of expression. Moreover, these

genes are more likely to approach real ESC-specificity; even the number of genes would continue to decrease if more differentiated cell types were examined.

B2. Dynamic Range of Expression

A stringent analysis of RNA-seq data sets not only help us to capture the complexity of the transcriptome but also allow us to perform a more accurate and descent approach to compare the transcript level among various cell types (Mutz et al., 2013). A careful quantitative delineation of gene classes through quantifying of the dynamic range of expression will bring further insight. For example, some genes are also expressed in one or two cell types yet indeed showing a large dynamic range of expression between ESC and a few other cells. Conventionally, these genes would never be considered as either ESC-specific or somatic-specific, which left the functions of Oct4/Sox2 binding sites for these genes unknow. However, to uncover the functional significance of transcription factors for pluripotency, we really need to carefully address this issue and cluster the Oct4/Sox2-associated genes into better classification. By carefully examining the dynamic range of expression, the approach not only identifies the “real” ESC-specific genes and broadly expressed genes with small fluctuations, but also separates the genes that are dynamically expressed in a few cell types. These dynamic gene categories will allow us to interrogate the functions of Oct4/Sox2 binding sites at the enhancer and unveil their regulatory roles in pluripotency.

C. Pluripotency Gene Regulatory Network

Thinking about the gene regulatory mechanisms that maintain and establish the pluripotent state, previous studies focused on three directions to address the question of pluripotency. First, how do the master regulators maintain their expression to stay in a pluripotent state? Second, with a limited number of TFs, how do they activate the other pluripotent genes and what are the precise targets of the master regulators. Third, to prevent differentiation, how do they repress the expression of differentiation-promoting genes (Orkin, 2005). Developing a comprehensive strategy to scrutinize the interaction between core regulators and their target genes is critical to augment the application of pluripotent cells. Employing the genome-wide ChIP-chip and ChIP-seq, a simplified deterministic gene hierarchy model described a highly-coordinated gene regulation network of pluripotency (Kim et al., 2008a). This gene regulation network converged on Oct4, Sox2, and Nanog, which acts as the core regulators. An autoregulatory circuit between Oct4, Sox2, and Nanog maintains its expression for pluripotency. These core regulators coordinately targeted the promoters of Klf4, c-Myc, Dax1, Rex1, Nac1, and Zfp281. Additionally, Oct4-regulatory module is associated with a number of other pluripotency factors such as Esrrb, Nr5a2, Tcfcp2l1, and Klf4 (Dejosez et al., 2010; Feng et al., 2009; Heng et al., 2010). These interactions revealed a coordinated cascade of pluripotent factors for governing pluripotency or guiding the developmental process (Chen et al., 2008). Interesting, when they evaluated the expression level of target genes in the pluripotency network, genes bound by more transcription factors tend to be more actively transcribed. In contrast, those genes with fewer transcription factors are primarily silenced in ESCs (Chen et al., 2008; Kim et al., 2008a).

By interacting with many intermediary transcription factors, Oct4/Sox2/Nanog can further coordinate a subsequent cascade of regulatory events. The coordinated regulatory network also suggests that Oct4/Sox2/Nanog can exert their function in gene transcription by the downstream regulation of other targets. For example, Oct4 regulates miR-290-295 and histone modifiers Jmjd1a and Jmjd2c (Hnisz et al., 2013; Loh et al., 2007). The miR-290-295 clusters promote pluripotency by targeting their downstream cell cycle regulators *Wee1* and *Fbxl5* (Lichner et al., 2011). *Jmjd1a* and *Jmjd2c* function as histone demethylase to prevent the accumulation of repressive histone mark H3K9me3 at the pluripotency genes (Loh et al., 2007). The interface between the pluripotency networks, epigenetic regulation, microRNA-network implement a highly connected mechanism in ESC to govern the process of reprogramming and pluripotency (Ng and Surani, 2011).

Moreover, a recent study delineated a substantial interaction of Oct4 and Sox2 with somatic and pluripotency enhancers, suggesting context-dependent regulators for both ESC self-renewal and lineage differentiation (Chronis et al., 2017). The somatic-enhancer inactivation and the pluripotency-enhancer activation were selected by the cooperation with stage-specific TFs and pluripotency TFs. In the model they proposed, Oct4/Sox2/Klf4 mediates the redistribution of the somatic transcription factors, loss of p300 / H3K27Ac, and gain of Hdac1 for ultimately transcriptional repression. In the context of the pluripotency enhancer, OSKM can engage at the early stage of 48 hrs induction or later stage of reprogramming. Both scenarios revealed extensive cooperative binding of Oct4, Sox2, Nanog, and Esrrb.

D. Transcriptional Regulation of Oct4 and Sox2

A transcriptional network, centering on Oct4 and Sox2, reveals target promoters or super-enhancers of pluripotent genes, suggesting a concrete mechanism toward pluripotency (Kim et al., 2008b; Whyte et al., 2013). The investigation of protein-interactome and protein-DNA interactions within this pluripotency gene regulatory network is often the first step to explore the functional roles of these master transcription factors in ESCs (Li and Belmonte, 2017; Li and Izpisua Belmonte, 2018).

Oct4 is a homeodomain transcription factor encoded by the *Pou5f1* gene, which contains a well-conserved Homeobox for DNA-binding (Reményi et al., 2003). Oct4 binds to a consensus octameric DNA nucleotide sequence ATTTGCAT (Petryniak et al., 2006). The level of Oct4 has a restricted expression to pluripotent and germ cells (Fuhrmann et al., 2001). Oct4 is a critical regulator for pluripotency through various interfaces of protein-protein interaction, epigenetic regulation, and directly transcriptional regulation (Esch et al., 2013; Hammachi et al., 2012). In *Oct4*$^{-/-}$ embryos, although the embryo could develop to the blastocyst stage, the inner cell mass was not pluripotent. Furthermore, the trophoblast did not proliferate adequately (Nichols et al., 1998). More quantitative research performed in mouse ESCs revealed that the precise level of Oct4 expression is required for governing the stem cell self-renewal and lineage commitment (Niwa et al., 2000). A moderate increase of Oct4 led to the differentiation into primitive endoderm and mesoderm, while the repression of Oct4 lost the pluripotency and dedifferentiated to trophectoderm. In mouse ESCs, Oct4 contributes to the cell fate decisions during the transition to a differentiated cell state (Thomson et al., 2011). Together, these studies

demonstrated the crucial role of Oct4 in maintaining pluripotency both *in vivo* and *in vitro*. As a master regulator for pluripotency, Oct4 controls the lineage commitment, and the requirement of the precise level illustrates a sophisticated gene regulatory program designated for pluripotency. In addition to direct transcriptional regulation of the target genes, Oct4 can recruit key epigenetic regulators to the target genes (Esch et al., 2013). The interactome of Oct4 discovered the unique protein interface of Oct4 to interact with Smarca4 and Chd4. The interaction improves the reprogramming efficiency and guard pluripotency by maintaining H3K27me3 in ESCs. Additionally, another study also characterized the endogenous association of Oct4 with proteins from multiple repression complexes such as NuRD, Sin3A, and Pm1 (Liang et al., 2008). The association with repression complexes suggests a repressive role of Oct4 gene regulation, yet the mechanisms and precise targets remain unclear.

Another master transcription factor, Sox2, is a member of the Sox family transcription factors, which contains a high mobility group (HMG) box for DNA binding (Schepers et al., 2002). Sox2 binds to a 6 to 7 nucleotides DNA sequence CTTTGTC through the recognition of a core motif sequence TTGT (Kamachi et al., 2000). Silencing of Sox2 compromised self-renewal and differentiated ESCs into multiple lineages (Ivanova et al., 2006). In the Sox2 null ESCs, the cells differentiate into trophectoderm like cells (Masui et al., 2007). Likewise, Sox2 is also necessary to activate multiple transcription factors that are essential for stabilizing ESCs in a pluripotent state. Sox2 depletion resulted in aberrant expression of multiple transcriptional regulators of Oct4, leading to the decrease in Oct4 expression and eventual inactivation of the Oct4/Sox2-regulated genes.

D1. POU/HMG/DNA Ternary Complex

In ESCs, Sox2 often dimerizes with Oct4 and acts synergistically to activate their target genes (Avilion et al., 2003; Reményi et al., 2003; Tapia et al., 2015). The regulatory regions of these genes contain an Oct4 octamer motif juxtaposed to the Sox2 elements by a spacer nucleotide either 0 bp or 3 bp. The crystal structures of Oct4/Sox2 composite elements on *Fgf4*, *Utf1*, and *Nanog* revealed heterodimer conformations with various nucleotide spacers and distinct functions (Jauch et al., 2008; Reményi et al., 2003; Tapia et al., 2015). Hence, one of the roles of Sox2 in maintaining pluripotency appears to regulate the expression of transcription factors that are necessary for the optimal expression of Oct4. In the pluripotency gene regulatory network, this Oct4/Sox2 cooperative binding acts as the core regulator coordinate the downstream cascade of pluripotency factors activation. The reciprocal transcriptional regulation between Oct4 and Sox2 further reinforce the ability to maintain ESC pluripotency via the Oct4/Sox2 complex (Chew et al., 2005).

D2. The Role of Pioneering Factor

Accumulating evidence suggests that pluripotency factors including Oct4/Sox2/Klf4/c-Myc interact with the promoter or enhancer for tissue-specific genes way earlier than the genes are transcribed (Smale, 2010; Soufi et al., 2014; Xu et al., 2007; Zaret and Carroll, 2011). As a consequence of pioneer factor binding, the chromatin configuration becomes permissive to activate gene expression upon lineage specification (Zaret and Mango, 2016). This epigenetic priming event highlights the interplay of pluripotency factors and epigenetic modification. The paradigm role of pioneer factor regulation is the hepatic-

specific albumin gene *Ab1*. The *Ab1* enhancer contains several transcription factor binding sites, including GATA-4, C/EBP, FoxA1, and NF1. A previous study in our lab performed by Xu et al. found an unmethylated CpG dinucleotide at the *Ab1* enhancer in ESCs (Xu et al., 2007, 2009). The unmethylated CpG dinucleotide in ESCs indicated that pioneer factors marked enhancers for at least some tissue-specific genes in ESCs. The pre-established mark establishes transcriptional competence in ESCs and iPSCs (Xu et al., 2009). The unmethylated CpG dinucleotide lies in the FoxA1 binding site. However, FoxA1 is not expressed in ESCs. Instead, FoxD3, another Fox family member, gains early access to the *Ab1* enhancer in ESCs, to be responsible for the unmethylated state, and hold the place for FoxA1 once the cell differentiated into endoderm (Smale, 2010). These studies support the pioneer factor model of establishing transcriptional competence for cell fate decisions.

Mechanistically, Oct4, Sox2, Klf4, and c-Myc are able to bind to their target sequences on nucleosomal DNA. The motif sequences for pioneer factor binding are unconventional, which can be partial or degenerate sequences (Soufi et al., 2014). Unlike the Oct4 canonical octamer motif of (ATGCAAAT), Oct4 targeted the hexameric motifs AT/AATGC or AAATAC at the nucleosomal-enriched DNA. On the other hand, Sox2 binding at the nucleosomal DNA showed less restriction at the sixth “G” nucleotide and targeted a degenerate motif CT/CT/ATTNT. Klf4 targeted a hexameric motif GGGT/AGG that was lacking the three-terminal nucleotides at nucleosome-enriched sites. Moreover, the c-Myc binding at nucleosomal DNA lost the restriction of two central nucleotides by recognizing degenerate sequences CANNTG. The unique binding sequences of pioneer factors

indicate their ability to initiate regulatory events in silent chromatin to control cell fate or induce reprogramming.

D3. Cooperative Binding of Transcription Factors and Histone Modifier with Oct4/Sox2

Oct4 and Sox2 are the core regulators of the pluripotency regulatory networks (Angie Rizzino1, 2013; Kim et al., 2008b). Aside from the core regulators, several interacting transcription factors or histone modifiers have been described (Marks and Stunnenberg, 2014; Yamanaka, 2008). Several studies have also employed ChIP-seq to profile genome-wide co-factor binding sites in both mouse and human ESCs genome (van den Berg et al., 2010; Boyer et al., 2005; Carninci et al., 2006; Hammachi et al., 2012; Jerabek et al., 2014; Kim et al., 2010; Loh et al., 2006; Ng and Surani, 2011; Orkin and Zon, 2008; Wang et al., 2006). By integrating the transcriptome profiling with the ChIP-seq data sets, these studies have revealed hundreds of target genes and many insights into the requirement of cooperative binding for pluripotency.

Notably, more than 90% of the Oct4 and Sox2 bindings overlap with the Nanog binding sites in both human and mouse (Boyer et al., 2005; Loh et al., 2006). Nanog is one of the ESC-specific genes and a well-studied target of Oct4/Sox2 (Jauch et al., 2008). As one of the main proteins in the transcriptional network for pluripotency of embryonic stem cells, Nanog cooperatively mediates gene activation with Oct4 and Sox2 in many pluripotency genes (Kim et al., 2008a). The study used biotinylation mediated ChIP-chip to identify genomic binding sites of multiple pluripotency factors. By the analysis of

pluripotency factors co-bind, they separated the target genes into two major categories based on the number of pluripotency factors co-binding: Oct4, Sox2, Nanog, Klf4, c-Myc, Dax1, Nac1, Zfp281, and Rex1. The first group is the genes bound by greater than four factors, and the second group contains those only bound by a few pluripotency factors. They discovered a correlation between the presence of pluripotency factors at the promoters and gene activity. The first gene group, bound by more factors, are most actively transcribed in the ESCs, where the second group promoters bound by few factors tend to be inactive. Although this ChIP-centric approach oversimplified the molecular mechanisms, the work is an initial attempt to separate the target genes with bivalent features to understand how the core regulators are necessary for the maintenance of pluripotency and self-renewal. Thus, a well-defined gene groups categorized by stringently evaluated transcriptome profiling should provide novel insights into the molecular mechanisms that govern ESC pluripotency.

E. Epigenetic Regulation of Pluripotency

In ESCs, the complex interrelationship between pluripotency and chromatin factors mediates the chromatin plasticity, which ultimately establishes the epigenetic barrier between pluripotency and differentiation (Becker et al., 2016; Meshorer et al., 2006). The interplay between the pluripotency network and histone modifications modify the chromatin accessibility to establish the competence of gene expression (Denholtz et al., 2013; Maherali et al., 2007). Distinct epigenetic characteristics at promoter and enhancer regions contribute to the expression of cell type-specific genes and mark the identities in the given cell types. In ESCs, electron microscopy and genome-wide chromatin profiling

revealed globally open and highly dynamic chromatin configuration (Azuara et al., 2006; Meissner et al., 2008; Park et al., 2004). In human ESCs, the unique chromatin states signified the expression of different gene classes characterized by RNA analysis and functional annotation (Rada-Iglesias et al., 2011). For genes annotated with pluripotent functions, enriched histone acetylation H3K27Ac regions overlapped with the enhancers of the ESC-expressing genes.

In contrast, repressive histone marks H3K9me3 and H3K27me3 occupied the enhancers to silence the developmental genes in ESCs. During the differentiation, the binding of lineage-determining TFs and the recruitment of nucleosome remodeling complexes delivered the signal and altered the chromatin landscape for the differentiated gene expression (Heinz et al., 2010). The induction of cell-type-specific genes and the chromatin modification established the transition barrier between the differentiated cells and pluripotent stem cells. Conversely, during the reprogramming, ectopic expression of OSKM also induced genome-wide chromatin remodeling to perpetuate the pluripotent state (Koche et al., 2011). In the early stage of reprogramming, although OSKM only activated a small subset of genes with the active promoters, a rapid and extensive chromatin remodeling established the active or poised states of the pluripotency-enhancers. These epigenetic mechanisms developed a productive engagement of OSKM and the activation of pluripotent genes in late reprogramming. In a recent study, during the reprogramming, a higher concentration of Oct4, Sox2, and Klf4 could recognize the incomplete consensus motifs in the nucleosomal DNA with lower affinity (Soufi et al., 2014). As pioneering TFs, Oct4, Sox2, and Klf4 may access the pluripotency-enhancers

in a very early stage of reprogramming and mediate the nucleosome displacement for priming the chromatin landscapes. These studies demonstrated that the collaborative mechanisms of TF bindings and chromatin dynamic are crucial for repressing differentiated genes and activating pluripotent genes.

F. Super-Enhancer

A recent study described the enhancer properties by distinguishing an unusual super-enhancer domain at most pluripotency gene regulated by master regulators (Whyte et al., 2013). These super-enhancers are featured with densely-occupied master regulators and Mediators and play prominent roles to define cell identity. In the context of gene expression, super-enhancer conferred higher activity to activate their target genes when compared with the typical-enhancer. The genomic sizes of super-enhancer are longer than the typical-enhancer, with a median size of 8.7 kbp. These regions showed enriched co-binding of tissue-specific or pluripotency transcription factors and mediator protein MED1. Additionally, the active histone mark H3K27Ac often displayed enriched deposition at separated enhancer clusters within the super-enhancer regions (Khan and Zhang, 2016). In ESCs, several master regulators densely co-bound at the super-enhancers and correlated with the active transcription of the pluripotency genes. In a subset of Oct4-occupied pluripotency genes, reduced Oct4 level led to a more significant decrease in super-enhancer activity. On the other hand, these super-enhancers also defined the cell identity in the differentiated cells. The high density of cell-type-specific transcription factors at the super-enhancer of differentiation-promoting genes suggested the control of mammalian cell identity.

Furthermore, these specialized enhancer domains also linked to human diseases such as cancer or type 2 diabetes (Hnisz et al., 2013; Lovén et al., 2013; Parker et al., 2013). The establishment of super-enhancer in tumor cells generated high transcriptional activity at critical tumor pathogenesis genes, which resulted in oncogene overexpression (Hnisz et al., 2013). The inhibition of transcriptional coactivator *BRD4* interfered bromodomain formation and disrupted the development of the super-enhancer region (Lovén et al., 2013). Thus, these cell-type-specific super-enhancers hold the potential to determine disease-associated biomarker for diagnosis and therapy. Although the genomic features of super-enhancer and their role in regulating transcription have been widely predicted in many cell types, few have rigorously investigated and compared their functions with the other typical-enhancers by deletion. It is still unclear if these predicted enhancer properties represent the paradigm in gene regulation.

Recently, a functional study employed the CRISPR/Cas9-mediated deletion approach to compare the functional differences between several super-enhancers and other enhancers in ESCs (Moorthy et al., 2017). They revealed that the activities of these super-enhancers in the context of gene expression are highly variable. By deleting the individual enhancer, they found a redundancy between individual enhancer subcluster within the super-enhancer regions. When simplifying the enhancer activity based on the bioinformatics prediction of super-enhancer, more than 80% of the potentially active regulatory elements at the highly-transcribed genes in ESCs are largely underestimated. These enhancers were also robust to the transcriptional activation of the ESCs genes

even they were not predicted as a super-enhancer. The robust transcriptional output of these individual enhancers and the redundant role of the enhancer clusters within super-enhancer highlighted the limited understanding of pluripotency gene regulation. A rigorous investigation of the mechanistic and functional insights into the pluripotency factor binding sites is, therefore, necessary to decipher the gene regulation required for pluripotency and reprogramming.

G. Gene-centric Approach to Uncover Mechanistic Insights of Gene Regulation

An overarching goal of pluripotent stem cell research is to unveil the molecular mechanisms by which OSN contribute to the activation of pluripotency and self-renewal genes. Genome-wide analyses of OSN occupancy identified thousands of or even more OSN peaks resided nearby not only the pluripotent genes but also the differentiation-promoting genes (Kim et al., 2008a; Loh et al., 2006; Orkin, 2005). The framework of previous transcriptional studies related to ESC pluripotency mostly initiated with ChIP-seq peak-centric analyses to examine occupancy by master regulators. When paying less attention to the degree of differential expression in ESCs, dealing with the tremendous amount of OSN bound genes come upon against the difficulties of determining the precise mechanisms. The real function of the transcription factor binding sites may thereby be challenging to reveal, and the possible differences in the regulatory mechanisms necessary for such dynamic regulation remain unclear. Previously, our lab described a gene-centric approach to investigate inflammatory gene regulation in a set of well-defined inducible gene clusters (Bhatt et al., 2012; Tong et al., 2016). The stringent system approach successfully uncovered mechanistic insights into inflammatory gene regulation

in Lipid A-stimulated macrophages. A similar gene-centric approach with high stringency criteria will help us to decipher the selection functions of Oct4/Sox2 binding sites.

In contrast to the peak-centric analysis, the gene-centric analysis places greater emphasis on quantitative aspects of gene classification before examining the Oct4/Sox2 ChIP-seq and interrogating the regulatory functions. The method is based on the assumption that the Oct4/Sox2 binding for the genes exhibiting a large dynamic range of expression between ESCs and other somatic cells function differently from those are non-dynamic. By investigating the genomic/genetic features and the functions of Oct4/Sox2 binding sites for the well-defined gene classes hold the potential to understand the molecular basis in different classes of pluripotent genes. Deciphering these selective regulation mechanisms is essential for understanding how pluripotency is established and maintained.

In conclusion, the maintenance and induction of pluripotency require fine control in pluripotency gene activation and differentiation-promoting gene inactivation. As the master regulators bind to thousands of genomic loci, clarifying the functional significance of these binding sites will improve our understanding of ESCs pluripotency. The knowledge will also lay the foundation for future efforts to develop stem cell therapies for devastating diseases. Because Oct4 and Sox2 cooperatively orchestrate the transcriptional cascade for ESCs pluripotency, we began our studies on the investigation of Oct4/Sox2 binding to composite enhancer motifs. The maintenance of ESC pluripotency requires proper activation of pluripotency genes and inactivation of

differentiation promoting genes. The precise mechanism for establishing pluripotency integrates cooperative binding of transcription factors and chromatin remodeler. This interplay between pluripotency gene regulation network and epigenetic regulation requires both common and gene-specific mechanism to define cell identity. Although many studies have investigated the transcriptional mechanisms of different master regulators, they relied on statistics and conventional system approaches to predict the functional significance of the transcription factor binding. The conventional system approaches focused on thousands of genomic loci with transcription factors binding, despite careful examination of binding characteristics, the precision and accuracy of the analyses may be compromised by an overwhelming number of genes lacking preferential expression in ESCs. Therefore, the real functions of these thousands of Oct4/Sox2 composite binding sites remain unclear. To gain more mechanistic insights into pluripotency, we need to use a different strategy to scrutinize the transcriptional regulation by Oct4 and Sox2 in embryonic stem cells. In chapter 2, we employ a combination of bioinformatic approach and functional deletion by CRISPR/Cas9 to interrogate the critical role of Oct4/Sox2 binding at the composite enhancer motif. Our research uncovered the critical role of Oct4/Sox2 composite binding in regulating ESC-specific / Dynamic gene activation and Silent gene inactivation, while the sites at the Non-Dynamic genes are non-functional. In chapter 3, we perform an in-depth genomic/genetic analysis of pluripotency gene regulation using a gene-centric approach that emphasized on quantitative aspects of RNA-seq, ChIP-seq, ATAC-seq, and genomic context. Taken together, our study led to understand better of how pluripotency is established and maintained by uncovering

critical role of Oct4/Sox2 composite binding in gene regulation and by elucidating more in-depth mechanistic details in complicated contexts.

REFERENCE

Angie Rizzino1, 2 (2013). The Sox2-Oct4 Connection: Critical players in a much larger interdependent network integrated at mul. Stem Cells.

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

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., et al. (2006). Chromatin signatures of pluripotent cell lines. Nat. Cell Biol. *8*, 532–538.

Becker, J.S., Nicetto, D., and Zaret, K.S. (2016). H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. Trends Genet. *32*, 29–41.

van den Berg, D.L.C., Snoek, T., Mullin, N.P., Yates, A., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R.A. (2010). An Oct4-Centered Protein Interaction Network in Embryonic Stem Cells. Cell Stem Cell.

Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell.

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

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell.

Chew, J.-L., Loh, Y.-H., Zhang, W., Chen, X., Tam, W.-L., Yeap, L.-S., Li, P., Ang, Y.-S., Lim, B., Robson, P., et al. (2005). Reciprocal Transcriptional Regulation of Pou5f1 and Sox2 via the Oct4/Sox2 Complex in Embryonic Stem Cells. Mol. Cell. Biol.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. Cell.

Cowan, C. a, Atienza, J., Melton, D.A., and Eggan, K. (2007). Nuclear Reprogramming of Somatic Cells After Fusion with Human Embryonic Stem Cells Nuclear Reprogramming of Somatic Cells After Fusion with Human Embryonic Stem Cells. Science (80-.).

Dejosez, M., Levine, S.S., Frampton, G.M., Whyte, W.A., Stratton, S.A., Barton, M.C., Gunaratne, P.H., Young, R.A., and Zwaka, T.P. (2010). Ronin/Hcf-1 binds to a hyperconserved enhancer element and regulates genes involved in the growth of embryonic stem cells. Genes Dev.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., Pellegrini, M., and Plath, K. (2013). Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization. Cell Stem Cell.

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D.J., Dash, C., Bazett-Jones, D.P., Le Grice, S., McKay, R.D.G., Buetow, K.H., et al. (2008). Global

Transcription in Pluripotent Embryonic Stem Cells. Cell Stem Cell.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature.

Esch, D., Vahokoski, J., Groves, M.R., Pogenberg, V., Cojocaru, V., Vom Bruch, H., Han, D., Drexler, H.C.A., Araúzo-Bravo, M.J., Ng, C.K.L., et al. (2013). A unique Oct4 interface is crucial for reprogramming to pluripotency. Nat. Cell Biol.

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

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

Fuhrmann, G., Chung, A.C.K., Jackson, K.J., Hummelke, G., Baniahmad, A., Sutter, J., Sylvester, I., Schöler, H.R., and Cooney, A.J. (2001). Mouse Germline Restriction of Oct4 Expression by Germ Cell Nuclear Factor. Dev. Cell.

Hammachi, F., Morrison, G.M., Sharov, A.A., Livigni, A., Narayan, S., Papapetrou, E.P., O'Malley, J., Kaji, K., Ko, M.S.H., Ptashne, M., et al. (2012). Transcriptional Activation by Oct4 Is Sufficient for the Maintenance and Induction of Pluripotency. Cell Rep.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589.

Heng, J.C.D., Feng, B., Han, J., Jiang, J., Kraus, P., Ng, J.H., Orlov, Y.L., Huss, M.,

Yang, L., Lufkin, T., et al. (2010). The Nuclear Receptor Nr5a2 Can Replace Oct4 in the Reprogramming of Murine Somatic Cells to Pluripotent Cells. Cell Stem Cell.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Hochedlinger, K., Blelloch, R., Brennan, C., Yamada, Y., Kim, M., Chin, L., and Jaenisch, R. (2004). Reprogramming of a melanoma genome by nuclear transplantation. Genes Dev.

Ito, K., and Suda, T. (2014). Metabolic requirements for the maintenance of self-renewing stem cells. Nat. Rev. Mol. Cell Biol.

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

Jauch, R., Ng, C.K.L., Saikatendu, K.S., Stevens, R.C., and Kolatkar, P.R. (2008). Crystal Structure and DNA Binding of the Homeodomain of the Stem Cell Transcription Factor Nanog. J. Mol. Biol. *376*, 758–770.

Jerabek, S., Merino, F., Schöler, H.R., and Cojocaru, V. (2014). OCT4: Dynamic DNA binding pioneers stem cell pluripotency. Biochim. Biophys. Acta - Gene Regul. Mech.

Kamachi, Y., Uchikawa, M., and Kondoh, H. (2000). Pairing SOX off: With partners in the regulation of embryonic development. Trends Genet.

Khan, A., and Zhang, X. (2016). DbSUPER: A database of Super-enhancers in mouse and human genome. Nucleic Acids Res.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008a). An Extended

Transcriptional Network for Pluripotency of Embryonic Stem Cells. Cell.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008b). An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. Cell *132*, 1049–1061.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I.R., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature.

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

Li, M., and Belmonte, J.C.I. (2017). Ground rules of the pluripotency gene regulatory network. Nat. Rev. Genet.

Li, M., and Izpisua Belmonte, J.C. (2018). Deconstructing the pluripotency gene regulatory network. Nat. Cell Biol.

Liang, J., Wan, M., Zhang, Y., Gu, P., Xin, H., Jung, S.Y., Qin, J., Wong, J., Cooney, A.J., Liu, D., et al. (2008). Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. Nat. Cell Biol.

Lichner, Z., Páll, E.A., Kerekes, A., Pállinger, É., Maraghechi, P., Bodouble acutesze, Z., and Gócza, E. (2011). The miR-290-295 cluster promotes pluripotency maintenance by regulating cell cycle phase distribution in mouse embryonic stem cells. Differentiation.

Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanog transcription network regulates

pluripotency in mouse embryonic stem cells. Nat. Genet.

Loh, Y.H., Zhang, W., Chen, X., George, J., and Ng, H.H. (2007). Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. Genes Dev. *21*, 2545–2557.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., et al. (2007). Directly Reprogrammed Fibroblasts Show Global Epigenetic Remodeling and Widespread Tissue Contribution. Cell Stem Cell *1*, 55–70.

Marks, H., and Stunnenberg, H.G. (2014). Transcription regulation and chromatin structure in the pluripotent ground state. Biochim. Biophys. Acta - Gene Regul. Mech.

Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc. Natl. Acad. Sci.

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

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766–770.

Meister, P., Towbin, B.D., Pike, B.L., Ponti, A., and Gasser, S.M. (2010). The spatial dynamics of tissue-specific promoters during C. elegans development. Genes Dev.

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

Moorthy, S.D., Davidson, S., Shchuka, V.M., Singh, G., Malek-Gilani, N., Langroudi, L., Martchenko, A., So, V., Macpherson, N.N., and Mitchell, J.A. (2017). Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. Genome Res.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods.

Mutz, K.O., Heilkenbrinker, A., Lönne, M., Walter, J.G., and Stahl, F. (2013). Transcriptome analysis using next-generation sequencing. Curr. Opin. Biotechnol.

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

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

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

Orkin, S.H. (2005). Chipping away at the embryonic stem cell network. Cell *122*, 828–830.

Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. Cell.

Park, S., Park, S.H., Kook, M.-C., Kim, E., Park, S., and Lim, J.H. (2004). Ultrastructure

of human embryonic stem cells and spontaneous and retinoic acid-induced differentiating cells. Ultrastruct. Pathol. *28*, 229–238.

Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., Van Bueren, K.L., Chines, P.S., Narisu, N., Black, B.L., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc. Natl. Acad. Sci. U. S. A.

Petryniak, B., Staudt, L.M., Postema, C.E., McCormack, W.T., and Thompson, C.B. (2006). Characterization of chicken octamer-binding proteins demonstrates that POU domain-containing homeobox transcription factors have been highly conserved during vertebrate evolution. Proc. Natl. Acad. Sci.

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

Reményi, A., Lins, K., Nissen, L.J., Reinbold, R., Schöler, H.R., and Wilmanns, M. (2003). Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. Genes Dev. *17*, 2048–2059.

Schepers, G.E., Teasdale, R.D., and Koopman, P. (2002). Twenty pairs of Sox: Extent, homology, and nomenclature of the mouse and human Sox transcription factor gene families. Dev. Cell.

Smale, S.T. (2010). Pioneer factors in embryonic stem cells and differentiation. Curr. Opin. Genet. Dev.

Song, Y., Ahn, J., Suh, Y., Davis, M.E., and Lee, K. (2013). Identification of Novel Tissue-Specific Genes by Analysis of Microarray Databases: A Human and Mouse

Model. PLoS One.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2014). Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. Cell *161*, 555–568.

Tada, M., Takahama, Y., Abe, K., Nakatsuji, N., and Tada, T. (2001). Nuclear reprogramming of somatic cells by in vitro hybridization with ES cells. Curr. Biol.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663–676.

Tapia, N., Maccarthy, C., Esch, D., Gabriele Marthaler, A., Tiemann, U., Araúzo-Bravo, M.J., Jauch, R., Cojocaru, V., and Schöler, H.R. (2015). Dissecting the role of distinct OCT4-SOX2 heterodimer configurations in pluripotency. Sci. Rep.

Thomson, M., Liu, S.J., Zou, L.N., Smith, Z., Meissner, A., and Ramanathan, S. (2011). Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell.

Tong, A.J., Liu, X., Thomas, B.J., Lissner, M.M., Baker, M.R., Senagolage, M.D., Allred, A.L., Barish, G.D., and Smale, S.T. (2016). A Stringent Systems Approach Uncovers Gene-Specific Mechanisms Regulating Inflammation. Cell *165*, 165–179.

Trounson, A., and DeWitt, N.D. (2016). Pluripotent stem cells progressing to the clinic. Nat. Rev. Mol. Cell Biol. *17*, 194–200.

Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D.N., Theunissen, T.W., and Orkin, S.H. (2006). A protein interaction network for pluripotency of embryonic stem cells. Nature.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl,

P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Xu, J., Pope, S.D., Jazirehi, A.R., Attema, J.L., Papathanasiou, P., Watts, J.A., Zaret, K.S., Weissman, I.L., and Smale, S.T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proc. Natl. Acad. Sci.

Xu, J., Watts, J.A., Pope, S.D., Gadue, P., Kamps, M., Plath, K., Zaret, K.S., and Smale, S.T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes Dev. *23*, 2824–2838.

Yamanaka, S. (2008). Pluripotency and nuclear reprogramming. Philos. Trans. R. Soc. B Biol. Sci.

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

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: Establishing competence for gene expression. Genes Dev.

Zaret, K.S., and Mango, S.E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. Curr. Opin. Genet. Dev.

CHAPTER 2

Critical Role for Oct4/Sox2 Binding to Composite Enhancer Motifs for the Establishment of Pluripotency

ABSTRACT

Despite the efforts to elucidate the molecular mechanisms underlying pluripotency, how Oct4 and Sox2x establish a pluripotent state remains elusive. As a first step, we examined ChIP-seq datasets and identified 1,000 sites within 15 kb of annotated genes at which Oct4 and Sox2 were strongly co-bound. By combining well-defined gene clusters, ChIP-seq data, and motif analysis, Oct4/Sox2 composite binding was highly prevalent near hundreds of genes expressed at high levels in ESC and a subset of somatic cell types. However, they account for only a small fraction of the 1,000 Oct4/Sox2 binding events. A large percentage of binding events were near silent or broadly expressed genes. This finding highlights our limited understanding of the critical role for Oct4/Sox2 binding to composite enhancer motifs for the establishment of pluripotency. Using our mouse secondary reprogramming cell model, we found the primary function of Oct4/Sox2 composite binding was to activate genes that are expressed in ESC but exhibit a large dynamic range of expression among somatic cell types. For the expression of these genes in a few differentiated cell types, other tissue-specific factors should support the transcription through an alternative mechanism. Notably, the Oct4/Sox2 binding at the enhancer composite motif of the ubiquitously expressed gene is not functional to either gene transcription or histone modification. Last, we confirmed the repressive role of the Oct4/Sox2 composite site for silencing the genes that are inactive in ESCs. These Oct4/Sox2 binding at the silent genes mediate transcriptional repression and contribute to a fraction of repressive histone marks H3K9me3 and H3K27me3 deposition. These findings uncovered the critical role for Oct4/Sox2 composite binding in pluripotency gene activation and ESC-silencing gene inactivation for the establishment of pluripotency.

INTRODUCTION

The discovery of embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs) has reshaped the fields of transplantation and regenerative medicine (Evans and Kaufman, 1981; Martin, 1981; Takahashi and Yamanaka, 2006). The maintenance and induction of pluripotency require distinct transitions in the transcriptional state to activate the pluripotency genes and inactivate the differentiation genes (Boyer et al., 2005; Ito and Suda, 2014; Orkin, 2005). The transcription factors that regulate gene expression in ESCs are critical to ESC identity and pluripotency. Master regulators including the Oct4, Sox2, and Nanog (OSN) transcription factors (TFs) are indispensable for governing ESCs identity through a complex hierarchy of gene regulation, which ensures the maintenance of pluripotency (Avilion et al., 2003; Mitsui et al., 2003; Nichols et al., 1998). Ectopic expression of Oct4, Sox2, Klf4, and c-Myc (OSKM) in mouse fibroblast induces pluripotency. These iPSCs give rise to tissue development both in vitro and in vivo (Maherali et al., 2007; Takahashi and Yamanaka, 2006). A transcriptional network, centering on Oct4 and Sox2, reveals target promoters or super-enhancers of pluripotent genes, suggesting a concrete mechanism toward pluripotency (Kim et al., 2008a; Whyte et al., 2013). The investigation of protein-interactome and protein-DNA interactions within this pluripotency gene regulatory network is often the first step to explore the functional roles of these master transcription factors in ESCs (Li and Belmonte, 2017; Li and Izpisua Belmonte, 2018).

Oct4 is a homeodomain transcription factor encoded by the *Pou5f1* gene, which contains a well-conserved Homeobox for DNA-binding (Reményi et al., 2003). Oct4 binds to a

consensus octameric DNA nucleotide sequence ATTTGCAT (Petryniak et al., 2006). The level of Oct4 has a restricted expression to pluripotent and germ cells (Fuhrmann et al., 2001). Oct4 is a critical regulator for pluripotency through various interfaces of protein-protein interaction, epigenetic regulation, and directly transcriptional regulation (Esch et al., 2013; Hammachi et al., 2012). In *Oct4*$^{-/-}$ embryos, although the embryo could develop to the blastocyst stage, the inner cell mass was not pluripotent. Furthermore, the trophoblast did not proliferate adequately (Nichols et al., 1998). More quantitative research performed in mouse ESCs revealed that the precise level of Oct4 expression is required for governing the stem cell self-renewal and lineage commitment (Niwa et al., 2000). A moderate increase of Oct4 led to the differentiation into primitive endoderm and mesoderm, while the repression of Oct4 lost the pluripotency and dedifferentiated to trophectoderm. In mouse ESCs, Oct4 contributes to the cell fate decisions during the transition to a differentiated cell state (Thomson et al., 2011). Together, these studies demonstrated the crucial role of Oct4 in maintaining pluripotency both *in vivo* and *in vitro*. As a master regulator for pluripotency, Oct4 controls the lineage commitment, and the requirement of the precise level illustrates a sophisticated gene regulatory program designated for pluripotency. In addition to direct transcriptional regulation of the target genes, Oct4 can recruit key epigenetic regulators to the target genes (Esch et al., 2013). The interactome of Oct4 discovered the unique protein interface of Oct4 to interact with Smarca4 and Chd4. The interaction improves the reprogramming efficiency and guard pluripotency by maintaining H3K27me3 in ESCs. Additionally, another study also characterized the endogenous association of Oct4 with proteins from multiple repression complexes such as NuRD, Sin3A, and Pm1 (Liang et al., 2008). The association with

repression complexes suggests a repressive role of Oct4 gene regulation, yet the mechanisms and precise targets remain unclear.

Another master transcription factor, Sox2, is a member of the Sox family transcription factors, which contains a high mobility group (HMG) box for DNA binding (Schepers et al., 2002). Sox2 binds to a 6 to 7 nucleotides DNA sequence CTTTGTC through the recognition of a core motif sequence TTGT (Kamachi et al., 2000). Silencing of Sox2 compromised self-renewal and differentiated ESCs into multiple lineages (Ivanova et al., 2006). In the Sox2 null ESCs, the cells differentiate into trophectoderm like cells (Masui et al., 2007). Likewise, Sox2 is also necessary to activate multiple transcription factors that are essential for stabilizing ESCs in the pluripotent state. Sox2 depletion resulted in aberrant expression of multiple transcriptional regulators for Oct4. The dysregulation leads to the decrease in Oct4 expression and eventual inactivation of the Oct4/Sox2-regulated genes. In ESCs, Sox2 often dimerizes with Oct4 and acts synergistically to activate their target genes (Avilion et al., 2003; Reményi et al., 2003; Tapia et al., 2015). The regulatory regions of these genes contain an Oct4 octamer motif juxtaposed to the Sox2 elements by a spacer nucleotide either 0 bp or 3 bp. The crystal structures of Oct4/Sox2 composite elements on *Fgf4*, *Utf1*, and *Nanog* revealed heterodimer conformations with various nucleotide spacers and distinct functions (Jauch et al., 2008; Reményi et al., 2003; Tapia et al., 2015). Hence, one of the roles of Sox2 in maintaining pluripotency appears to regulate the transcription factors that are necessary for the optimal expression of Oct4. In the pluripotency gene regulatory network, this Oct4/Sox2 cooperative binding acts as the core regulator to coordinate downstream cascade of the

pluripotency factor activation. The reciprocal transcriptional regulation between Oct4 and Sox2 reinforce their ability to maintain ESC pluripotency via the Oct4/Sox2 complex (Chew et al., 2005).

Previous computational and biochemical studies have demonstrated a hierarchy gene regulatory network to explore the complex mechanisms regulated by the master regulators in pluripotency. Employing genome-wide ChIP-chip/ChIP-seq profiling techniques, a number of groups defined the binding sites of Oct4 and Sox2 in both mouse or human ESCs (Apostolou et al., 2013; van den Berg et al., 2010; Chronis et al., 2017; Jerabek et al., 2014; Kim et al., 2008a; Orkin et al., 2008; Wang et al., 2006). On the basis of these studies, hundreds of Oct4/Sox2 target genes and mechanistic insight began to arise. Oct4 and Sox2 appear to bind cooperatively and target the genes known for their essential role in pluripotency. In the early stage of reprogramming, Oct4 and Sox2 occupied the somatic-enhancer to direct the inactivation of differentiation-promoting genes (Chronis et al., 2017). Together, the Oct4/Sox2 ChIP-seq and functional validation have validated several gene regulatory circuits required for pluripotency and reprogramming. The previous Oct4/Sox2 ChIP-seq analyses have revealed many characteristics of the Oct4 and Sox2 binding sites. However, the peak-centric approach focusing on TFs occupancy and statistical correlation is restricted by large sample sizes and biased towards genes with a small magnitude of variance. While most studies have relied on the low threshold to define the differential expression, they easily grouped these genes with small variances into ESC-specific or differentiation genes. The gene annotation of genome-wide Oct4/Sox2 binding sites revealed that Oct4 and Sox2 are

associated with genes with various degree of dynamic range of expression. However, the real functions of these thousands of Oct4/Sox2 composite binding sites remain unclear. Therefore, we need to use a different strategy to unveil the selective transcriptional regulation of Oct4 and Sox2.

In this study, we describe the gene-specific mechanisms and selective functions of Oct4/Sox2 in the pluripotency using a combination of bioinformatic and experimental approaches. By stringently comparing the nascent transcript levels in mouse ESC with a small number of primary somatic cell types, we classified the targets of Oct4/Sox2 based on their dynamic range of expression. This classification allowed us to interrogate the functions of the Oct4/Sox2 composite bindings further and examine the characteristics of Oct4/Sox2 binding sites. We analyzed the the binding strength, motif sequences, TF co-binding, chromatin accessibility, and histone modification at the Oct4/Sox2-bound enhancers. Furthermore, these analyses enabled us to initiate CRISPR/Cas9 experiments to delete Oct4/Sox2 composite elements in well-defined locations near genes within each gene class. We uncovered that Oct4/Sox2 selectively mediates ESC-specific / Dynamic gene activation and Silent gene inactivation, while the sites at the Non-Dynamic genes are non-functional.

RESULTS

General Features of Oct4/Sox2 ChIP-seq in ESCs

To understand the molecular mechanisms by which Oct4 and Sox2 contribute to the activation of pluripotency and self-renewal genes, we analyzed Oct4 and Sox2 ChIP-seq in ESC line V6.5 (GEO: GSE90895) (Chronis et al., 2017). We combined two biological replicates of Oct4 and Sox2 ChIP-seq and retained the reproducible peaks called by HOMER findPeaks algorithm (Benner et al., 2017). From the analysis of two biological replicates, we identified 15,506 Oct4 peaks and 11,207 Sox2 peaks (FDR < 0.01) genome-wide. Strikingly, only 499 Oct4 peaks (3.3%) and 341 Sox2 peaks (3.0%) bind to promoter regions (-500 bp to +150 bp relative to TSS). More than 90% of Oct4 and Sox2 peaks fall in intergenic or intronic regions (Figure 2-1A and B). The genomic distribution of called peaks indicates that the majority of Oct4 and Sox2 regulate gene transcription through the binding of enhancer regions.

Master transcription factors of pluripotency often form unusual enhancer domains with multiple co-binding regulators. A previous study uncovered a crystal structure of HOU/HMG/DNA ternary complex with Oct4/Sox2 heterodimer (Reményi et al., 2003; Whyte et al., 2013). To examine how many Oct4 and Sox2 co-bind genome-wide, we analyzed Oct4/Sox2 ChIP-seq to identify the composite binding sites. Among all 15,506 Oct4 peaks and 11,207 Sox2, we discovered 8,100 composite binding sites of which the distance between Oct4 and Sox2 is less than 100 bp (Figure 2-1C, left). Because the weaker peaks are less reproducible and are more likely to represent technical artifacts, we set a threshold of peak score > 20 to include only strongly reproducible Oct4 and Sox2

binding. After imposing the peak score cutoff, we found more than 80% of peaks (n = 3,092) exhibit composite binding. Furthermore, 1,035 strong Oct4/Sox2 composite binding sites are within 15 kbps of an annotated TSS (Figure 2-1C). The tendency of Oct4 and Sox2 to co-bind is striking. This suggests these composite sites may play a more fundamental role in regulating pluripotency.

Oct4/Sox2 composite elements exhibit two major heterodimer conformations (Reményi et al., 2003). Previous studies demonstrate, distinct Oct4/Sox2 heterodimer conformations have functional significance in pluripotency. A canonical composite element with juxtaposed Oct4 and Sox2 sites is more critical in reprogramming than those with three spacer nucleotides (Tapia et al., 2015). MEME de novo motif analysis at genomic loci with the top Oct4/Sox2 peak scores found almost 98% (791 out of 810) of these sites show the two proteins arranged into a juxtaposed composite element (Figure 2-1D). Overall, when we examined transcription factor binding by ChIP-seq, thousands of genomic loci were bound by Oct4 and Sox2. Although the majority of these sites display composite Oct4/Sox2 elements and have conserved composite recognition sequences, they are unlikely to function equivalently in pluripotency gene regulation. Mainly, The functional significance of Oct4/Sox2 binding to thousands of sites genome wide is still ambiguous.

Compare Nascent Transcript Profiles Between ESC and Selected Somatic Cell Types

It is possible that the thousands of Oct4/Sox2 composite binding sites regulate a small subset of genes in ESCs. While the statistical tests return the tissue-specific genes with more than 2-fold differential expression, we are still unable to distinguish between ESC genes that are constitutively expressed in many cell types and “real” ESC-specific genes. To understand the functional significance of Oct4 and Sox2, we used nascent transcript RNA-seq. This allows us to evaluate the dynamic range of expression profiles of ESCs in comparison to three other primary cell types (NEUR, BMDM, DP). Surprisingly, we found there are only a small number of genes with more than 5 RPKM that are approaching ESC specificity. Only 91 genes (3.01%) are expressed 20-fold or higher in ESC relative to the three cell types (Figure 2-2A). Most genes expressed in ESC show variance between 0.2 to 5-fold. Therefore, it is not a much ESC specificity to account for the function of thousands of Oct4/Sox2 composite sites.

Using stringent criteria, enabled us to separated genes based on their nascent transcript profiles (Figure 2-2B). For a gene to be classified as ESC-specific, it must meet a minimum expression of 5 RPKMs and have over 20-fold higher expression in ESCs compared to the other cell types. Overall, we identified 91 ESC-specific genes (e.g., *Pla2g1b*) with more than 20-fold higher expression in ESC relative to the somatic cell types (Figure 2-2C, left). 1,931 Non-Dynamic genes (e.g., *Hnrnpr*, *Pds5a*) are broadly expressed in all cell types with a small variance of 0.2 – 5-fold (Figure 2-2C, middle). We also identified 248 Dynamic genes with a 20-fold dynamic range of expression in at least one or two cell types. For example, *Zfp57* represents a Dynamic gene with 20-fold higher expression in BMDMs and DPs but is also expressed in NEUR (Figure 2-2C, right).

Interrogate Composite Oct4/Sox2 Bindings in Gene Groups with Different Dynamic Range of Expression

To identify the targets of Oct4/Sox2 in distinct gene classes, we compared the occupancy of Oct4 and Sox2 between ESC-specific and Non-Dynamic genes with various thresholds of ChIP-seq peak score. When included all peaks without setting any threshold, nearly 80% of ESC-specific genes were bound by either Oct4 or Sox2 (Figure 2-3A and B). Compare to Non-Dynamic genes (20%), both Oct4 and Sox2 show around 3 to 4-fold enrichment in ESC-specific genes. The fold enrichments greatly increase when adding more stringency onto peak score. While 40% of ESC-specific genes contain strong Oct4 binding (peak score > 20), only 7% of genes in Non-Dynamic class show similar strong binding (Figure 2-3A). With the same peak threshold, 47% of ESC-specific genes contain strong Sox2 binding, but only 7% of Non-Dynamic genes are bound (Figure 2-3B). Oct4 and Sox2 also display a higher tendency to form composite elements at ESC-specific genes. When focusing on peaks with a score greater than 20, 90% of Oct4 and 85% of Sox2 peaks present composite elements (n = 35) at ESC-specific genes. In contrast, one-fourth of Oct4 or Sox2 binds solely to a Non-Dynamic gene (Figure 2-3C). These demonstrated a greatly enriched Oct4/Sox2 composite binding in the vicinity of ESC-specific genes, providing the evidence of functional significance in pluripotency. However, these 35 composite elements binding at ESC-specific genes only account for a small fraction among all 1,035 strong Oct4/Sox2 sites within 15 kbps of annotated TSS.

To gain more insights into these 1,035 Oct4/Sox2 composite sites, we further interrogated the binding events in all three classes: ESC-specific, Dynamic, and Non-Dynamic genes. In addition, we also included those genes with transcript level less than 5 RPKM and grouped them into Silent genes in ESCs. As previously described, 35 out of 91 ESC-specific genes (38.46%) contain a strong Oct4/Sox2 binding. In contrast, only 101 out of 1,931 Non-Dynamic genes (5.23%) and 711 out of 17,434 Silent genes (4.08%) present a strong Oct4/Sox2 peak within 15 kbps (Figure 2-S1). The random distribution of a strong Oct4/Sox2 throughout the mouse genome is 4.47%, which is similar to the percentage in Non-Dynamic or Silent genes. Low percentage of Oct4/Sox2 binding at these two classes might represent non-functional binding or play an unusual role in pluripotency. Notably, genes grouped in Dynamic class also exhibit a range of 15% - 31% Oct4/Sox2 occupancy, which is close to ESC-specific genes. Based on the nascent transcript profiles, these Dynamic genes is also expressed in a few cell types and may not exhibit ESC-specificity or functional significance in pluripotency. However, the enrichment of strong binding at Dynamic genes might suggest another potential function of Oct4/Sox2 binding in pluripotency.

Although there is only 4 – 5% of Non-Dynamic and Silent genes contain a strong Oct4/Sox2 binding, it is surprising that 80% of 1,035 strong Oct4/Sox2 binds nearby the genes of these two classes. To compare the genomic features of Oct4/Sox2 peaks at different classes, we evaluated the motif sequence, Nanog co-binding, and the distance to the annotated TSS. MEME de novo motif analysis discovered the conserved recognition sequence of a juxtaposed Oct4/Sox2 composite motif with no spacer

nucleotide in all four classes (Figure 2-3D). Regardless of the gene classes, almost every genomic region under the peak contains a composite sequence that contributes to the de novo motif. The protein weight matrix also reveals a similar sequence of core motif and contiguous sequence. Nanog is one of the ESC-specific genes and a well-studied target of Oct4/Sox2 (Jauch et al., 2008). As one of the main proteins in the transcriptional network for pluripotency of embryonic stem cells, Nanog cooperatively mediates gene activation with Oct4 and Sox2 in many pluripotency genes (Kim et al., 2008a). Because Nanog binding could be a prominent co-factor to distinguish those functional binding from non-functional, we processed and analyzed Nanog ChIP-seq (GEO: GSE90895 / GSE44288) to examine Nanog binding in the targets of Oct4/Sox2 (Chronis et al., 2017; Whyte et al., 2013). However, among all 1,035 Oct4/Sox2 peaks, more than 80% of target genes also present a Nanog co-binding near the Oct4/Sox2 composite binding site (Figure 2-3E). When we set the different threshold of Nanog ChIP-seq peak score (20, 30, and 40), Non-Dynamic and Silent genes regularly show around 20% less Nanog co-binding at the Oct4/Sox2 composite binding sites. The higher percentage of Nanog co-binding at the ESC-specific and Dynamic genes might indicate these Oct4/Sox2 peaks are crucial for pluripotency. Although we only focused on annotating Oct4/Sox2 peaks within in 15 kbps of the TSS, the distribution of enhancer to the TSS is varied. We then evaluated the distance of enhancer to the annotated TSS (dTSS) among 1,035 Oct4/Sox2 peaks. Intriguingly, when plotting the distribution of dTSS against all genes in each class, most Oct4/Sox2 peaks at ESC-specific genes demonstrate a more proximal binding to the TSS (Figure 2-3F). In contrast, Oct4/Sox2 tend to locate at a more distal enhancer in other classes.

Oct4/Sox2 Preferentially Recognize H3K27Ac-marked Enhancers in ESC-specific Genes

The priming epigenetic conditions, the binding of pioneering factors, and the histone modification denote different activities of the DNA regulatory elements (Ernst and Kellis, 2010). In human ESCs, H3K27Ac is greatly enriched at the proximal enhancer of active genes in ESCs (Rada-Iglesias et al., 2011). To assess the whether the histone modification and chromatin accessibility varied between gene classes, we measured the RPKM level of H3K27Ac ChIP-seq (GSE 56138), H3K4me3 ChIP-seq (GSE 62380), and ATAC-seq (GSE 52397) at the 1,035 Oct4/Sox2 composite sites (Buecker et al., 2014; Ji et al., 2015; Di Stefano et al., 2016). Consistent with the previous observation, the composite elements at ESC-specific genes preferentially bind to the enhancers marked with a higher level of H3K27Ac (Figure 2-4A and B, left). Interestingly, although genes in Dynamic group are barely approaching ESC-specificity, the enhancers still show a moderate enrichment of H3K27Ac. Unlike histone H3 acetylation, H3K4me3 does not show significant enrichment in any gene classes (Figure 2-4A and B, middle). Previously, ENCODE H3K4me3 ChIP-seq revealed that the tri-methylation predominantly marks promoter regions (Calo and Wysocka, 2013). However, the majority of Oct4 and Sox2 peaks locate at intergenic or intronic regions (Figure 2-1A and B). The genomic distributions might explain why the RPKM level of H3K4me3 is low and does not distinguish the bindings at ESC-specific genes from other classes. Finally, we plotted the RPKM level of ATAC-seq at all Oct4/Sox2 composite binding sites to different gene classes. The reads intensity of ATAC-seq correlates with the genomic regions with open

configuration (Buenrostro et al., 2013). However, the ATAC-seq data do not necessarily show enrichment in any gene classes, suggesting Oct4/Sox2 composite elements do not preferentially bind to open configuration in any gene classes (Figure 2-4A and B, right). Notably, a large portion of Oct4 and Sox2 peaks are also targeting genomic regions with low accessibility. As the pioneer factors, some master transcription factors may recognize incomplete motifs at nucleosomes (Soufi et al., 2014; Xu et al., 2007). Previous analysis of Oct4 genomic distribution also found that OCT4 occupied low-accessible chromatin in undifferentiated human ESCs (Simandi et al., 2016).The binding at compacted chromatin regions may play an important role to initiate reprogramming and establish pluripotency state. The results also indicated that the Oct4/Sox2 binding at other classes could also be functional significance in pluripotency even the target genes are not even approaching ESC-specificity.

To quantitatively compare the histone modification across four gene classes, we separated H3K27Ac and H3K4me3 ChIP-seq data at the 1,035 Oct4/Sox2 sites into ten bins on descending order. Next, we examined how these bins were distributed in different gene classes. In both ESC-specific and Dynamic genes, almost 70% of the enhancers exhibit high H3K27Ac (Figure 2-4C) of bin 1 to 3. In contrast, 40% and 55% of Oct4/Sox2 bindings at Non-Dynamic and Silent genes locate at enhancers with low H3K27Ac (bin 6 to 10). In terms of H3K4me3, more than 35% of ESC-specific genes reside in bin 1, showing a strong enrichment when compared to only 6.75% in Silent genes (Figure 2-4D). Nevertheless, H3K4me3 ChIP-seq does not show a dramatic difference between gene classes. In Dynamic and Non-Dynamic genes, the Oct4/Sox2 binding sites display

similar distributions of H3K4me3 (Figure 2-4D). In summary, the Oct4/Sox2 composite bindings at ESC-specific and Dynamic genes are enriched with active histone modifications, demonstrating a functional significance of gene activation in pluripotency.

Disruption of Oct4/Sox2 Composite Binding in the ESC-specific Gene *Pla2g1b* Does Not Eliminate its Transcription

Through our previous characterization of the Oct4/Sox2 peaks, only a few dozen strong Oct4/Sox2 composite sites are nearby an ESC-specific gene. Although the genomic and epigenetic features show enrichment in ESC-specific genes, the function of the remaining 3,000 binding sites in maintaining pluripotency remains unclear.

Both Oct4/Sox2 binding and active enhancer histone mark H3K27Ac are enriched in Dynamic genes, this led us to hypothesize that composite binding is a critical regulator for not only ESC-specific genes but also genes in the Dynamic class (Figure 2-4 and S1). Based on the gene annotation of Oct4/Sox2 peaks, we hypothesized Oct4/Sox2 binding may also contribute to the small transcription changes in broadly-expressed Non-Dynamic genes. However, another explanation could be that Oct4/Sox2 binding is essential for broadly-expressed gene activation, with other tissue/cell-specific factors activating these genes in somatic cell types. While there are more than 711 peaks associated with Silent genes, we are uncertain of the role Oct4/Sox2 might play in repressing these genes (Figure 2-S1). Beyond the role of gene activation, some binding sites may contribute to nuclear organization, chromatin configuration, or could be redundant, or non-functional.

Because Oct4/Sox2 play a substantial role in both the maintenance of pluripotency and the induction of reprogramming, we first evaluated its contribution to the transcription of ESC-specific genes. We mutated composite motifs in CCE ESC lines using CRISPR. *Pla2g1b* is a classic ESC-specific gene. Both its nascent transcript profile and ENCODE polyA mRNA-seq datasets display an ESC-specificity of more than 20-fold differential expression between ESC and all other cell types (Figure 2-2C and 2-5A). Oct4/Sox2 composite element binds 1.3 kbps away from the TSS of *Pla2g1b*, which is the only strong Oct4/Sox2 binding in the 20 kbps window (Figure 2-5B). This enhancer displays an open configuration of chromatin and is marked with both H3K27Ac and H3K4me3. MEME motif analysis identified a conserved composite motif TTGTAATGCAAA with no spacer nucleotide at the Oct4/Sox2 peak region. To disrupt Oct4/Sox2 binding at the *Pla2g1b* enhancer, we generated mutant clones replacing the motif sequences wtih GGATCCGAATTC by CRISPR/Cas9 and homology-directed repair (HDR) mutation (Figure 2-5B). For each CRISPR-HDR mutation, we selected two mutant clones (C9 and H7) as biological replicates. Oct4/Sox2 motif mutation strongly diminished Oct4 and Sox2 binding at the *Pla2g1b* enhancer (Figure 2-5C and D). However, the transcription of *Pla2g1b* only reduced by 60% when compared to the wild-type (Figure 2-5E). The residual transcription of *Pla2g1b* was not even close to the inactive level in macrophages. Considering Oct4/Sox2 are the critical regulators for pluripotency, the residual level of transcription leads to an intriguing question of why the expression of an ESC-specific gene is not fully-dependent on Oct4/Sox2. One possibility is that this Oct4/Sox2 binding at the *Pla2g1b* enhancer exhibits a redundancy with other Oct4/Sox2 sites. However, we

could not locate other strong Oct4/Sox2 composite bindings any closer than 90 kbps away from the TSS of *Pla2g1b*. Another possibility is that this Oct4/Sox2 binding is indeed critical for activating gene transcription during the establishment of pluripotency and reprogramming, but it does not necessarily maintain the transcription in established ESC lines.

DOX-Inducible Secondary iPSCs

To examine the role of Oct4/Sox2 bindings throughout the establishment of pluripotency, we established a DOX-inducible system for mouse secondary reprogramming of TetO-OSKM iPSCs (Figure 2-S2). Previously, primary mouse embryonic fibroblasts (MEFs) harboring a single dox-inducible polycistronic cassette coding from OSKM were efficiently converted to a pluripotent state upon the addition of DOX (Chronis et al., 2017; Ho et al., 2013; Sridharan et al., 2013; Wernig et al., 2008). A similar DOX-inducible system in human cells also demonstrated a unique platform to differentiate these DOX-induced "primary" iPSCs into secondary fibroblasts, and subsequently culture with the presence of DOX to reprogram to pluripotent "secondary" iPSCs (Hockemeyer et al., 2008). In order to establish the secondary reprogramming model in mouse, we differentiated the primary iPSCs into embryoid bodies (EB) and subsequently cultured under neural progenitor cells (NPCs) growth condition with the presence of 0.5 μM retinoic acid (Lee et al., 2000; Sagner et al., 2018). To assure the elimination of the primary iPSCs, we cultured the differentiated NPCs using trypsin for at least four passages (Figure 2-S2A). EB displays a classic spheroid-like colony in suspension culture, and NPCs acquired short spindle shape morphology (Figure 2-S2B). To induce pluripotent state, we plated the cells in the

presence of 2 μg/ml DOX and LIF under mouse ESC culture conditions. After 14 to 20 days of DOX treatment, colonies with typical ESC-like morphology emerged (Figure 2-S2B). The selected single-cell colony could be further expanded into secondary iPSCs and well-maintained in the absence of DOX. Both primary and secondary iPSCs expressed pluripotency markers *Oct4*, *Sox2*, *Nanog*, *Ssea1*, *Klf4*, *Rex1*, and *Nr0b1*, which were utterly inactive in NPCs (Figure 2-S2C). In contrast, neural lineage-specific markers *Sox1*, *Nes*, *Pax6*, and *Pax3* were only expressed in NPCs (Figure 2-S2D). This secondary reprogramming assay demonstrates couple fundamental transitions between the pluripotent state and the differentiated state, which allows us to evaluate the gene transcription during the establishment of pluripotency.

Oct4/Sox2 Composite Binding is Critical for *Pla2g1b* Transcription During the Secondary Reprogramming

Next, we performed the same CRISPR-HDR mutation in the primary iPSCs and examined the changes in gene transcription. In the *Pla2g1b* mutants (A1-1 and A1-2), mutating the composite motif sequences diminished the Oct4/Sox2 bindings in both primary and secondary iPSCs (Figure 2-6A). Likewise, the disruption of Oct4/Sox2 binding only reduced approximately 60% of *Pla2g1b* transcription in primary iPSCs (Figure 2-6B). Another possible explanation for the residual transcription of *Pla2g1b* could be the transcription heterogeneity of the cell populations. However, an additional single-colony expansion of the original primary iPSCs showed that most mutant colonies still transcribe *Pla2g1b* at a moderate level (Figure 2-S3A). Extended the subculture passages also showed a consistent ~ 60% of reduced transcription in the mutants (Figure 2-S4A). These

results suggest that the residual expression of *Pla2g1b* is neither caused by transcription heterogeneity nor the transcription retention in the first couple passages of secondary iPSCs. When differentiated into NPCs, *Pla2g1b* showed substantial low transcription in both wild-type and the mutants. With the presence of Oct4/Sox2 binding, secondary reprogramming reactivated the *Pla2g1b* transcription to the equivalent level as primary iPSCs. Notably, the mutants reduced more than 95% of the wild-type *Pla2g1b* transcription and remained inactive in secondary iPSCs. The reduced gene transcription suggests that the Oct4/Sox2 composite binding at the enhancer is critical for the *Pla2g1b* gene activation during the establishment of pluripotency.

In the previous chromatin analysis, most of the ESC-specific gene enhancers possess strong enrichment of active histone marks H3K27Ac and H3K4me3 (Figure 2-4C and D). Likewise, the Oct4/Sox2-regulated *Pla2g1b* enhancer shows open chromatin configuration and H3K27Ac/H3K4me3 marking (Figure 2-6C, top). It was proposed that the master transcription factors established ESC-specificity, while the epigenetics barrier locked the cell status in a range of cell lineage (Smale, 2010; Zaret and Carroll, 2011). However, it is still obscure if Oct4/Sox2 directly or indirectly alters chromatin landscapes. To determine whether the Oct4/Sox2 binding contributes to H3K27Ac/H3K4me3 deposition, we measured H3K27Ac/H3K4me3 enrichment in CRISPR mutants (blue shades). In primary iPSCs, mutating Oct4/Sox2 composite motif reduced half H3K27Ac/H3K4me3 at peak 1 but did not show a significant difference at peak 2 (Figure 2-6C). The partially reduced H3K27Ac/H3K4me3 recapitulate the residual transcription of *Pla2g1b* in the mutants. Previous in fibroblast and bone marrow-derived macrophage,

IFNβ stimulation rapidly changed the transcription factor binding and RNA polymerase II recruitment and established epigenetic memory for faster and greater transcription upon restimulation (Kamada et al., 2018).

When differentiated into EB and NPCs, H3K27Ac and H3K4me3 level diminished simultaneously, indicating an ESC-specificity of the Oct4/Sox2-regulated enhancer. Upon the addition of DOX, H3K27Ac and H3K4me3 restored enrichment at the enhancer and correlated well with the kinetic changes of *Pla2g1b* transcription in wild-type (Figure 2-6B and C). Conversely, in the mutants, the level of H3K27Ac and H3K4me3 at both loci did not restore after the secondary reprogramming. These epigenetic changes indicate that the Oct4/Sox2 composite elements are also critical for H3K27Ac and H3K4me3 deposition at the *Pla2g1b* enhancer. Thus, by investigating the ESC-specific gene, *Pla2g1b*, we demonstrated that Oct4/Sox2 composite binding plays an essential role to activate gene transcription and optimize histone modification in pluripotency.

Oct4/Sox2 Composite Bindings are Essential for the Dynamic Gene Transcription, Yet Other Factors Are Required for the Transcription in Lineage Differentiation

Next, we began to investigate whether the Oct4/Sox2 composite binding is also a critical regulator of Dynamic genes that exhibit high expression in ESC and deficient expression in at least some somatic cell types (Figure 2-2B, right). First, we mutated the Oct4/Sox2 composite motif at the enhancer of *Zfp57* by the same CRISPR-HDR mutation approach. *Zfp57* is an example of Dynamic gene expressed in ESCs and neural lineage, but the transcription is mostly inactive in other tissue/cell types (Figure 2-1C, right, and Figure 2-

7A). In ESCs, Oct4/Sox2 binds at the *Zfp57* enhancer exhibiting open chromatin configuration and high H3K27Ac (Figure 2-7B). However, the genomic region is neither enriched with H3K27Ac nor H3K4me3 in NPCs (E1 and E2). Alternatively, a proximal high H3K27Ac and H3K4me3 marked the first intron of *Zfp57* (E3 and E4). The distinct chromatin landscapes indicate a possible lineage-specific mechanism to activate *Zfp57* transcription in the absence of Oct4/Sox2 binding.

In the *Zfp57* mutants (B4-1 and B4-2), mutating the composite motif sequences diminished the Oct4/Sox2 bindings in both primary and secondary iPSCs (Figure 2-7C). In the absence of Oct4/Sox2 binding, the *Zfp57* mutants reduced two-third of transcription in primary iPSCs (Figure 2-7D). Both single-colony expansion and extended subculture passages of the CRISPR-mutated primary iPSCs showed a consistent 60% reduction of *Zfp57* transcription (Figure 2-S3B and 2-S4B). This consistency indicates that the population heterogeneity does not contribute to the residual *Zfp57* transcription. The level decreased further in EBs but then reactivated to ~ 75% in NPCs, which is consistent with the wild-type. The reactivation of *Zfp57* in both clones indicates an alternatively regulatory element for the transcription in NPCs. Moreover, this lineage-specific mechanism does not regulate through the Oct4/Sox2 composite motif. However, the transcription of *Zfp57* dropped significantly and remained silent in the mutants after the reprogramming. The reduction suggests that Oct4/Sox2 is critical for *Zfp57* transcription in the pluripotent state, while an alternative mechanism is meanwhile responsible for active transcription in NPCs. Since two lineage-specific mechanisms are responsible for the *Zfp57* transcription in ESCs and NPCs, we then questioned if the Oct4/Sox2 motif mutation affects the histone

marks deposition differently. Centered on the Oct4/Sox2 composite binding site, both E1 and E2 showed significant loss of H3K27Ac enrichment in the mutants after the secondary reprogramming (Figure 2-7E). H3K4me3 level reduced ~ 60% when compared to the wild-type. Since H3K4me3 did not show strong enrichment even in the wild-type, the fold changes of the H3K4me3 enrichment could, therefore, be moderate. Unlike E1 and E2 sites, disruption of Oct4/Sox2 binding did not affect H3K27Ac and H3K4me3 level at E3 and E4 sites (Figure 2-7F). Both histone marks demonstrated similar deposition kinetics throughout the differentiation and secondary reprogramming. Although H3K4me3 did not fluctuate much at E3 site, both H3K27Ac and H3K4me3 summited at the stage of NPCs. The dynamic changes of H3K27Ac and H3K4me3 deposition support that the Oct4/Sox2-regulated enhancer is ESC-specific, while another enhancer at the first intron of *Zfp57* is responsible for transcription in the differentiated neural lineage.

In the vicinity of the Dynamic class, some genes may show large dynamic range of expression between ESCs and the others but are still broadly expressed in more than one somatic tissue/cell types. For example, *Epb4.1l5* is a Dynamic gene expressed in both ESCs and NEUR in our nascent transcript profiles (data not shown). When compared with a broader range of tissue/cell types, in addition to neural tissues, *Epb4.1l5* is also expressed in thymus, fat pad, and kidney (Figure 2-8A). Oct4/Sox2 also bind strongly to a distal *Epb4.1l5* enhancer with high ATAC signal and H3K27Ac enrichment only in ESCs (Figure 2-8B). At the proximal end of the *Epb4.1l5* TSS, both ESCs and NPCs show the enrichment of H3K27Ac and H3K4me3, indicating an active transcription in both cell types. To interrogate the function of Oct4/Sox2 composite binding, we mutated

the composite motif sequences and validated the binding with Oct4- / Sox2-ChIP-qPCR (Figure 2-8C). Similar to the observation in *Zfp57*, although the mutants (F3 and G11) only reduced 70% of transcription in the primary iPSCs, *Epb4.1l5* became silent and remained inactive after the secondary reprogramming (Figure 2-8D). In the CRISPR-mutated primary iPSCs, the reduction of *Epb4.1l5* was consistent in all expanded single colonies and remained unchanged for extended subcultures (Figure 2-S3C and 2-S4C). The *Epb4.1l5* transcription went down in EBs and reactivated in NPCs regardless of the intact of Oct4/Sox2 composite site, suggesting a different mechanism for gene activation in the neural lineage. Consistent with the function in another Dynamic gene *Zfp57*, Oct4/Sox2 composite binding is essential for *Epb4.1l5* activation during the secondary reprogramming.

In terms of the changes in chromatin landscapes, centered on the Oct4/Sox2 site, E1 and E2 lost H3K27Ac enrichment in the mutants (Figure 2-8E). H3K4me3 deposition remained low and did not change at both E1 and E2 sites. The H3K4me3 ChIP-seq in ESCs also reveals very few reads are aligned to E1 and E2 sites, suggesting H3K4me3 is less likely to mark this Oct4/Sox2-regulated enhancer (Figure 2-8B). Interestingly, at E3 and E4 sites, most H3K27Ac and H3K4me3 level remained unchanged and summited in NPCs except the mutants of secondary iPSCs (Figure 2-8F). After the secondary reprogramming, both H3K27Ac and H3K4me3 level reduced by 50% to 60% in the mutants. It is unclear why both histone mark depositions near the *Epb4.1l5* TSS is altered. One possible explanation could be attributed to the transcriptional inactivation in the absence of Oct4/Sox2 binding after the reprogramming. H3K27Ac and H3K4me3 often

mark the enhancer and promoter of genes with active transcription (Calo and Wysocka, 2013). Since the *Epb4.1l5* transcription remained silent, the E3/E4 site near the TSS might not be able to recruit necessary chromatin remodeler to retain the active histone marks deposition.

Previously, genes with transcription like *Zfp57* or *Epb4.*1l5 would never be considered as ESC-specific genes. When evaluated the function of master transcription factors, most studies often paid less attention to the functions of the Oct4/Sox2 associated with them. These Dynamic genes are indeed highly expressed in the ESCs but also expressed in a handful of somatic cells. When compared the transcript profiles with a broad range of tissue/cell types, the Dynamic genes still exhibit a large dynamic range of expression between ESCs and the others (Figure 2-1C, Figure 2-7A, and Figure 2-8A). Together, by investigating the Dynamic genes, *Zfp57* and *Epb4.1l5*, we demonstrated that Oct4/Sox2 composite binding is an essential role in transcription of the Dynamic genes, with other factors activating these genes in a few somatic cell types. Although the Dynamic genes are not even ESC-specific, the Oct4/Sox2-regulated enhancers exhibit ESC-specificity and functional significance. These Oct4/Sox2 bindings also optimize histone modification for proper transcription in pluripotency.

Oct4/Sox2 Composite Binding is not Functional to the Transcription of the Non-Dynamic Genes

One intriguing discovery is, when examined the Oct/Sox2 ChIP-seq, is that a hundred of Non-Dynamic genes contain a strong composite binding within 15 kbps of the TSS (Figure

2-S1). These Non-Dynamic genes are actively transcribed in a broad range of tissue/cell types (Figure 2-1B, middle), but Oct4 and Sox2 are exclusively expressed in the pluripotent state. Logically, a different mechanism rather than Oct4/Sox2 is responsible for the gene activation in the somatic lineage tissue/cell types. The main question left unsolved is whether these Oct4/Sox2 bindings exclusively activate the Non-Dynamic genes in the ESCs (similar to the Dynamic genes), or they are not functional to the transcription at all. In the analysis above, we found that 99 out of 101 Oct4/Sox2 sites at Non-Dynamic genes also exhibit a conserved composite motif sequence (Figure 2-3D). Although Oct4/Sox2 preferentially binds to a more distal enhancer of Non-Dynamic genes with less H3K27Ac enrichment and Nanog co-biding (Figure 2-3E and F, 2-4A and B), little we have learned about the functions of these Oct4/Sox2 bindings.

To understand the function of the strong Oct4/Sox2 composite bindings at the Non-Dynamic genes, we mutated the motif sequences at the enhancer of *Pds5a* or *Hnrnpr* by the same CRISPR-HDR mutation as described. *Pds5a* and *Hnrnpr* are both classic Non-Dynamic genes with active transcription in ESCs and exhibit 0.2 to 5-fold variance between all other cell types (Figure 2-S5A and B). In ESCs, Oct4/Sox2 bind to the upstream intergenic region of *Pds5a* and the third intron of *Hnrnpr* (Figure 2-9A and D). Both composite bindings locate at the chromatin with open configuration but deficient with H3K27Ac or H3K4me3 deposition (Pds5a_E1, Hnrnpr_E1). When we zoomed out to look for additionally accessible chromatin regions, we identified another open chromatin with enriched H3K27Ac and H3K4me3 in both further upstream of *Pds5a* and the promoter of *Hnrnpr* (Pds5a_E2, Hnrnpr_E2). The mutation of the composite motif diminished Oct4

and Sox2 bindings in both the *Pds5a* mutants (E7 and H11) and the *Hnrnpr* mutants (B5 and E8) (Figure 2-9B and E). After the secondary reprogramming, Oct4 and Sox2 remained unbound in the mutants. Strikingly, in the absence of Oct4/Sox2 binding, the transcription of *Pds5a* did not change and remained consistent with the wild-type (Figure 2-9C). Over the differentiation and secondary reprogramming, the mRNA levels of *Pds5a* fluctuated with small variances but were steady between the wild-type and the mutants (E7 and H11). Similarly, Oct4/Sox2 motif mutation did not inactivate the transcription of *Hnrnpr* either (Figure 2-9F). The *Hnrnpr* transcription exhibited bigger kinetic changes throughout the differentiation and reprogramming, which showed two-fold induction in the NPCs regardless of the Oct4/Sox2 binding. The unchanged *Pds5a* and *Hnrnpr* transcription suggest that the Oct4/Sox2 composite bindings are not required for their transcription.

Besides the transcriptional changes of *Pds5a* and *Hnrnpr*, we also evaluated the potential roles of the Oct4/Sox2 bindings in histone modification. Centered on the Oct4/Sox2 site, depletion of Oct4/Sox2 binding neither modified the H3K27Ac nor the H3K4me3 level at the Pds5a_E1 site (Figure 2-9G). The H3K27Ac and H3K4me3 enrichment at the distal upstream Pds5a_E2 site also remained unchanged over the differentiation and the reprogramming. In the Hnrnpr mutants, the proximal promoter Hnrnpr_E1 demonstrated a comparable level of H3K27Ac and H3K4me3 enrichment as the wild-type, indicating the active transcription of *Hnrnpr* (Figure 2-9H). The intronic Oct4/Sox2 binding site remained unmarked by H3K27Ac and H3K4me3 regardless of the Oct4/Sox2 binding. Together,

the Oct4/Sox2 bindings at the Non-Dynamic genes *Pds5a* and *Hnrnpr* neither activate the transcription nor modify histone modification when establishing pluripotency.

In the vicinity of Non-Dynamic class, the majority of the Oct4/Sox2 composite sites lack the active histone marking. However, there still are 10% of the Oct4/Sox2 peaks locating at the chromatin regions with enriched H3K27Ac and H3K4me3 (Figure 2-4C and D). Unlike the bindings at *Pds5a* or *Hnrnpr*, these Oct4/Sox2 binding at an active enhancer of the Non-Dynamic gene could function differently as H3K27Ac/H3K4me3 often marked the genes with active transcription. To further understand the function of Oct4/Sox2 bindings at the Non-Dynamic genes, we mutated the Oct4/Sox2 motif sequences at *Dido1* and *Ift52*. Nascent transcript profiles and ENCODE polyA mRNA-seq datasets both showed high RPKM level of *Dido1* and *Ift52* with the small variance between ESCs and other tissue/cell types (Figure 2-S5C and D). Focusing on *Dido1*, Oct4/Sox2 binds to the upstream genomic locus with high ATAC signal and enriched H3K27Ac/H3K4me3, which represent an active enhancer (Figure 2-10A, dashed box). Centered on *Dido1*, three other genes *Tcfl5* (blue), *Gid8* (red), and *Slc17a9* (green) reside within 200 kbps window of the *Dido1* neighborhood. *Slc17a9* is another Dynamic gene with no strong Oct4/Sox2 binding near 15 kbps of the TSS (Figure 2-S6A). *Gid8* does not belong to any gene groups as the dynamic range of expression is 6.2-fold and 6.3-fold when compared to NEUR and BMDM, respectively (Figure 2-S6B). *Tcfl5* is considered not expressed because the nascent transcript does not meet the 5 RPKM threshold for ESC-expressing gene (Figure 2-S6C). CRISPR-HDR mutation (A12 and F9) of the motif sequences disrupted the bindings of Oct4 and Sox2 in both primary and secondary iPSCs (Figure 2-10B). However, the

mutation did not affect the transcription of *Dido1* or any neighboring genes within 200 kbps (Figure 2-10C). Over the differentiation and the reprogramming, the mRNA level of *Dido1, Tcfl5*, *Gid8*, and *Slc17a9* were unchanged between the wild-type and the mutants. Given the fact that histone modification or chromatin accessibility does not always correspond to gene expression, this Oct4/Sox2 could indeed non-functional to the pluripotency gene transcription. Although it is still possible that this Oct4/Sox2 binding is mediating another gene transcription further away through the long-range chromatin interaction, the Oct4/Sox2 binding does not regulate any gene transcription within this 200 kbps window.

In *Ift52*, the Oct4/Sox2 composite element targets at the intronic region 14.4 kbps away from the TSS. This binding site also exhibits high ATAC signal and H3K27Ac enrichment, indicating an accessible active enhancer (Figure 2-10D, dashed box). In the 120 kbps neighborhood, *Sgk2* locates 30 kbps upstream of *Ift52* (blue). Two other genes, *Mybl2* and *Gtsf1l,* are 38 kbps and 75 kbps downstream of *Ift52,* respectively (red and green). The nascent transcript profiles of ESCs and the three cell types revealed that *Skg2* and *Gtsf1l* are both inactive with less than 2 RPKM expression in ESCs (Figure 2-S6D and F). Conversely, *Mybl2* is actively transcribed in ESCs and show large dynamic range of expression with NEUR and DP, which is, therefore, a Dynamic-gene (Figure 2-S6E). Considering the binding proximity, our previous ChIP-seq analysis annotated these Oct4 and Sox2 peaks to *Ift52* rather than *Mybl2*. To investigate whether the Oct4/Sox2 binding at *Ift52* could be a distal enhancer for the neighboring genes, we mutated the composite motif at the *Ift52* intron. The Oct4/Sox2 binding was similarly curtailed by mutating the

composite motif sequences in primary and secondary iPSCs (G5 and G11) (Figure 2-10E). Consistent with our previous discoveries in the Oct4/Sox2 bindings at the Non-Dynamic genes, mutating the intronic Oct4/Sox2 site did not alter the transcription of *Ift52*. There were also no changes to the mRNA level of two neighboring genes *Sgk2* and *Gtsf1l* (Figure 2-10F). Interestingly, the Oct4/Sox2 binding at the *Ift52* intron had moderate regulation to the *Mybl2* transcription. The level of *Mybl2*, like most of the Dynamic genes, diminished during the differentiation but restored after the secondary reprogramming. Mutating the intronic Oct4/Sox2 site reduced 80% of the *Mybl2* transcription in the secondary iPSCs. The reduced transcriptional reactivation in the reprogramming suggests that this distal Oct4/Sox2-regulated enhancer contributes to the *Mybl2* gene activation in ESCs. However, none of the mutants reduced the gene expression to completely silent level, suggesting this Oct4/Sox2 site is not sufficient for fully activating *Mybl2* during reprogramming. The residual 20% transcription might be explained by the enhancer activity of another proximal intronic Oct4/Sox2 site 8,000 bp away from the *Mybl2* TSS. This *Mybl2* intronic Oct4/Sox2 site also contains Nanog co-binding and shows enriched H3K27Ac/H3K4me3 at open chromatin configuration (data not shown). It is likely that these two Oct4/Sox2 sites (*Ift52* intronic and *Mybl2* intronic) demonstrate synergistic function to activate *Mybl2* transcription. Interestingly, both Oct4/Sox2 sites located at the super-enhancer regions (data discussed in Chapter 3), supporting that more than one Oct4/Sox2 bindings might be required for pluripotency gene transcription (Hnisz et al., 2013; Whyte et al., 2013).

In summary, although these Oct4/Sox2 peaks are strong and harbor conserved composite motif, the composite sites are not required for the transcription of the annotated Non-Dynamic genes in ESCs. There is a possibility that they interact with another gene hundreds of kbps away. One example we have investigated is the *Ift52* intronic Oct4/Sox2 site, which moderately regulates the *Mybl2* transcription from 22 kbps upstream of the TSS. This distal Oct4/Sox2-regulated enhancer may not be as critical as those exclusive Oct4/Sox2 sites binding within 15 kbps of ESC-specific or Dynamic genes. We did not detect any change of histone mark for active enhancer (H3K27Ac) in any mutants (data not shown). This might result from the synergistic role with another proximal *Mybl2* Oct4/Sox2 site. This might also imply that histone mark deposition is dependent on different mechanisms on *Mybl2*.

Oct4/Sox2 Composite Binding Mediates Transcriptional Repression in the Silent Genes

The maintenance of pluripotency and induction of reprogramming require the activation of pluripotency genes that are entirely inactive in differentiated cells and the silencing of genes that are active in differentiated states. A few studies have discovered that Oct4 and Sox2 may associate with both activator and repressor complex (Ang et al., 2011; Pardo et al., 2010; Pasini et al., 2010). Meanwhile, we identified a strikingly high number (n= 711) of ESC-silent genes contained a strong Oct4/Sox2 composite binding within 15 kbps of the annotated TSS (Figure 2-S1). This finding has led us to hypothesize that, in ESCs, Oct4/Sox2 not only activates genes of the pluripotency network but also simultaneously represses the differentiation-specific genes.

Because these Oct4/Sox2 binds in the vicinity of the silent genes, we first evaluated the binding at the enhancer lacking active histone marks (H3K27Ac and H3K4me3) but with repressive histone marks (H3K9me3 and H3K27me3) deposition. *Oxgr1* is a silent gene in ESCs with very low RPKM level in both nascent transcript and polyA mRNA-seq (Figure 2-S7A). ChIP-seq datasets demonstrated a strong Oct4/Sox2 composite binding at the *Oxgr1* intronic region moderately enriched with both repressive histone marks H3K9me3 and H3K27me3 (Figure 2-11A, E1). Although there is an ATAC-seq peak at the Oct4/Sox2 site, neither H3K27Ac nor H3K4me3 is marking. To assess whether Oct4/Sox2 composite binding represses *Oxgr1* transcription, we mutated the motif sequences at the intronic enhancer. Mutating the composite motif strongly compromised the Oct4/Sox2 binding at the enhancer (Figure 2-11B). The enhancer of mutants (C9-1 and C9-2) remained unbound by Oct4/Sox2 after the secondary reprogramming. Notably, mutating the Oct4/Sox2 site induced almost thirty-fold of the transcription in secondary iPSCs (Figure 2-11C). In the primary iPSCs, depletion of Oct4/Sox2 binding did not alter the expression. Two likely explanations may account for the unchanged *Oxgr1* expression. First, the stably maintained iPSCs have recruited necessary components during the induction of pluripotent state. Removing the binding of Oct4/Sox2 is not sufficient to dissociate the repression complex from the enhancer. Second, the epigenetic memory establishes the barrier to avoid differentiation-specific gene expression (Papp and Plath, 2013). When the primary iPSCs began to differentiate, external stimulation promoted the dissociation of the repression complex and reshaped the chromatin landscapes. After the

secondary reprogramming, the mutants could not recruit the necessary repression complex or optimize histone modification for the transcriptional repression of *Oxgr1*.

To determine whether the Oct4/Sox2 binding optimizes the deposition of active and repressive histone marks, we measured H3K27Ac, H3K4me3, H3K9me3, and H3K27me3 in the CRISPR mutants. Consistent with the ChIP-seq data, the low enrichment level of H3K27Ac and H3K4me3 did not change over the differentiation and reprogramming at E1 site (Figure 2-11D). Although mutating the Oct4/Sox2 site activated the *Oxgr1* transcription in secondary reprogramming, H3K27Ac/H3K4me3 deposition did not change significantly. This result indicates that Oct4/Sox2 does not optimize H3K27Ac/H3K4me3 at the *Oxgr1* enhancer. On the other hand, E2 and E3 sites are enriched with both repressive histone marks, corresponding with the inactive transcription of *Oxgr1* (Figure 2-11A). Interestingly, mutating the Oct4/Sox2 site reduced half of the H3K9me3 and H3K27me3 at E2 site and E3 site, respectively (Figure 2-11E). This indicates that the Oct4/Sox2 binding alone is insufficient for optimal H3K9me3 and H3K27me3 binding. The other half of H3K9me3 and H3K27me3 signal might depend on other mechanisms. For example, the Polycomb repressive complex 2 (PRC2) is responsible for catalyzing the trimethylation of H3 residue at lysine 27 (Boyer et al., 2006; Ezhkova et al., 2009; Lee et al., 2006; Xu et al., 2014). Also, another methyltransferase SETDB1 contributes to H3K9me3 (Bilodeau et al., 2009; Rea et al., 2000; Schultz et al., 2002).

To validate the repressor role of the Oct4/Sox2 binding at Silent genes, we mutated the composite motif sequences in another Silent gene *Gnrhr* (Figure 2-S7B). Oct4/Sox2 composite element targets the *Gnrhr* enhancer with ATAC signal and H3K9me3 enrichment (Figure 2-12A, E1). We did not observe significant enrichment of active histone marks (H3K27Ac and H3K4me3). Conversely, H3K9me3 and H3K27me3 marking extend over the *Gnrhr* gene body with the strongest peak at E2 and E3 sites. Mutating the enhancer Oct4/Sox2 motif sequences impaired both factors binding in iPSCs (Figure 2-12B). Likewise, in the absence of the Oct4/Sox2 binding, the transcription of *Gnrhr* induced for more than forty-fold after the reprogramming (Figure 2-12C). This result again confirms that the Oct4/Sox2 composite element mediates transcriptional repression at the enhancer of Silent genes. However, mutating the Oct4/Sox2 site did not change H3K27Ac and H3K4me3 signal at E1 site (Figure 2-12D). H3K9me3 also stayed enriched at E1, E2, and E3 site (Figure 2-12E, top), suggesting the binding is not necessarily responsible for the histone marks deposition at the enhancer. The mutants reduced less than one-third of H3K27me3 level at E2 and E3 after the secondary, yet the changes were not significant (Figure 2-12E, bottom).

In summary, we demonstrated that Oct4/Sox2 composite element mediated transcriptional repression when associated with the Silent genes *Oxgr1* and *Gnrhr*. As a transcriptional repressor, Oct4/Sox2 might directly or indirectly contribute to repressive histone marks deposition with gene-specific mechanisms. However, Oct4/Sox2 does not contribute to the entire H3K9me3 or H3K27me3 marking. Other Oct4/Sox2-independent histone modifiers should regulate the repressive histone marks deposition more directly.

The residual H3K9me3/H3K27me3 might also indicate that *Oxgr1* and *Gnrhr* are not fully-activated in the absence of the Oct4/Sox2 bindings, despite the fact of a 30 to 40-fold induced transcription. It is also likely that other pluripotency transcription factors or histone modifiers can contribute to the silencing of *Oxgr1* and *Gnrhr,* and act independently with Oct/Sox2 mediated transcriptional repression.

Besides the majority of Silent genes, about another 50 genes contain an Oct4/Sox2 composite site marked with H3K27Ac and H3K4me3 (Figure 2-4C and D). Because these Oct4/Sox2 bind at an active enhancer, they may not act as a repressor for the annotated Silent genes. Instead, they may function differently as H3K27Ac and H3K4me3 often correspond to the genes with active transcription. To gain more insights into the Oct4/Sox2 bindings at the Silent genes, we mutated the Oct4/Sox2 motif sequences at *Uba7* and *Lax1*. *Uba7* is silent in ESCs but actively transcribed in the immune lineage cells (BMDM, DP, B cell, T cell, and Thymus) (Figure 2-S7C). *Lax1* is also inactive in ESCs but expressed in lymphocyte lineage cells (DP, B cell, T cell, and Thymus) (Figure 2-S7D). Focusing on *Uba7*, Oct4/Sox2 binds to the upstream site with high ATAC signal and enriched H3K27Ac, which is also the intronic region of *Traip* (Figure 2-13A, dashed box). Besides *Traip* (red), *Camkv* (blue) also located at 30 kbps upstream of *Uba7* while another gene *Ip6k1* (green) is 25 kbps downstream (Figure 2-13A). All three genes reside within 120 kbps window of the *Uba7* neighborhood. *Camkv* is another Silent gene with low expression in ESCs but highly expressed in NUER (Figure 2-S8A). No strong Oct4/Sox2 binding is found in 15 kbps of the *Camkv* TSS (Figure 2-13A). *Traip* and *Ip6k1* also belong to the Silent gene class since the nascent transcripts do not meet the 5 RPKM

cutoff for ESC-expressing genes (Figure 2-S8B and C). Because this Oct4/Sox2 binding is within 15 kbps to both the TSS of *Uba7* and *Traip*, we annotated this site to both genes in our previous ChIP-seq analysis. In the mutants (A1 and G9), Oct4 and Sox2 did not bind to the enhancer in both primary and secondary iPSCs (Figure 2-13B). However, the mutation did not affect the transcription of *Uba7* or upstream genes *Traip* and *Camkv* (Figure 2-13C). These two genes remained inactive with high Ct value above 33 cycles (data not shown), suggesting Oct4/Sox2 did not mediate repression in any of them.

Interestingly, mutating the Oct4/Sox2 site slightly induced *Ip6k1* for 4 to 5-fold after the secondary reprogramming (Figure 2-13C). Although a 4-fold induction is only two cycles difference with the wild-type, the Oct4/Sox2 binding is possibly involved in *Ip6k1* inactivation when establishing pluripotency. Over the differentiation and the reprogramming, *Ip6k1* induced fifteen-fold in differentiation cells but is subsequently inactivated upon the induction of reprogramming. The dynamic change and the induced transcription of *Ip6k1* in the mutants support our hypothesis that Oct4/Sox2 also mediate the silencing of genes that are active in the differentiated states. Since the Oct4/Sox2 site is about 40 kbps away for the *Ip6k1* TSS, the site could be less critical for the repression. When we examined the Oct4/Sox2 ChIP-seq at the vicinity of *Ip6k1*, we found another Oct4/Sox2 site at the *Ip6k1* promoter, but the peak score is low (Figure 2-13A, green). These two Oct4/Sox2 sites could act synergistically to mediate *Ip6k1* inactivation in the ESCs. Mutating the distal Oct4/Sox2 site only did not induce *Ip6k1* dramatically when the promoter Oct4/Sox2 retained.

Last, we mutated the Oct4/Sox2 site at the upstream intergenic region of *Lax1*. Likewise, Oct4/Sox2 site binds to open chromatin with high ATAC signal and H3K27Ac enrichment (Figure 2-13D, dashed box). In the 120 kbps window of *Lax1* neighborhood, *Atp2b4* locates 64 kbps upstream of *Lax1* (Figure 2-13D, green). Two other genes, *Zc3h11a* and *Zbed6,* are 28 kbps and 29 kbps downstream of *Lax1*, respectively (blue and red). The nascent transcript profiles of ESCs and the three cell types revealed that *Zc3h11a* and *Zbed6* are both Non-Dynamic genes expressed in ESCs (Figure 2-S8D and E). In the Oct4/Sox2 ChIP-seq, we found another Oct4/Sox2 site near *Zc3h11a* and *Zbed6* (Figure 2-13D). However, the peak score of Sox2 did not meet the cutoff threshold of peak score > 21, which is, therefore, not considered as a strong Oct4/Sox2 composite binding. Conversely, *Atp2b4* is another Silent gene with only 1 RPKM in ESCs and is also inactive in three other cell types (Figure 2-S8F). Consistent with the previous findings, mutating the Oct4/Sox2 motif impaired the binding at the *Lax1* enhancer (Figure 2-13E). However, the mutation did not change the transcription of *Lax1* and two downstream gene *Zc3h11a* and *Zbed6* (Figure 2-13F). Both *Zc3h11a* and *Zbed6* remained active after the secondary reprogramming, while the level of *Lax1* is still undetectable. Instead, the upstream gene *Atp2b4* showed a three-fold induction in the mutants (F12 and H11). Considered Atp2b4 was initially inactive; a three-fold induction was less likely to activate the transcription in the mutants fully. Moreover, the *Atp2b4* expression still went down when inducing reprogramming from the NPCs, suggesting the repression of *Atp2b4* was not entirely dependent on the Oct4/Sox2 site. Since Oct4/Sox2 binds 60 kbps away from the *Atp2b4* TSS, the regulation the site might be less significant for the repression. Other transcription

factors might play a more dominant role, synergistically with the Oct4/Sox2 site, to inactivate *Atp2b4* in ESCs. However, the mechanism is still obscure.

DISCUSSION

By carefully interrogating the functions of Oct4/Sox2 peaks within 15 kbps of the Oct4/Sox2 targets, we discovered that Oct4/Sox2 functions differently between ESC-specific, Dynamic, Non-Dynamic, and Silent genes (Figure 2-S9). The unique Oct4/Sox2 functions at each gene class refined the functional significance of these transcription factors in pluripotency. Our finding supports the view that Oct4 and Sox2 are critical activators of many genes that exhibit the largest dynamic range of expression between ESC and somatic cells. Additionally, Oct4/Sox2 is also essential for the transcriptional activation of constitutively expressed genes (Dynamic genes) in ESC. However, in the differentiated cell types, other tissue-specific factors should support to the transcription through an alternative mechanism. Oct4/Sox2 peaks associated with the genes with minimal differential expression (Non-Dynamic genes) are not functional to the gene transcription. However, if the enhancer presents H3K27Ac deposition, the site can be a moderate regulator for a distal gene (e.g., *Mybl2*). Notably, the large number of Oct4/Sox2 binding sides at the Silent genes are likely to mediate the transcriptional repression in ESCs.

A previous phenotypic study demonstrated that Oct4-mediated pluripotency gene activation was necessary and sufficient for inducing the pluripotent state and blocking the differentiation in vitro and in vivo (Hammachi et al., 2012). When we moved beyond the

identification of common features of large clusters of Oct4/Sox2-regulated genes, we began to appreciate the unique molecular mechanisms used to regulate individual genes classified by the dynamic range of expression. Especially when there is a large number of the Oct4/Sox2 binding sites located at the Silent genes. Today, we conclude an essential repressive function of Oct4/Sox2 binding sites at the genes that are silent in ESCs but actively transcribed in the differentiated cells. The Silent gene class composes more than 70% of the strong Oct4/Sox2 composite sites (711 out of 1035). Such a large number of the Silent genes highlighted the importance of the repressive role played by Oct4/Sox2 in maintenance of pluripotency maintenance and induction of reprogramming. The repressive role of the Oct4/Sox2 bindings at Silent genes supports the view that several OSK-induced mechanisms mediate both somatic enhancer silencing and pluripotency enhancer selection (Chronis et al., 2017).

Besides the gene transcription, we also found that Oct4/Sox2 composite binding optimizes histone modification at the Oct4/Sox2-regulated enhancers. By mutating the Oct4/Sox2 motif sequences, we confirmed that the Oct4/Sox2 composite element is required for optimal H3K27Ac and H3K4me3 deposition, and contributes partially to the repressive histone H3K27me3/H3K9me3 marking at the Silent genes. In the ESC-specific gene *Pla2g1b*, deleting the Oct4/Sox2 composite element removed both H3K27Ac and H3K4me3 marking after the secondary reprogramming. Likewise, the Oct4/Sox2 composite elements at the Dynamic genes *Zfp57* and *Epb4.1l5* are both required for the optimal H3K27Ac/H3K4me3 marking. However, the same mechanism does not mediate the histone marks deposition at an alternative NEUR-specific enhancer. In the Non-

Dynamic genes *Hnrnpr* and *Pds5a*, neither the H3K27Ac nor H3K4me3 marking has changed in the absence of the Oct4/Sox2 binding. In the Silent genes *Oxgr1* and *Gnrhr*, Oct4/Sox2 binding did not alter the level of active histone marks. In contrast, they partially contributed to the deposition of repressive histone marks H3K27me3 and H3K9me3. The underlying mechanisms are still unclear. Many studies have highlighted the interplay of the pluripotency factors and histone modifier to alter the chromatin accessibility or histone modification for the competence of gene expression (Angie Rizzino1, 2013; Denholtz et al., 2013; Maherali et al., 2007). Oct4/Sox2 may recruit histone modifiers directly or through the interaction with other lineage-determining TFs to assemble the nucleosome remodeling complexes (Heinz et al., 2010; Koche et al., 2011). It is also possible that the loss of Oct4/Sox2 composite binding fails to recruit necessary histone modifiers, which ultimately affects the nucleosome structure and histone modifications.

By comparing the Nanog co-bind, distance to the annotated TSS, chromatin accessibility, and H3K27Ac/H3K4me3 marking in the well-defined gene sets, we were able to characterize a few features that preferentially enriched in the functional bindings contributing to gene activity. First, the proximity of the Oct4/Sox2 binding sites to the TSS is significantly shorter in ESC-specific genes. Second, H3K27Ac and H3K4me3 deposition are significantly enriched in the vicinities of ESC-specific and Dynamic genes. The correlation of binding proximity and active histone marks deposition at the active genes suggest that Oct4/Sox2-mediated gene activation is likely to reside in a proximal regulatory element marked with active histone modification. The previous study has found that the unique chromatin states signify the expression of different gene classes

characterized by RNA analysis and functional annotation (Rada-Iglesias et al., 2011). For genes annotated with pluripotent functions, enriched histone acetylation H3K27Ac regions show overlapping with the enhancers of the ESC-expressing genes. In contrast, repressive histone marks H3K9me3 and H3K27me3 occupy the enhancers to silence the differentiation genes in ESCs. Because of the quality of the H3K9me3 and H3K27me3 ChIP-seq data sets is not sufficient for a convincing quantitative analysis, we only measured the H3K27Ac/H3K4me3 deposition at the Oct4/Sox2 binding at the Silent genes. The low enrichment of H3K27Ac/H3K4me3 marking supports our discovery that Oct4/Sox2 acts as a repressor to inactivate these Silent genes. However, the frequency of Nanog co-bind and the chromatin accessibility are comparable between gene groups. The previous study found that Oct4 and Sox2, as pioneer factors, were able to recognize partial motif on a nucleosomal-occupied region (Soufi et al., 2014). In our study, we did not address the role of pioneer factor further. However, the distribution of chromatin accessibility at the Oct4/Sox2 sites across all gene groups reveals that targeting inaccessible chromatin is ubiquitous at the Oct4/Sox2 binding regardless of the gene transcription features.

Importantly, we also found that the majority of Oct4/Sox2 composite bindings are associated with the Silent genes. By mutating the motif sequences at the presentative genes, we confirmed that the Oct4/Sox2 sites mediated the transcriptional repression of *Oxgr1*, *Gnrhr, Ip6k1, and Atp2b4.* Additionally, the Oct4/Sox2 binding also contributes to a fraction of H3K27me3/H3K9me3 deposition when establishing the pluripotent state. However, the mechanism is obscure. Only a few studies have discussed the role of

transcriptional repression mediated by Oct4 or Sox2. Interestingly, Oct4 has been characterized to interact with both activator and repressor complexes (Ang et al., 2011; Bilodeau et al., 2009; Boyer et al., 2006; P. et al., 2009; Pardo et al., 2010; Pasini et al., 2010). Endogenous Nanog and Oct4 interacted with multiple repression proteins from NuRD, Sin3A, and Pml complexes (Liang et al., 2008). Moreover, this unique Hdac1/2 and Mta1/2 containing complex NODE (Nanog and Oct4 associated deacetylase) correlated with the expression of differentiation-promoting genes and ESC differentiation. In contrast, another study concluded that the activity of Oct4 as an activator is sufficient for the induction of iPSC formation. However, a latest genome-wide study discovered that OSK collaborated with stage-specific TFs to mediate somatic-enhancer inactivation to drive reprogramming (Chronis et al., 2017; Hammachi et al., 2012). Our findings agree with the previous view that Oct4/Sox2 and their partner Nanog activate the crucial gene components of the pluripotency network and, simultaneously, repress the differentiation-promoting genes (Boyer et al., 2005; Chronis et al., 2017; Kim et al., 2008b; Loh et al., 2006; Orkin, 2005). It will be interesting to investigate the underlying mechanisms that direct the role of activator or repressor of the Oct4/Sox2 composite binding.

Thus, by refining the ESC transcriptome analysis and identifying precise Oct4/Sox2 targets, we attempted to rigorously evaluate the function of Oct4/Sox2 for better understanding of how pluripotency is established and maintained. In summary, we revealed a selective Oct4/Sox2 transcriptional regulation mediates gene activation and inactivation when establishing pluripotency. Although this reductionist approach only covers a subset of Oct4/Sox2 binding sites, the method established with stringent criteria

is valuable to precisely examine the functional regulation and propose mechanistic insight of pluripotent regulation. Moreover, we proved the repressive role of Oct4/Sox2 for the genes that are entirely inactive in ESCs. The requirement of transcriptional repression when establishing the pluripotent state explains why a large fraction of Oct4/Sox2 are bound at an inactive gene. Together, this study deepens our understanding of transcriptional networks of pluripotency and provides mechanistic insight focusing on Oct4 and Sox2 regulation.

EXPERIMENTAL PROCEDURES

Cell Culture and Reagents

CCE ES cells and primary iPSCs were cultured and maintained in standard ES growth media. ES growth media were made by ES KnockOut™ DMEM (Gibco™, ThermoFisher Scientific) containing 15% ES certified fetal bovine serum (Omega Scientific) plus 1x L-glutamine, 1x penicillin/streptomycin, 1% non-essential amino acid, 100μM β-mercaptoethanol, and 1,000 units/ml ESGRO® Recombinant Mouse LIF Protein (Millipore Sigma). CCE ES cell line was cultured on gelatin (Stem Cell Technologies) coated tissue culture flasks under feeder-independent condition. Primary iPSCs were grown on gelatin-coated tissue culture flasks with a layer of mouse embryonic fibroblasts (feeders) mitotically inactivated with mytomycin C. The cells were feeder-depleted and grown overnight before experiments. These primary iPSCs were gifts from Dr. Kathrin

Plath's lab at UCLA and were induced reprogramming by adding 2μg/ml doxycycline to the tetO-OSKM MEFs.

Tet-on iPSCs Neural Differentiation and Secondary Reprogramming Model

The primary tetO-OSKM MEFs derived from day 13.5 mouse embryos harbor a heterozygous R26-M2rtTA allele and a single dox-inducible polycistronic cassette encoding Oct4, Sox2, Klf4, and cMyc in the *Col1A* locus (Chronis et al., 2017). For the induction of reprogramming for primary iPSCs, tetO-OSKM MEFs were grown in ES growth media with 2μg/ml doxycycline to induce the expression of OSKM and split onto pre-seeded mitomycin C-inactivated mouse feeders. These primary iPSCs were cultured and split for at least 10 passages to enrich iPSCs and analyze pluripotency gene expression (*Oct4*, *Sox2*, *Nanog*, *Ssea1*, *Klf4*, *Rex1*, and *Nr0b1*). To differentiate into neural progenitor cells (NPCs), the primary iPSCs were counted (1.0 x 10^5/ml) and subsequently plated in Corning® Ultra-low attachment culture dish for embryoid body formation. Embryoid bodies were formed in EB formation media (DMEM, 7.5% fetal bovine serum, 1x L-glutamine, 1x penicillin/streptomycin, 1% non-essential amino acid, and 100μM β-mercaptoethanol) for 3 to 5 days and subsequently plated onto adherent tissue culture dishes with 0.5 μM retinoic acid (Millipore Sigma) to induce neural differentiation (Sagner et al., 2018). Forty-eight hours later, neural-committed embryoid bodies were collected and grown in NPC culture media (DMEM plus 10% fetal bovine serum). NPCs were cultured using trypsin for at least four passages and validated based on the expression of *Sox1*, *Nes*, *Pax6*, and *Pax3* before the start of secondary reprogramming. For the derivation of secondary iPSCs, NPCs were plated at densities of

1.0 x 10^4 per 35 mm on gelatin-coated dishes with monolayer mouse feeders. Forty-eight hours later, NPC culture media were replaced by ESC culture media supplemented with 2μg/ml doxycycline and cultured as previously described. The identities of cell lineage were validated by the kinetic changes of pluripotency genes (*Oct4*, *Sox2*, *Nanog*, *Ssea1*, *Klf4*, *Rex1*, and *Nr0b1*) and neural progenitor genes (*Sox1*, *Nes*, *Pax6*, and *Pax3*) expression.

CRISPR/Cas9 Mutagenesis

Single guide RNA and homology direct repair (HDR) template targeting Oct4/Sox2 composite binding sites were designed using MIT CRISPR Designer (http://crispr.mit.edu/) and Benchling CRISPR Guide Design (https://www.benchling.com/crispr/). HDR template was designed to substitute Oct4/Sox2 composite motif with two enzymatic sequence EcoRI- GAATTC, BamHI- GGATCC to interfere Oct4/Sox2 binding. Target sequence was cloned into pSpCas9(BB)-2A-Puro (PX459) V2.0 (Addgene, #62988) to express both Cas9 and sgRNA (Cong and Zhang, 2014). Cas9/gRNA plasmid and HDR template were co-transfected into CCE ESCs or iPSCs. Forty-eight hours after the transfection, CCE ESCs or iPSCs were cultured in media containing 1.25 μg/ml puromycin and expanded in ES media after two days of puromycin selection. Puromycin-resistant cells were collected, diluted, and plated in 96-well plates to obtain single cell colonies. Single cell-derived colonies were expanded for genotyping. To determine the genomic sequence after CRISPR mutagenesis, cells were lysed, and proteins were degraded using 1 mg/ml proteinase K at 55℃ overnight. Isopropanol was added to precipitate genomic DNA. Precipitate DNA pellets were washed by 80% ethanol and resuspended in 10mM Tris pH

7.9. Genomic regions flanking Oct4/Sox2 composite motif were PCR-amplified and sequenced to confirm the HDR mutation. Two mutants with homozygous Oct4/Sox2 binding sites depletion and EcoRI- GAATTC, BamHI- GGATCC substitution were selected for secondary reprogramming, qRT-PCR, and ChIP experiments.

Chromatin Immunoprecipitation and ChIP-qPCR

Forty million cells (ESCs, iPSCs, EBs, NPCs) were crosslinked with 1% formaldehyde (ThermoFisher Scientific) and lysed to collect nuclei pellets. Nuclei pellets were sonicated with Misonix 3000 sonicator for major fragments between 200bp to 500bp. Fragmentized chromatin lysates were incubated with ChIP grade antibodies Oct4 (R&D, AF1759), Sox2 (R&D, AF2018), H3K4me3 (Millipore Sigma, 05-745R), H3K27Ac (Active Motif, 39133), H3K27me3 (Active Motif, 39155), or H3K9me3 (Abcam, ab8898) overnight. The immunoprecipitated complex was pulled down by Protein G Dynabeads (Invitrogen, 10004D) and reverse-crosslinked with Proteinase K (ThermoFisher Scientific, EO0491) at 60℃ overnight. Immunoprecipitated DNA was purified and quantified by phenol-chloroform (Sigma, P3803) and Qubit (Thermo Fisher, Q32854), respectively. ChIP-qPCR was utilized to measure the enrichment of transcription factors or histone modification at genomic loci of interest. Primer pairs were designed using ± 200 bp genomic sequence specific to the target loci to generate 100 bp to 125 bp amplified products. Quantification of fold-enrichment was calculated based on the fold change of the percentage of input between target genomic loci and negative control region (*Hbb-b2* or *Actb*). Primers used for ChIP-qPCR were summarized in Table 1.

RNA-seq

Chromatin-associated RNA was fractioned and isolated as previously described (Bhatt et al., 2012). Ribosomal RNA (rRNA) was depleted using RiboMinus™ Transcriptome Isolation Kit (ThermoFisher Scientific). After rRNA depletion, 200 ng RNA was subjected to prepare strand-specific cDNA libraries using Illumina TruSeq RNA Sample Prep Kit v2 (Illumina) with dUTP second-strand method (Levin et al., 2010). All cDNA libraries were single-end sequenced (50 bp) on Illumina HiSeq 2000.

Reads were mapped to mouse NCBI37/mm9 reference genome by HISAT2 v2.1.0 and only those uniquely mapping reads with no more than two mismatches were retained (Kim et al., 2015). RPKM values were calculated as previously described (Mortazavi et al., 2008) and based on the gene annotation of NCBI37/mm9 reference genome. Chromatin RNA RPKM was calculated by counting all mapped reads within each transcription unit and dividing by the length of each locus. mRNA RPKM was calculated by counting all mapped exonic reads and dividing by the length of the spliced product. SeqMonk's RNA-seq quantification pipeline was used to calculate RPKM value (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). All RPKM represents an average from two to three biological replicates in each tissue/cell type. Published mRNA sequencing datasets were obtained from Mouse ENCODE Project (http://www.mouseencode.org/) and summarized in Table 2.

ChIP-seq Read Mapping and Processing

Published transcription factor ChIP-seq datasets were obtained from GEO (Buecker et al., 2014; Chronis et al., 2017) and summarized in Table 3. Reads from ChIP-seq were

mapped to mouse NCBI37/mm9 reference genome using Bowtie2 software (Langmead and Salzberg, 2012). Uniquely aligned reads were used for peak calling and gene annotation using HOMER (Benner et al., 2017). Peaks with false discovery rate (FDR) < 0.01 and enriched over input were called. Only reproducible peaks from replicates were retained for downstream analyses. Called peaks were annotated to nearest TSS of genes. Composite bindings of multiple factors were determined by the distance between two peaks. Only the distance of peak summits less than 100 bp were considered a composite binding.

Motif Analyses

HOMER called ChIP-seq peak regions were used for motif analyses. To define a consensus sequence that may be utilized by transcription factors, ± 100 bp, ± 50 bp, or ± 25 bp genomic sequence information from the center position of peak loci were analyzed by MEME-ChIP (Bailey et al., 2009; Machanick and Bailey, 2011). Oct4 and Sox2 composite bindings were determined by the genomic loci with the presence of Oct4 and Sox2 peaks and the peak summits are within 100 bp. For *de novo* motif analyses at Oct4/Sox2 composite binding regions, the center position was made by the midpoint of two peak summits. A minimum window of 10 bp and a maximum window of 30 bp are set as parameter for identifying composite motif.

Histone Marks ChIP-seq and ATAC-seq Datasets

Published histone mark ChIP-seq and ATAC-seq datasets were obtained from GEO (Becker et al., 2016; Chronis et al., 2017; Ji et al., 2015; Di Stefano et al., 2016) and

summarized in Table 3. Reads from ChIP-seq and ATAC-seq were mapped to mouse NCBI37/mm9 reference genome using Bowtie2 software (Langmead and Salzberg, 2012). Reads were removed if they were duplicated, mapped to mitochondrial genome, or aligned to unmapped contiguous sequences. Enrichment of histone marks H3K27Ac and H3K4me3 were analyzed by calculating the RPKM values of 1.5 kbps flanking region centered on the Oct4/Sox2 peaks. Chromatin accessibility (ATAC sensitivity) was analyzed by calculating the RPKM values of 1.0 kbps flanking region centered on the Oct4/Sox2 peaks.

RNA Extraction and qRT-PCR

Cells grown in 6-well plate with a confluency of 80 – 90% were lysed in TRI reagent (Molecular Research Center, TR118). RNA was extracted by Qiagen RNeasy kit following the manufacturer's instructions and reverse-transcribed into cDNA by SuperScript™ III Reverse Transcriptase (ThermoFisher Scientific). Levels of cDNA were quantified by real-time PCR using PowerUp™ SYBR™ Green Master Mix (ThermoFisher Scientific) on a BioRad CFX384 Real-Time PCR system. Gene expression levels were calculated relative to a standard curve and normalized to housekeeping gene *Gapdh* in duplicates. Primers used for qRT-PCR were summarized in Table 1.

FIGURE LEGENDS

Figure 2-1. General Features of Oct4 and Sox2 ChIP-Seq in Embryonic Stem Cells

ChIP-seq datasets of Oct4 and Sox2 in ESC line V6.5 (GEO: GSE90895) were processed and analyzed to study Oct4/Sox2 composite binding regions.

(A) The pie chart displays the genomic distribution of 15,506 reproducible Oct4 peaks. Peaks were called by HOMER and retained with false discovery rate < 0.01. Promoter region was defined as -500 bp ~ +150 bp relative to the TSS. Exonic, intronic, intergenic, non-coding, and TTS were annotated based on genomic location using HOMER (Benner et al., 2017). (B) The pie chart displays the genomic distribution of 11,207 reproducible Sox2 peaks. Peaks were called by HOMER and retained with false discovery rate < 0.01. Promoter region was defined as -500 bp ~ +150 bp relative to the TSS. Exonic, intronic, intergenic, non-coding, and TTS were annotated based on genomic location using HOMER (Benner et al., 2017). (C) Composite binding sites were analyzed according to the distance of Oct4 and Sox2 peaks. The distance of Oct4 and Sox2 was calculated based on the center of called peaks. Distance less then 100 bp was considered to be an Oct4/Sox2 composite binding. The Venn diagrams indicate the number of Oct4/Sox2 composite, Oct4-only, and Sox2 only peaks with various stringencies. Different thresholds were added at genome-wide (left), Oct4/Sox2 peak score > 20 (middle), and within 15 kbps of annotated TSS (right). (D) Oct4/Sox2 composite motif was identified by MEME de novo motif analysis based on the top 10% called peaks from Oct4 and Sox2 ChIP-seq. Right column: observed motif frequency and statistical significance of the identified motif.

Figure 2-2. Compare Nascent Transcript Profiles Between ESC and Three Somatic Cells NEUR, BMDM, and DP

Chromatin-associated transcripts from CCE ESC lines, E14.5 cortical neurons, bone marrow derived macrophages, and CD4+ CD8+ thymocytes were analyzed by RNA-seq. (A) The distribution of minimum fold change values of CCE ESC lines over NEUR, BMDM or DP is shown for 3,030 genes with expression level higher than five RPKM. Three colored dashed lines indicate represent 2- (orange), 20- (red), and 100-fold (dark red) thresholds. (B) The 3,030 expressed genes in ESCs were grouped based on the dynamic range of expression and minimum fold changes between ESC and three somatic cell types. Genes exhibited at least 20-fold specific in ESC are categorized into ESC-specific genes (n = 91). Broadly-expressed genes (0.2 – 5-fold) showed no dynamic range of expression among all cell types are grouped into Non-Dynamic genes (n = 1,931). Genes with dynamic range of expression (20-fold) between ESCs and only one or two cell types are considered to be Dynamic genes (n = 248). Values were RPKM and clustered by cell types and descending fold change values. Values were color-coded based on expression percentile. (C) The bar graphs demonstrated representative examples of each gene group. (Left) *Pla2g1b* (10.19 RPKM) is an ESC-specific gene with 200-fold differential expression in ESC when compared to all three somatic cell types. (Middle) *Hnrnpr* (17.19 RPKM) and *Pds5a* (38.86 RPKM) are both Non-Dynamic genes with 0.95-fold and 1.4-fold changes, respectively. (Right) *Zfp57* (82.95 RPKM) is a Dynamic gene with more than 20-fold differential expression when compared to BMDM and DP, but not NEUR (24.01 RPKM, 3.4-fold).

Figure 2-3. Characterize Oct4/Sox2 Peaks by Peak Strength, Nanog Co-binding, and Distance to the Transcription Starting Site of Annotate Targets

(A) The dot chart and line graph plot the percentage of genes with Oct4 binding (y-axis, left) and fold enrichment of Oct4 occupancy (y-axis, right) between 91 ESC-specific and 1,931 Non-Dynamic genes with different peak score stringencies. Black dot (ESC-specific) and grey cross (Non-Dynamic) indicate the percentage of genes with Oct4 peaks within 15 kbps of TSS at particular peak score threshold. Line with blue triangles highlights the fold enrichment of Oct4 occupancy according to the ratio between ESC-specific and Non-Dynamic. (B) The dot chart and line graph plot the percentage of genes with Sox2 binding (y-axis, left) and fold enrichment of Sox2 occupancy (y-axis, right) between 91 ESC-specific and 1,931 Non-Dynamic genes with different peak score stringencies. Black dot (ESC-specific) and grey cross (Non-Dynamic) indicate the percentage of genes with Sox2 peaks within 15 kbps of TSS at particular peak score threshold. Line with red triangles highlights the fold enrichment of Sox2 occupancy according to the ratio between ESC-specific and Non-Dynamic. (C) The bar graphs show the percentage of Oct4 and Sox2 peaks with composite binding in the vicinity of ESC-specific genes (left) and Non-Dynamic genes (right). (D) MEME de novo motif analysis compares strong Oct4/Sox2 composite motif sequences between ESC-specific genes (35 Oct4/Sox2 peaks), Dynamic genes (53 Oct4/Sox2 peaks), Non-Dynamic genes (101 Oct4/Sox2 peaks) and Silent genes (711 Oct4/Sox2 peaks). Right column: observed motif frequency and statistical significance of the identified motif. (E) The bar graph displays the percentage of Nanog co-binding at Oct4/Sox2 composite binding sites in ESC-specific, Dynamic, Non-Dynamic, and Silent genes. Nanog co-occupancy was examined with different peak score cutoff called from

Nanog ChIP-seq (GEO: GSE90895). (F) The distribution of distance between Oct4/Sox2 peaks to the TSS of annotated genes is shown for Oct4/Sox2 targets in different gene groups.

Figure 2-4. Characterize Oct4/Sox2 Peaks by Histone Modification and Chromatin Accessibility

(A) The RPKM distribution of H3K27Ac ChIP-seq (left), H3K4me3 ChIP-seq (middle), and ATAC-seq (right) at Oct4/Sox2 composite binding sites is shown for ESC-specific, Dynamic, Non-Dynamic, and Silent genes. Enrichments of histone mark H3K27Ac and H3K4me3 were analyzed by calculating the RPKM values in a 1.5 kbps window centered on the Oct4/Sox2 peaks. ATAC signals (chromatin accessibility) were analyzed by quantifying the RPKM values of 1.0 kbps window centered on the Oct4/Sox2 peaks.
(B) Box plots display the distribution (minimum, 1st quartile, median, 3rd quartile, and maximum) of H3K27Ac (left), H3K4me3 (middle), and ATAC (right) among four gene groups. (C) All 1,035 Oct4/Sox2 composite binding sites were ranked on the descending order of H3K27Ac level and binned into ten groups. Bin1 exhibits the highest H3K27Ac level and bin 10 is the lowest. The bar graph and table compare the percentage of Oct4/Sox2 peaks distributed in each H3K27Ac enrichment bin between gene groups. (D) All 1,035 Oct4/Sox2 composite binding sites were ranked on the descending order of H3K4me3 level and binned into ten groups. Bin 1 exhibits the highest H3K4me3 level and bin 10 is the lowest. The bar graph and table compare the percentage of Oct4/Sox2 peaks distributed in each H3K4me3 enrichment bin between gene groups.

Figure 2-5. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of ESC-specific Gene *Pla2g1b* by CRISPR in CCE ESCs

(A) The bar graph shows *Pla2g1b* expression profiles of ESC and twelve mouse tissues/cells. The RPKM values were calculated by analyzing mouse ENCODE RNA-seq datasets. (B) Genome browser snapshot displays the Oct4/Sox2 binding at the enhancer of *Pla2g1b*. The composite binding site shows active enhancer characteristics of ATAC sensitive, high H3K27Ac, and high H3K4me3. The wild-type genomic sequences reveal Oct4/Sox2 motif (TTGTAATGCAAA) discovered by MEME. The mutant sequences indicate the disruption of Oct4/Sox2 motif mediated by CRISPR/Cas9 HDR mutation. (C) Bar graph validates Oct4 binding at the *Pla2g1b* enhancer in control, two independent clones lacking the Oct4/Sox2 composite motif (C9 and H7). Fold enrichment of Oct4 ChIP-qPCR was analyzed by calculating the fold change of percentage of input between Oct4 binding site and negative control region (*Hbb-b2*). (D) Bar graph validates Sox2 binding at the *Pla2g1b* enhancer in control, two independent clones lacking the Oct4/Sox2 composite motif (C9 and H7). Fold enrichment of Sox2 ChIP-qPCR was analyzed by calculating the ratio of fold change of input between Oct4 binding site and negative control region (*Hbb-b2*). (E) Bar graph shows the normalized expression of the *Pla2g1b* mRNA in control and mutants. Expression level of *Pla2g1b* in BMDM are included to indicate silent level of mRNA.

Figure 2-6. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of ESC-specific Gene *Pla2g1b* by CRISPR in Secondary Reprogramming Model

(A) Bar graphs show Oct4 binding (left) and Sox2 binding (right) at the *Pla2g1b* enhancer in control and two independent mutants (A1-1, A1-2) for primary iPSCs and dox-induced day 14 secondary iPSCs. Fold enrichment of Oct4 and Sox2 ChIP-qPCR was analyzed by calculating the fold change of the percentage of input between Oct4/Sox2 binding site and negative control region (*Hbb-b2*). (B) The line chart shows normalized expression of *Pla2g1b* mRNA in control and two independent clones lacking Oct4/Sox2 composite motif. Kinetic changes of *Pla2g1b* mRNA expression were measured by qRT-PCR in tet-on primary iPSCs, embryoid bodies, neural progenitors, dox-induced day 7, and dox-induced day 14 (secondary iPSCs). Relative expression level was normalized to *Pla2g1b* mRNA expression in control tet-on iPSCs. (C) Genome browser snapshot displays the Oct4/Sox2 peaks, histone marks enrichment, and ATAC peak at the enhancer of *Pla2g1b*. Blue shades highlighted two enhancer regions with enriched H3K27Ac and H3K4me3 in ESCs. Bar graphs show H3K27Ac ChIP-qPCR (top, blue) and H3K4me3 ChIP-qPCR (bottom, yellow) of peak 1 (left) and peak 2 (right) regions at the enhancer of *Pla2g1b* in control and two independent mutants. Kinetic changes of H3K27Ac and H3K4me3 were examined by ChIP-qPCR in specific pluripotent and differentiated states throughout the secondary reprogramming model.

Figure 2-7. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of Dynamic Gene *Zfp57* by CRISPR

(A) The bar graph shows *Zfp57* expression profiles of ESC and twelve mouse tissues/cells. The RPKM values were calculated by analyzing mouse ENCODE RNA-seq datasets. (B) Genome browser snapshot displays the Oct4/Sox2 binding at the enhancer

of *Zfp57*. Bedgraph visualization of H3K27Ac ChIP-seq, H3K4me3 ChIP-seq, and ATAC-seq is shown for ESC and neural progenitor cells. The blue shades highlighted two genomic loci (E1 and E2) with ATAC signal and H3K27Ac centered on Oct4/Sox2-targeted enhancer. Green shades highlighted two additional genomic loci (E3 and E4) with enriched H3K27Ac and H3K4me3 in neural progenitor cells. (C) Bar graphs show Oct4 binding (left) and Sox2 binding (right) at the *Zfp57* enhancer in control and two independent mutants (B4-1, B4-2) lacking Oct4/Sox2 motif for primary iPSCs and dox-induced day 14 secondary iPSCs. Fold enrichment of Oct4 and Sox2 ChIP-qPCR was analyzed by calculating the fold change of the percentage of input between Oct4/Sox2 binding site and negative control region (*Hbb-b2*). (D) The line chart shows normalized expression of *Zfp57* mRNA in control and mutants. Kinetic changes of *Zfp57* mRNA expression were measured by qRT-PCR in tet-on primary iPSCs, embryoid bodies, neural progenitors, dox-induced day 7, and dox-induced day 14 (secondary iPSCs). Relative expression level was normalized to *Zfp57* mRNA expression in control tet-on primary iPSCs. (E) Bar graphs display H3K27Ac and H3K4me3 enrichment at E1 and E2 region in control and two mutants. Kinetic changes of H3K27Ac and H3K4me3 were examined by ChIP-qPCR in specific pluripotent and differentiated states throughout the secondary reprogramming model. (F) Bar graphs display H3K27Ac and H3K4me3 enrichment at E3 and E4 region in control and two mutants. Kinetic changes of H3K27Ac and H3K4me3 were examined by ChIP-qPCR in specific pluripotent and differentiated states throughout the secondary reprogramming model.

Figure 2-8. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of Broadly Expressed Dynamic Gene *Epb4.1l5* by CRISPR

(A) The bar graph shows *Epb4.1l5* expression profiles of ESC and twelve mouse tissues/cells. The RPKM values were calculated by analyzing mouse ENCODE RNA-seq datasets. (B) Genome browser snapshot displays the Oct4/Sox2 binding at the enhancer of *Epb4.1l5*. Bedgraph panels of H3K27Ac ChIP-seq, H3K4me3 ChIP-seq, and ATAC-seq is shown for ESC and neural progenitor cells. The blue shades highlighted two genomic loci (E1 and E2) with ATAC signal and H3K27Ac centered on the *Epb4*.1l5 enhancer with Oct4/Sox2 peaks. Green shades highlighted two additional genomic loci (E3 and E4) with enriched H3K27Ac and H3K4me3 in neural progenitor cells. (C) Bar graphs show Oct4 binding (left) and Sox2 binding (right) at the *Epb4.1l5* enhancer in control and two independent clones (F3, G11) lacking Oct4/Sox2 motif for primary iPSCs and dox-induced day 14 secondary iPSCs. (D) The line chart shows normalized expression of *Zfp57* mRNA in control and mutants. Kinetic changes of *Epb4.1l5* mRNA expression were measured by qRT-PCR in tet-on primary iPSCs, embryoid bodies, neural progenitors, dox-induced day 7, and dox-induced day 14 (secondary iPSCs). Relative expression level was normalized to the expression in control tet-on primary iPSCs. (E) Bar graphs show kinetic changes of H3K27Ac and H3K4me3 at E1 and E2 in control and two mutants by ChIP-qPCR. (F) Bar graphs show kinetic changes of H3K27Ac and H3K4me3 at E3 and E4 in control and two mutants by ChIP-qPCR.

Figure 2-9. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancers of Broadly Expressed Non-Dynamic Genes *Pds5a and Hnrnpr* by CRISPR

(A) Genome browser snapshot displays the bedgraph visualization of Oct4/Sox2 ChIP-seq, H3K27Ac ChIP-seq, H3K4me3, and ATAC-seq at the enhancer of *Pds5a*. Blue shades highlighted Pds5a_E1 genomic location centered on Oct4/Sox2 peaks and a distal Pds5a_E2 region showing H3K27Ac, H3K4me3, and ATAC signal in ESCs. (B) Bar graphs display Oct4 (left) and Sox2 (right) binding in control and two independent mutants (E7 and H11) lacking Oct4/Sox2 motif. Bindings of Oct4 and Sox2 were shown in both primary iPSCs and dox-induced day 14 secondary iPSCs. (C) Kinetic changes of *Pds5a* mRNA expression across the secondary reprogramming model were shown in control and two mutants. (D) Genome browser snapshot displays the bedgraph visualization of Oct4/Sox2 ChIP-seq, H3K27Ac ChIP-seq, H3K4me3, and ATAC-seq at the enhancer of *Hnrnpr*. Blue shades highlighted Hnrnpr_E1 genomic location centered on Oct4/Sox2 peaks and a distal Hnrnpr_E2 region showing H3K27Ac, H3K4me3, and ATAC signal in ESCs. (E) Bar graphs display Oct4 (left) and Sox2 (right) binding in control and two independent mutants (B5 and E8) lacking Oct4/Sox2 motif. Bindings of Oct4 and Sox2 were examined in both primary iPSCs and dox-induced day 14 secondary iPSCs. (F) Kinetic changes of *Hnrnpr* mRNA expression across the secondary reprogramming model were shown in control and two mutants. (G) Bar graphs show kinetic changes of H3K27Ac and H3K4me3 at Pds5a_E1 and Pds5a_E2 in control and two mutants by ChIP-qPCR. (H) Bar graphs show kinetic changes of H3K27Ac and H3K4me3 at Hnrnpr_E1 and Hnrnpr_E2 in control and two mutants by ChIP-qPCR.

Figure 2-10. Evaluate the Role of Oct4/Sox2 Composite Binding at the Active Enhancers of Non-Dynamic Genes *Dido1* and *Ift52* by CRISPR

(A) Genome browser snapshot highlights three nearby genes: *Tcfl5* (blue), *Gid8* (red), and *Slc17a9* (green), within ±100 kbps of *Dido1*. Bedgraph panels show Oct4 ChIP-seq, Sox2 ChIP-seq, ATAC-seq, H3K27Ac ChIP-seq, and H3K4me3 ChIP-seq centered on Oct4/Sox2 targeting the *Dido1* enhancer (dashed box). (B) Bar graphs compare Oct4 binding (left) and Sox2 binding (right) at the upstream enhancer of *Dido1* in control and two independent clones (A12 and F9). (C) Line charts compare the relative expression level of *Tcfl5*, *Dido1*, *Gid8*, and *Slc17a9* in control and two mutants (A12 and F9) lacking Oct4/Sox2 motif. Relative expression level was normalized to the expression in control tet-on primary iPSCs. (D) Genome browser snapshot highlights three nearby genes: *Sgk2* (blue), *Mybl2* (red), and *Gtsf1l* (green), within ±50 kbps of *Ift52*. Bedgraph panels show Oct4 ChIP-seq, Sox2 ChIP-seq, ATAC-seq, H3K27Ac ChIP-seq, and H3K4me3 ChIP-seq centered on Oct4/Sox2 targeting the intronic region of *Ift52* (dashed box). (E) Bar graphs compare Oct4 binding (left) and Sox2 binding (right) at the intronic enhancer of *Ift52* in control and two independent clones (G5 and G11). (F) Line charts compare the relative expression level of *Sgk2*, *Ift52*, *Mybl2*, and *Gtsf1l* in control and two mutants (G5 and G11) lacking Oct4/Sox2 motif. Relative expression level was normalized to the expression in control tet-on primary iPSCs.

Figure 2-11. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of Silent Gene *Oxgr1* by CRISPR

(A) Genome browser snapshot displays Oct4/Sox2 composite binding associated with silent gene *Oxgr1*. Bedgraph tracks show ATAC-seq, H3K27Ac ChIP-seq, H3K4me3 ChIP-seq, H3K9me3, and H3K27me3 centered on Oct4/Sox2 peaks. Yellow shade

highlights genomic location centered on Oct4/Sox2 peaks (E1). Red shades highlight downstream E2 region and upstream E3 region. (B) Bar graphs show Oct4 binding (left) and Sox2 binding (right) in control and mutant clones (C9-1 and C9-2) for primary iPSCs and dox-induced day 14 secondary iPSCs. (C) The line chart compares the fold change of the *Oxgr1* mRNA expression in mutant clones with control tet-on primary iPSCs. (D) The bar graphs show the kinetic changes of H3K27ac and H3K4me3 in control and mutant clones across the secondary reprogramming model at E1 region. (E) The bar graphs show the kinetic changes of H3K9me3 and H3K27me3 in control and mutant clones across the secondary reprogramming model at E2 and E3 regions.

Figure 2-12. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of Silent Gene *Gnrhr* by CRISPR

(A) Genome browser snapshot displays Oct4/Sox2 composite binding associated with silent gene *Gnrhr*. Bedgraph panels display ATAC-seq, H3K27Ac ChIP-seq, H3K4me3 ChIP-seq, H3K9me3, and H3K27me3 centered on Oct4/Sox2 peaks. Yellow shade highlights genomic location centered on Oct4/Sox2 peaks (E1). Red shades highlight downstream E2 region and upstream E3 region with enriched H3K9me3 or H3K27me3. (B) Bar graphs show Oct4 binding (left) and Sox2 binding (right) in control and mutant clones (G3-1 and G3-2) for primary iPSCs and dox-induced day 14 secondary iPSCs. (C) The line chart compares the fold change of the *Gnrhr* mRNA expression in mutant clones with control tet-on primary iPSCs. (D) The bar graphs show the kinetic changes of H3K27ac and H3K4me3 in control and mutant clones throughout the secondary reprogramming model at E1 region. (E) The bar graphs show the kinetic changes of

H3K9me3 and H3K27me3 in control and mutant clones throughout the secondary reprogramming model at E2 and E3 regions.

Figure 2-13. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Active Enhancers of Silent Genes *Uba7* and *Lax1* by CRISPR

(A) Genome browser snapshot highlights three nearby genes: *Camkv* (blue), *Traip* (red), and *Ip6k1* (green), within ±50 kbps of *Uba7*. Bedgraph panels show Oct4 ChIP-seq, Sox2 ChIP-seq, ATAC-seq, H3K27Ac ChIP-seq, and H3K4me3 ChIP-seq centered on Oct4/Sox2 targeting the *Uba7* enhancer (dashed box). (B) Bar graphs compare Oct4 binding (left) and Sox2 binding (right) at the upstream enhancer of *Uba7* in control and two independent clones (A1 and G9). (C) Line charts compare the relative expression level of *Camkv, Traip*, *Uba7*, and *Ip6k1* in control and two mutants (A1 and G9) lacking Oct4/Sox2 motif. Relative expression level was normalized to the expression in control tet-on primary iPSCs. (D) Genome browser snapshot highlights three nearby genes: *Zc3h11a* (blue), *Zbed6* (red), and *Atp2b4* (green), within ±60 kbps of *Lax1*. Bedgraph panels show Oct4 ChIP-seq, Sox2 ChIP-seq, ATAC-seq, H3K27Ac ChIP-seq, and H3K4me3 ChIP-seq centered on Oct4/Sox2 targeting the upstream enhancer of *Lax1* (dashed box). (E) Bar graphs compare Oct4 binding (left) and Sox2 binding (right) at the upstream enhancer of *Lax1* in control and two independent clones (F12 and H7). (F) Line charts compare the relative expression level of *Zc3h11a*, *Zbed6*, *Lax1*, and *Atp2b4* in control and two mutants (F12 and H7) lacking Oct4/Sox2 motif. Relative expression level was normalized to the expression in control tet-on primary iPSCs.

Figure 2-S1. ESC Gene Groups and Oct4/Sox2 Binding

Description of ESC gene groups features and Oct4/Sox2 occupancy. Grouping of ESC expressed genes was done by quantifying the fold change of nascent transcript expression between ESC and three somatic cell types. Features of dynamic range of expression and fold change cutoff are shown in table for each gene group. Silent genes are defined by < 5 RPKM of the nascent transcripts in ESC. The last column shows the percentage of genes with strong Oct4/Sox2 binding (peak score > 20) within 15 kbps of annotated TSS. Columns are color-coded from the maximum percentage (red) to the minimum percentage (green).

Figure 2-S2. A DOX-inducible System for Mouse Secondary Reprogramming of TetO-OSKM iPSCs

(A) The schematic diagram describes the experimental designs of mouse secondary reprogramming of tetO-OSKM iPSCs and the implementation of CRISPR/Cas9 HDR mutation. HDR template and Cas9/sgRNA plasmid were co-transfected into primary iPSCs for the selection of single cell colonies and genotyping for the mutant clones lacking Oct4/Sox2 composite motif. Culture condition, growth supplement, and serum concentration are described for the specific stages of differentiation and re-reprogramming. (B) Representative cell culture morphologies of primary iPSC, embryoid body, neural progenitor cell, and secondary iPSC. (C) The line chart shows the kinetic changes of the mRNA expression of pluripotency specific genes: *Pou5f1, Sox2, Nanog, Ssea1, Klf4, Rex1, and Nr0b1*. (D) The line chart shows the kinetic changes of the mRNA expression of neural lineage specific genes: *Nes*, *Sox1*, *Pax6*, and *Pax3*. mRNA

expression was analyzed by comparing with endogenous *Gapdh* expression and shown as percentage of *Gapdh*.

Figure 2-S3. Single Colony Expansion of CRISPR-mutated Primary iPSCs

(A) The bar graph shows the fold change of primary ΔPla2g1b_A1 mutant clone and secondary single colony expansion clones 1 – 10 with the expression of *Pla2g1b* mRNA in control cells. (B) The bar graph shows the fold change of primary ΔZfp57_B4 mutant clone and secondary single colony expansion clones 1 – 10 with the expression of *Zfp57* mRNA in control cells. (C) The bar graph shows the fold change of primary ΔEpb4.1l5_F3 mutant clone and secondary single colony expansion clones 1 – 10 with the expression of *Epb4.1l5* mRNA in control cells. The data shown represent an average of three biological replicates. Error bars indicate the standard error.

Figure 2-S4. Extended Subculture of CRISPR Mutated Primary iPSCs

(A) The line chart shows the mRNA expression of *Pla2g1b* and kinetic changes from passage 0 (primary iPSC) to passage 15 in control and ΔPla2g1b_A1 mutant clone. (B) The line chart shows the mRNA expression of *Zfp57* and kinetic changes from passage 0 (primary iPSC) to passage 15 in control and ΔZfp57_B4 mutant clone. (C) The line chart shows the mRNA expression of *Epb4.1l5* and kinetic changes from passage 0 to passage 15 in control and ΔEpb4.1l5_F3 mutant clone. Relative expression level was normalized to the expression in control primary iPSCs.

Figure 2-S5. Properties of Oct4/Sox2-bound Non-Dynamic Genes

(A) The bar graphs indicate RNA-seq RPKM values measuring the nascent transcript levels (left) and polyA mRNA levels (right) of *Pds5a* in mouse ESC lines and selected tissues or cell types. (B) The bar graphs indicate RNA-seq RPKM values measuring the nascent transcript levels (left) and polyA mRNA levels (right) of *Hnrnpr* in mouse ESC lines and selected tissues or cell types. (C) The bar graphs indicate RNA-seq RPKM values measuring the nascent transcript levels (left) and polyA mRNA levels (right) of *Dido1* in mouse ESC lines and selected tissues or cell types. (D) The bar graphs indicate RNA-seq RPKM values measuring the nascent transcript levels (left) and polyA mRNA levels (right) of *Ift52* in mouse ESC lines and selected tissues or cell types.

Figure 2-S6. Properties of ESC Genes In the Neighborhood of *Dido1* and *Ift52*

The bar graphs indicate RNA-seq RPKM values measuring the nascent transcript levels of *Slc17a9* (A), *Gid8* (B), *Tcfl5* (C), *Sgk2* (D), *Mybl2* (E), and *Gtsf1l* (F) in mouse ESC lines and three somatic cell types: NEUR, BMDM, and DP.

Figure 2-S7. Properties of Oct4/Sox2-bound Silent Genes

(A) The bar graphs indicate RNA-seq RPKM values measuring the nascent transcript levels (left) and polyA mRNA levels (right) of *Oxgr1* in mouse ESC lines and selected tissues or cell types. (B) The bar graphs indicate RNA-seq RPKM values measuring the nascent transcript levels (left) and polyA mRNA levels (right) of *Gnrhr* in mouse ESC lines and selected tissues or cell types. (C) The bar graphs indicate RNA-seq RPKM values measuring the nascent transcript levels (left) and polyA mRNA levels (right) of *Uba7* in mouse ESC lines and selected tissues or cell types. (D) The bar graphs indicate RNA-

seq RPKM values measuring the nascent transcript levels (left) and polyA mRNA levels (right) of *Lax1* in mouse ESC lines and selected tissues or cell types.

Figure 2-S8. Properties of ESC Genes In the Neighborhood of *Uba7* and *Lax1*

The bar graphs indicate RNA-seq RPKM values measuring the nascent transcript levels of *Camkv* (A), *Traip* (B), *Ip6k1* (C), *Zc3h11a* (D), *Zbed6* (E), and *Atp2b4* (F) in mouse ESC lines and three somatic cell types: NEUR, BMDM, and DP.

Figure 2-S9. Summary of Gene-Specific Functions of Oct4/Sox2 Composite Sites

The table summarizes the Oct4/Sox2 composite sites-regulated gene groups, chromatin status, and the functional role validated by CRISPR/Cas9 experiments.

Table 2-1. Primer Sequences for qRT-PCR

The table lists the primers used for qRT-PCR

Table 2-2. PolyA mRNA-seq Datasets from Mouse ENCODE

The table lists the mRNA-seq datasets downloaded and analyzed from Mouse ENCODE.

Table 2-3. Transcription Factors ChIP-seq, Histone Marks ChIP-seq, and ATAC-seq Datasets

The table lists the transcription factor ChIP-seq, histone marks ChIp-seq, and ATAC-seq datasets utilized for bioinformatic analysis.

Figure 2-1. General Features of Oct4 and Sox2 ChIP-Seq in Embryonic Stem Cells

Figure 2-2. Compare Nascent Transcript Profiles Between ESC and Three Somatic Cells: NEUR, BMDM, DP

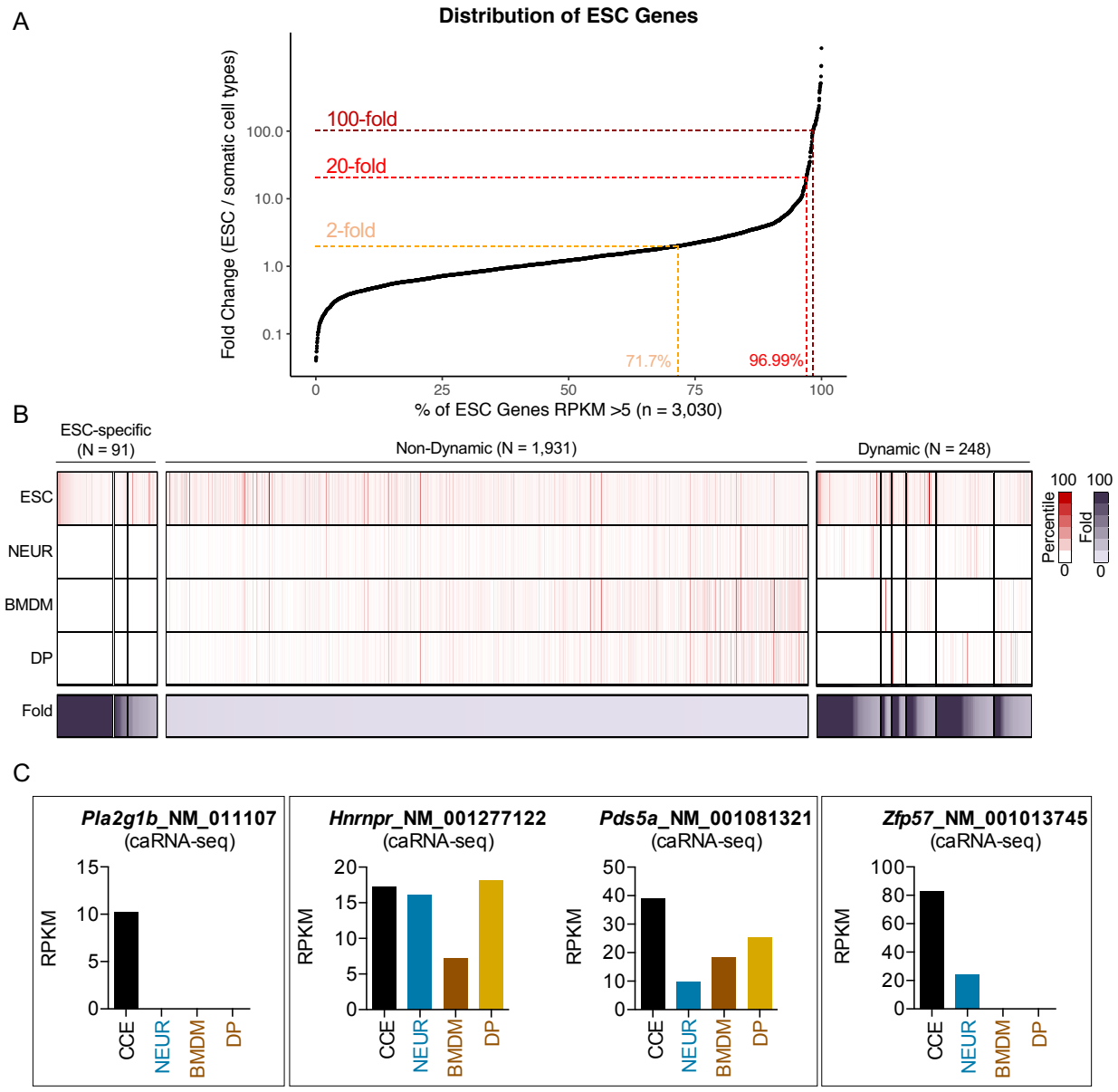


Figure 2-3 Characterize Oct4/Sox2 Peaks by Peak Strength, Nanog Co-binding, and Distance to the Transcription Starting Site of Annotated Targets

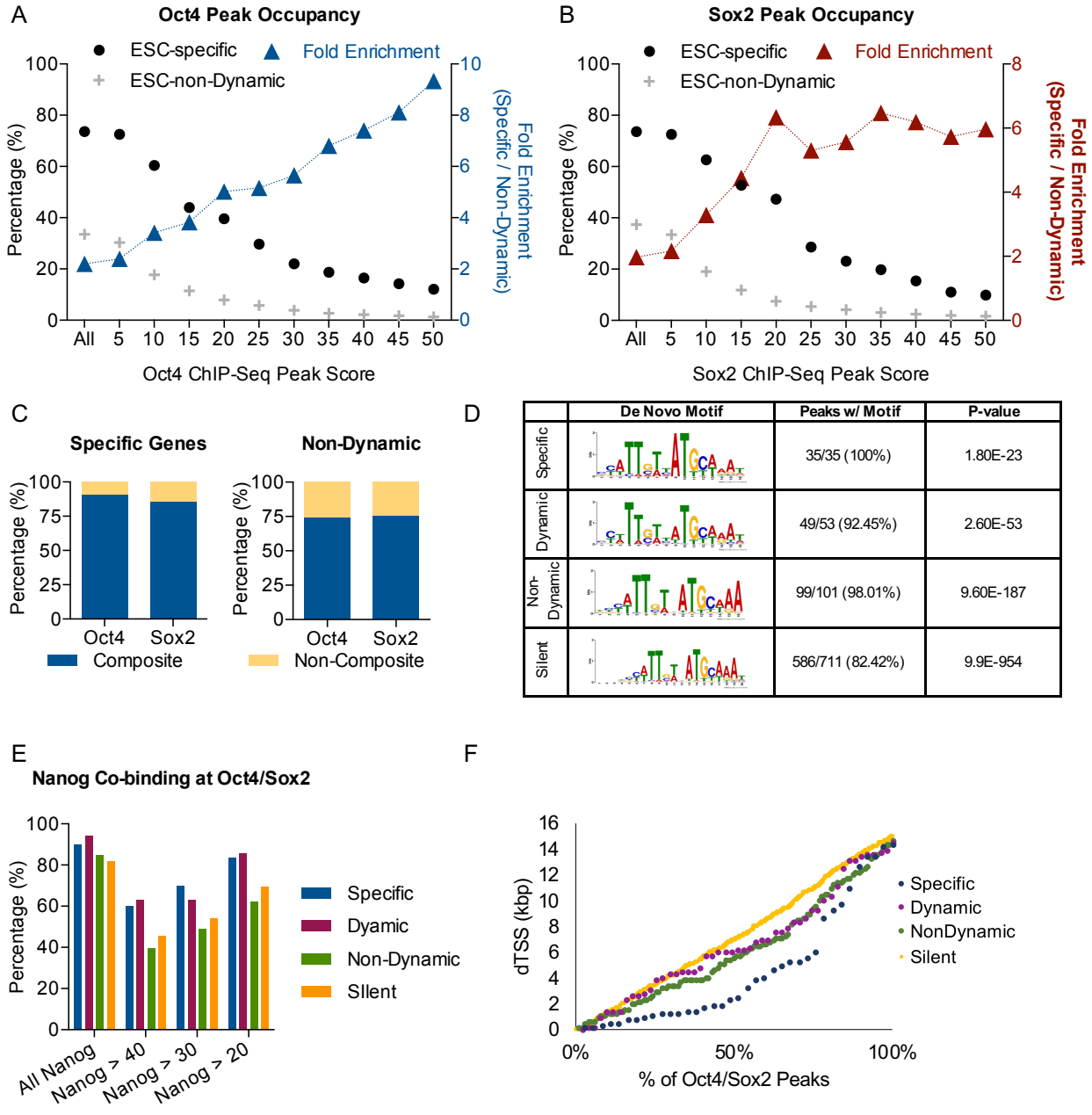
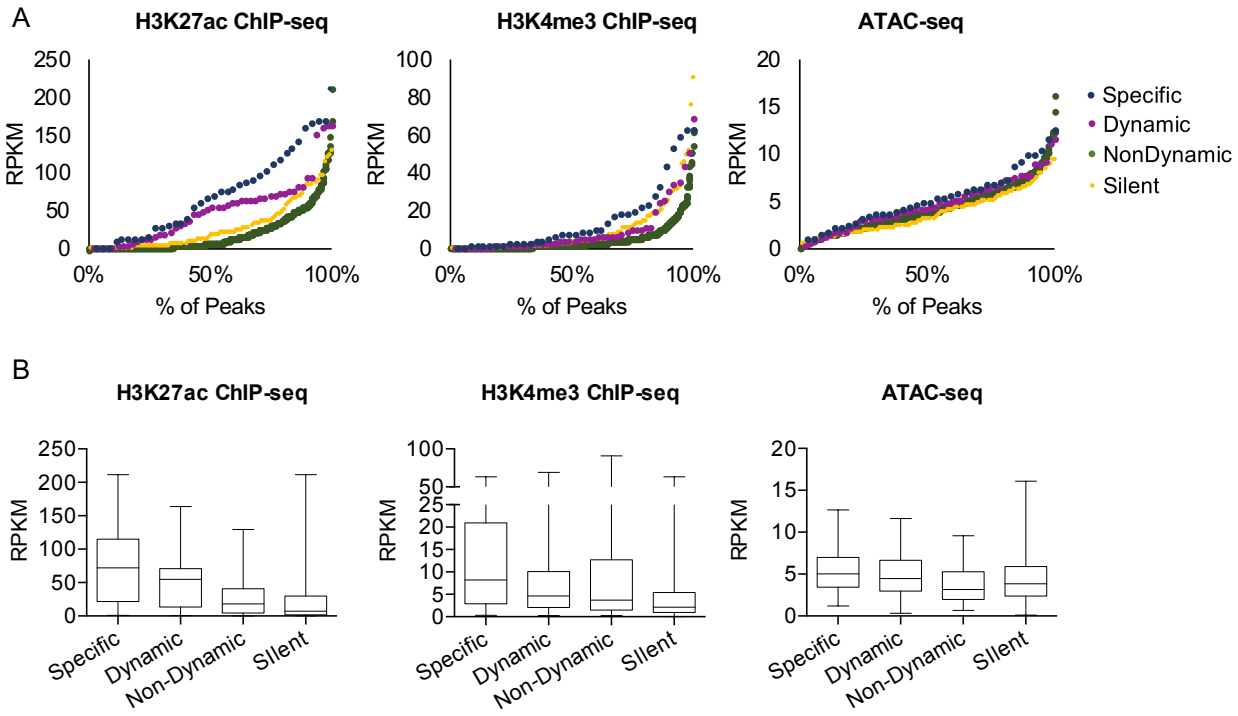
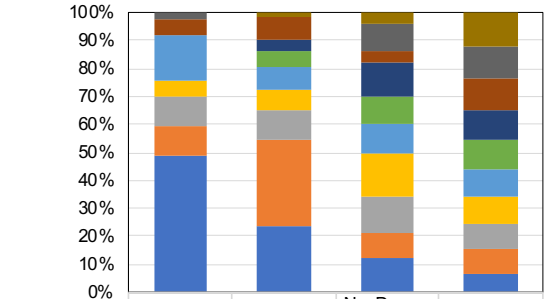


Figure 2-4 Characterize Oct4/Sox2 Peaks by Histone Modification and Chromatin Accessibility



C

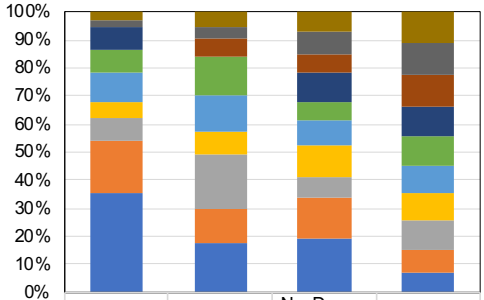
H3K27ac ChIP-seq



H3K27ac	Specific	Dynamic	NonDynamic	Silent
Bin 10	0.00%	1.96%	3.88%	12.24%
Bin 9	2.70%	0.00%	9.71%	11.11%
Bin 8	5.41%	7.84%	3.88%	11.25%
Bin 7	0.00%	3.92%	12.62%	10.55%
Bin 6	0.00%	5.88%	9.71%	10.83%
Bin 5	16.22%	7.84%	10.68%	9.70%
Bin 4	5.41%	7.84%	15.53%	9.56%
Bin 3	10.81%	9.80%	12.62%	9.56%
Bin 2	10.81%	31.37%	8.74%	8.58%
Bin 1	48.65%	23.53%	12.62%	6.61%

D

H3K4me3 ChIP-seq



H3K4me3	Specific	Dynamic	NonDynamic	Silent
Bin 10	2.70%	5.88%	6.80%	11.39%
Bin 9	2.70%	3.92%	8.74%	10.97%
Bin 8	0.00%	5.88%	5.83%	11.39%
Bin 7	8.11%	0.00%	10.68%	10.69%
Bin 6	8.11%	13.73%	6.80%	10.27%
Bin 5	10.81%	13.73%	8.74%	9.85%
Bin 4	5.41%	7.84%	11.65%	10.13%
Bin 3	8.11%	19.61%	6.80%	9.85%
Bin 2	18.92%	11.76%	14.56%	8.72%
Bin 1	35.14%	17.65%	19.42%	6.75%

Figure 2-5. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of ESC-specific Gene *Pla2g1b* by CRISPR in CCE ESCs

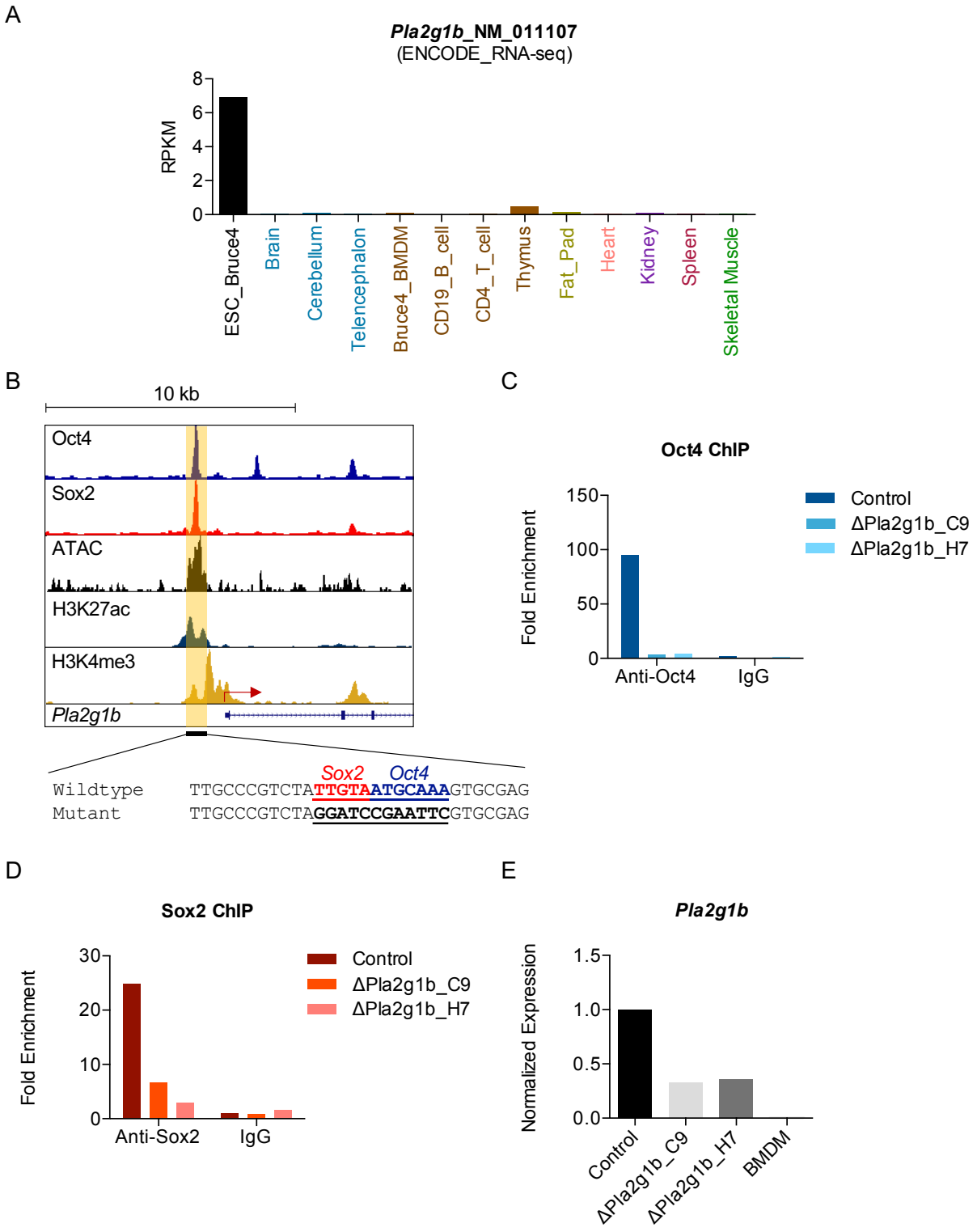


Figure 2-6. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of *Pla2g1b* by CRISPR in Secondary Reprogramming Model

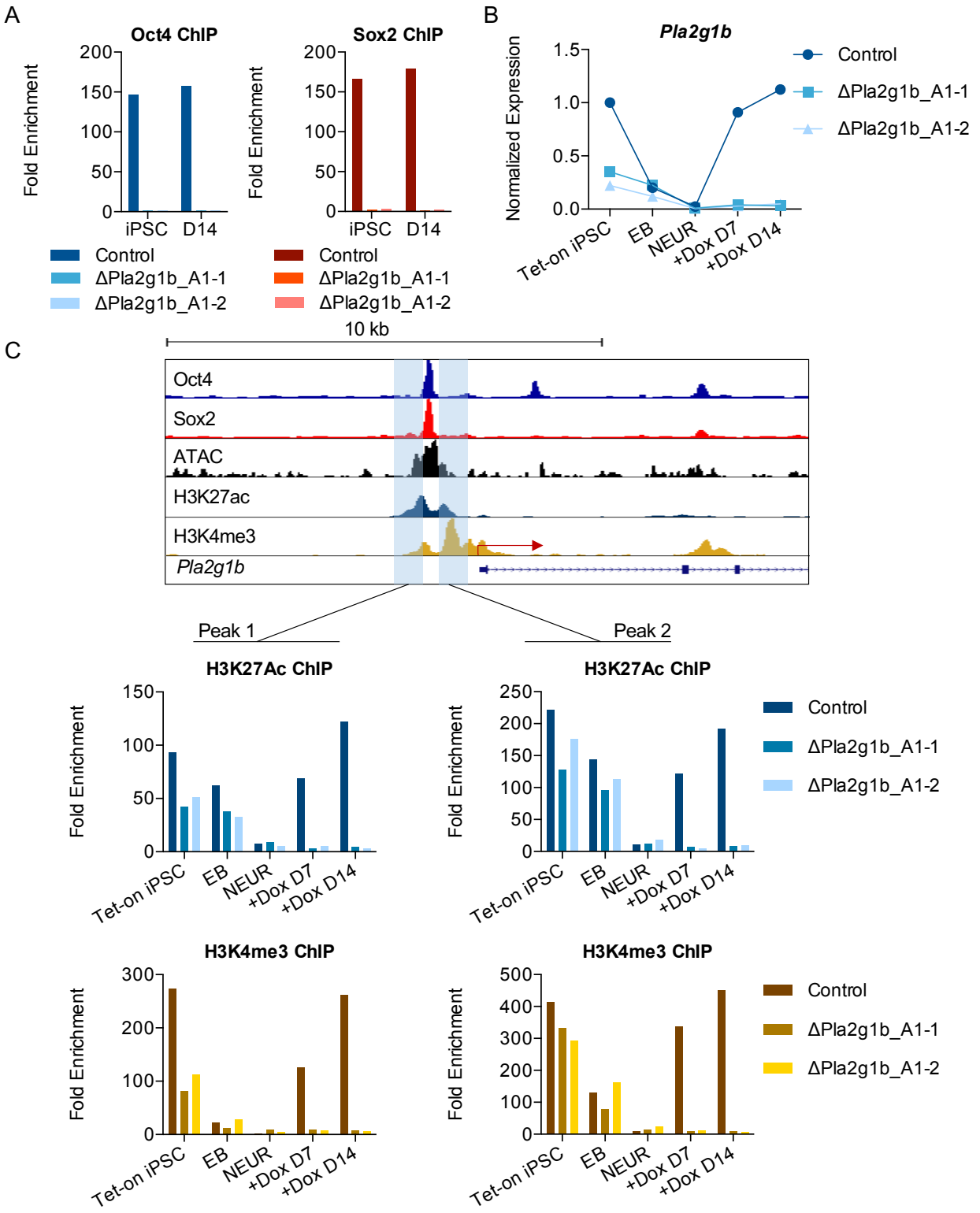


Figure 2-7. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of Dynamic Gene *Zfp57* by CRISPR

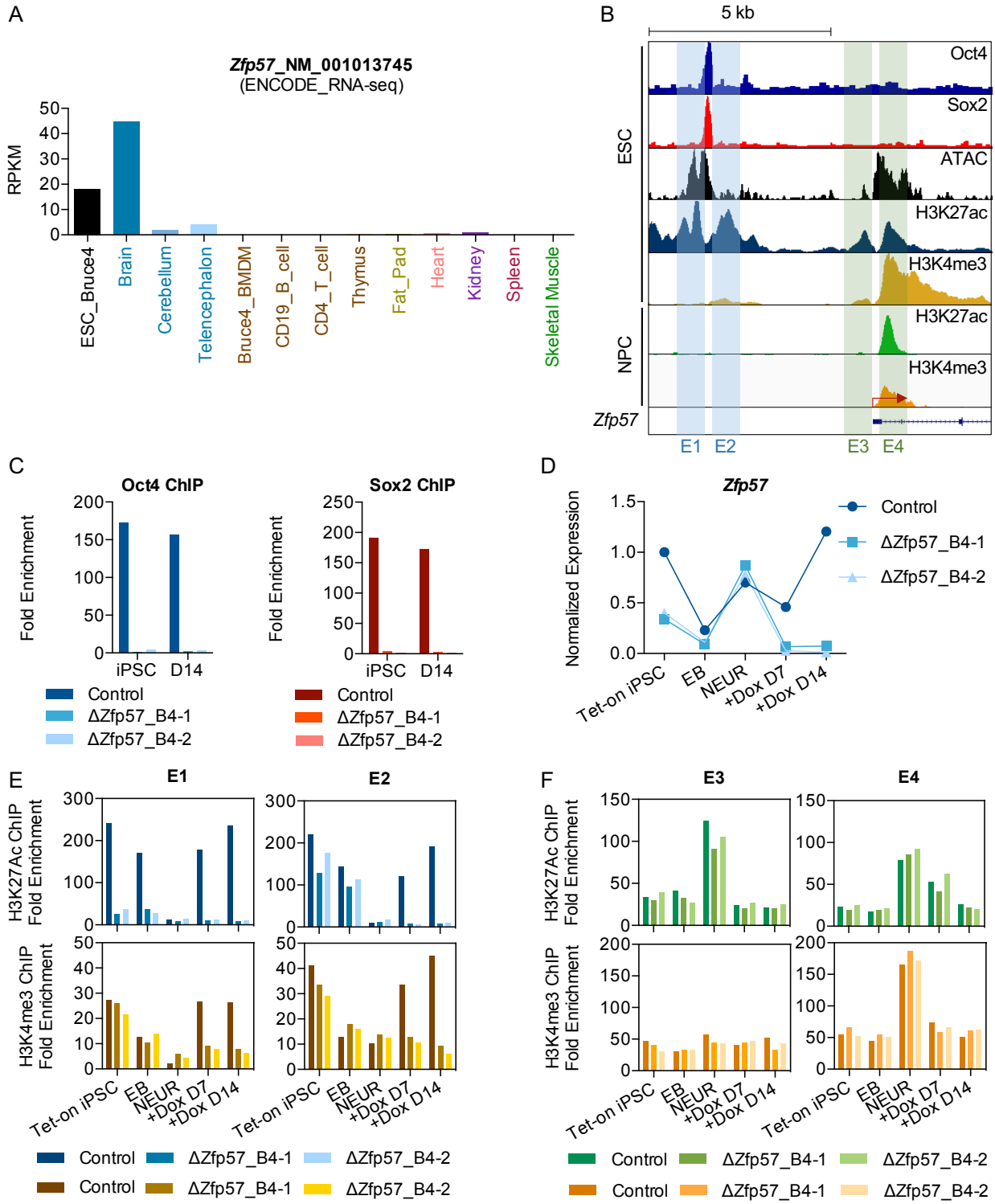


Figure 2-8. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of Broadly Expressed Dynamic Gene *Epb4.1l5* by CRISPR

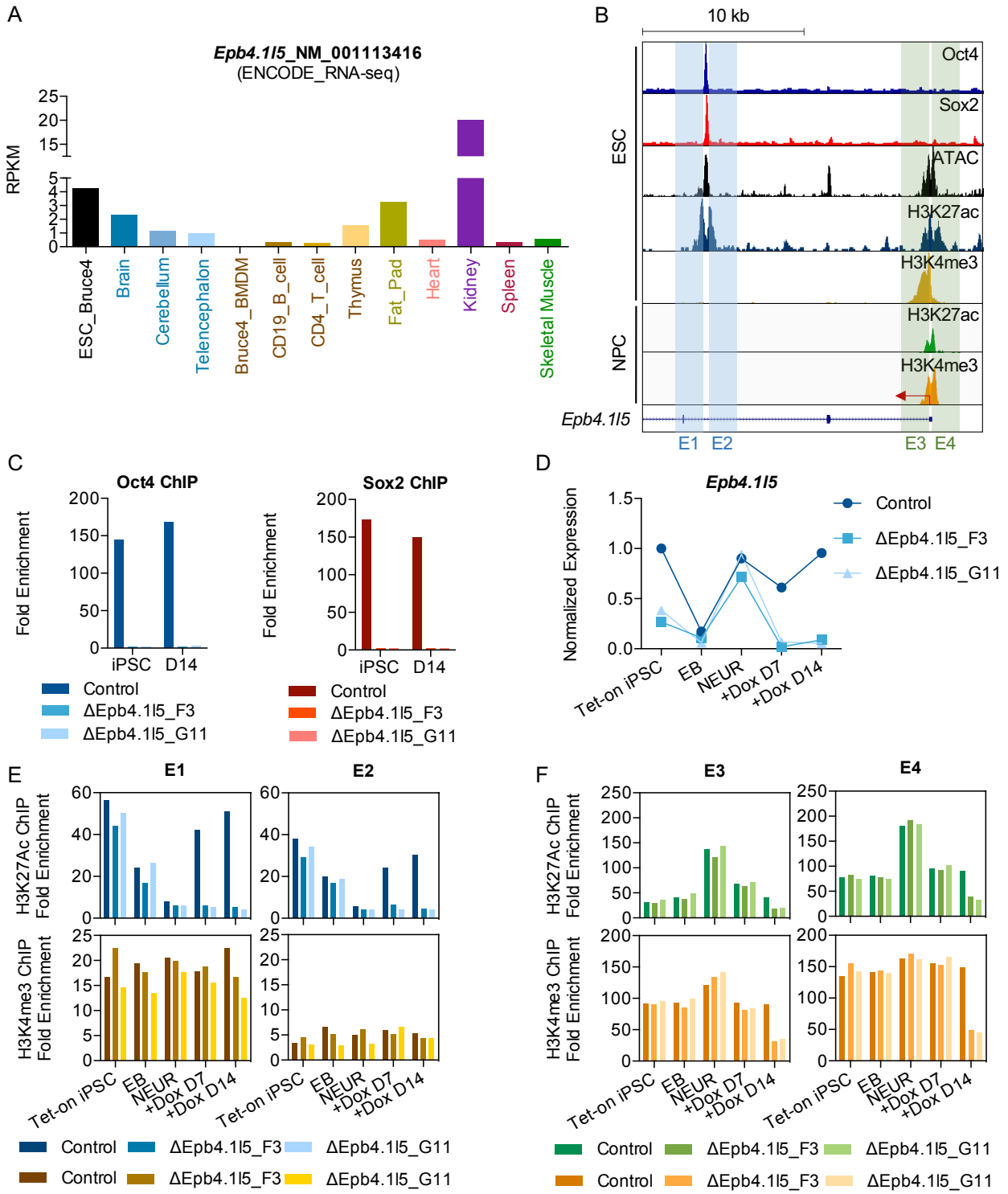


Figure 2-9. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancers of Broadly Expressed Non-Dynamic Genes *Pds5a and Hnrnpr* by CRISPR

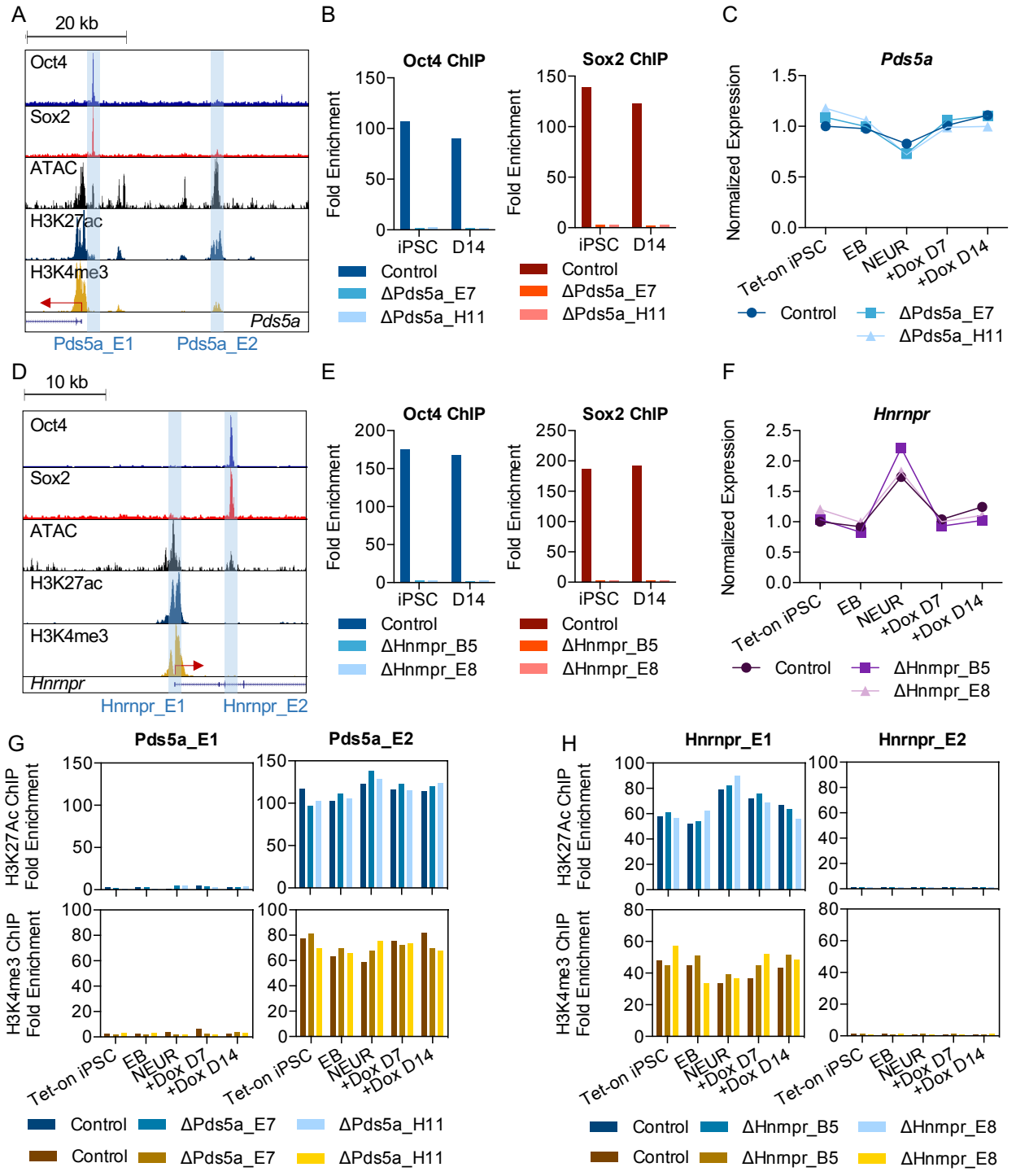


Figure 2-10. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Active Enhancers of Non-Dynamic Genes *Dido1* and *Ift52* by CRISPR

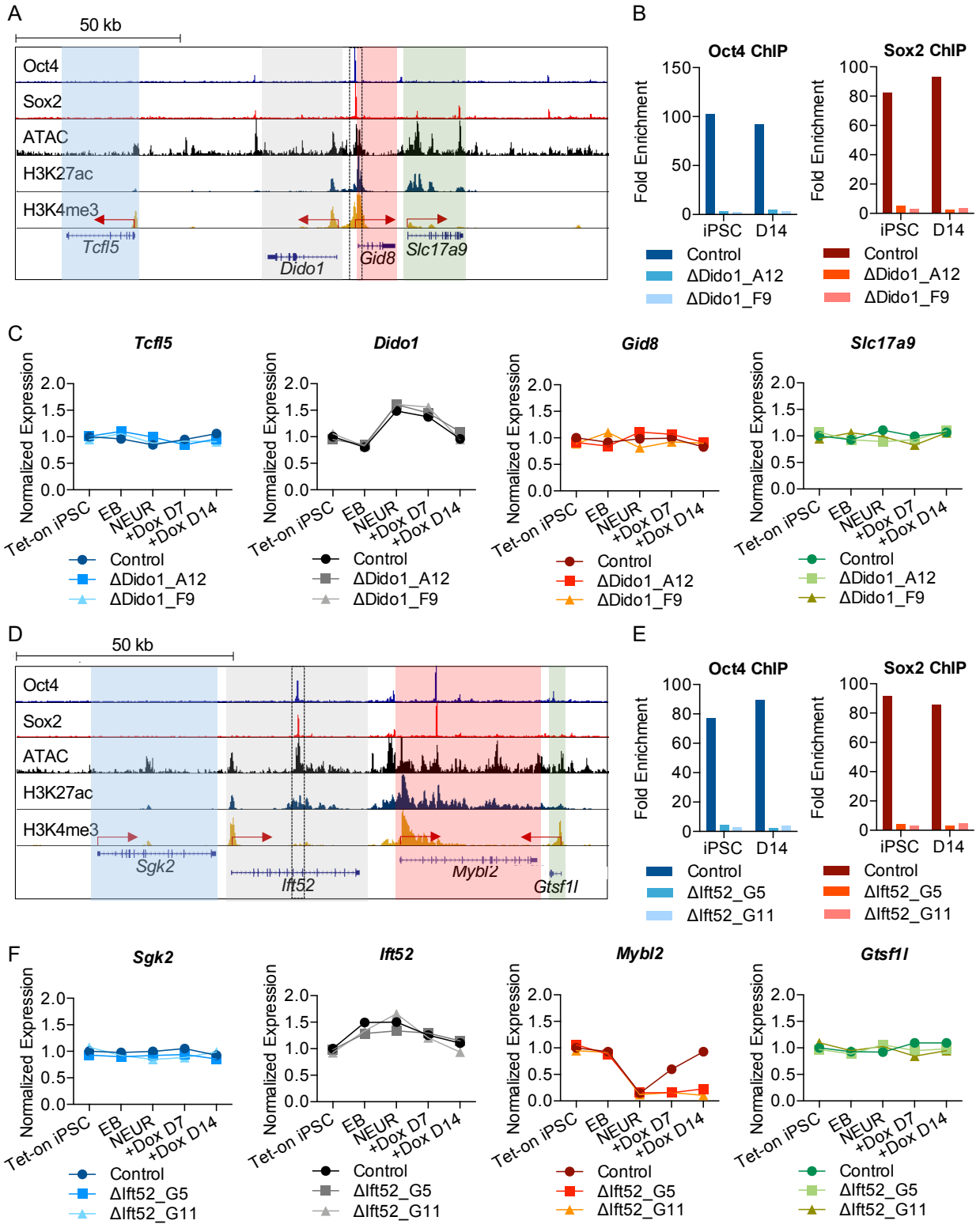


Figure 2-11. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of Silent Gene *Oxgr1* by CRISPR

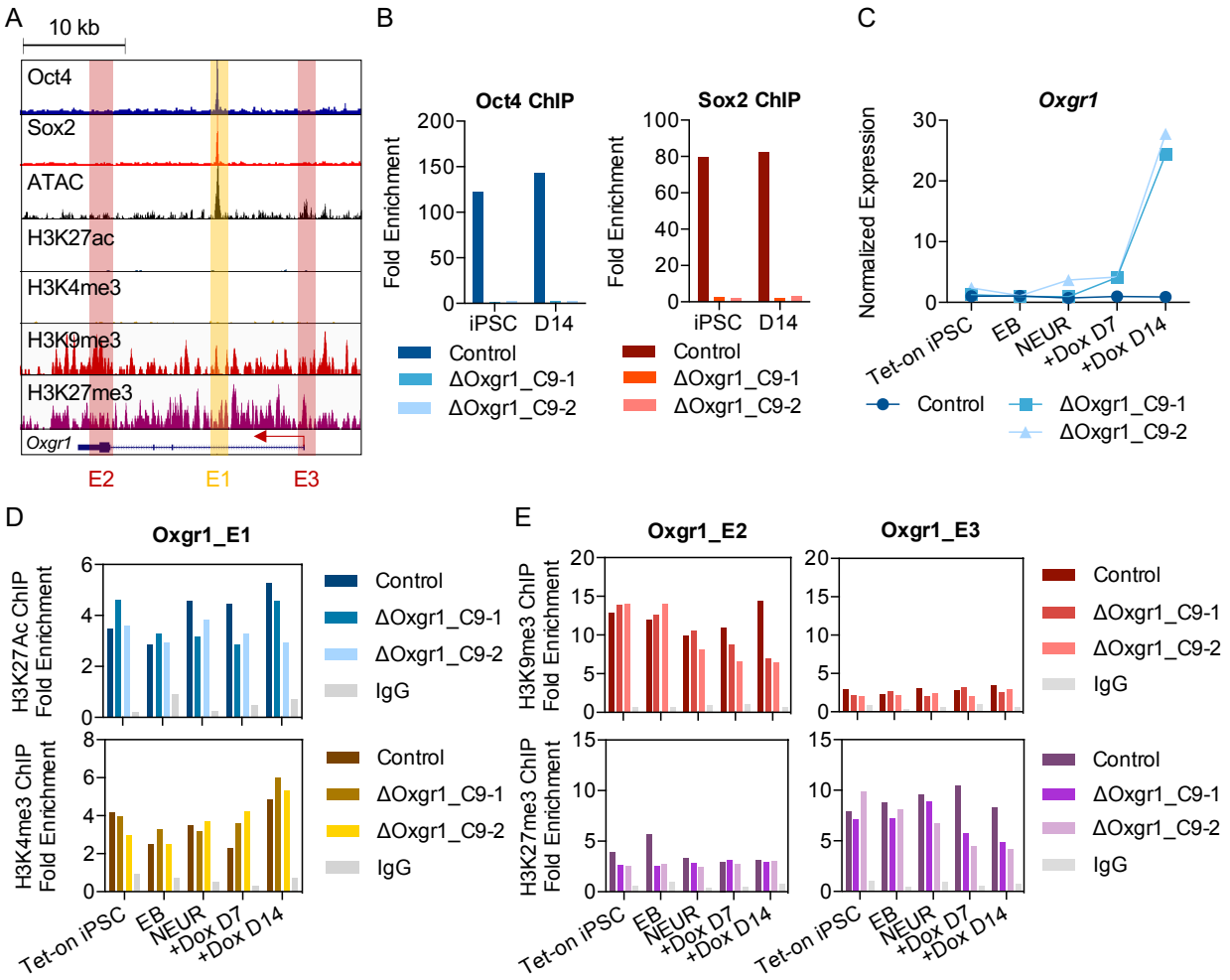


Figure 2-12. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Enhancer of Silent Gene *Gnrhr* by CRISPR

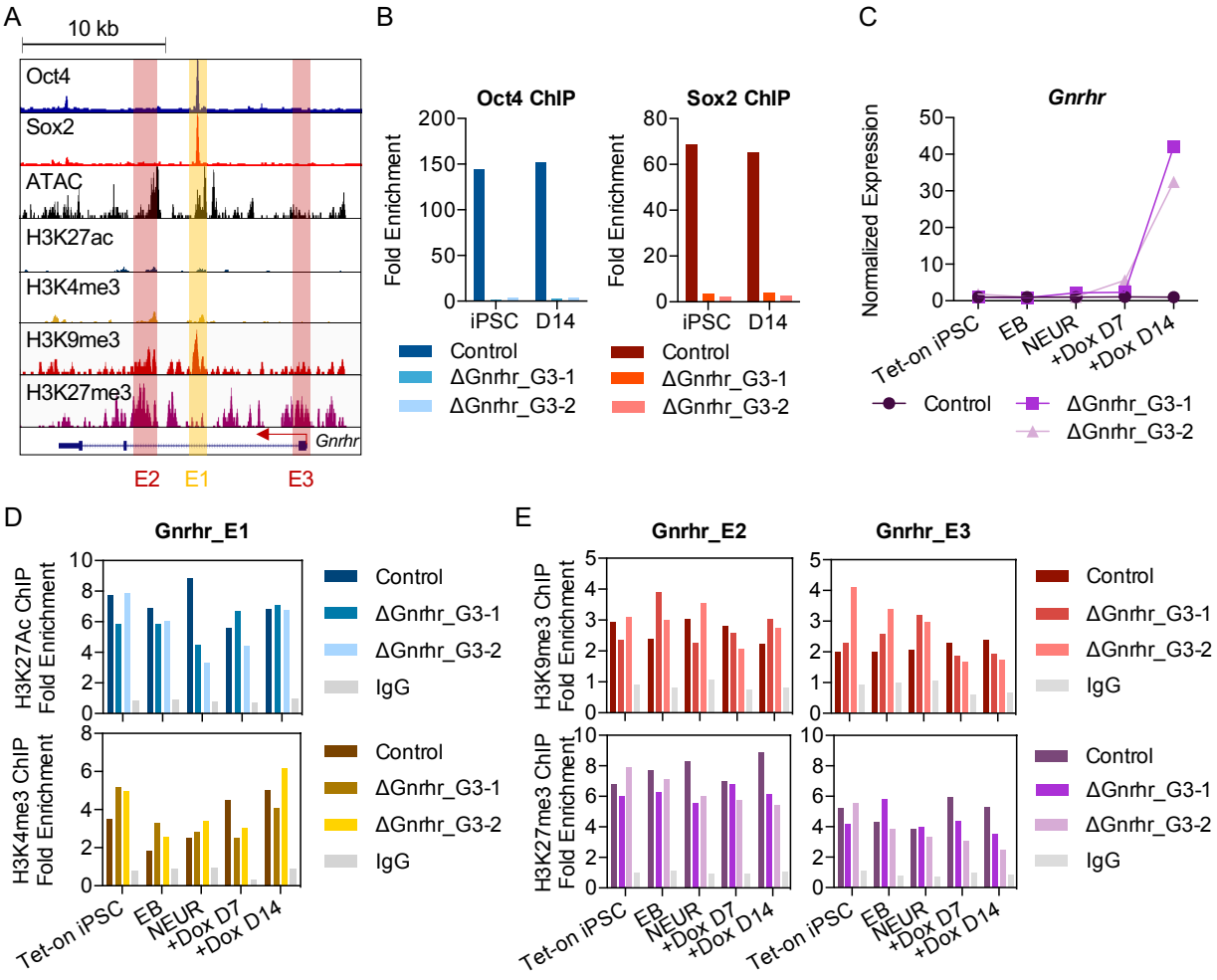


Figure 2-13. Evaluate the Roles of Oct4/Sox2 Composite Binding at the Active Enhancers of Silent Genes *Uba7* and *Lax1* by CRISPR

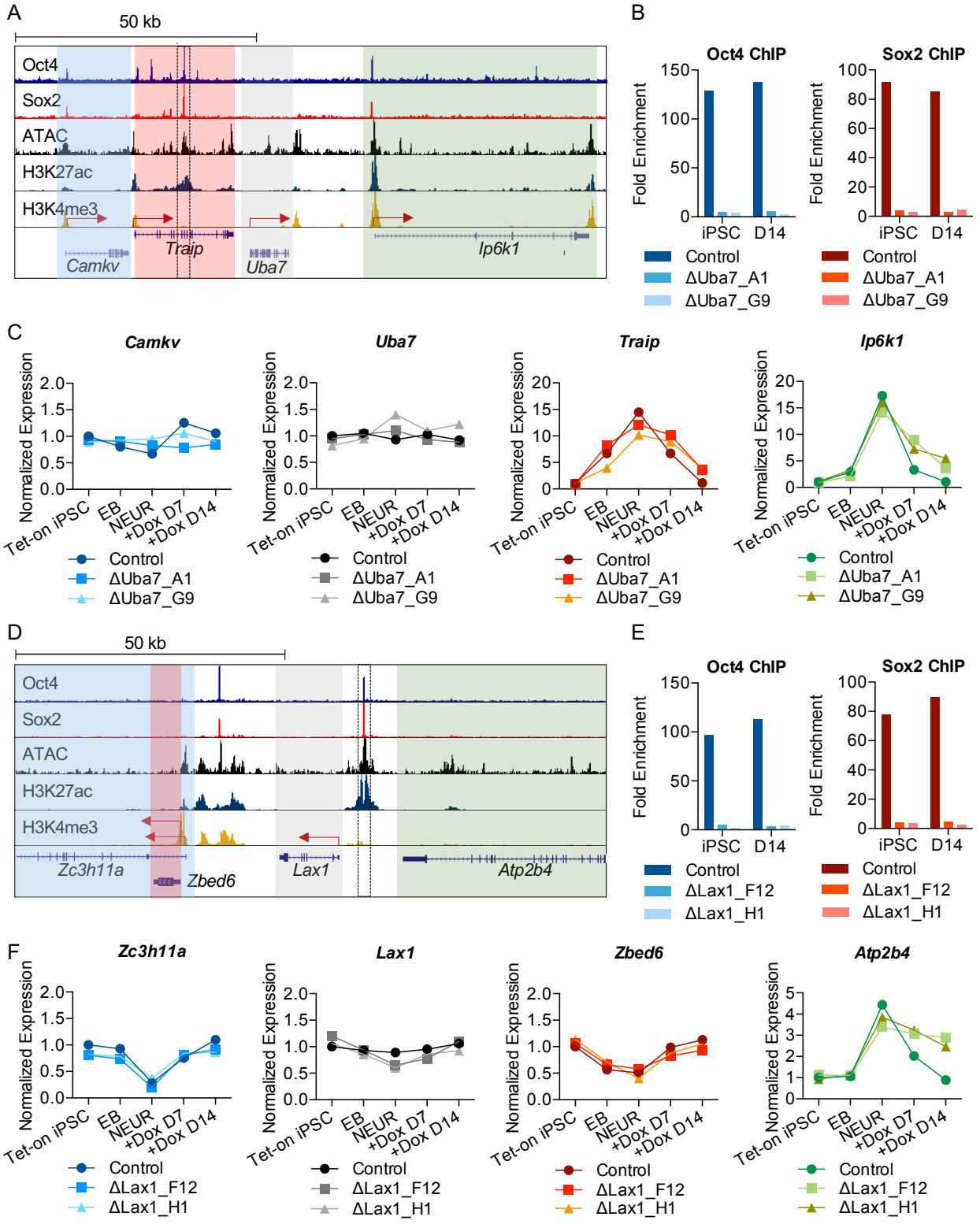


Figure 2-S1. ESC Gene Groups and Oct4/Sox2 Binding

Gene Groups	BMDM	DP	NEUR	# of Genes	# with Oct4/Sox2 >20	%
ESC-Speific	20X	20X	20X	91	35	38.46%
ESC-Dynamic	20X	0.2 - 5X	20X	12	3	25.00%
	20X	20X	0.2 - 5X	75	17	22.67%
	0.2 - 5X	20X	20X	16	5	31.25%
	0.2 - 5X	0.2 - 5X	20X	44	7	15.91%
	0.2 - 5X	20X	0.2 - 5X	67	10	14.93%
	20X	0.2 - 5X	0.2 - 5X	34	8	23.53%
ESC-NonDynamic	0.2 - 5X	0.2 - 5X	0.2 - 5X	1931	101	5.23%
ESC-Silent (RPKM < 5)				17434	711	4.08%

Figure 2-S2. A DOX-inducible System for Mouse Secondary Reprogramming of TetO-OSKM iPSCs

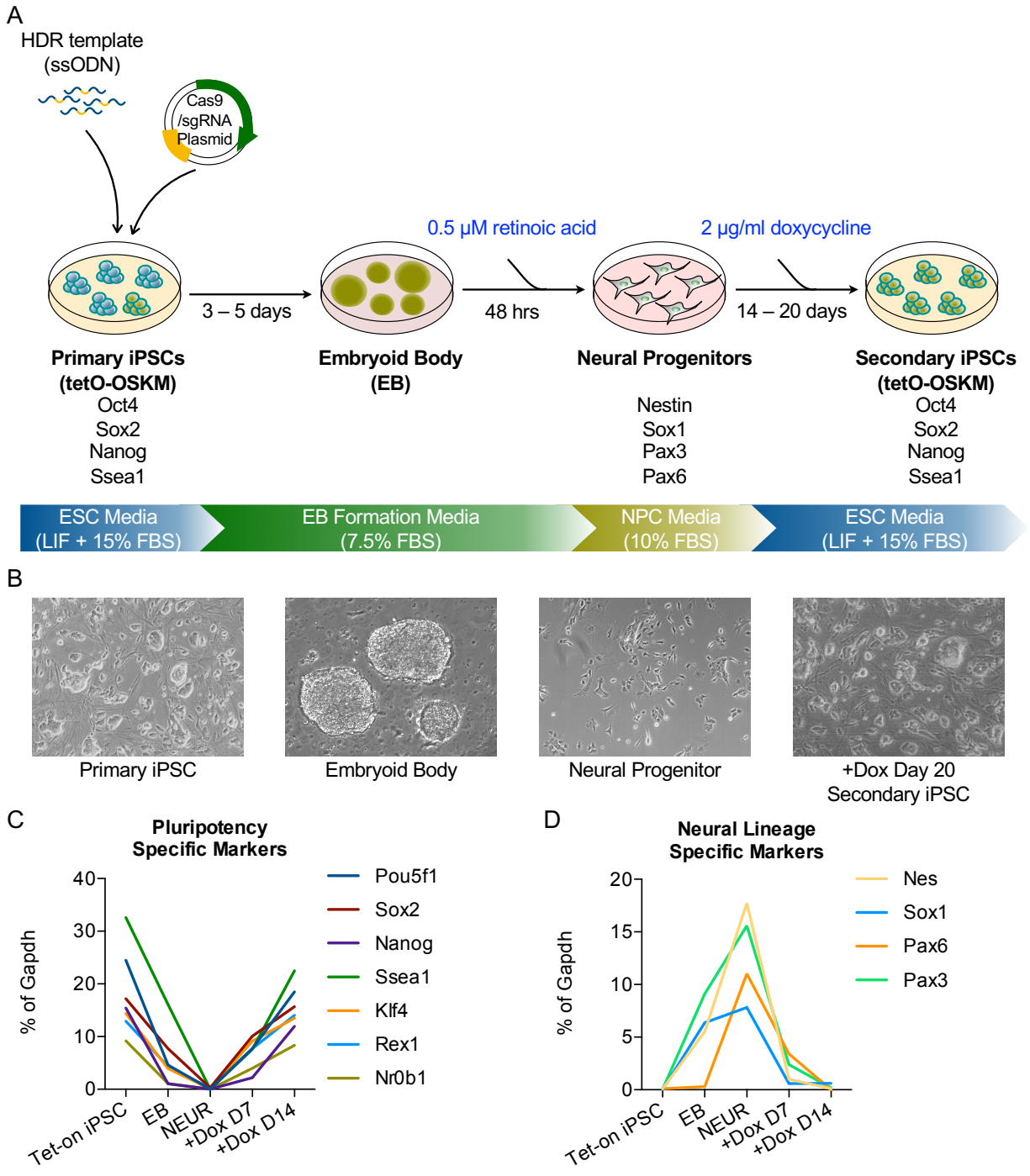


Figure 2-S3. Single Colony Expansion of CRISPR-mutated Primary iPSCs

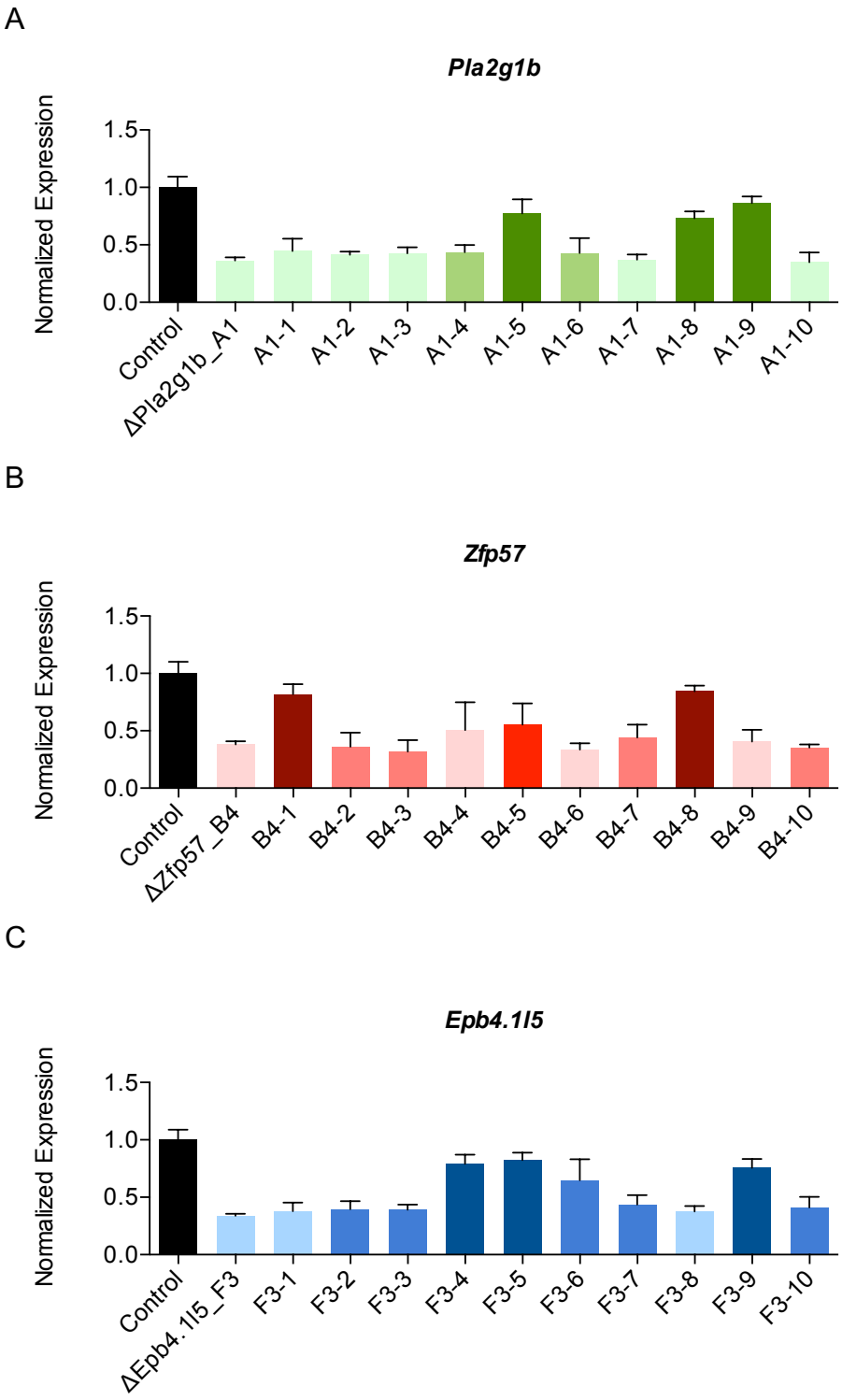
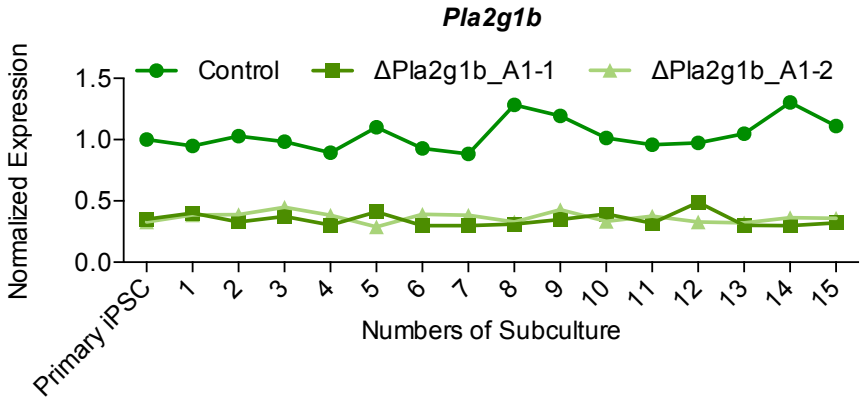
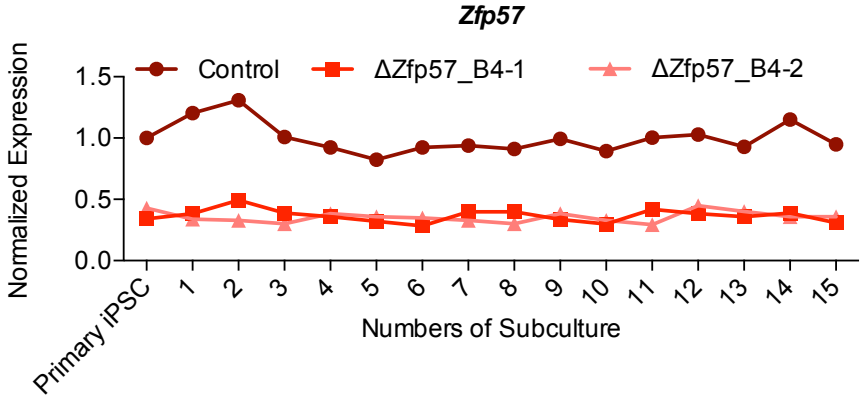


Figure 2-S4. Extended Subculture of CRISPR Mutated Primary iPSCs

A



B



C

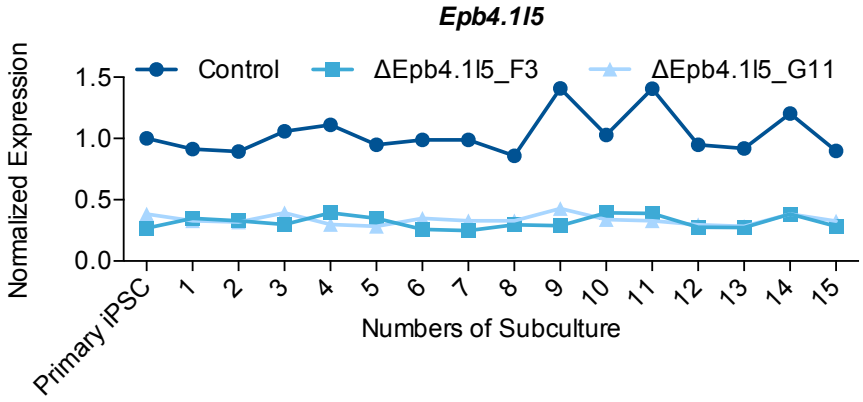


Figure 2-S5. Properties of Oct4/Sox2-bound Non-Dynamic Genes

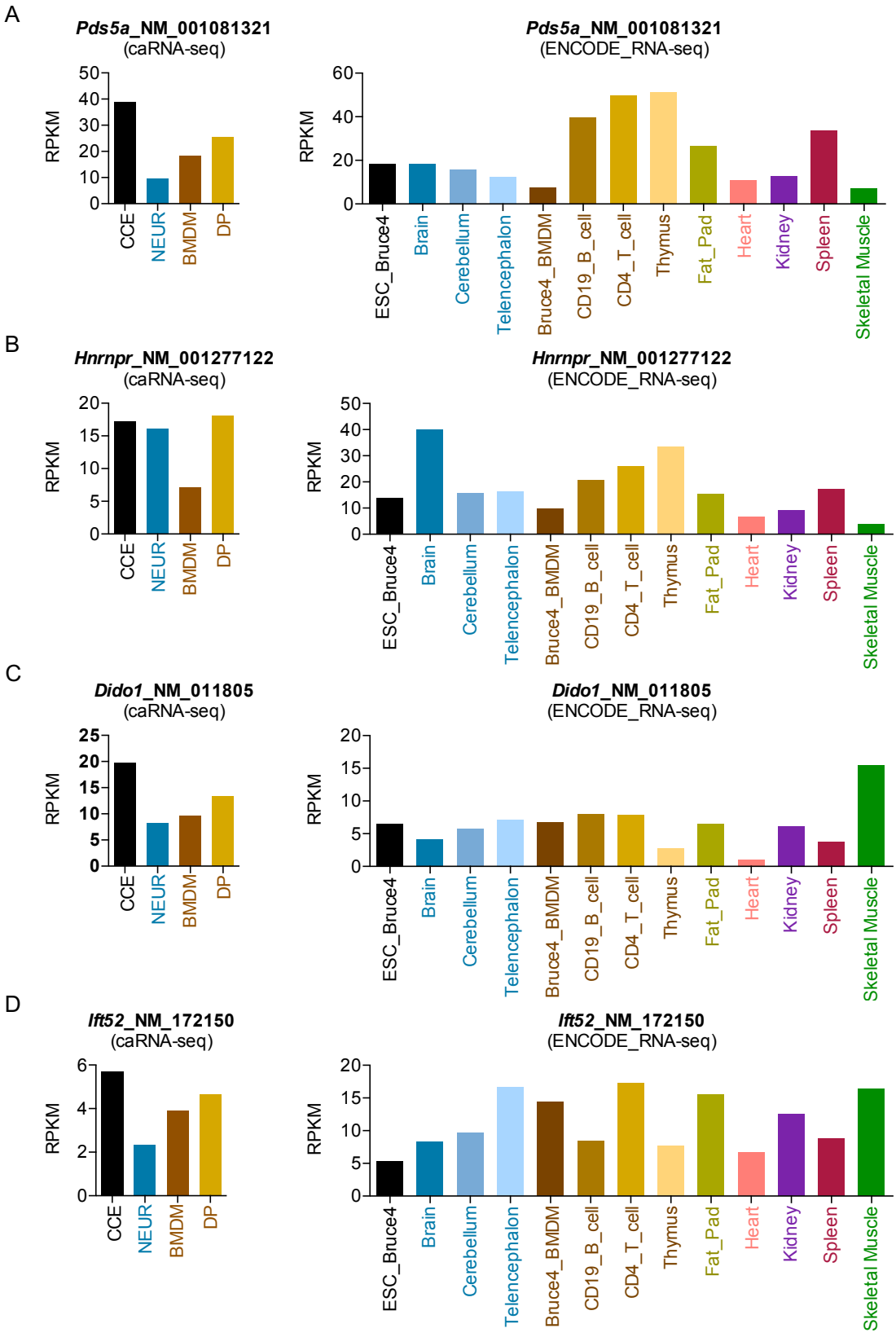


Figure 2-S6. Properties of ESC Genes In the Neighborhood of *Dido1* and *Ift52*

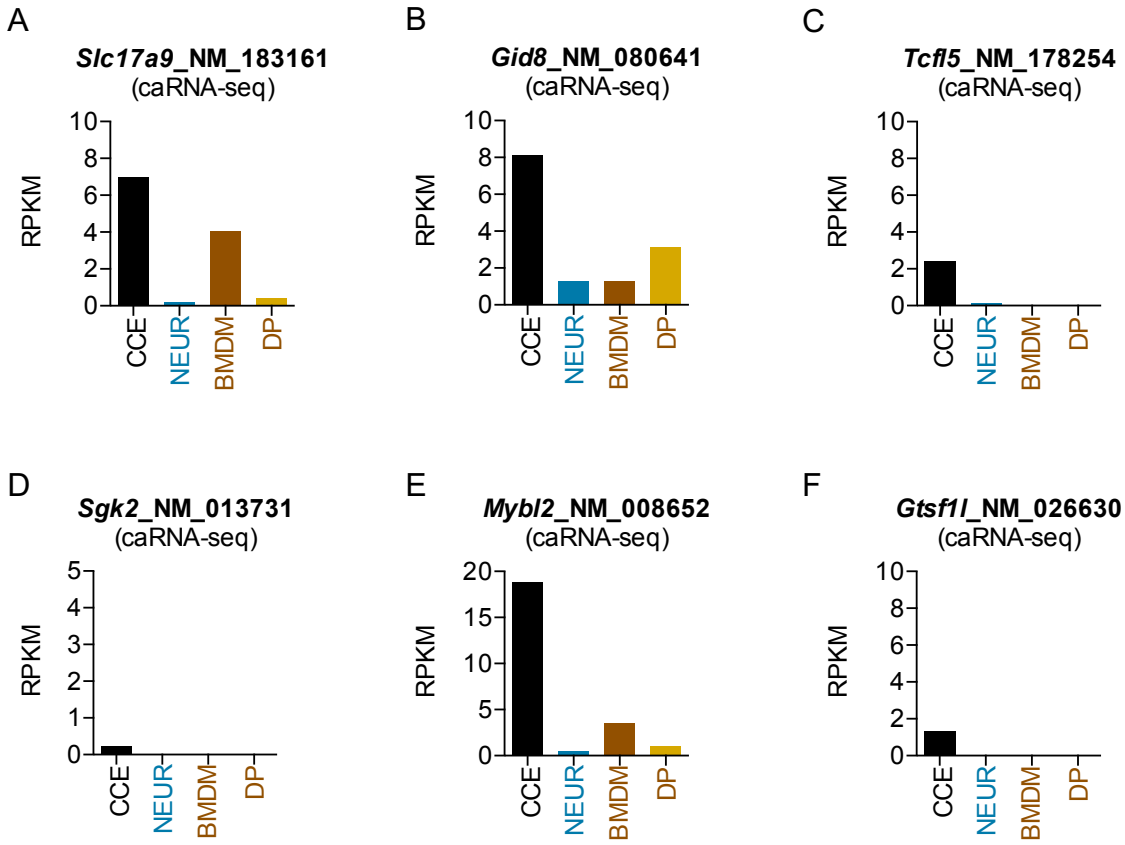


Figure 2-S7. Properties of Oct4/Sox2-bound Silent Genes

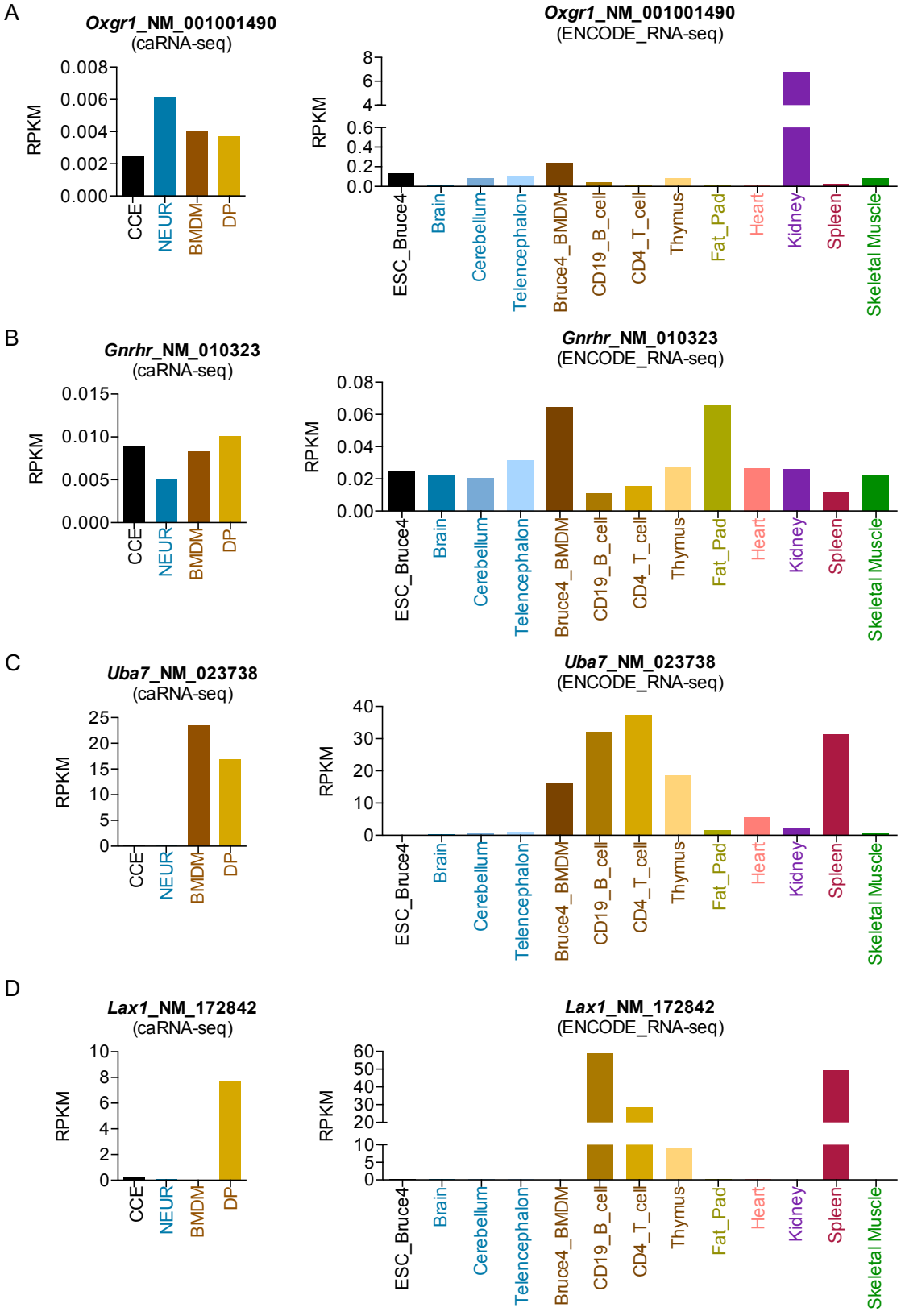


Figure 2-S8. Properties of ESC Genes In the Neighborhood of *Uba7* and *Lax1*

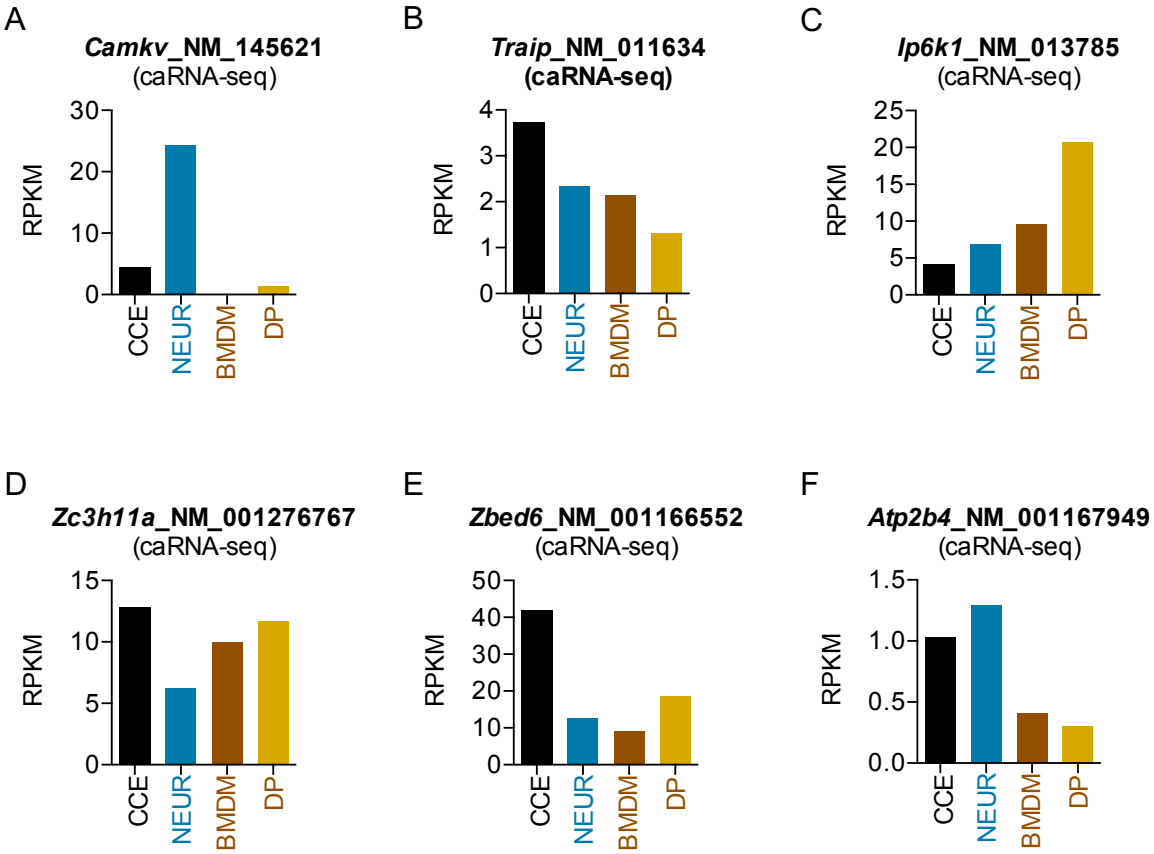


Figure 2-S9 Summary of Gene-Specific Functions of Oct4/Sox2 Composite Sites

Gene Group	Chromatin Status	Example	Oct4/Sox2 Transcription Regulation
ESC-specific	High ATAC, High H3K27ac	*Pla2g1b*	ESSENTIAL for the initial activation during reprogramming
Dynamic, expressed in a few cell types	High ATAC, High H3K27ac	*Zfp57*	ESSENTIAL, with other factors activating the genes in somatic cell types
Dynamic, broadly-expressed	High ATAC, High H3K27ac	*Epb4.1l5*	ESSENTIAL, with other factors activating the genes in somatic cell types
Non-dynamic, broadly-expressed	Low ATAC, Low H3K27ac	*Hnrnpr, Pds5a*	Non-functional toward transcription
	High ATAC, High H3K27ac	*Ift52, Dido1*	Non-functional toward transcription of annotated genes. A distal enhancer moderately acvtivates nearby gene (*Mybl2*).
Silent	Low ATAC, Low H3K27ac	*Oxgr1, Gnrhr*	Transcription repression
	High ATAC, High H3K27ac	*Uba7, Lax1*	Non-funcitonal toward transcription of annotated genes. A distal enhancer moderatley represses nearby genes (*Ip6k1*, *Atp2b4*)

Table 2-1. Primer Sequences for qRT-PCR

	Foreward	Reverse
Pou5f1	TCAGGTTGGACTGGGCCTAGT	GGAGGTTCCCTCTGAGTTGCTT
Sox2	GAGGGCTGGACTGCGAACT	TTTGCACCCCTCCCAATTC
Nanog	GAAATCCCTTCCCTCGCCATC	CTCAGTAGCAGACCCTTGTAAGC
Ssea1	GTGACGCTAACTGGCAAAGC	GGAGGGCGATTCGAAGTTCA
Klf4	GCTTGCAGCAGTAACAACCC	GGTGGGTTAGCGAGTTGGAA
Rex1	GGGTACGAGTGGCAGTTTCT	CATTTCTCTAATGCCCACAGCG
Nr0b1	GCTCTTTAACCCAGACCTGC	GGATCTGCTGGGTTCTCCAC
Nes	TTGGCTTTCCTGACCCCAAG	ATAGGTGGGATGGGAGTGCT
Sox1	CACAACTCGGAGATCAGCAA	GTCCTTCTTGAGCAGCGTCT
Pax6	CACGTACAGTGCTTTGCCAC	GCGGAGGGGTGTAGGTATCA
Pax3	AAACCCAAGCAGGTGACAAC	CTAGATCCGCCTCCTCCTCT
Pla2g1b	GGAGTGATCCCCTGAAGGAT	TGAAGTCCTCGCATTTGTTG
Zfp57	CTCCAGTTGCACAGGGGTAT	TCACGGTAAGTCTTGCCACA
Epb4.1l5	GACACCAGCACAAGCAGAAA	CTGGTTATCTTGGGCCAGAA
Hnrnpr	CCAGAAGTCATGGCAAAGGT	CTCCCCTGTCCTCAAAATGA
Pds5a	TCATCATGGAAGGTGATGGA	GACAGGTCGCTGACTGATGA
Ift52	GGAAGCTCTGGTTTCAGACG	AGCAGACAGAGCCTGTGGAT
Dido1	GATGGCCTTACGTTGAAGGA	GGGACACTGGTCCCTGACTA
Oxgr1	TTGACAGCCACCACTTTCTG	GATCCGAATGACCCTCAAGA
Gnrhr	TGCCTTCAATGCTTCCTTCT	AACTCCCCAGCATACCACTG
Lax1	TAACCCCAGCATTTCTTCCA	CCTCCTTCAGCAAACTCCAG
Uba7	TGCTCACTGCCTACATCAGG	AGGAGAGCCTCATCCAGTGA
Tcfl5	GTGGGAGAAGAAGCGCTATG	TCCATTCGGTTATGCCTCTC
Gid8	GCAGAGAAATTTCGGATGGA	CTGTCTCACGCTGACGGATA
Slc17a9	CAGTTGTGCTCTGCTTGCTC	CGGAGATGAACCCACTGAAT
Sgk2	CGCCATTGGTTACCTTCACT	TAGAGGACTGCCCCTAAGCA
Mybl2	AGGGACTGCAAGCCTGTCTA	GCAGCTATGGCAATCTCCTC
Gtsf1l	ACGTGGTTCCCATCAGAAAG	CGAACATTGGGTGACAGTTG
Camkv	GCTCAAGATTGTGCACAGGA	ATGGCCCAACAGTCTACAGG
Traip	TGTACTGCGTGTCCCTCAAG	TCTGGGCTGACCTCAGTTCT
Ip6k1	ACTGCACAGCCACTCAGATG	CCATCTTCAAGTCCAGCACA
Zc3h11a	CATAAAGCTGGGGAGATCCA	CTTTTCAGCCAGGACCTCAG
Zbed6	CTGACCCTCAGCACATCTCA	CCGTTTCCAATAGCACCACT
Atp2b4	AGATGTCGGGTTTGCTATGG	CTTTCAGTGGGGAATCCTGA
Gapdh	GGTGCTGAGTATGTCGTGGA	GTGGTTCACACCCATCACAA

Table 2-2. Mouse ENCODE Project RNA-seq Datasets

Tissue / Cell	Replicates	Library	Accession
ESC_Bruce4	1	ENCLB443JOH	ENCFF001LCA
	2	ENCLB331QIU	ENCFF001LCC
CD4-positive, alpha-beta T primary cell	1	ENCLB412BPE	ENCFF001QVP - QVQ
CD19+ B-cell	1	ENCLB243CWC	ENCFF001QJP - QJU
Bone marrow macrophage	1	ENCLB619FLT	ENCFF001LBG
	2	ENCLB193VSG	ENCFF001LBL
Brain	1	ENCLB333FBI	ENCFF001QWP - QWQ
	2	ENCLB130CTN	ENCFF001QWR - QWS
Cerebellum	1	ENCLB796LUO	ENCFF001QLJ - QLK
	2	ENCLB036XUG	ENCFF001QLL - QLO
Skeletal Muscle	1	ENCLB091FFS	ENCFF001QTC - QTH
	2	ENCLB223ZBY	ENCFF001QTI - QTT
Fat Pad	1	ENCLB692FPW	ENCFF001QMH - QMI
Thymus	1	ENCLB702UHC	ENCFF001QUN - QUT
	2	ENCLB797ZQK	ENCFF001QUX - QVK
Telencephalon	1	ENCLB564WVJ	ENCFF001QLZ - QMA
	2	ENCLB378YVY	ENCFF001QNK
Spleen	1	ENCLB078JRX	ENCFF001QTZ - QUE
Heart	1	ENCLB649LBK	ENCFF001QNR - QNV
	2	ENCLB620OMQ	ENCFF001QNW - QOB
Kidney	1	ENCLB505CAN	ENCFF001QOK - QOL

Table 2-3. Transcription Factors ChIP-seq, Histone Marks ChIP-seq, and ATAC-seq Datasets

Sequencing Datasets	Cell Type	GEO Accession
Oct4 ChIP-seq	ESC	GEO: GSE90895
Sox2 ChIP-seq	ESC	GEO: GSE90895
Nanog ChIP-seq	ESC	GEO: GSE90895
H3K27Ac ChIP-seq	ESC	GEO: GSE 56138
H3K4me3 ChIP-seq	ESC	GEO: GSE 62380
ATAC-seq	ESC	GEO: GSE 52397
H3K27Ac ChIP-seq	NPC	GEO: GSE61874
H3K4me3 ChIP-seq	NPC	GEO: GSE61874
ATAC-seq	NPC	GEO: GSE84646

REFERENCE

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

Angie Rizzino1, 2 (2013). The Sox2-Oct4 Connection: Critical players in a much larger interdependent network integrated at mul. Stem Cells.

Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., Borowsky, M.L., et al. (2013). Genome-wide chromatin interactions of the nanog locus in pluripotency, differentiation, and reprogramming. Cell Stem Cell.

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

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res. *37*.

Becker, J.S., Nicetto, D., and Zaret, K.S. (2016). H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. Trends Genet. *32*, 29–41.

Benner, C., Heinz, S., and Glass, C.K. (2017). HOMER - Software for motif discovery and next generation sequencing analysis.

van den Berg, D.L.C., Snoek, T., Mullin, N.P., Yates, A., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R.A. (2010). An Oct4-Centered Protein Interaction Network in Embryonic Stem Cells. Cell Stem Cell.

Bhatt, D.M., Pandya-Jones, A., Tong, A.J., Barozzi, I., Lissner, M.M., Natoli, G., Black, D.L., and Smale, S.T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279–290.

Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B., and Young, R.A. (2009). SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes Dev.

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

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

Buecker, C., Srinivasan, R., Wu, Z., Calo, E., Acampora, D., Faial, T., Simeone, A., Tan, M., Swigut, T., and Wysocka, J. (2014). Reorganization of enhancer patterns in transition from naive to primed pluripotency. Cell Stem Cell.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods.

Calo, E., and Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? Mol. Cell.

Chew, J.-L., Loh, Y.-H., Zhang, W., Chen, X., Tam, W.-L., Yeap, L.-S., Li, P., Ang, Y.-S., Lim, B., Robson, P., et al. (2005). Reciprocal Transcriptional Regulation of Pou5f1 and Sox2 via the Oct4/Sox2 Complex in Embryonic Stem Cells. Mol. Cell. Biol.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. Cell.

Cong, L., and Zhang, F. (2014). Genome engineering using crispr-cas9 system. In Chromosomal Mutagenesis: Second Edition, p.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., Pellegrini, M., and Plath, K. (2013). Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization. Cell Stem Cell.

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.

Esch, D., Vahokoski, J., Groves, M.R., Pogenberg, V., Cojocaru, V., Vom Bruch, H., Han, D., Drexler, H.C.A., Araúzo-Bravo, M.J., Ng, C.K.L., et al. (2013). A unique Oct4 interface is crucial for reprogramming to pluripotency. Nat. Cell Biol.

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

Ezhkova, E., Pasolli, H.A., Parker, J.S., Stokes, N., Su, I. hsin, Hannon, G., Tarakhovsky, A., and Fuchs, E. (2009). Ezh2 Orchestrates Gene Expression for the Stepwise Differentiation of Tissue-Specific Stem Cells. Cell.

Fuhrmann, G., Chung, A.C.K., Jackson, K.J., Hummelke, G., Baniahmad, A., Sutter, J., Sylvester, I., Schöler, H.R., and Cooney, A.J. (2001). Mouse Germline Restriction of Oct4 Expression by Germ Cell Nuclear Factor. Dev. Cell.

Hammachi, F., Morrison, G.M., Sharov, A.A., Livigni, A., Narayan, S., Papapetrou, E.P.,

O'Malley, J., Kaji, K., Ko, M.S.H., Ptashne, M., et al. (2012). Transcriptional Activation by Oct4 Is Sufficient for the Maintenance and Induction of Pluripotency. Cell Rep.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Ho, R., Papp, B., Hoffman, J.A., Merrill, B.J., and Plath, K. (2013). Stage-Specific Regulation of Reprogramming to Induced Pluripotent Stem Cells by Wnt Signaling and T Cell Factor Proteins. Cell Rep.

Hockemeyer, D., Soldner, F., Cook, E.G., Gao, Q., Mitalipova, M., and Jaenisch, R. (2008). A Drug-Inducible System for Direct Reprogramming of Human Somatic Cells to Pluripotency. Cell Stem Cell.

Ito, K., and Suda, T. (2014). Metabolic requirements for the maintenance of self-renewing stem cells. Nat. Rev. Mol. Cell Biol.

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

Jauch, R., Ng, C.K.L., Saikatendu, K.S., Stevens, R.C., and Kolatkar, P.R. (2008). Crystal Structure and DNA Binding of the Homeodomain of the Stem Cell Transcription Factor Nanog. J. Mol. Biol. *376*, 758–770.

Jerabek, S., Merino, F., Schöler, H.R., and Cojocaru, V. (2014). OCT4: Dynamic DNA binding pioneers stem cell pluripotency. Biochim. Biophys. Acta - Gene Regul. Mech.

Ji, X., Dadon, D.B., Abraham, B.J., Lee, T.I., Jaenisch, R., Bradner, J.E., and Young, R.A. (2015). Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. Proc. Natl. Acad. Sci.

Kamachi, Y., Uchikawa, M., and Kondoh, H. (2000). Pairing SOX off: With partners in the regulation of embryonic development. Trends Genet.

Kamada, R., Yang, W., Zhang, Y., Patel, M.C., Yang, Y., Ouda, R., Dey, A., Wakabayashi, Y., Sakaguchi, K., Fujita, T., et al. (2018). Interferon stimulation creates chromatin marks and establishes transcriptional memory. Proc. Natl. Acad. Sci.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: A fast spliced aligner with low memory requirements. Nat. Methods.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008a). An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. Cell *132*, 1049–1061.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008b). An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. Cell.

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

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods.

Lee, S.H., Lumelsky, N., Studer, L., Auerbach, J.M., and McKay, R.D. (2000). Efficient

generation of midbrain and hindbrain neurons from mouse embryonic stem cells. Nat. Biotechnol.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K. ichi, et al. (2006). Control of Developmental Regulators by Polycomb in Human Embryonic Stem Cells. Cell.

Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods *7*, 709–715.

Li, M., and Belmonte, J.C.I. (2017). Ground rules of the pluripotency gene regulatory network. Nat. Rev. Genet.

Li, M., and Izpisua Belmonte, J.C. (2018). Deconstructing the pluripotency gene regulatory network. Nat. Cell Biol.

Liang, J., Wan, M., Zhang, Y., Gu, P., Xin, H., Jung, S.Y., Qin, J., Wong, J., Cooney, A.J., Liu, D., et al. (2008). Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. Nat. Cell Biol.

Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat. Genet.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: Motif analysis of large DNA datasets. Bioinformatics.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., et al. (2007). Directly Reprogrammed Fibroblasts Show Global Epigenetic Remodeling and Widespread Tissue Contribution.

Cell Stem Cell *1*, 55–70.

Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc. Natl. Acad. Sci.

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

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

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods.

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

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

Orkin, S.H. (2005). Chipping away at the embryonic stem cell network. Cell *122*, 828–830.

Orkin, S.H., Wang, J., Kim, J., Chu, J., Rao, S., Theunissen, T.W., Shen, X., and Levasseur, D.N. (2008). The transcriptional network controlling pluripotency in ES cells. In Cold Spring Harbor Symposia on Quantitative Biology, p.

P., Y., J., H., J., G., Y.L., O., M., H., Y.-H., L., L.-P., Y., P., R., B., L., and H.-H., N.

(2009). Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. Genes Dev.

Papp, B., and Plath, K. (2013). Epigenetics of reprogramming to induced pluripotency. Cell.

Pardo, M., Lang, B., Yu, L., Prosser, H., Bradley, A., Babu, M.M., and Choudhary, J. (2010). An Expanded Oct4 Interaction Network: Implications for Stem Cell Biology, Development, and Disease. Cell Stem Cell.

Pasini, D., Cloos, P.A.C., Walfridsson, J., Olsson, L., Bukowski, J.P., Johansen, J. V., Bak, M., Tommerup, N., Rappsilber, J., and Helin, K. (2010). JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature.

Petryniak, B., Staudt, L.M., Postema, C.E., McCormack, W.T., and Thompson, C.B. (2006). Characterization of chicken octamer-binding proteins demonstrates that POU domain-containing homeobox transcription factors have been highly conserved during vertebrate evolution. Proc. Natl. Acad. Sci.

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

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtier, K., Ponting, C.P., Allis, C.D., et al. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature.

Reményi, A., Lins, K., Nissen, L.J., Reinbold, R., Schöler, H.R., and Wilmanns, M. (2003). Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. Genes Dev. *17*, 2048–2059.

Sagner, A., Gaber, Z.B., Delile, J., Kong, J.H., Rousso, D.L., Pearson, C.A., Weicksel, S.E., Melchionda, M., Mousavy Gharavy, S.N., Briscoe, J., et al. (2018). Olig2 and Hes regulatory dynamics during motor neuron differentiation revealed by single cell transcriptomics. PLoS Biol.

Schepers, G.E., Teasdale, R.D., and Koopman, P. (2002). Twenty pairs of Sox: Extent, homology, and nomenclature of the mouse and human Sox transcription factor gene families. Dev. Cell.

Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G., and Rauscher, F.J. (2002). SETDB1: A novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes Dev.

Simandi, Z., Horvath, A., Wright, L.C., Cuaranta-Monroy, I., De Luca, I., Karolyi, K., Sauer, S., Deleuze, J.F., Gudas, L.J., Cowley, S.M., et al. (2016). OCT4 Acts as an Integrator of Pluripotency and Signal-Induced Differentiation. Mol. Cell.

Smale, S.T. (2010). Pioneer factors in embryonic stem cells and differentiation. Curr. Opin. Genet. Dev.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2014). Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. Cell *161*, 555–568.

Sridharan, R., Gonzales-Cope, M., Chronis, C., Bonora, G., McKee, R., Huang, C., Patel, S., Lopez, D., Mishra, N., Pellegrini, M., et al. (2013). Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1γ in reprogramming to pluripotency. Nat. Cell Biol.

Di Stefano, B., Collombet, S., Jakobsen, J.S., Wierer, M., Sardina, J.L., Lackner, A., Stadhouders, R., Segura-Morales, C., Francesconi, M., Limone, F., et al. (2016). C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. Nat. Cell Biol.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663–676.

Tapia, N., Maccarthy, C., Esch, D., Gabriele Marthaler, A., Tiemann, U., Araúzo-Bravo, M.J., Jauch, R., Cojocaru, V., and Schöler, H.R. (2015). Dissecting the role of distinct OCT4-SOX2 heterodimer configurations in pluripotency. Sci. Rep.

Thomson, M., Liu, S.J., Zou, L.N., Smith, Z., Meissner, A., and Ramanathan, S. (2011). Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell.

Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D.N., Theunissen, T.W., and Orkin, S.H. (2006). A protein interaction network for pluripotency of embryonic stem cells. Nature.

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

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Xu, C.-R., Li, L.-C., Donahue, G., Ying, L., Zhang, Y.-W., Gadue, P., and Zaret, K.S. (2014). Dynamics of genomic H3K27me3 domains and role of EZH2 during pancreatic

endocrine specification. EMBO J.

Xu, J., Pope, S.D., Jazirehi, A.R., Attema, J.L., Papathanasiou, P., Watts, J. a, Zaret, K.S., Weissman, I.L., and Smale, S.T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *104*, 12377–12382.

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: Establishing competence for gene expression. Genes Dev.

CHAPTER 3

In-Depth Genomic/Genetic Analyses of Transcriptional Regulation by Oct4 and Sox2 in Embryonic Stem Cells

ABSTRACT

The transcriptional regulation mediated by Oct4 and Sox2 targets distinct gene groups with various transcription features. A well-defined gene classification is required for scrutinizing the regulatory function of Oct4 and Sox2 binding in embryonic stem cells (ESCs). The discrepancy in the transcriptional state of different gene groups indicates that the Oct4 and Sox2 function differently for the transcriptional output. However, the mechanisms that distinguish different functions of Oct4 and Sox2 transcriptional regulation remain unclear. In this study, we described a gene-centric approach that emphasized on quantitative aspects of RNA-seq, ChIP-seq, ATAC-seq, and genomic context to examine possible selective mechanisms for the Oct4 and Sox2 transcriptional regulation. Using a comparative and integrative approach for genomic and genetic features in ESCs, we analyzed the motif conservation, pluripotency TFs / somatic TFs co-bind, epigenetic properties, enhancer properties, and CpG content, among the 1,035 strong Oct4/Sox2 composite bindings within the 15 kbp of annotated transcription start sites. Additionally, as Oct4 and Sox2 are critical for the establishment of pluripotency, we investigated the prevalence and strength of cooperative binding with the cell-type-specific transcription factors, pluripotency factors, and histone modifiers at the Oct4/Sox2x composite sites. These analyses enabled us to scrutinize the potential mechanisms that distinguished the selective Oct4/Sox2 transcriptional regulation in pluripotency. As a first step to uncover the selective mechanisms for the Oct4 and Sox2 transcriptional regulation, this gene-centric approach allowed us to quantitatively evaluate each genomic/genetic feature at the Oct4/Sox2 binding sites with different functional significance.

INTRODUCTION

Master regulators including the Oct4, Sox2, and Nanog (OSN) transcription factors (TFs) are indispensable for governing ESCs identity through a complex hierarchy of gene regulation, which ensures the maintenance of pluripotency (Avilion et al., 2003; Mitsui et al., 2003; Nichols et al., 1998). In differentiated cells, ectopic expression of four transcription factors Oct4, Sox2, Klf4, and c-Myc (OSKM), can induce pluripotency, and these induced pluripotent stem cells (iPSCs) give rise to tissue development both in vitro and in vivo (Maherali et al., 2007; Takahashi and Yamanaka, 2006; Takahashi et al., 2007). The induction of pluripotency requires fundamental transitions in the gene transcription. A successful reprogram to pluripotent state results in the silence of somatic genes and the activation of ESC-transcribed genes (Jerabek et al., 2014). These ESC-transcribed genes could be entirely inactive in the differentiated cells or constitutively expressed in multiple lineages. Most studies have relied on low stringency criteria to define the differential expression. The vague definition of these ESC-transcribed genes often compromises the real functions and precise targets of the master regulators. Hence, a careful interrogation of the transcript profiles of ESCs and differentiated cells with high stringency criteria will be the first step to dissect the pluripotency network. By integrating the master regulators binding profiles (ChIP-seq), epigenetic characteristics, and the DNA context, an in-depth genomic/genetic analysis will provide insights into the principle of pluripotency gene regulation.

A recent study delineated a substantial interaction of Oct4 and Sox2 with somatic and pluripotency enhancers, suggesting context-dependent regulators for both ESC self-

renewal and lineage differentiation (Chronis et al., 2017). The somatic-enhancer inactivation and the pluripotency-enhancer activation were selected by the cooperation with stage-specific TFs and pluripotency TFs. This cooperative binding reiterated the previous perspective of gene regulatory circuit in ESCs. The central regulators appear to promote the expression of other pluripotency/self-renewal genes and simultaneously prevent the differentiation-promoting genes (Orkin, 2005). In human ESCs, analysis of OCT4 genome-wide occupancy revealed that, in addition to pluripotency factors co-occupied enhancers, OCT4 also bound with RAR: RXR or β-catenin at the enhancers activated by retinoic acid (RA) or canonical Wnt / β-catenin signal (Simandi et al., 2016). The unexpected collaboration between OCT4 and the differentiation transcription factors demonstrated an integrative role of OCT4 in pluripotency and signal-induced differentiation. Given the cooperative binding of other TFs at the Oct4/Sox2 sites, we must ask how the TF co-bind might distinguish the functions of those Oct4/Sox2 sites genome-wide. Furthermore, it is unclear how the genomic/genetic context such as conservation, CpG content, or epigenetic properties controls their transcriptional activities.

In ESCs, the complex interrelationship between pluripotency and chromatin factors mediates the chromatin plasticity, which establishes the epigenetic barrier between pluripotency and differentiation (Becker et al., 2016; Meshorer et al., 2006). The interplay between pluripotency factors and histone modifier alters the chromatin accessibility for the competence of gene expression (Denholtz et al., 2013; Maherali et al., 2007). Distinct epigenetic characteristics at promoter and enhancer regions contribute to the expression of cell type-specific genes and mark the identities in the given cell types. In ESCs, electron

microscopy and genome-wide chromatin profiling revealed globally open and highly dynamic chromatin configuration (Azuara et al., 2006; Meissner et al., 2008; Park et al., 2004). In human ESCs, the unique chromatin states signified the expression of different gene classes characterized by RNA analysis and functional annotation (Rada-Iglesias et al., 2011). For genes annotated with pluripotent functions, enriched histone acetylation H3K27Ac regions overlapped with the enhancers of the ESC-expressing genes.

In contrast, repressive histone marks H3K9me3 and H3K27me3 occupied the enhancers to silence the developmental genes in ESCs. During the differentiation, the binding of lineage-determining TFs and the recruitment of nucleosome remodeling complexes delivered the signal and altered the chromatin landscape for the differentiated gene expression (Heinz et al., 2010). The induction of cell-type-specific genes and the chromatin modification established the transition barrier between the differentiated cells and pluripotent stem cells. Conversely, during the reprogramming, ectopic expression of OSKM also induced genome-wide chromatin remodeling to perpetuate the pluripotent state (Koche et al., 2011). In the early stage of reprogramming, although OSKM only activated a small subset of genes with the active promoters, a rapid and extensive chromatin remodeling established the active or poised states of the pluripotency-enhancers. These epigenetic mechanisms developed a productive engagement of OSKM and the activation of pluripotent genes in late reprogramming. In a recent study, during the reprogramming, a higher concentration of Oct4, Sox2, and Klf4 could recognize the incomplete consensus motifs in the nucleosomal DNA with lower affinity (Soufi et al., 2014). As pioneering TFs, Oct4, Sox2, and Klf4 may access the pluripotency-enhancers

in a very early stage of reprogramming and mediate the nucleosome displacement for priming the chromatin landscapes. These studies demonstrated that the collaborative mechanisms of TF bindings and chromatin dynamic are crucial for repressing differentiated genes and activating pluripotent genes. Whether these epigenetic properties denote the selective functions of Oct4/Sox2 composite in pluripotency and reprogramming, however, remains elusive.

In this study, we perform an in-depth genomic/genetic analysis of pluripotency gene regulation using gene-centric approach that emphasized on quantitative aspects of RNA-seq, ChIP-seq, ATAC-seq, and genomic context. By identifying Oct4/Sox2 targets and interrogating the conservation, pluripotency TFs / somatic TFs co-bind, epigenetic properties, enhancer properties, and CpG content, we carefully addressed the genetic/genomic features of the 1,035 Oct4/Sox2 composite sites. We also compared these genetic/genomic features within the representative genes selected for CRISPR/Cas9 experiments. This analysis enabled us to scrutinize the potential mechanisms that distinguished the selective Oct4/Sox2 transcriptional regulation in pluripotency. Given the complexity of the pluripotency gene network, characterizing the characteristics of the Oct4/Sox2 sites focused on the well-defined gene sets demonstrates an example to interrogate the selective mechanisms regulating transcription.

RESULTS

Classification of ESC Gene Groups Based on Mouse Encode Database

In chapter 2, we categorized ESC gene groups according to the dynamic range of the nascent transcripts between CCE ESC line and three somatic cell types. Even we only compared the transcript level with three somatic cell types, only a small number of genes are even approaching ESC-specificity (ESC-specific genes, N = 91). It is reasonable to foresee that this number could further reduce if we include more somatic tissue/cell types to determine the dynamic range of expression and ESC-specificity. To obtain a more comprehensive picture of gene classification, we evaluated the mRNA transcript profiles of ESC and twelve somatic tissue/cell types on Mouse Encode Database (http://www.mouseencode.org/). Representative somatic tissue/cell types include CD4+ Primary T cell, CD19+ B cell, BMDM, Brain, Cerebellum, Skeletal Muscle, Fat Pad, Thymus, Telencephalon, Spleen, Heart, and Kidney. Using stringent criteria include fold change between individual tissue/cell types and minimum expression threshold of 4.9 RPKM in ESC, we separated genes based on their mRNA transcript profiles (Figure 3-1A). The 4.9 RPKM cutoff returned the genes with top 20% mRNA transcript level. To classify ESC genes with different level of the dynamic range of expression, we calculated the fold change of gene expression between ESCs and individual tissue/cell types. Next, we cataloged the gene numbers based on the binary characteristics of > 20-fold or 0.2 – 5-fold mRNA transcription in ESCs (Figure 3-1A, second and third columns). Overall, including ESC-specific genes (Group 1), we classified 789 ESC genes into twelve Dynamic gene groups with various degree of dynamic range of expression. In Group 1, these 40 ESC-specific genes show more than 20-fold expression over twelve tissue/cell

types. *Utf1* is an example of the Group 1 ESC-specific gene, which is highly expressed in only ESCs (Figure 3-1B). Starting from Group 2, the number of tissue/cell types with 20-fold differential expression begin to drop from eleven to only one tissue/cell type in group 12. For example, *Phlda2* represents a Group 3 Dynamic gene with 20-fold differential expression in 10 tissue/cell types but is also expressed in Fat Pad and Kidney (Figure 3-1C). On the other hand, *Itgb5* is a Group 12 gene exhibiting 20-fold expression in CD4 T cell but is broadly expressed in other 11 tissue/cell types (Figure 3-1D).

Strong Oct4/Sox2 Binding is Greatly Enriched within Genes that are Highly Expressed in ESC and Poorly Expressed in a Substantial Number of Somatic Cell Types

To compare the Oct4/Sox2 bindings in different gene groups, we analyzed Oct4/Sox2 ChIP-seq data and evaluated the occupancy within 15 kbp of the annotated gene TSS. Because the weaker peaks are less reproducible and are more likely to represent technical artifacts, we set a threshold of peak score > 20 to include only strongly reproducible Oct4 and Sox2 binding (Figure 3-1A). Among all 789 ESC genes with various degree of the dynamic range of expression in twelve somatic cell types, we identified 115 genes with strong Oct4/Sox2 binding. Consistent with the gene classification based on the nascent transcript profiles, more than one-third of Group 1 ESC-specific genes (37.50%, 15 out of 40) contain a composite Oct4/Sox2 site (Figure 3-1A, blue shade). In contrast, only about 6 – 7% genes in Group 10, 11, and 12 present a strong Oct4/Sox2 peak within 15 kbp (Figure 3-1A, green shades). These genes are expressed in ESCs but showing a dynamic range of expression in only a few cell types

(3 for Group 10; 2 for Group 11; 1 for Group 12). A 6 – 7% of Oct4/Sox2 occupancy is similar to the random distribution of a strong Oct4/Sox2 throughout the mouse genome (4.47%). The low percentage of Oct4/Sox2 binding at these three groups corroborate our discovery in Non-Dynamic genes classified by the nascent transcript profiles. Oct4/Sox2 sites at Group 10 to 12 might represent non-functional binding to the gene transcription in pluripotency. Notably, genes grouped in Group 2, Group 3, and Group 4 also exhibit a range of 20% - 33% Oct4/Sox2 occupancy, which is close to ESC-specific genes (Figure 3-1A, purple shades). These genes may not approach ESC-specificity but exhibit large dynamic range of expression in more than nine different somatic cell types. In Chapter 2, CRISPR-HDR mutation of the Oct4/Sox2 sites at these of Dynamic genes proved the functional significance in activating transcription at pluripotent state. In the differentiation states lacking Oct4/Sox2 expression, a lineage-specific mechanism takes the place of Oct4/Sox2 for gene activation. The analyses of strong Oct4/Sox2 binding in gene groups classified by mRNA transcript profiles in ESCs and twelve somatic cell types demonstrate a consistent trend as our nascent transcript classification. Strong Oct4/Sox2 binding is significantly enriched in the genes exhibiting the properties of high expression in ESCs and inactive in a substantial number of somatic cells. The enrichment persists when we include more somatic cell types for gene classification. Even these genes will never approach ESC-specificity, the transcription is predominantly dependent on Oct4/Sox2 binding in the pluripotent state, with other mechanisms required for the transcription in somatic cells.

Human OCT4/SOX2 Composite Binding is Enriched Only in ESC-specific Genes

To assess if the Oct4/Sox2 composite binding is conserved and exhibits similar properties between mouse and human, we analyzed OCT4 and SOX2 ChIP-seq in naïve human ESC and human ESC line H9 (GEO: GSE69479, GSE69647) (Ji et al., 2016; Zhou et al., 2016). We combined two biological replicates of OCT4, and only retained the reproducible peaks called by HOMER findPeaks algorithm (Benner et al., 2017). For the SOX2 ChIP-seq, only one replicate of the experiment is accessible on GEO. From the analysis of HOMER peak calling, we identified 14,106 Oct4 peaks and 44,147 Sox2 peaks (FDR < 0.01) genome-wide. Similar to the observation in mouse, only a small fraction of OCT4 (3.9%) and SOX2 peaks (7.4%) locates at the promoter region (-500 bp to +150 bp relative to TSS). Nearly 90% OCT4 and SOX2 peaks fall into the intergenic or intronic region (Figure 3-2A and B). The genomic distribution of called peaks indicates that the majority of Oct4 and Sox2 regulate gene transcription through the binding of enhancer regions.

To classify gene groups with various degree of dynamic range of expresion, we evaluated the mRNA transcript profiles of ESC and fourteen somatic tissue/cell types on Human Encode Database (http://www.encodeproject.org/). Representative somatic tissue/cell types include Adipose tissue, Adrenal gland, B cell, Brain, CD14+ monocyte, Sigmoid colon, Heart, Human endothelial cell of umbilical vein, Small intestine, Liver, Ovary, Foreskin fibroblast, Spleen, and Skeletal muscle myoblast. Using stringent criteria include fold change between individual tissue/cell types and minimum expression threshold of 15 RPKM in human H1 ESC, we separated genes based on their mRNA transcript profiles (Figure 3-2C). The 15 RPKM cutoff returned the genes with top 20% mRNA transcript

level. To classify ESC genes with different level of the dynamic range of expression, we calculated the fold change and cataloged the gene numbers as previously described for Mouse RNA-seq datasets (Figure 3-2C, second and third columns). Overall, including ESC-specific genes (Group 1), we classified 1,761 ESC genes into fifteen gene groups. In Group 1, these 44 ESC-specific genes show more than 20-fold expression over fifteen tissue/cell types. On the other hand, in Group 15 Non-Dynamic class, the genes are broadly expressed with small variance of 0.2 – 5-fold in all cell types.

By examining OCT4 and SOX2 ChIP-seq, we identified 120 composite binding sites within 15 kbp of these 1,751 ESC genes. In Group 1, 19 out of 44 ESC-specific genes (43.18%) contain a strong OCT4/SOX2 composite site (Figure 3-2C, blue shade). In contrast, only 8 out of 268 Group 15 Non-Dynamic genes (2.99%) present an OCT4/SOX2 binding within 15 kbp (Figure 3-2C, red shade). A 14-fold enrichment of strong OCT4/SOX2 in the vicinity of ESC-specific genes versus Non-Dynamic is consistent and more significant than the finding in the mouse. One major difference between human and mouse is that OCT4/SOX2 binding does not exhibit enriched binding in the dynamic genes (Group 2 to 14). There are two possible explanations for the difference. First, human ESCs display a primed state of pluripotency (Brons et al., 2007; Warrier et al., 2017). OCT4/SOX2 ChIP-seq in human might only capture the OCT4/SOX2 binding at the ESC-specific genes in the primed state rather than the binding in a more naïve state. The other possibility could be that human OCT4/SOX2 has a more discriminating mechanism to regulate a refined population of pluripotency genes.

Strikingly, when we intersected the lists of Oct4/Sox2-regulated ESC-specific genes, there were only four genes overlapped between human and mouse (Figure 3-2D and E). *POU5F1*, *SOX2*, *NANOG*, and *SALL4* are critical components in the previously identified pluripotency gene regulatory network (Kim et al., 2008). The Oct4/Sox2 sites at these four genes exhibit high PhasCon conservation score in both human and mouse, indicating that the bindings are functionally significant to the transcriptional activation in pluripotency. However, besides these four genes, most of the binding sites are less conserved in mouse (Figure 3-2E). The comparative genomic analysis of Oct4/Sox2 composite sites highlight that the distinct mechanisms for pluripotency gene transcription could vary between species.

PhasCon Conservation Analysis of Oct4/Sox2 Peaks and Composite Motif in Mouse

It is surprising that only a handful number of ESC-specific genes and the Oct4/Sox2 sites are conserved between human and mouse. The poor conservation of the Oct4/Sox2 sites led us to evaluate the conservation of the bindings at different gene groups. Continue with the analysis from Chapter 2, we compared the conservation score of the Oct4/Sox2 composite binding at ESC-specific, Dynamic, Non-Dynamic, and Silent genes. To better characterize the conservation feature, we separately evaluated the score at the ± 100 bp regions of Oct4/Sox2 peaks or the identified composite motif sequences. Interestingly, the 200 bp window at Oct4/Sox2 peaks showed a higher conservation score distribution in the Silent genes while there were no differences between ESC-specific, Dynamic, and Non-Dynamic genes (Figure 3-3A). The distribution suggests that Oct4/Sox2 targets a

more conserved enhancer in the vicinity of Silent genes. Since we focus on a 200 bp window centered on Oct4/Sox2 peaks, it is possible that other sequences dominate these 200 bp regions and biased the results. To avoid the bias, we retrieved the 12-bp Oct4/Sox2 composite motif sequences contributing to the MEME de novo motif for the conservation analysis (Figure 3-3B). Likewise, the Oct4/Sox2 sites are more conserved in the Silent genes than other groups, with 23 motifs reaching conservation score = 1 (open red circles).

The conservation analysis suggests that Oct4/Sox2-mediated transcriptional repression might be a universal mechanism broadly implemented in different species. While there are 23 Oct4/Sox2 sites at Silent genes showing conservation score = 1, we then examined the chromatin status and histone modification at these sites (Figure 3-3C). Among these 23 Oct4/Sox2 bindings, 16 of them bind to the chromatin with open configuration. Only four sites contain a shifted ATAC signal near the proximity, and three Oct4/Sox2 peaks reside in the close chromatin. For example, at the *Hobx1* enhancer, Oct4/Sox2 recognizes a conserved region despite the chromatin data shows close chromatin configuration (lacks ATAC signal) and the absence of H3k27Ac enrichment (Figure 3-3D). The ATAC-seq data reiterate that Oct4/Sox2 do not preferentially recognize different chromatin configurations for gene-specific functions. Interesting, only 6 out of 23 Oct4/Sox2 sites contain H3K27Ac deposition (Figure 3-3C, last column). For example, at the *Zic5* enhancer, Oct4/Sox2 site is enriched with H3K27Ac and ATAC signal (Figure 3-3E). Majority of these conserved Oct4/Sox2 bindings at Silent genes lack

the active enhancer marking H3K27Ac, indicating the bindings are not necessarily associated with active transcription.

The Co-binding of Other Transcription Factors Does Not Distinguish Oct4/Sox2 Sites in Different Gene Groups

Master transcription factors of pluripotency often form unusual enhancer domains with multiple co-binding regulators (Whyte et al., 2013). To compare the genomic features of Oct4/Sox2 peaks at different classes, we examined the co-binding of Nanog, c-Myc, p300, Brg1, Esrrb, and Hdac1 at the Oct4/Sox2 sits (Figure 3-4A). Since weaker peaks are less reproducible and are more likely to represent technical artifacts, we set a cutoff to only include the peaks with top 20% peak score in each TF ChIP-seq. We started with examining the co-binding of Nanog at the Oct4/Sox2 sites. Nanog is one of the ESC-specific genes and a well-studied target of Oct4/Sox2 (Jauch et al., 2008). As one of the main proteins in the transcriptional network for pluripotency of embryonic stem cells, Nanog cooperatively mediates gene activation with Oct4 and Sox2 in many pluripotency genes (Kim et al., 2008). However, when we examined the co-binding of Nanog at the Oct4/Sox2 composite sites, more than 80% of the Oct4/Sox2 sites present a strong Nanog peak with equivalent distribution in all gene groups (Figure 3-4B, top left). The high percent of co-binding might indicate that Nanog is functionally relevant to Oct4/Sox2 regulation does not contribute to the selective functions of Oct4/Sox2 in different gene groups.

As one of the Yamanaka factors, c-Myc is sufficient to reprogram differentiated cells to pluripotent state when combined with Oct4, Sox2, and Klf4 (Takahashi and Yamanaka, 2006). Interestingly, the number of the c-Myc co-binding event is strikingly low in across four gene groups (Figure 3-4B, top middle). Only 37 strong Oct4/Sox2 composite sites present a c-Myc co-binding (Figure 3-4A). The low co-binding rate is possibly attributed to distinct genomic distributions of c-Myc and Oct4/Sox2. During the reprogramming, c-Myc tends to target promoter regions, while the majority of Oct4/Sox2 locates at the enhancers (Chronis et al., 2017). The distinct genomic distribution might indicate that c-Myc regulates pluripotency gene transcription through an Oct4/Sox2-indepdent mechanism.

Previously in Chapter 2, we discovered the enrichment of H3K27Ac deposition in ESC-specific and Dynamic classes. We also demonstrated the role of Oct4/Sox2 composite binding in optimizing H3K27Ac level for gene reactivation after the secondary reprogramming. At the enhancer regions, CBP/P300 bromodomain inhibition reduces the levels of H3K27Ac histone modification, suggesting P300 is required for the histone acetylation and transcription of enhancer-proximal genes (Raisner et al., 2018). Notably, more than 80% of ESC-specific (31 out of 37) and Dynamic genes (43 out of 51) present a P300 peak at the Oct4/Sox2 sites, while only half of the Non-Dynamic (52 out of 103) and Silent genes (368 out of 711) show co-binding (Figure 3-4A and B, top right). The high percentage of P300 co-binding at ESC-specific and Dynamic genes correlate with the enriched H3K27Ac deposition and the functional significance of the Oct4/Sox2 in pluripotency gene activation.

Next, we examined the co-binding of Brg1 at the Oct4/Sox2 sites in four groups. Brg1 is a chromatin remodeler supporting gene regulatory functions of Oct4 in mouse ESCs (King and Klose, 2017a). The previous study demonstrated that Oct4, as a pioneer factor, required the cooperation with Brg1 to optimize the chromatin accessibility for the binding of additional transcription factors. However, we only observed a moderate difference in Brg1 co-binding rate between gene groups (Figure 3-4A). In ESC-specific and Dynamic genes, about 80% of the Oct4/Sox2 sites contained a Brg1 peak (specific: 31 out of 37; Dynamic: 41 out of 51) (Figure 3-4B, bottom left). The Brg1 co-binding rate slightly decreased to 66% in Non-Dynamic and 61.6% in Silent genes. In Chapter 2, we found that the ATAC signal was equivalently distributed in four gene groups, suggesting that Oct4/Sox2 does not preferentially bind to open or close chromatin configuration. The Brg1 co-binding rate also indicates that this Oct4-associated chromatin remodeler might play a ubiquitous function to mediate chromatin accessibility in any gene groups.

Esrrb is another crucial component of pluripotency gene networks and activates Oct4 transcription to sustain self-renewal and pluripotency (van den Berg et al., 2010; Zhang et al., 2008). However, Esrrb co-binding rate only revealed a moderate difference between gene groups (Figure 3-4A). In ESC-specific and Dynamic genes, about 80% of the Oct4/Sox2 sites contained an Esrrb peak (specific: 33 out of 37; Dynamic: 42 out of 51) (Figure 3-4B, bottom middle). The Esrrb co-binding rate slightly decreased to 66% in Non-Dynamic and 61.6% in Silent genes, suggesting Esrrb co-binding did not distinguish the selective functions of the Oct4/Sox2 sites in different gene groups.

Last, another chromatin remodeler, Hdac1, also plays critical functions in ESCs pluripotency and differentiation. Hdac1 is a histone deacetylase which removes the acetyl groups on the histone and functions as a transcriptional repressor during embryonic development (Kidder and Palmer, 2012). Interestingly, we still found a substantial number of Oct4/Sox2 bindings at ESC-specific and Dynamic genes existing a Hdac1 peak (Figure 3-4A). The Hdac1 co-binding rates at these two gene groups (67.5% and 72.5%) are even higher than Non-Dynamic (51.4%) and Silent genes (49%) (Figure 3-4B, bottom right). It is obscure of such high Hdac1 co-binding rate at the vicinity of genes that are ESC-specific. Additionally, considered the low levels of H3K27Ac deposition in Non-Dynamic and Silent genes, Hdac1 does not correlate with the chromatin features in either group.

Except for P300, most of the co-factors do not distinguish the Oct4/Sox2 sites with selective functions across four groups. These co-factors may bind similarly, but the ChIP-seq peak score may be varied. To compare the binding strength of the co-factors at the Oct4/Sox2 sites, we included all peaks called by HOMER and set the sites with no peak calling as zero (Figure 3-4C). Consistent with the findings in co-binding rate, the ChIP-seq peak scores of P300 were higher in ESC-specific and Dynamic genes with two-fold enrichment of the average in Non-Dynamic or Silent genes. However, the peak score distributions of Nanog, c-Myc, Brg1, Esrrb, or Hdac1 did not exhibit significant differences between gene groups. In summary, we could not find the co-factor bindings, in the context of co-binding rate or peak strength, able to distinguish the selective functions of Oct4/Sox2 sites in different gene groups.

Initial Examination of Oct4/Sox2-regulated Enhancer Properties

The priming epigenetic conditions and the histone modification confer different activities of the DNA regulatory elements (Ernst and Kellis, 2010). In human ESCs, H3K27Ac is greatly enriched at the proximal enhancer of active genes in ESCs (Rada-Iglesias et al., 2011). To assess the whether the histone modification and chromatin accessibility varied between gene classes, we measured the RPKM level of H3K27Ac ChIP-seq (GSE 56138), H3K4me3 ChIP-seq (GSE 62380), and ATAC-seq (GSE 52397) at the 1,035 Oct4/Sox2 composite sites (Buecker et al., 2014; Ji et al., 2015; Di Stefano et al., 2016). Additionally, based on the epigenetic landscapes, co-factors binding, mediator occupancy, and the size of the enhancer regions, previous studies characterized the enhancers into super-enhancer and typical-enhancer (Whyte et al., 2013). The super-enhancer contains clusters of enhancers with high densities of transcription factor binding and extensive occupancy of the mediator as well as active histone marks such as H3K27Ac. One major functional significance of these super-enhancers is that they display superior ability to activate transcription than the typical-enhancers. To understand the enhancer properties among all 1,035 Oct4Sox2 sites, we performed a systematic analysis to characterize the Oct4/Sox2-regulated enhancers with the enrichment of H3K27Ac/H3K4me3, chromatin accessibility (ATAC-seq signal), and the types of enhancer (Figure 3-5).

First, we focused on the 37 ESC-specific genes (Figure 3-5A). Among the 37 ESC-specific genes, the majority of Oct4/Sox2 composite sites present the enrichment of H3K27Ac/H3K4me3 marking at an open chromatin configuration. The high levels of active histone marks deposition correlate with the transcription features of these ESC-specific

genes. Next, we annotated the composite Oct4/Sox2 sites with the super-enhancer or typical-enhancer previously defined by the Young lab (Whyte et al., 2013). In the 37 Oct4/Sox2 bindings at ESC-specific genes, 15 of them are super-enhancer, 18 of them are typical-enhancer, and 4 sites are unclassified regions (Figure 3-5E). As super-enhancers confer higher activity in gene transcription, 40.54% of Oct4/Sox2 binding at the super-enhancers support our finding that these Oct4/Sox2 sites are essential for gene activation in pluripotency.

The enhancer properties of the Oct4/Sox2 bindings at Dynamic genes display similar features as ESC-specific genes (Figure 3-5B). Most Oct4/Sox2-regulated enhancers are enriched with H3K27Ac/H3K4me3 modification and present an ATAC peak co-localization. The annotation of enhancer types revealed 18 super-enhancer, 29 typical enhancer, and 4 unclassified regions (Figure 3-5E). Importantly, these four unclassified Oct4/Sox2 binding sites often lack H3K27Ac/H3K4me3 marking and ATAC signal. Likewise, 35.29% of Oct4/Sox2 binding at the super-enhancers support our finding that these Oct4/Sox2 sites are essential for gene activation at the pluripotent state.

Shifting our attention to Non-Dynamic and Silent genes, we first examined their H3K27Ac marking at the Oct4/Sox2 sites (Figure 3-5C and D). Unlike ESC-specific or Dynamic genes, H3K27Ac deposition significantly reduced in the vicinity of Non-Dynamic and Silent gene groups. Previously in our CRISPR experiments of mutating Oct4/Sox2 composite motif sequences, we discovered these composite Oct4/Sox2 sites were non-functional to the transcription in Non-Dynamic genes or mediated transcriptional

repression in the Silent genes. The absence of H3K27Ac deposition also indicates that the bindings do not necessarily activate the gene transcription in these two groups. A substantial number of Oct4/Sox2 binding at Non-Dynamic and Silent genes still contain the enrichment ATAC signal, suggesting that Oct4/Sox2 do not preferentially target specific chromatin configurations or H3K4me3 marked regions across the four gene groups. Notably, the annotation of enhancer types demonstrated significantly fewer numbers of super-enhancers in Non-Dynamic and Silent genes. Only 5 Non-Dynamic (4.85%) and 47 Silent genes (6.61%) contain an Oct4/Sox2 site at the super-enhancer (Figure 3-5E). More than 80% of Oct4/Sox2 bindings at Non-Dynamic (87 out of 103) and Silent genes (575 out of 711) are typical-enhancers. The dramatic reduction of Oct4/Sox2-targeted super-enhancers in these two gene groups denotes the different functions of Oct4/Sox2 in pluripotency gene regulation. Interestingly, most of these Oct4/Sox2-targeted super-enhancers in Non-Dynamic and Silent genes show a high level of H3K27Ac deposition. In our previous H3K27Ac ChIP-seq analysis in Chapter 2, we did observe a few numbers of Non-Dynamic and Silent genes displaying enriched H3K27Ac at the Oct4/Sox2 sites. Mutating the motif sequences in the selected enhancers (Non-Dynamic: *Ift52*, *Dido1*; Silent: *Lax1*, *Uba7*) did not alter the transcription of the annotated genes. Instead, some of these Oct4/Sox2 bindings are activating other genes dozens kilo base-pair away, possibly through the long-range chromatin interaction.

Quantitative Examination of Oct4/Sox2-regulated Enhancer Properties

To interpret the enrichment of histone mark deposition and ATAC signal, we set a cutoff threshold of 30 RPKM for H3K27Ac ChIP-seq, 8 RPKM for H3K4me3 ChIP-seq, and peak

score = 5 for ATAC-seq. By their enhancer types, we separated the Oct4/Sox2-regulated enhancers into super-enhancer, typical-enhancer, and unclassified region. We compared the enrichment of H3K27Ac, H3K4me3, and ATAC signal using the cutoff thresholds under different types of enhancer (Figure 3-6). In ESC-specific genes, both the Oct4/Sox2 bindings at super-enhancer and typical-enhancer exhibit great enrichment of H3K27Ac deposition (Figure 3-6A). Only two super-enhancers and six typical-enhancers do not meet the RPKM 30 threshold for H3K27Ac. Although the H3K27Ac levels are comparable between super- and typical-enhancers, nearly 70% of the Oct4/Sox2 bindings at ESC-specific genes (26 out of 37) present enriched H3K27Ac marking. On the other hand, neither the Oct4/Sox2 sites at super-enhancers nor typical-enhancers skew toward the enrichment of H3K4me3 and ATAC signal. Only the super-enhancers contain a 2-fold enrichment of H3K4me3 > 8 RPKM.

Interestingly, in Dynamic gene group, while H3K27Ac levels remain enriched in both super-enhancers and typical-enhancers, enriched H3K4me3 marking and ATAC signal also correlates with the Oct4/Sox2 binding at the enhancers (Figure 3-6B). For the bindings at super-enhancers, 32.69% of them contain H3K4me3 > 8 RPKM, and 34.62% of the regions present a strong ATAC peak. Only one super-enhancer is marked with weaker H3K4me3. In the typical-enhancers, the numbers of weak H3K4me3 or ATAC peak increase slightly (H3K4me3 < 8 RPKM: 11.54%; ATAC peak score < 5: 9.62%), but the majority of typical-enhancers remain enriched in strong H3K4me3 (44.23%) and ATAC peak (46.15%). The enrichments of strong H3K4me3 and ATAC peak in the vicinity of Dynamic genes indicate that the regions are more accessible and tend to present more

active histone marks deposition. Notably, the numbers of ATAC peak score < 5 in much less than ESC-specific genes, suggesting that the Oct4/Sox2 sites at Dynamic genes might be more accessible when entering the pluripotent state.

Shifting our attention to Non-Dynamic and Silent genes, the most dramatic change is the ratio of super-enhancer and typical-enhancer. More than 80% of the Oct4/Sox2 sites locate at the typical-enhancer regions, and 56.31% of Non-Dynamic genes and 62.45% of Silent genes do not show a strong H3K27Ac marking (Figure 3-6C and D). The percentage is at least two-fold more than the weak H3K27Ac in ESC-specific or Dynamic genes. Moreover, 54.37% of Non-Dynamic genes and 71.03% of Silent genes lack the H3K4me3 deposition. The deficiency of active histone marks associated with active gene transcription corroborates the non-functional binding of Oct4/Sox2 at Non-Dynamic genes and the repressive role at Silent genes. Besides histone modification, the typical-enhancers in Non-Dynamic or Silent genes barely coincide with a strong ATAC peak. More than half of the regions display a feature of closed chromatin configuration (ATAC peak score < 5).

Analysis of CpG Content at the Oct4/Sox2-regulated Enhancers

In the promoter regions, CpG content often dictates the nucleosome stability. Since the CpG-island promoters are too rigid to form stable nucleosomes, the regions tend to be nucleosome-free and exhibit higher accessibility for transcription factor bindings (Ramirez-Carrozzi et al., 2009; Tazi and Bird, 1990). The accessibility of the high CpG content regions do not require further chromatin remodeling; these regulatory elements

often associate with ubiquitously expressed genes. The prevalence of high CpG content at the Oct4/Sox2 sites might indicate the different functions of these Oct4/Sox2 bindings. To examine the CpG content at the Oct4/Sox2-regulated enhancers, we assessed their chromatin states by comparing the CpG content and histone marks deposition or ATAC peaks. First, we calculated the obs/exp CpG ratio at the 200 bp windows of Oct4/Sox2 composite bindings and compared across the four gene groups (Figure 3-7A and B). The distribution of obs/exp CpG ratio revealed that the majority of Oct4/Sox2 sites are targeting the regions with obs/exp CpG ratio less than 0.6. Only a small fraction of Oct4/Sox2 peaks contain a CpG-island with obs/exp CpG ratio greater than 0.6 in Non-Dynamic and Silent genes. The average of CpG content between gene groups is comparable, suggesting that the CpG content might not contribute to the gene-specific mechanisms regulated by Oct4/Sox2 binding at different gene groups. The consistently low CpG content distributions at the Oct4/Sox2-regulated enhancers reiterate the role of pioneer factors in mediating enhancer accessibility to regulate cell-specific genes (Buecker et al., 2014; Soufi et al., 2014; Zaret and Carroll, 2011).

Next, we compare the CpG content with the level of histone modification and chromatin accessibility. Because more than 60% of Oct4/Sox2 target the regions with obs/exp CpG ratio less than 0.2, we separated the property of CpG content by the cutoff threshold of 0.2 obs/exp CpG ratio. By the CpG content threshold, we further separated the gene groups with the previously defined H3K27Ac > 30 RPKM, H3K4me3 > 8 RPKM, and ATAC peak score > 5. H3K27Ac deposition did not lean toward higher obs/exp CpG ratio in any gene groups (Figure 3-7C). Instead, the levels of H3K27Ac remain enriched in

ESC-specific and Dynamic genes regardless of the CpG content. In contrast, H3K27Ac barely deposit at either high CpG (> 0.2) or low CpG (0 – 0.2) Oct4/Sox2-regulated enhancers in Non-Dynamic and Silent genes. Similarly, H3K4me3 markings did not show different prevalence between high CpG or low CpG Oct4/Sox2 sites in four gene groups (Figure 3-7D). Oct4/Sox2-regulated enhancers are enriched with H3K4me3 deposition regardless of the CpG content. Conversely, both high and low CpG Oct4/Sox2 sites in Non-Dynamic and Silent genes are deficient with H3K4me3 marking.

Last, the analysis of ATAC peaks with distinct CpG content groups further demonstrated that the CpG contents do not necessarily contribute to the chromatin accessibility at the composite Oct4/Sox2 sites (Figure 3-7E). The only minor difference between high CpG and low CpG Oct4/Sox2 sites is Non-Dynamic and Silent genes. The low chromatin accessibility correlates with low CpG content (Non-Dynamic: 40.78%; Silent: 35.44%). However, the differences are not significant.

Characterization of Enhancer Properties and Chromatin States of the Representative Genes in Different Gene Groups

Previously in Chapter 2, we utilized CRISPR-HDR to independently disrupt Oct4/Sox2 composite binding and determined the gene-specific functions in regulating transcription of the Oct4/Sox2 sites. These strong Oct4/Sox2 composite binding sites function differently in transcription and histone modification in ESC-specific, Dynamic, Non-Dynamic, and Silent gene groups. To gain more genomic insights into these functional

and non-functional bindings, we evaluated co-factor bindings, enhancer properties, and chromatin states of the Oct4/Sox2 sites we mutated (Figure 3-8).

Pla2g1b is an ESC-specific gene in which Oct4/Sox2 composite binding is essential for its transcriptional activation during secondary reprogramming. In the vicinity of the Oct4/Sox2 site, only c-Myc, Brg1, and Esrrb are co-bound. Active enhancer histone mark H3K27Ac almost meets the 30 RPKM cutoff (27.4 RPKM) but the region is enriched with H3K4me3 deposition which correlates with active transcription in ESCs. Based on the previous characterization of enhancer types, the Oct4/Sox2 binds to a super-enhancer region with the CpG content of 0.47. The *Pla2g1b*-associated Oct4/Sox2 binding is likely to function more independently, without the cooperation of other co-factors. However, the active histone marks deposition and the super-enhancer characteristic support the finding of an essential role in transcriptional activation in ESCs.

Zfp57 and *Epb4.1l5* are highly expressed in ESC and NEUR. As Dynamic genes, they exhibit a 20-fold dynamic range of expression in BMDM and DP. The functional studies using CRISPR-HDR mutation confirmed the role of Oct4/Sox2 as transcriptional activators of Dynamic genes. At both enhancers, Nanog binds strongly within ± 100 bp from the center position of the Oct4/Sox2 sites. Additionally, P300, Brg1, Esrrb, and Hda1 bind near to Oct4/Sox2. The peak score is strong in *Zfp57* and weaker in Epb4.1l5. However, neither the Oct4/Sox2 binding at *Zfp57* nor *Epb4.1l5* contains c-Myc co-binding. The Oct4/Sox2 site at *Zfp57* is highly enriched with H3K27Ac deposition (75.7 RPKM), while the H3K27Ac level at the *Epb4.1l5* enhancer is moderate (23.5 RPKM). However,

we did not observe enriched H3K4me3 marking or strong ATAC signal in either Dynamic genes. These two Oct4/Sox2 sites represent the different types of the enhancer. The Oct4/Sox2-regulated *Zfp57* enhancer is characterized as a super-enhancer, and the *Epb4.1l5* enhancer is a typical enhancer. The discrepant enhancer properties could be attributed to the strength of the co-factor bindings and the level of H3K27Ac deposition. Although the Oct4/Sox2-regulated enhancers represent different types of enhancer, they are both essential for the transcription of *Zfp57* or *Epb4.1l5* at the pluripotent state. The results indicate that the functional significance or necessity of transcription factor binding should not merely rely on whether the region is a super-enhancer or not.

Shifting our attention to the Non-Dynamic genes, we included two subgroups of Non-Dynamic genes with a distinct level of active histone marks deposition. The genes in the first subgroup, *Pds5a, and Hnrnpr,* are deficient with H3K27Ac and H3K4me3 marking. Both Oct4/Sox2 sites contain Nanog, P300, Brg1, Esrrb, and Hdac1 peaks in the 200 bp regions but their binding is ubiquitously weaker at the *Hnrnpr* enhancer. The weaker co-binding TF peak scores are correlated with the weaker Oct4/Sox2 binding strength at *Hnrnpr,* indicating that these transcription factors might also increase the binding strength of other factors. Additionally, both Oct4/Sox2 composite sites do not associate with strong ATAC peaks and are located at typical-enhancers. In the second subgroups of Non-Dynamic genes (*Dido1* and *Ift52*), Oct4/Sox2 binds to the enhancers enriched with either H3K27Ac or H3K4me3 deposition. Likewise, these two enhancers present co-bindings of Nanog, P300, Brg1, Esrrb, and Hdac1 peaks in the 200 bp regions of the Oct4/Sox2 peaks. In the *Ift52* enhancer, c-Myc also co-binds with the Oct4/Sox2 peaks. Both

Oct4/Sox2 sites co-localize with strong ATAC peaks, indicating that the bindings occur at accessible chromatin regions. Notably, the Oct4/Sox2-bound *Ift52* enhancer is characterized as a super-enhancer. This super-enhancer characteristic supports the function we discovered for this Oct4/Sox2 site at the *Ift52* enhancer. One of the most exciting findings by CRISPR mutation in Non-Dynamic genes is that these Oct4/Sox2 sites are not functionally-relevant to the transcription of annotated Non-Dynamic genes. The only exception is the Oct4/Sox2 binding at the *Ift52* enhancer, which contributes to the transcription of a downstream Dynamic gene *Mybl2.* Thus, the active transcription of *Mybl2* in ESCs explains why a Non-Dynamic gene associated Oct4/Sox2 binding is locating at the super-enhancer.

The last group of genes are silent in ESCs. Likewise, we separated the Oct4/Sox2 binding sites into two subgroups based on the presence of H3K27Ac deposition. The first group is deficient with H3K27Ac marking at *Oxgr1* and *Gnrhr*. The second group is enriched with H3K27Ac at the active enhancers nearby *Uba7* and Lax1. In the CRISPR-HDR mutation experiments, we demonstrated that these Oct4/Sox2 sites mediated the transcriptional repression of these Silent genes in ESCs. At the *Oxgr1* enhancer, Oct4 and Sox2 are associated with weak Nanog, Brg1, Esrrb, and Hdac1 co-binding at the typical enhancer. However, we did not observe particular enhancer properties that distinguished the repressive role of the Oct4/Sox2 binding. On the other hand, at the *Gnrhr* enhancer, only Esrrb shows very weak co-binding with Oct4/Sox2 and the region is unclassified. Notably, the CpG content ratio of the *Gnrhr* enhancer is 0.72, which is the only enhancer exhibiting the feature of CpG island (ratio > 0.6). *Uba7* and *Lax1* are also

the silent genes, but the Oct4/Sox2-regulated enhancers display enriched H3K27Ac deposition. The Oct4/Sox2 only co-binds with weak peak scores of Brg1, Esrrb, and Hdac1 at the *Uba7* enhancer, while the *Lax1* enhancer also contains a cooperative binding with Nanog. H3K27Ac enrichment and strong ATAC peaks also denote both Oct4/Sox2-bound enhancers.

In summary, the examination of the enhancer properties including transcription factors co-binding, histone modifications, ATAC signal, enhancer types, and CpG content, help us understand genomic characteristics of these Oct4/Sox2 bindings. However, we did not observe particular features that participate in the selective Oct4/Sox2 transcriptional regulation in pluripotency. The results also highlight the limitation of the data analysis to characterize the different functions of Oct4/Sox2 sites in distinct gene groups. The mechanistic studies of the Oct4/Sox2 sites by CRISPR-HDR mutation are therefore valuable to evaluate the functional significance in pluripotency gene transcription.

Examination of Cooperative Transcription Factor Binding at the Oct4/Sox2-bound Enhancers During Reprogramming

In the previous study in reprogramming, Oct4, Sox2, Klf4, and c-Myc (OSKM) cooperate with many stage-specific transcription factors to mediate the activation of pluripotency-enhancer and the inactivation of somatic-enhancer (Chronis et al., 2017). This study generated a dataset of OSKM and stage-specific transcription factors ChIP-seq at the stages of MEFs, 48hrs OSKM induction, and pre-iPSCs during reprogramming. Although our study did not focus on the functional changes of Oct4/Sox2 sites during

reprogramming, the mechanistic insights of the ESC composite site could emerge with the evaluation of the TF cooperative bindings at different stages of reprogramming. To examine Oct4/Sox2 with stage-specific transcription factor binding at the 1,035 ESC Oct4/Sox2 sites, we processed and analyzed ChIP-seq datasets of Oct4, Sox2, Brg1, Cebpb, Cebpa, Fra1, Runx1, Hdac1, and P300 in mouse embryonic fibroblasts (MEF), 48hr OSKM induction, pre-iPSC lines (GEO: GSE90895) (Figure 3-9A and B and C and D). To identify the cooperative binding, we intersected the called peak regions of Oct4/Sox2 and the stage-specific transcription factors within the 1.5 kbp window of the ESC Oct4/Sox2 sites. The heatmap illustrated the cooperative bindings of Oct4, Sox2, Brg1, Cebpb, Cebpa, Fra1, Runx1, Hdac1, and P300 at the ESC Oct4/Sox2 sites during different stages of reprogramming.

First, we assessed the loci of Oct4/Sox2 peaks in MEF, 48hrs OSKM induction, and pre-iPSCs in ESC-specific, Dynamic, Non-Dynamic, and Silent genes. To visualize the binding strength in the heatmap, we ranked the peak scores in ascending order and color-coded based on the percentile of peak score. Interestingly, we found that only half of the Oct4/Sox2-regulated enhancers, regardless of the gene groups, did not exhibit an Oct4/Sox2 binding at the early stage of reprogramming (Figure 3-10A). Some gene enhancers lacking the Oct4/Sox2 binding in 48hrs OSKM induction or pre-iPSCs might exhibit another Oct4/Sox2 peak at a different enhancer designed explicitly for early reprogramming. This result also complies with the previous understanding that Oct4/Sox2 could occupy a different enhancer required for remodeling chromatin configuration or nuclear organization, which do not necessarily contribute to transcriptional regulation in

the pluripotent state. At the later stage of reprogramming, Oct4/Sox2 relocates to another functional enhancer to mediate proper transcription of the pluripotency genes. However, neither the prevalence of Oct4/Sox2 bindings nor the peak scores display a significant difference between ESC-specific, Dynamic, Non-Dynamic, and Silent genes (Figure 3-10B). The similar level of the early Oct4/Sox2 bindings at the ESC Oct4/Sox2 sites also suggests that these early binding sites do not contribute to the selective transcriptional regulation in the pluripotency genes.

Next, we evaluated the cooperative bindings of Cebpa, Cebpb, Fra1, and Runx1 at the Oct4/Sox2 sites in MEF and 48hrs OSKM induction (Figure 3-9). All these transcription factors are highly expressed in MEFs. Centering on the ESC Oct4/Sox2 sites, we barely found co-occupancy of these MEF transcription factors in MEFs (Figure 3-10A and B). In 48hrs OSKM induction, the co-occupancy of Cebpa, Fra, and Runx1 increase slightly, but the prevalence is not significant and does not enrich in specific gene groups. The peak scores of Cebpb, Fra1, and Runx1 are slightly higher in Dynamic or Non-Dynamic genes. However, the increase peak score might not be real as the average of peak score are only calculated by a few numbers of the binding. Thus, the deficiency of the MEF transcription factors occupied at the ESC Oct4/Sox2 sites indicates that these Oct4/Sox2 bindings might participate in neither the removal of somatic-lineage transcription factors nor the inactivation of somatic-enhancer.

Last, we examined the cooperative bindings of Brg1, P300, and Hdac1 at the Oct4/Sox2 sites in MEF, 48hrs OSKM induction, and pre-iPSCs (Figure 3-9). The prevalence of co-

occupancy and peak scores do not distinguish the composite Oct4/Sox2 bindings in different gene groups. Instead, the bindings of Brg1, P300, and Hdac1 display a stage-specific induction during the reprogramming (Figure 3-10A). The prevalence of Brg1 and P300 co-occupancy is both induced at the Oct4/Sox2-regulated enhancers in the pre-iPSCs despite the lower peak scores. The binding sites are stronger in MEF and 48hrs OSKM induction. However, the increase peak score might not be real as the average of peak score are only calculated by a few numbers of the binding. Hdac1 co-occupancy also increase moderately in pre-iPSCs, but the changes are not as significant as Brg1 and P300.

DISCUSSION

By carefully documenting the genetic and genomic characteristics at the Oct4/Sox2 composite sites, we provided a quantitative view of the co-factor binding, histone modification, chromatin accessibility, and DNA context in pluripotency and reprogramming. We also compared the discrepancy and conservation of pluripotency gene regulation between human and mouse. This study scrutinized the potential mechanisms that distinguished the selective Oct4/Sox2 transcriptional regulation in pluripotency. Given the complexity of the pluripotency gene network, our study also demonstrated that our gene-centric approach has the potential to uncover mechanistic details of transcriptional regulation.

In this study, we extended our gene classification method to calculate the dynamic range of expression between ESCs and somatic tissue/cell types. We included twelve polyA

mRNA-seq data sets from different somatic tissue/cell types to cover gene with various degree of the dynamic range of expression. The correlation between strong binding, Oct4/Sox2 occupancy, and gene classes support the discovery that Oct4/Sox2 composite binding is a critical activator of large numbers of genes exhibiting high expression in ESC and deficient expression in at least some somatic cell types. We also compared the trends of Oct4/Sox2 binding enrichment between mouse and human. Surprisingly, although the Oct4/Sox2 binding rate are both enriched in the ESC-specific genes, the Oct4/Sox2 targets are poorly conserved between human and mouse. Among the ESC-specific genes, only the bindings at the enhancer/promoter of *Nanog*, *Sall4*, *Sox2*, and *Pou5f1* are conserved. This poor conservation is likely due to the discrepant pluripotent states between two species. Human ESCs display the characteristics of the primed state of pluripotency, while the mouse ESCs typically remain in the naïve pluripotent state (Nichols and Smith, 2009, 2010; Ying et al., 2008). Comparative transcriptomic analysis between naïve and primed human/mouse ESCs also suggested that the substantial differences in gene expression may lead to the false interpretation of pluripotency gene regulation (Ernst et al., 2015). Still, the poor conservation highlights the awareness of the discrepancy of pluripotency gene regulation between species.

Besides the Oct4/Sox2 bindings, we also examined the co-binding frequency and binding strength of other pluripotency factors or histone modifier. Consistent with the previous studies, these strong Oct4/Sox2 sites displays many cooperative binding with Nanog, Esrrb, and Brg1 (van den Berg et al., 2008; Descalzo et al., 2012; King and Klose, 2017b; Singhal et al., 2014). The interaction of Oct4 and these transcription factors typically

secure the ESC identity, suggesting a potential functional relevance in pluripotency. However, we did not find significant differences in their bindings biased toward particular gene groups. While the underlying mechanism to select Oct4/Sox2 functions in different gene groups remains unknown, these co-binding factors are less likely to distinguish the functions. Moreover, we also observed a fraction of Oct4/Sox2 composite sites associated with histone modifier P300 or Hdac1. Both P300 and Hdac1 are two-fold enriched in the Oct4/Sox2 bindings at ESC-specific and Dynamic genes. Considered the opposite functions of P300 and Hdac1 in histone acetylation, the equivalent enrichment of P300 and Hdac1 in the ESC-specific and Dynamic genes is intriguing. As the enrichment of P300 correlates with H3K27Ac deposition in these two gene groups, the Hdac1 co-binding may play an unconventional function.

A previous study described the enhancer properties by distinguishing an unusual super-enhancer domain at most pluripotency gene regulated by master regulators (Whyte et al., 2013). These super-enhancers are featured with densely-occupied master regulators and Mediators and play prominent roles to define cell identity. Here, we found that both super-enhancer and typical-enhancer are present in all gene groups. The percentage of Oct4/Sox2 located at the super-enhancer is higher in ESC-specific and Dynamic genes, while most Oct4/Sox2 target the typical-enhancer in Silent genes. However, many highly transcribed genes in ESCs are not associated with a super-enhancer. The distribution of super-enhancer and typical enhancer reminds the fact that the enhancer types do not delineate much functional significance of the Oct4/Sox2 binding. An Oct4/Sox2 binding at any enhancer types can still be functionally required for the gene activation or

inactivation. The super-enhancer characteristics can be potential features to identify cell-type-specific enhancers but are not sufficient to denote the real functions of master regulator in pluripotency. Our discovery also support the view that enhancers and super-enhancers have an equivalent regulatory role in ESCs (Moorthy et al., 2017). Using CRISPR/Cas9-mediated deletion, they found that target gene expression reductions are variable and ranging from 12% to as much as 92%. They enhancer clusters within the super-enhancer region also function partially redundant mediate the transcription of target genes. This result highlights the importance to identify all functionally regulatory regions in particular cell types and accurately assign the enhancers to the precise targets.

We also quantitatively assessed the H3K27Ac/H3K4me3 deposition and CpG content at the Oct4/Sox2 composite binding sites and compared across gene groups. Our in-depth analysis of histone modification revealed enrichment of both active histone marks at the Oct4/Sox2 binding essential for transcriptional activation. Because of the quality of the H3K9me3 and H3K27me3 ChIP-seq data sets is not sufficient for a convincing quantitative analysis, we only measured the H3K27Ac/H3K4me3 deposition at the Oct4/Sox2 binding at the Silent genes. The low enrichment of H3K27Ac/H3K4me3 marking supports our discovery that Oct4/Sox2 acts as a repressor to inactivate these Silent genes. On the other hand, the CpG contents are comparable between gene groups. Although the high CpG genomic regions are too rigid to form stable nucleosomes hence the chromatin will be more permissive to transcription factors, the features are likely to be more influential at the promoters (Ramirez-Carrozzi et al., 2009). Since the majority of

Oct4/Sox2 binds at the enhancer, the CpG content may have less influence on the accessibility of Oct4 and Sox2.

Key findings early on suggested that OSK collaborated among themselves and interacted with stage-specific TFs to mediate pluripotency-enhancer selection and somatic-enhancer silencing during the reprogramming (Chronis et al., 2017). In Chapter 2, we reported the bivalent functions of Oct4/Sox2 composite binding as activator or repressor in distinct gene classes. Here we further examined if the pluripotency factors or the somatic TFs co-occupancy determined the selective functions of Oct4/Sox2 bindings. Our analysis provides a comprehensive view of the cooperative bindings at the Oct4/Sox2-regulated enhancer. However, we did not identify potential transcription factors that distinguish the selective Oct4/Sox2 transcriptional regulation among the 1,035 sites. The underlying mechanisms remain unclear. Nevertheless, this analysis serves as a first step trying to investigate how Oct4/Sox2 functions differently to regulate the gene transcription in pluripotency.

In summary, we integrated TF ChIP-seq data sets, histone marks ChIP-seq data sets, ATAC-seq, enhancer properties, and CpG content, to examine the genomic and genetic features at the Oct4/Sox2 binding sites in pluripotency and reprogramming. By comparing the dynamic range of expression between ESCs and several somatic tissue/cell types, we refined the gene classification to examine the Oct4/Sox2 occupancy between gene groups better. This gene-centric approach allowed us to quantitatively evaluate each genomic/genetic feature at the Oct4/Sox2 binding sites with different functional

significance. This analysis is just the first step to uncover the underlying mechanisms to select Oct4/Sox2 functions. We can also use the same strategy to explore other gene components in the pluripotency network to scrutinize the factors that account for the selective Oct4/Sox2 transcriptional regulation in pluripotency.

EXPERIMENTAL PROCEDURES

Mouse/Human ENCODE PolyA mRNA-seq Read Mapping and Processing

Published mRNA sequencing datasets were obtained from ENCODE project (http://www.encodeproject.org/) or mouse ENCODE Project (http://www.mouseencode.org/) and summarized in Chapter 2 Table 2 and chapter 3 Table 1. Reads were mapped to mouse NCBI37/mm9 reference genome or human GRCh37/hg19 reference genome by HISAT2 v2.1.0 (Kim et al., 2015). Aligned reads were restricted to uniquely mapping with up to two mismatches per read. RPKM values were calculated as previously described (Mortazavi et al., 2008) and based on the gene annotation of NCBI37/mm9 or human GRCh37/hg19 reference genome. mRNA RPKM was calculated by counting all mapped exonic reads and dividing by the length of the spliced product. SeqMonk's RNA-seq quantification pipeline was used to calculate RPKM value (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). All RPKMs represent an average from two biological replicates if they are available on ENCODE transcriptome datasets. Organ or cell-specific transcript profiles were retained only if the correlation scores between two replicates were higher than 0.9.

ChIP-seq Read Mapping and Processing

Published transcription factor ChIP-seq datasets were obtained from GEO (Buecker et al., 2014; Chronis et al., 2017; Ji et al., 2016; Zhou et al., 2016) and summarized in Table 2. MEFs, 48hrs OSKM induction MEFs, pre-iPSCs, and ESCs ChIP-seq were processed and analyzed equally with the same pipelines. Reads from ChIP-seq were mapped to mouse NCBI37/mm9 or human GRCh37/hg19 reference genome using Bowtie2 software (Langmead and Salzberg, 2012). Uniquely aligned reads were used for peak calling and gene annotation using HOMER (Benner et al., 2017). Peaks with false discovery rate (FDR) < 0.01 and enriched over input were called. Only reproducible peaks from replicates were retained for downstream analyses if biological replicate data were available. For Oct4 and Sox2 ChIP-seq, called peaks were annotated to nearest TSS of genes. Composite bindings of multiple factors were determined by the distance between two peaks. Only the distance of peak summits less than 100 bp were considered a composite binding.

ATAC-seq Read Mapping and Processing

Published histone mark ChIP-seq and ATAC-seq datasets were obtained from GEO (Becker et al., 2016; Chronis et al., 2017; Ji et al., 2015; Di Stefano et al., 2016) and summarized in Table 2. Reads from histone mark ChIP-seq and ATAC-seq were mapped to mouse NCBI37/mm9 reference genome using Bowtie2 software (Langmead and Salzberg, 2012). Reads were removed from the subsequent analysis if they were duplicated, mapped to mitochondrial genome, or aligned to unmapped contiguous

sequences. For ATAC-seq, peak calling was performed by MACS2 using parameter callpeak --nomodel -g mm --keepdup all -q .01 -- llocal 10000. Chromatin accessibility (ATAC sensitivity) was analyzed by calculating the RPKM values of 1.0 kbp flanking region centered on the Oct4/Sox2 peaks. Genomic loci with ATAC peak scores higher than 5 were considered open chromatin configuration. For H3K27Ac and H3K4me3 ChIP-seq, the enrichments were analyzed by calculating the RPKM values of 1.5 kbp flanking region centered on the Oct4/Sox2 peaks. Genomic loci with H3K27Ac higher than 30 RPKM or H3K4me3 higher than 8 RPKM were considered active histone marking.

Conservation Analysis

Conservation Score Analysis was performed using UCSC PhasCons placental mammal data (Siepel et al., 2005). Conservation score was quantified as the average PhastCons score over the 200 bp Oct4/Sox2 peak regions or the 12 bp Oct4/Sox2 composite motif. The 200 bp Oct4/Sox2 peak regions were determined by the ± 100 bp sequences from the center position of Oct4/Sox2 peak summits. The Oct4/Sox2 sites contributing to the MEME *de novo* Oct4/Sox2 composite motif (12 bp) were subjected to conservation analysis. Enhancer Oct4/Sox2 motifs in different gene groups were analyzed by MEME-ChIP to define a consensus sequence centered on ± 100 bp, ± 50 bp, or ± 25 bp from the center position of peak loci.

Transcription Factor Co-Binding Analysis

Published transcription factor ChIP-seq datasets were obtained from GEO (Buecker et al., 2014; Chronis et al., 2017) and summarized in Table 2. ChIP-seq data mapping and

processing were performed as previously described. Co-binding of transcription factors was determined by the distance between the peak summits. We generated sets of Oct4/Sox2 sites co-bound by other factors (Nanog, cMyc, p300, Brg1, Esrrb, and Hdac1) at ESCs, by extending peak summits called by HOMER by 100 bp in each direction and intersecting with 200 bp regions centered on Oct4/Sox2 peaks. Only the peaks reside with the ± 100 bp of Oct4/Sox2 center position were considered a co-binding event in ESCs. For the transcription factors co-binding during the reprogramming, we extended the peak summits of Oct4, Sox2, Brg1, Cebpb, Cebpa, Fra1, Runx1, Hdac1, and pe00 called by HOMER in MEF, 48hrs OSKM induction, and pre-iPSCs, by 750 bp in each direction. These 1.5 kbp regions were then intersected with the Oct4/Sox2 sites in ESCs. The intersection of genomic sites was performed by BEDtools intersect algorithm (Quinlan and Hall, 2010)

CpG Content Analysis

Two hundred bp region over the center position of Oct4/Sox2 peaks were used to calculate CpG content at the enhancer. CpG content was calculated by dividing the number of observed CpG by the number of expected CpG.

Super-Enhancer Analysis

The genomic region of a super-enhancer or a typical-enhancer was based on the previous definition characterized by Whyte et al. of R. Young's lab (Whyte et al., 2013). Two hundred and thirty-one super-enhancers were defined in mouse ESCs based on the following criteria: 1) Median enhancer size ~8,000 bp; 2) Enrichment of Mediator 1; 3)

Enriched H3K27Ac and H3K4me3 histone modification; 4) Co-binding of Oct4/Sox2/Nanog/Klf4/Esrrb. An Oct4/Sox2 binding at super-enhancer, typical-enhancer, or unclassified region was evaluated by extending the center position of Oct4/Sox2 peaks by 100 bp in each direction and intersecting with identified enhancers using BEDtools intersect algorithm (Quinlan and Hall, 2010).

FIGURE LEGENDS

Figure 3-1. ESC Gene Groups and Oct4/Sox2 Occupancy Based On ENCODE Mouse Tissue/Cell RNA-seq Datasets

(A) Selected mouse tissue/cell types polyA mRNA-seq on ENCODE database represents the transcript profiles of pluripotency genes (ESC_Bruce4) and differentiation genes (CD4+ Primary T cell, CD19+ B cell, BMDM, Brain, Cerebellum, Skeletal Muscle, Fat Pad, Thymus, Telencephalon, Spleen, Heart, Kidney). The table describes the ESC gene groups features and Oct4/Sox2 occupancy. Grouping of ESC expressed genes (> 4.9 RPKM) was done by quantifying the fold change of transcript expression between ESC and 12 somatic tissue/cell types. Numbers of tissue/cell types exhibiting > 20-fold dynamic range of expression or showing 0.2 – 5-fold non-dynamic changes are shown in table for each gene group. The last column shows the percentage of genes with strong Oct4/Sox2 binding (peak score > 20) within 15 kbp of annotated TSS. Columns are color-coded from the maximum percentage (red) to the minimum percentage (green). Genes exhibiting ESC-specificity and Dynamic range of expression in more than nine tissue/cell types are highlighted in blue and purple, respectively. Genes broadly expressed in most cell types (9 out of 12) are highlighted in green. (B) *Utf1* (NM_009482) is an example of group 1 ESC-specific gene. The bar graph indicates RNA-seq RPKM values measuring polyA mRNA levels of *Utf1* in Bruce4 ESC lines and twelve tissue/cell types. (C) *Phlda2* (NM_009434) is an example of group 3 (dynamic in 10 out of 12 tissues/cells). The bar graph indicates RNA-seq RPKM values measuring polyA mRNA levels of *Phlda2* in Bruce4 ESC lines and twelve tissue/cell types. (D) *Itgb5* (NM_001277122) is an example of group 12 (broadly-expressed, dynamic in 1 out of 12 tissues/cells). The bar graph

indicates RNA-seq RPKM values measuring polyA mRNA levels of *Itgb6* in Bruce4 ESC lines and twelve tissue/cell types.

Figure 3-2. Identify Human Oct4/Sox2 Targets by Human ENCODE RNA-seq and ChIP-seq Datasets

ChIP-seq datasets of Oct4 and Sox2 in naïve human ESC and human ESC line H9 (GEO: GSE69479, GSE69647) were processed and analyzed to study Oct4/Sox2 composite binding regions (Ji et al., 2016; Zhou et al., 2016).

(A) The pie chart displays the genomic distribution of 14,160 reproducible Oct4 peaks. Peaks were called by HOMER and retained with false discovery rate < 0.01. Promoter region was defined as -500 bp ~ +150 bp relative to the TSS. Exonic, intronic, intergenic, non-coding, and TTS were annotated based on genomic location using HOMER (Benner et al., 2017). (B) The pie chart displays the genomic distribution of 44,147 Sox2 peaks. Peaks were called by HOMER and retained with false discovery rate < 0.01. Promoter region was defined as -500 bp ~ +150 bp relative to the TSS. Exonic, intronic, intergenic, non-coding, and TTS were annotated based on genomic location using HOMER (Benner et al., 2017). (C) Selected human tissue/cell types polyA mRNA-seq on ENCODE database represents the transcript profiles of pluripotency genes (H1 hESCs) and differentiation genes (Adipose tissue, Adrenal gland, B cell, Brain, CD14+ monocyte, Sigmoid colon, Heart, Human endothelial cell of umbilical vein, Small intestine, Liver, Ovary, Foreskin fibroblast, Spleen, Skeletal muscle myoblast). The table describes the ESC gene groups features and Oct4/Sox2 occupancy. Grouping of ESC expressed genes (> 15 RPKM) was done by quantifying the fold change of transcript expression between

human ESC and 14 somatic tissue/cell types. Numbers of tissue/cell types exhibiting > 20-fold dynamic range of expression or showing 0.2 – 5-fold non-dynamic changes are shown in table for each gene group. The last column shows the percentage of genes with strong Oct4/Sox2 binding (peak score > 20) within 15 kbp of annotated TSS. Columns are color-coded from the maximum percentage (red) to the minimum percentage (green). Genes exhibiting ESC-specificity when compared to fourteen tissue/cell types are highlighted in blue. Genes broadly expressed in human ESCs and all fourteen tissue/cell types are highlighted in red. (D) The distribution of PhasCon conservation score is shown for Oct4/Sox2-bound human ESC-specific genes. Highlighted genes (*SOX2, SALL4, NANOG, and POU5F1)* are the only four Oct4/Sox2 targets shared between human and mouse. (E) The distribution of PhasCon conservation score is shown for Oct4/Sox2-bound mouse ESC-specific genes. Highlighted genes (*Sox2, Pou5f1, Sall4, and Nanog)* are the only four Oct4/Sox2 targets shared between mouse and human.

Figure 3-3. PhasCon Conservation Analysis of Oct4/Sox2 Peaks and Composite Motif

(A) The distribution of PhasCon conservation score at the 200 bp Oct4/Sox2-bound enhancer is shown for Oct4/Sox2 targets in ESC-specific, Dynamic, Non-Dynamic, and Silent genes. (B) The distribution of PhasCon conservation score at the 12 bp Oct4/Sox2 composite motif is shown for Oct4/Sox2 targets in ESC-specific, Dynamic, Non-Dynamic, and Silent genes. Silent genes with conserved motif sequences (PhasCon conservation score = 1) are highlighted with open red circles. (C) The table lists all 23 Silent genes with conserved Oct4/Sox2 motif sequences (PhasCon conservation score = 1). The last two

columns show the colocalization of ATAC signal and H3K27Ac enrichment at the Oct4/Sox2 sites. ATAC signals were analyzed by the peak strength and ATAC co-binding. Peak scores > 5 are annotated with positive (green), and peak scores < 5 or no peaks called are annotated with negative (red). ATAC peaks near Oct4/Sox2 but farther than 100 bp are annotated with shift (yellow). Enrichments of histone mark H3K27Ac were analyzed by calculating the RPKM values in a 1.5 kbp window centered on the Oct4/Sox2 peaks. H3K27Ac RPKMs > 30 are annotated with Active (green), and RPKMs < 30 are annotated with Inactive (red). (D) Example of Silent gene with conserved Oct4/Sox2 motif at an enhancer lacking ATAC signal and H3K27Ac. Genome browser snapshot displays the Oct4/Sox2 peaks, ATAC peak, and H3K27Ac enrichment at the enhancer of *Hobx1*. (E) Example of Silent gene with conserved Oct4/Sox2 motif at an enhancer with positive ATAC signal and enriched H3K27Ac. Genome browser snapshot displays the Oct4/Sox2 peaks, ATAC peak, and H3K27Ac enrichment at the enhancer of *Zic5*.

Figure 3-4. The Frequency and Strength of Nanog, c-Myc, p300, Brg1, Esrrb, Hdac1 Co-Binding Neary Oct4/Sox2 Composite Binding Sites

ChIP-seq datasets of Nanog, c-Myc, p300, Brg1, Esrrb, and Hdac1 in ESC line V6.5 (GEO: GSE90895) were processed and analyzed to study Oct4/Sox2 composite binding regions. (A) The table lists the number of Oct4/Sox2 sites with Nanog, c-Myc, p300, Brg1, Esrrb, and Hdac1 co-binding in ESC-specific, Dynamic, Non-Dynamic, and Silent genes. Only the top 20% Nanog, c-Myc, p300, Brg1, Esrrb, and Hdac1 peaks were included in the co-binding analysis. The number in each column indicates the number of co-binding over the number of different gene classes. Co-binding events were evaluated by extending TF

peak summits called by HOMER by 100 bp in each direction and intersecting with 200 bp regions centered on Oct4/Sox2 peaks. (B) The bar graphs show the percentage of Oct4/Sox2 sites with or without the co-binding of Nanog (purple), c-Myc (orange), p300 (green), Brg1 (browb), Esrrb (blue), and Hdac1 (yellow) peaks among four gene classes. Only the top 20% Nanog, c-Myc, p300, Brg1, Esrrb, and Hdac1 peaks were included in the co-binding analysis. Co-binding events were evaluated by extending TF peak summits called by HOMER by 100 bp in each direction and intersecting with 200 bp regions centered on Oct4/Sox2 peaks. (C) The ChIP-seq peak score distribution of Nanog, c-Myc, p300, Brg1, Esrrb, and Hdac1 in different gene groups. The red horizontal lines indicate the average of the TFs ChIP-seq peak scores. All peak scores of co-binding Nanog, c-Myc, p300, Brg1, Esrrb, and Hdac1 were included, with the score = 0 for the sites without TFs co-binding. No peak score threshold is set.

Figure 3-5. Properties of the Enhancers with Composite Oct4/Sox2 Binding Sites

Chromatin-associated transcripts from CCE ESC lines, E14.5 cortical neurons (NEUR), bone marrow derived macrophages (BMDM), and CD4+ CD8+ thymocytes (DP) were analyzed by RNA-seq and grouped based on the dynamic range of expression and minimum fold changes between ESC and three somatic cell types. ChIP-seq datasets of H3K27Ac in R1 ESC line and H3K4me3 in ESC line V6.5 (GEO: GSE56138 and GSE62380) were processed and analyzed to study Oct4/Sox2 composite binding regions. ATAC-seq done in ESC line V6.5 (GEO: GSE67298) were processed and analyzed to study Oct/Sox2 composite binding regions. The genomic region of a super-enhancer or

a typical-enhancer was based on the previous definition characterized by Whyte et al. of R. Young's lab (Whyte et al., 2013).

The heat map shows the read density of H3K27Ac ChIP-seq and H3K4me3 in a 1.5 kbp window, colocalized ATAC peak score, and enhancer types, centered on Oct4/Sox2 bindings at (A) 37 ESC-specific genes, (B) 51 Dynamic genes, (C) 103 Non-Dynamic genes, and (D) 711 Silent genes. Shades of red indicate percentile values of nascent transcript. For H3K27Ac (dark blue), H3K4me3 (orange), and ATAC (green), the colors indicate the read value of peak score. (E) The tables summarize the number (top) and percentage (bottom) of super-enhancer, typical-enhancer, unclassified enhancer in four gene groups.

Figure 3-6. Properties of the Enhancers with Composite Oct4/Sox2 Binding Sites (Continued)

Tables display the number and percentage of composite Oct4/Sox2 binding sites divided by enhancer types and various chromatin features (H3K27Ac or H3K4me3 or ATAC) in ESC-specific (A), Dynamic (B), Non-Dynamic (C), and Silent (D) genes. Threshold was set to distinguish active/inactive enhancers or open/close chromatin configurations: H3K27Ac read value threshold = 30; H3K4me3 read value threshold = 8; ATAC peak score threshold = 5.

Figure 3-7. Examine the Properties of CpG Content at the Enhancer with Composite Oct4/Sox2 Binding

(A) The dot plots compare the CpG contents between ESC-specific (blue), Dynamic (purple), Non-Dynamic (green), and Silent (yellow) genes. The y-axis shows CpG content of the Oct4/Sox2-bound enhancers, which equals to the number of observed CpG divided by the number of expected CpG in a 200 bp window. The red horizontal lines indicate the average of the CpG content in four gene groups. (B) The distribution CpG contents in ESC-specific (blue), Dynamic (purple), Non-Dynamic (green), and Silent (yellow) genes. The y-axis shows CpG content of the Oct4/Sox2-bound enhancers, which equals to the number of observed CpG divided by the number of expected CpG in a 200 bp window. The x-axis shows the percent of Oct4/Sox2-bound enhancers in each category. (C) Tables display the number (top) and percentage (bottom) of composite Oct4/Sox2 binding sites divided by CpG content and H3K27Ac read value in each category. (D) Tables display the number (top) and percentage (bottom) of composite Oct4/Sox2 binding sites divided by CpG content and H3K4me3 read value in each category. (E) Tables display the number (top) and percentage (bottom) of composite Oct4/Sox2 binding sites divided by CpG content and ATAC peak score in each category. A threshold of 0.2 was set for CpG content.

Figure 3-8. Enhancer Properties of Representative Genes in Different Gene Groups

The heat map shows the TF co-binding, histone modification, chromatin configuration, enhancer properties, and CpG content of the representative genes for Chapter 2 CRISPR experiments. All analyses were performed centered on the composite Oct4/Sox2 sites. Shades of dark red indicate percentile values of nascent transcript and purple color codes the dynamic range of expression (fold) between ESCs and three somatic cell types.

Transcription factors ChIP-seq of Oct4, Sox2, Nanog, c-Myc, p300, Brg1, Esrrb, and Hdac1 are color-coded on descending order according to the percentile values of peak score. The number in the columns indicates the peak score called in each TF ChIP-seq. Shades of dark blue and orange indicate read values of H3K27Ac and H3K4me3 ChIP-seq, while the green shades indicate peak scores of ATAC-seq. The properties of each enhancer are shown to the right of the heat map: super-enhancer (blue), typical-enhancer (yellow), unclassified (white), and CpG contents (color-coded from the maximum (red) to the minimum (blue)).

Figure 3-9. Transcription Factors Binding at the Oct4/Sox2-bound Enhancers in MEF, 48hrs OSKM Induction, and pre-iPSCs

ChIP-seq datasets of Oct4, Sox2, Brg1, Cebpb, Cebpa, Fra1, Runx1, Hdac1, and p300 in mouse embryonic fibroblasts (MEF), 48hr OSKM induction, pre-iPSC lines (GEO: GSE90895) were processed and analyzed to study Oct4/Sox2 composite binding regions during the reprogramming.

The heat map shows the peak score of Oct4, Sox2, Brg1, Cebpb, Cebpa, Fra1, Runx1, Hdac1, and p300 with 1.5 kbp window of the Oct4/Sox2 sites at (A) 37 ESC-specific genes, (B) 51 Dynamic genes, (C) 103 Non-Dynamic genes, and (D) 711 Silent genes. Shades with different color indicate percentile values of peak score of different TFs ChIP-seq.

Figure 3-10. The Frequency and Strength of Transcription Factors Binding at the Oct4/Sox2-bound Enhancers in MEF, 48hrs OSKM Induction, and pre-iPSCs

(A) The table displays the frequency of Oct4, Sox2, Brg1, Cebpb, Cebpa, Fra1, Runx1, Hdac1, and p300 binding within 1.5 kbp of the ESC Oct4/Sox2 composite sites in MEF, 48hrs OSKM induction, and pre-iPSC lines. Four gene classes are separated to compare the percent of other transcription factor bindings. The columns are color-coded from the maximum percentage (red) to the minimum percentage (green). (B) The table displays the average peak scores of Oct4, Sox2, Brg1, Cebpb, Cebpa, Fra1, Runx1, Hdac1, and p300 binding within 1.5 kbp of the ESC Oct4/Sox2 composite sites in MEF, 48hrs OSKM induction, and pre-iPSC lines. Four gene classes are separated to compare the percent of other transcription factor bindings. The columns are color-coded from the maximum score (red) to the minimum score (blue).

Table 3-1. Lists of PolyA mRNA-seq Datasets from Human ENCODE

The table lists the RNA-seq datasets downloaded and analyzed from human ENCODE.

Table 3-2. Transcription Factors ChIP-seq, Histone Marks ChIP-seq, ATAC-seq, and Human Oct4/Sox2 ChIP-seq Datasets

The table lists the mouse and human ChIP-seq, histone marks ChIP-seq, and ATAC-seq analyzed in this study.

Figure 3-1. ESC Gene Groups and Oct4/Sox2 Occupancy Based On ENCODE Mouse Tissue/Cell RNA-seq Datasets

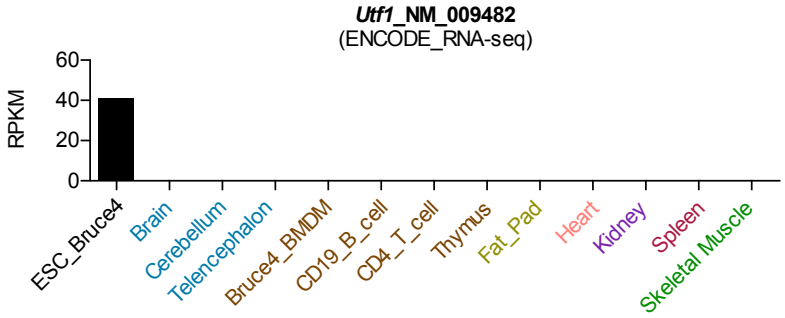
A

Embryonic Stem Cells:
ESC_Bruce4

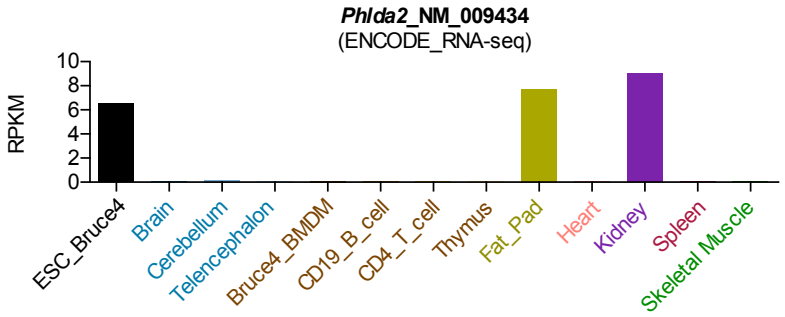
Somatic Cells / Tissue:
CD4+ Primary T cell, CD19+ B cell, BMDM, Brain, Cerebellum, Skeletal Muscle, Fat Pad, Thymus, Telencephalon, Spleen, Heart, Kidney

Categories	ESC vs 12 Cells/Tissues		# of Genes	# with Oct4/Sox2 > 20	% of StrongOS
	> 20X	0.2 - 5 X			
(1) ESC_20X_Specific	12	0	40	15	37.50%
(2) ESC_20X_Dynamic_11	11	1	35	7	20.00%
(3) ESC_20X_Dynamic_10	10	2	9	3	33.33%
(4) ESC_20X_Dynamic_9	9	3	33	7	21.21%
(5) ESC_20X_Dynamic_8	8	4	27	4	14.81%
(6) ESC_20X_Dynamic_7	7	5	36	7	19.44%
(7) ESC_20X_Dynamic_6	6	6	32	4	12.50%
(8) ESC_20X_Dynamic_5	5	7	94	22	23.40%
(9) ESC_20X_Dynamic_4	4	8	103	19	18.45%
(10) ESC_20X_Dynamic_3	3	9	77	5	6.49%
(11) ESC_20X_Dynamic_2	2	10	119	9	7.56%
(12) ESC_20X_Dynamic_1	1	11	184	13	7.07%

B



C



D

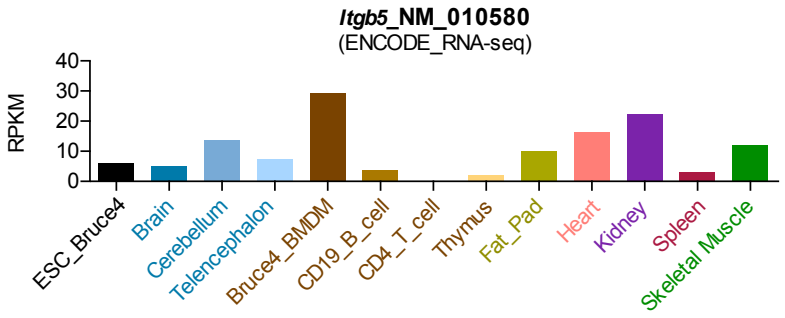
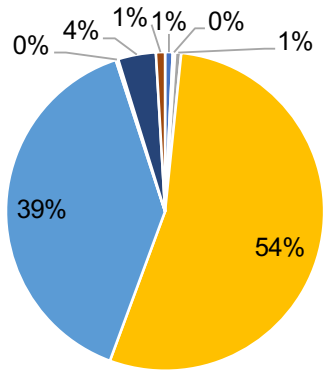
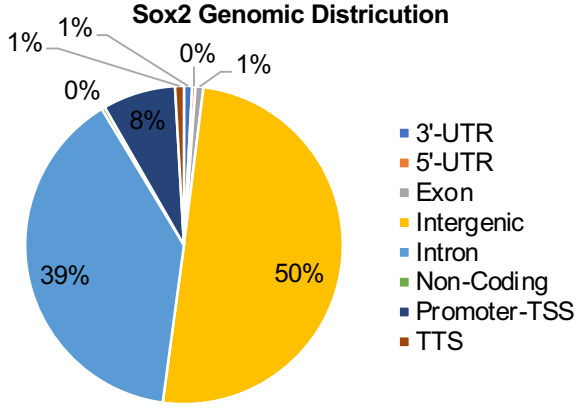


Figure 3-2. Identify Human Oct4/Sox2 Targets by Human ENCODE RNA-seq and ChIP-seq Datasets

A



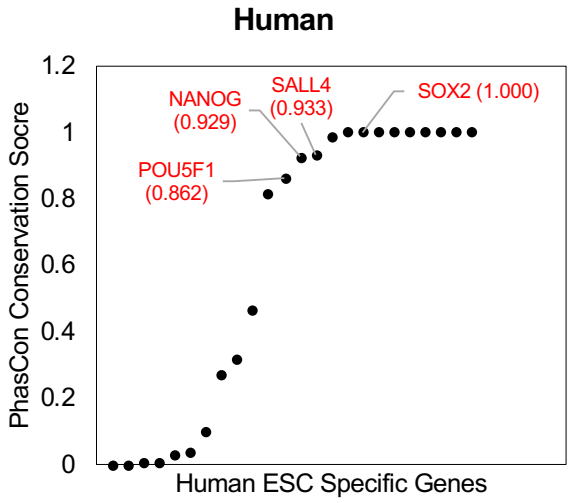
B



C

Categories	ESC vs 14 Cells/Tissues		# of Genes	# with Oct4/Sox2 > 20	% of StrongOS
	> 20X	0.2 - 5 X			
(1) ESC_20X_Specific	14	0	44	19	43.18%
(2) ESC_20X_Dynamic_13	13	1	37	5	13.51%
(3) ESC_20X_Dynamic_12	12	2	46	3	6.52%
(4) ESC_20X_Dynamic_11	11	3	36	3	8.33%
(5) ESC_20X_Dynamic_10	10	4	59	4	6.78%
(6) ESC_20X_Dynamic_9	9	5	66	5	7.58%
(7) ESC_20X_Dynamic_8	8	6	72	6	8.33%
(8) ESC_20X_Dynamic_7	7	7	58	2	3.45%
(9) ESC_20X_Dynamic_6	6	8	61	7	11.48%
(10) ESC_20X_Dynamic_5	5	9	58	5	8.62%
(11) ESC_20X_Dynamic_4	4	10	113	8	7.08%
(12) ESC_20X_Dynamic_3	3	11	150	10	6.67%
(13) ESC_20X_Dynamic_2	2	12	198	13	6.57%
(14) ESC_20X_Dynamic_1	1	13	495	22	4.44%
(15) ESC_Non-Dynamic	0	14	268	8	2.99%

D



E

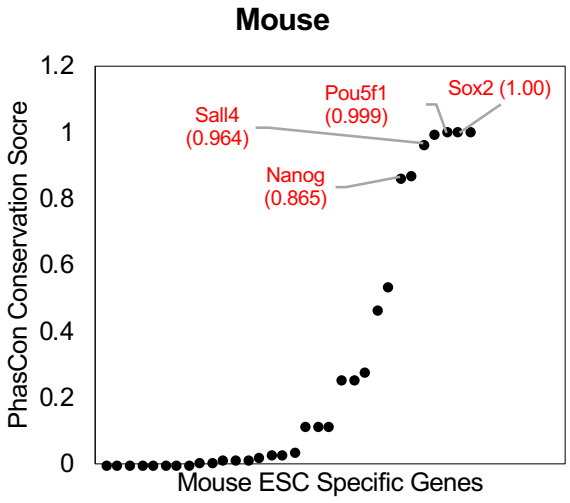
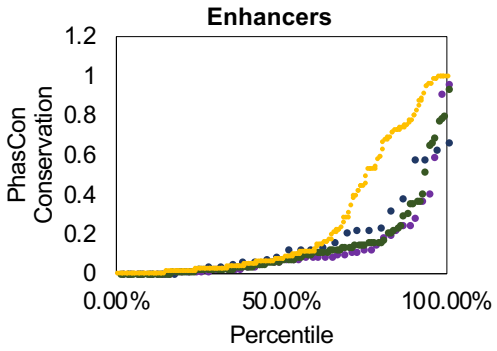
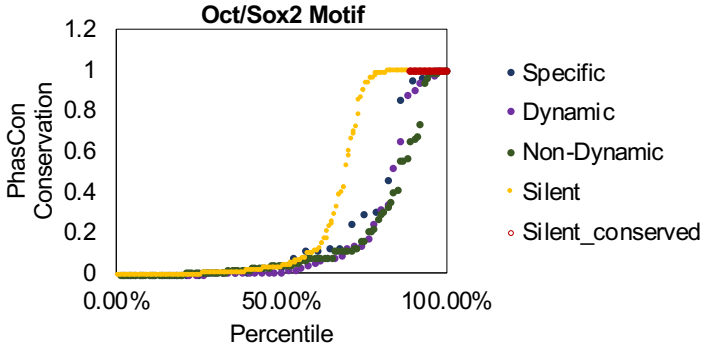


Figure 3-3. PhasCon Conservation Analysis of Oct4/Sox2 Peaks and Composite Motif

A



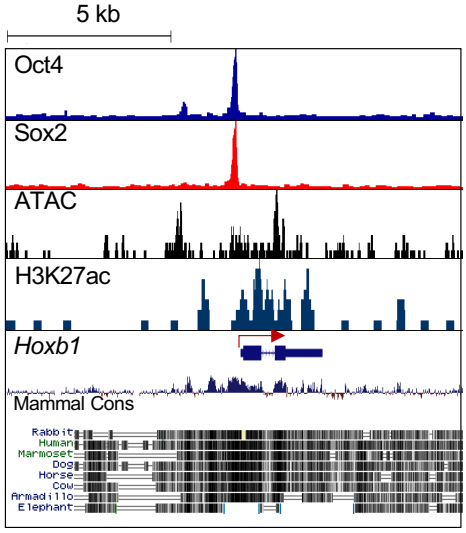
B



C

Total = 23		ESC	
RefID	Gene	ATAC	H3K27Ac
NM_028275	1700112E06Rik	Positive	Inactive
NM_001146198	Nkx2-1	Positive	Inactive
NM_001024918	Rfx4	Positive	Inactive
NM_010095	Ebf2	Positive	Inactive
NM_008262	Onecut1	Positive	Inactive
NM_001013385	Grm4	Positive	Inactive
NM_024124	Hdac9	Positive	Inactive
NM_011440	Sox14	Positive	Inactive
NR_033490	2610100L16Rik	Positive	Inactive
NM_009865	Cdh10	Positive	Inactive
NM_001012765	Adcy5	Positive	Active
NM_144810	Klhdc8a	Positive	Active
NM_001159569	Meis2	Positive	Active
NM_022987	Zic5	Positive	Active
NM_001033250	Lemd1	Positive	Active
NM_013833	Rax	Positive	Active
NM_008086	Gas1	Shift	Inactive
NM_001042617	Cadps	Shift	Inactive
NM_177753	Sox21	Shift	Inactive
NM_008242	Foxd1	Shift	Inactive
NM_008266	Hoxb1	Negative	Inactive
NM_019446	Barhl1	Negative	Inactive
NR_015561	C130071C03Rik	Negative	Inactive

D



E

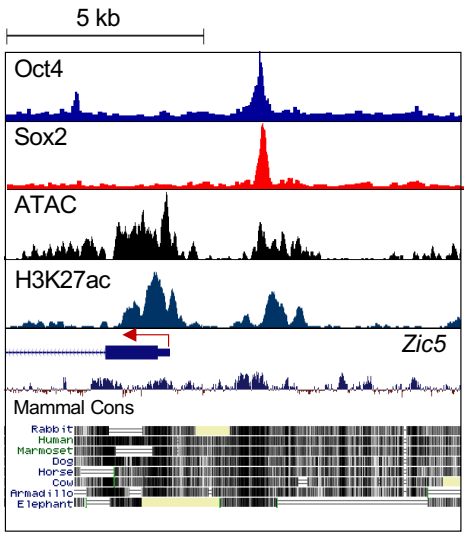
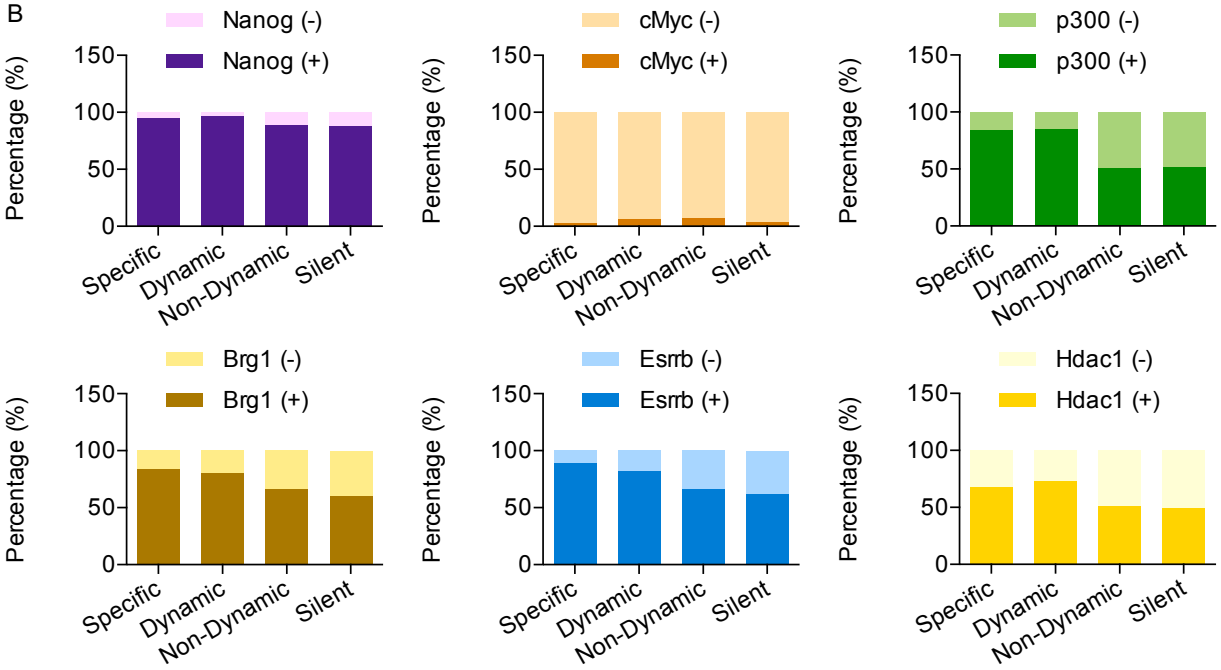


Figure 3-4. The Frequency and Strength of Nanog, c-Myc, p300, Brg1, Esrrb, Hdac1 Co-Binding Neary Oct4/Sox2 Composite Binding Sites

A

	Nanog	cMyc	p300	Brg1	Esrrb	Hdac1
Specific	35 / 37	1 / 37	31 / 37	31 / 37	33 / 37	25 / 37
Dynamic	49 / 51	3 / 51	43 / 51	41 / 51	42 / 51	37 / 51
Non-Dynamic	91 / 103	7 / 103	52 / 103	68 / 103	68 / 103	53 / 103
Silent	694 / 711	26 / 711	368 / 711	438 / 711	438 / 711	349 / 711

B



C

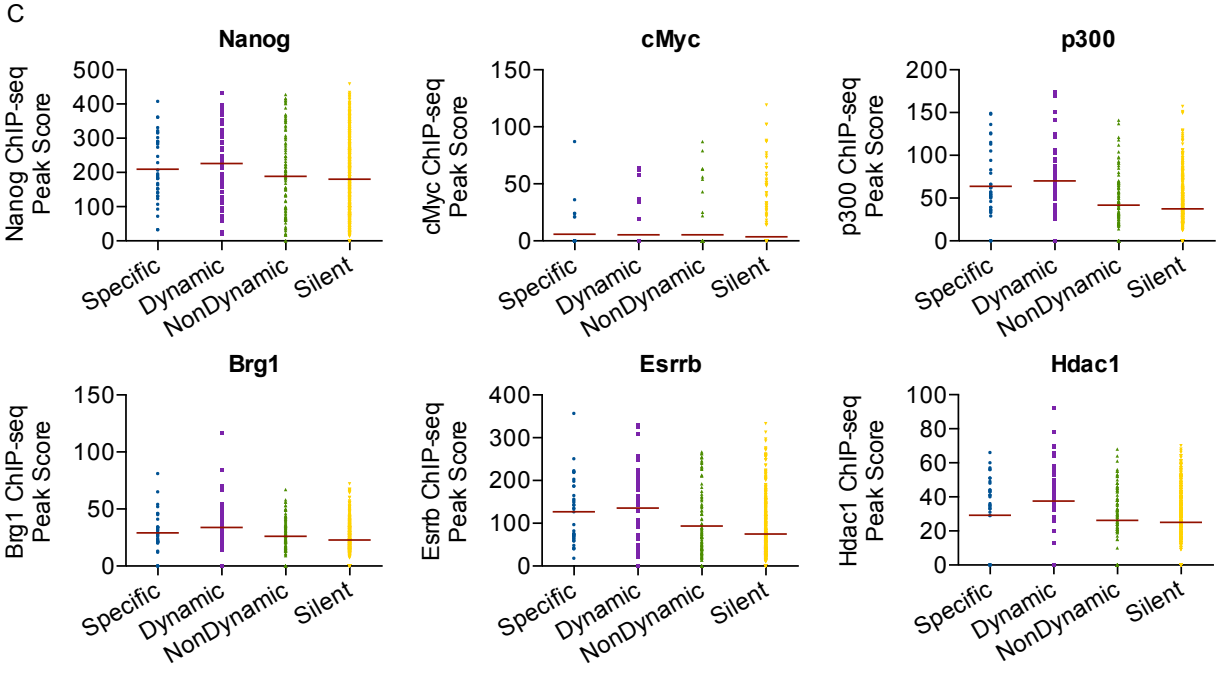
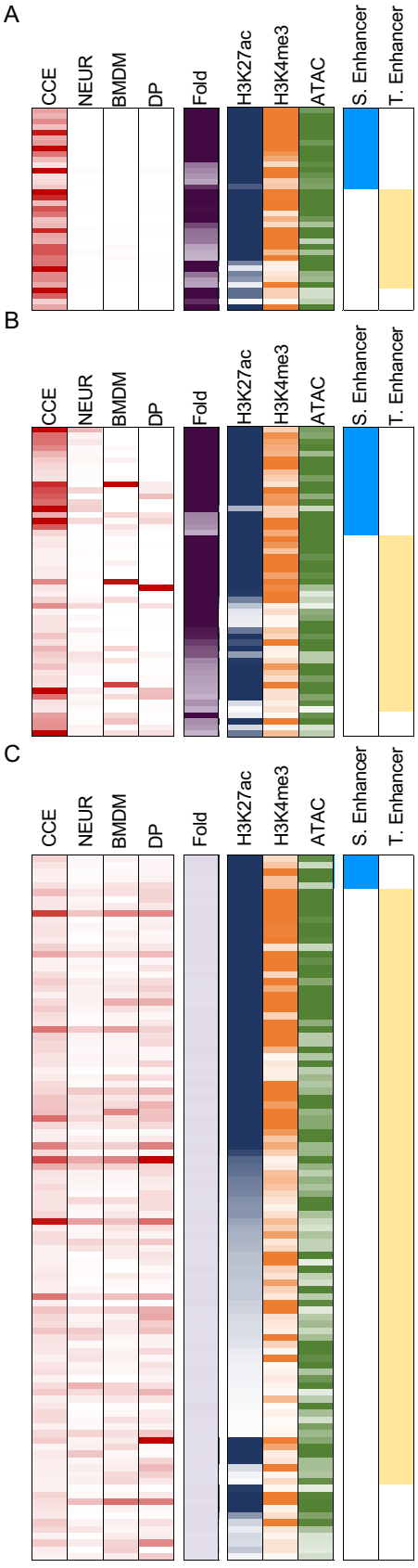
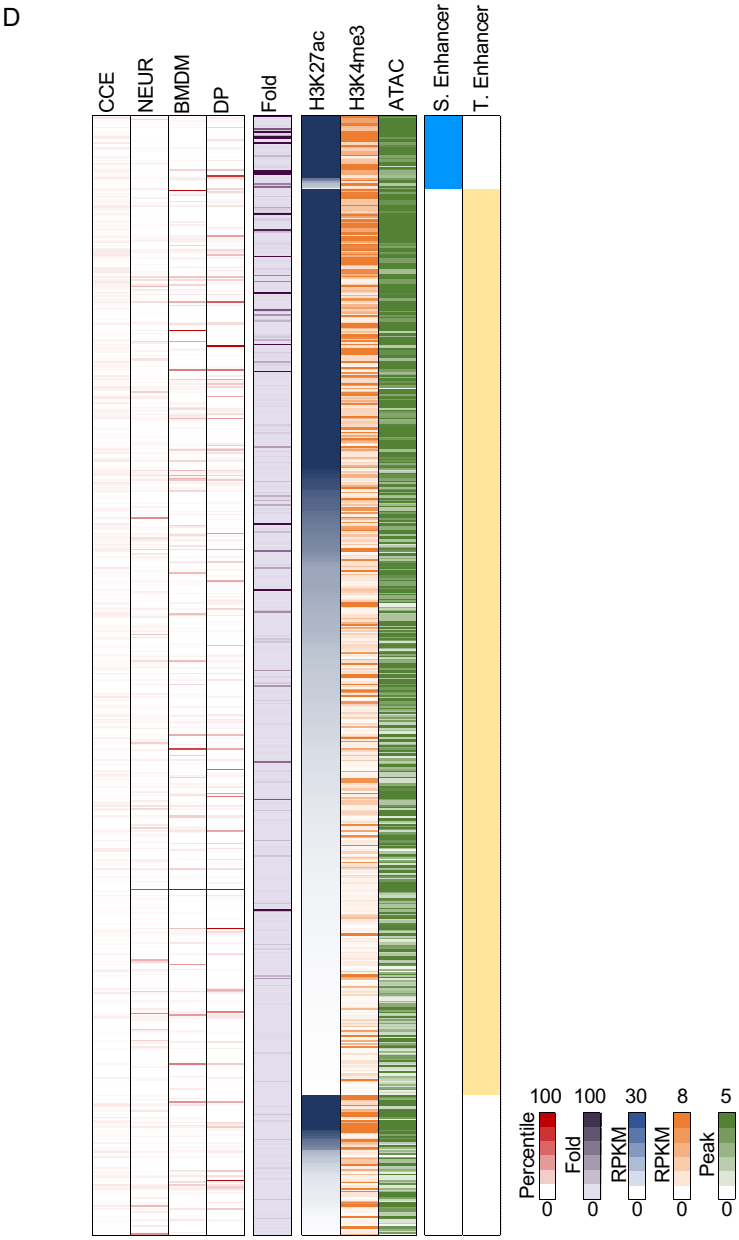


Figure 3-5. Properties of the Enhancers with Composite Oct4/Sox2 Binding Sites





E

	Super Enhancer	Typical Enhancer	Unclassified	Total
Specific	15	18	4	37
Dynamic	18	29	4	51
Non-Dynamic	5	87	11	103
Silent	47	575	89	711

Specific	40.54%	48.65%	10.81%	100%
Dynamic	35.29%	56.86%	7.84%	100%
Non-Dynamic	4.85%	84.47%	10.68%	100%
Silent	6.61%	80.87%	12.52%	100%

Figure 3-6. Properties of the Enhancers with Composite Oct4/Sox2 Binding Sites (Continued)

A

Specific

		Super Enhancer	Typical Enhancer	Unclassified
H3K27ac	> 30	13	12	1
	< 30	2	6	3
H3K4me3	> 8	10	9	1
	< 8	5	9	3
ATAC	> 5	8	10	1
	< 5	7	8	3

		Super Enhancer	Typical Enhancer	Unclassified
H3K27ac	> 30	35.14%	32.43%	2.70%
	< 30	5.41%	16.22%	8.11%
H3K4me3	> 8	27.03%	24.32%	2.70%
	< 8	13.51%	24.32%	8.11%
ATAC	> 5	21.62%	27.03%	2.70%
	< 5	18.92%	21.62%	8.11%

B

Dynamic

		Super Enhancer	Typical Enhancer	Unclassified
H3K27ac	> 30	16	16	1
	< 30	2	13	4
H3K4me3	> 8	17	23	2
	< 8	1	6	3
ATAC	> 5	18	24	2
	< 5	0	5	3

		Super Enhancer	Typical Enhancer	Unclassified
H3K27ac	> 30	30.77%	30.77%	1.92%
	< 30	3.85%	25.00%	7.69%
H3K4me3	> 8	32.69%	44.23%	3.85%
	< 8	1.92%	11.54%	5.77%
ATAC	> 5	34.62%	46.15%	3.85%
	< 5	0.00%	9.62%	5.77%

C

Non-Dynamic

		Super Enhancer	Typical Enhancer	Unclassified
H3K27ac	> 30	5	29	1
	< 30	0	58	10
H3K4me3	> 8	1	31	3
	< 8	4	56	8
ATAC	> 5	1	29	2
	< 5	4	58	9

		Super Enhancer	Typical Enhancer	Unclassified
H3K27ac	> 30	4.85%	28.16%	0.97%
	< 30	0.00%	56.31%	9.71%
H3K4me3	> 8	0.97%	30.10%	2.91%
	< 8	3.88%	54.37%	7.77%
ATAC	> 5	0.97%	28.16%	1.94%
	< 5	3.88%	56.31%	8.74%

D

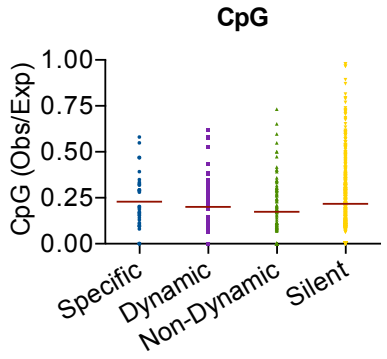
Silent

		Super Enhancer	Typical Enhancer	Unclassified
H3K27ac	> 30	33	131	14
	< 30	14	444	75
H3K4me3	> 8	18	70	22
	< 8	29	505	67
ATAC	> 5	25	202	28
	< 5	22	373	61

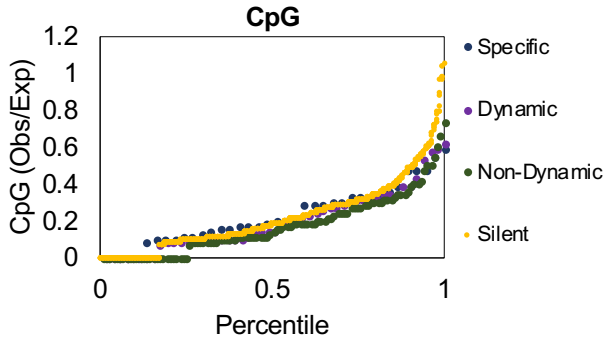
		Super Enhancer	Typical Enhancer	Unclassified
H3K27ac	> 30	4.64%	18.42%	1.97%
	< 30	1.97%	62.45%	10.55%
H3K4me3	> 8	2.53%	9.85%	3.09%
	< 8	4.08%	71.03%	9.42%
ATAC	> 5	3.52%	28.41%	3.94%
	< 5	3.09%	52.46%	8.58%

Figure 3-7. Examine the Properties of CpG Content at the Enhancer with Composite Oct4/Sox2 Binding

A



B



C

Number of Gene

H3K27Ac RPKM	CpG (Obs / Exp) > 0.2	CpG (Obs / Exp) 0 - 0.2
Specific		
> 30	12	14
< 30	6	5
Dynamic		
> 30	16	17
< 30	7	11
Non-Dynamic		
> 30	16	19
< 30	22	46
Silent		
> 30	76	102
< 30	248	285

Percentage of Gene

H3K27Ac RPKM	CpG (Obs / Exp) > 0.2	CpG (Obs / Exp) 0 - 0.2
Specific		
> 30	32.43%	37.84%
< 30	16.22%	13.51%
Dynamic		
> 30	31.37%	33.33%
< 30	13.73%	21.57%
Non-Dynamic		
> 30	15.53%	18.45%
< 30	21.36%	44.66%
Silent		
> 30	10.69%	14.35%
< 30	34.88%	40.08%

D

Number of Gene

H3K4me3 RPKM	CpG (Obs / Exp) > 0.2	CpG (Obs / Exp) 0 - 0.2
Specific		
> 8	9	11
< 8	9	8
Dynamic		
> 8	21	20
< 8	2	8
Non-Dynamic		
> 8	16	19
< 8	22	46
Silent		
> 8	73	37
< 8	251	350

Percentage of Gene

H3K4me3 RPKM	CpG (Obs / Exp) > 0.2	CpG (Obs / Exp) 0 - 0.2
Specific		
> 8	24.32%	29.73%
< 8	24.32%	21.62%
Dynamic		
> 8	41.18%	39.22%
< 8	3.92%	15.69%
Non-Dynamic		
> 8	15.53%	18.45%
< 8	21.36%	44.66%
Silent		
> 8	10.27%	5.20%
< 8	35.30%	49.23%

E

Number of Gene

ATAC Peak	CpG (Obs / Exp) > 0.2	CpG (Obs / Exp) 0 - 0.2
Specific		
> 5	11	8
< 5	7	11
Dynamic		
> 5	20	22
< 5	3	6
Non-Dynamic		
> 5	9	23
< 5	29	42
Silent		
> 5	120	135
< 5	204	252

Percentage of Gene

ATAC Peak	CpG (Obs / Exp) > 0.2	CpG (Obs / Exp) 0 - 0.2
Specific		
> 5	29.73%	21.62%
< 5	18.92%	29.73%
Dynamic		
> 5	39.22%	43.14%
< 5	5.88%	11.76%
Non-Dynamic		
> 5	8.74%	22.33%
< 5	28.16%	40.78%
Silent		
> 5	16.88%	18.99%
< 5	28.69%	35.44%

Figure 3-8. Enhancer Properties of Representative Genes in Different Gene Groups

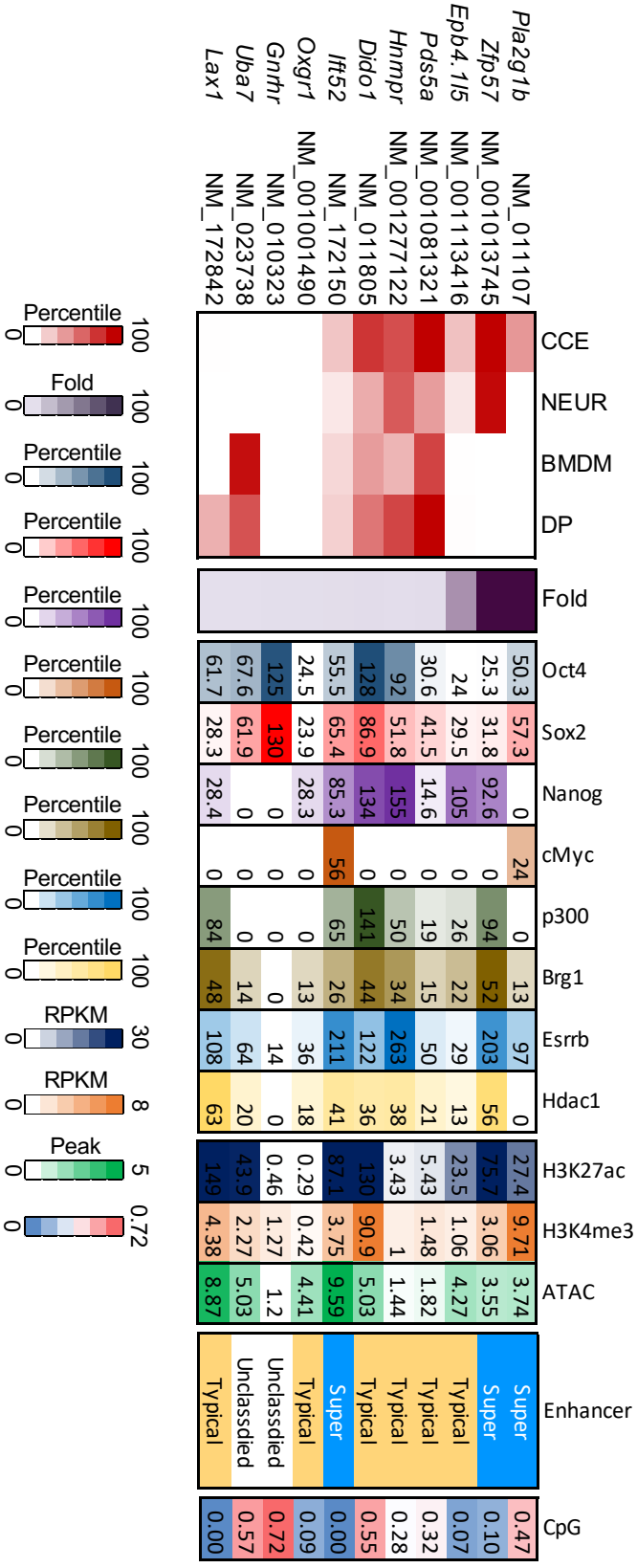


Figure 3-9. Transcription Factors Binding at the Oct4/Sox2-bound Enhancers in MEF, 48hrs OSKM Induction, and pre-iPSCs

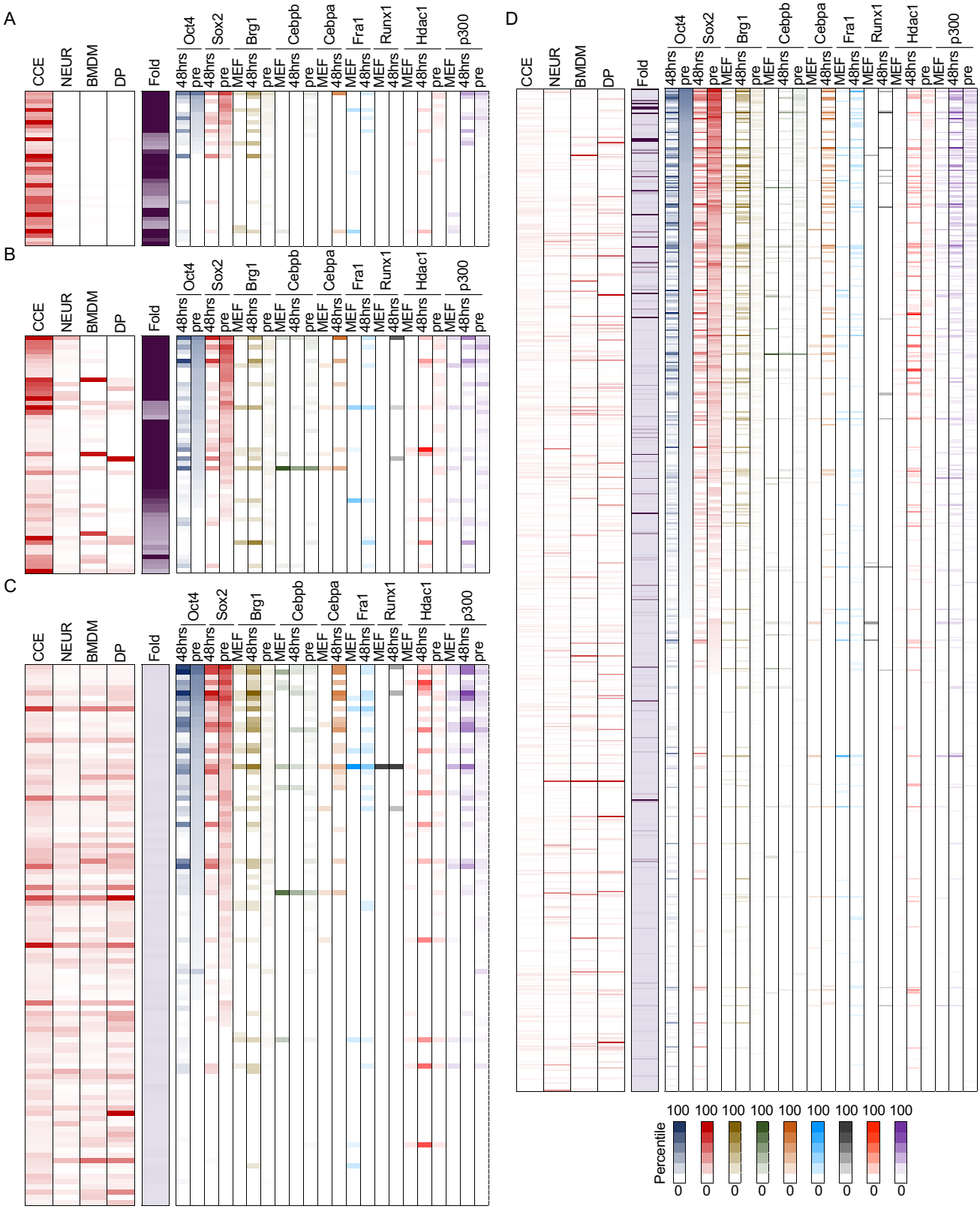


Figure 3-10. The Frequency and Strength of Transcription Factors Binding at the Oct4/Sox2-bound Enhancers in MEF, 48hrs OSKM Induction, and pre-iPSCs

A

	Oct4		Sox2		Brg1			Cebpb		
	48hrs	Pre	48hrs	Pre	MEF	48hrs	Pre	MEF	48hrs	Pre
Specific	43.24%	40.54%	29.73%	43.24%	13.51%	27.03%	43.24%	2.70%	0.00%	24.32%
Dynamic	66.67%	72.55%	43.14%	76.47%	17.65%	29.41%	62.75%	7.84%	7.84%	31.37%
Non-Dynamic	54.37%	62.14%	38.83%	64.08%	17.48%	33.01%	51.46%	6.80%	7.77%	28.16%
Silent	52.04%	52.88%	36.29%	54.57%	14.91%	24.89%	41.63%	5.34%	8.30%	23.91%

	Cebpa		Fra1		Runx1		Hdac1			p300		
	MEF	48hrs	MEF	48hrs	MEF	48hrs	MEF	48hrs	Pre	MEF	48hrs	Pre
Specific	0.00%	5.41%	8.11%	10.81%	0.00%	0.00%	5.41%	10.81%	21.62%	8.11%	18.92%	32.43%
Dynamic	7.84%	11.76%	5.88%	15.69%	0.00%	5.88%	13.73%	19.61%	39.22%	13.73%	29.41%	62.75%
Non-Dynamic	4.85%	18.45%	6.80%	17.48%	0.97%	3.88%	13.59%	17.48%	23.30%	15.53%	26.21%	45.63%
Silent	3.23%	12.10%	5.20%	12.24%	0.84%	2.25%	10.69%	16.46%	16.60%	11.11%	22.36%	37.27%

B

	Oct4		Sox2		Brg1			Cebpb		
	48hrs	Pre	48hrs	Pre	MEF	48hrs	Pre	MEF	48hrs	Pre
Specific	90.0625	85.33333	146.8182	186.875	88	124.6	26.25	36	0	14.33333
Dynamic	90.32353	90.86486	157.5909	202.5897	87.11	127.87	25.94	94.25	117.5	34.75
Non-Dynamic	102.125	82.84375	178.35	171.8636	70.56	133.14	21.9	95.85714	58.75	19.17241
Silent	82.01351	74.41223	136.3876	160.9433	59.29	110.39	21.71	55.94737	35.59322	18.19412

	Cebpa		Fra1		Runx1		Hdac1			p300		
	MEF	48hrs	MEF	48hrs	MEF	48hrs	MEF	48hrs	Pre	MEF	48hrs	Pre
Specific	0	49.5	46.667	28.5	0	0	13.5	47	25.25	64	126.29	52.083
Dynamic	16	42.667	65.667	32.75	0	41.333	15.571	93.5	25.3	70.286	102.8	47
Non-Dynamic	18.6	39.158	59.857	35.833	85	47.75	19.5	111.5	25.167	79.125	151.07	43.319
Silent	18.348	34.023	39.027	29.368	46	36.438	11.961	73.325	24.288	66.063	105.11	38.838

Table 3-1 Lists of PolyA mRNA-seq Datasets from Human ENCODE

Tissue / Cell	Replicates	Library	Accession
H1-hESC	1	ENCLB555AMA	SRR5048077
	2	ENCLB555AMB	SRR5048078
Endothelial cell of umbilical vein	1	ENCLB555AVJ	SRR3192477
	2	ENCLB555AVK	SRR3192476
CD14-positive monocyte	1	ENCLB555AWV	SRR5048177 - 179
	2	ENCLB555AWY	SRR5048180 - 182
B Cell	1	ENCLB555AUP	SRR5048160 - 162
	2	ENCLB555APR	SRR5048157 - 159
Skeletal muscle myoblast	1	ENCLB555AWI	SRR3192475
	2	ENCLB555AWH	SRR3192474
Foreskin fibroblast	1	ENCLB555ANG	SRR5048073
	2	ENCLB555ANH	SRR5048074
Adrenal gland	1	ENCLB603REP	ENCFF028DUO, ENCFF470RWW
	2	ENCLB981BIW	ENCFF709FHN, ENCFF681HNP
Ovary	1	ENCLB178YZR	ENCFF419GVS, ENCFF135CVY
Liver	1	ENCLB490UAX	ENCFF650JAM, ENCFF803DXA
	2	ENCLB828MEI	ENCFF187OKV, ENCFF283RUU
Heart	1	ENCLB171YLN	ENCFF464TEM, ENCFF221QNJ
	2	ENCLB797GKI	ENCFF770NYA, ENCFF076IRZ
Spleen	1	ENCLB138JMP	ENCFF058MGQ, ENCFF058MGQ
	2	ENCLB680CYA	ENCFF926YPC, ENCFF111IRS
Adipose Tissue	1	ENCLB236EKW	ENCFF170RHF, ENCFF437XFH
	2	ENCLB564EVI	ENCFF592VVB, ENCFF359HIQ
Brain	1	ENCLB187ZUS	ENCFF850ZLY, ENCFF897IUQ
	2	ENCLB306ITM	ENCFF456MMS, ENCFF716WNR
Sigmoid Colon	1	ENCLB521TPI	ENCFF734ZAD, ENCFF261RWK
	2	ENCLB679EPU	ENCFF322RPT, ENCFF782AHJ
Small Intestine	1	ENCLB486ZUR	ENCFF540SNP, ENCFF540SNP
	2	ENCLB670VDL	ENCFF338DKW, ENCFF338DKW

Table 3-2. Lists of Transcription Factors ChIP-seq, Histone Marks ChIP-seq, ATAC-seq, and Human Oct4/Sox2 ChIP-seq

Sequencing Datasets	Cell Type	GEO Accession
Oct4 ChIP-seq	ESC	GEO: GSE90895
Sox2 ChIP-seq	ESC	GEO: GSE90895
Nanog ChIP-seq	ESC	GEO: GSE90895
cMyc ChIP-seq	ESC	GEO: GSE90895
P300 ChIP-seq	ESC	GEO: GSE90895
Brg1 ChIP-seq	ESC	GEO: GSE90895
Esrrb ChIP-seq	ESC	GEO: GSE90895
Hdac1 ChIP-seq	ESC	GEO: GSE90895
Brg1 ChIP-seq	MEF	GEO: GSE90895
Cebpa ChIP-seq	MEF	GEO: GSE90895
Cebpb ChIP-seq	MEF	GEO: GSE90895
Fra1 ChIP-seq	MEF	GEO: GSE90895
Runx1 ChIP-seq	MEF	GEO: GSE90895
Hdac1 ChIP-seq	MEF	GEO: GSE90895
P300 ChIP-seq	MEF	GEO: GSE90895
Oct4 ChIP-seq	48hrs OSKM	GEO: GSE90895
Sox2 ChIP-seq	48hrs OSKM	GEO: GSE90895
Brg1 ChIP-seq	48hrs OSKM	GEO: GSE90895
Cebpa ChIP-seq	48hrs OSKM	GEO: GSE90895
Cebpb ChIP-seq	48hrs OSKM	GEO: GSE90895
Fra1 ChIP-seq	48hrs OSKM	GEO: GSE90895
Runx1 ChIP-seq	48hrs OSKM	GEO: GSE90895
Hdac1 ChIP-seq	48hrs OSKM	GEO: GSE90895
P300 ChIP-seq	48hrs OSKM	GEO: GSE90895
Oct4 ChIP-seq	pre iPSCs	GEO: GSE90895
Sox2 ChIP-seq	pre iPSCs	GEO: GSE90895
Brg1 ChIP-seq	pre iPSCs	GEO: GSE90895
Cebpb ChIP-seq	pre iPSCs	GEO: GSE90895
Hdac1 ChIP-seq	pre iPSCs	GEO: GSE90895
P300 ChIP-seq	pre iPSCs	GEO: GSE90895

REFERENCE

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

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., et al. (2006). Chromatin signatures of pluripotent cell lines. Nat. Cell Biol. *8*, 532–538.

Becker, J.S., Nicetto, D., and Zaret, K.S. (2016). H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. Trends Genet. *32*, 29–41.

Benner, C., Heinz, S., and Glass, C.K. (2017). HOMER - Software for motif discovery and next generation sequencing analysis.

van den Berg, D.L.C., Zhang, W., Yates, A., Engelen, E., Takacs, K., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R.A. (2008). Estrogen-Related Receptor Beta Interacts with Oct4 To Positively Regulate Nanog Gene Expression. Mol. Cell. Biol.

van den Berg, D.L.C., Snoek, T., Mullin, N.P., Yates, A., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R.A. (2010). An Oct4-Centered Protein Interaction Network in Embryonic Stem Cells. Cell Stem Cell.

Brons, I.G.M., Smithers, L.E., Trotter, M.W.B., Rugg-Gunn, P., Sun, B., Chuva De Sousa Lopes, S.M., Howlett, S.K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R.A., et al. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature.

Buecker, C., Srinivasan, R., Wu, Z., Calo, E., Acampora, D., Faial, T., Simeone, A., Tan, M., Swigut, T., and Wysocka, J. (2014). Reorganization of enhancer patterns in

transition from naive to primed pluripotency. Cell Stem Cell.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. Cell.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., Pellegrini, M., and Plath, K. (2013). Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization. Cell Stem Cell.

Descalzo, S.M., Rué, P., Garcia-Ojalvo, J., and Arias, A.M. (2012). Correlations between the levels of Oct4 and Nanog as a signature for naïve pluripotency in mouse embryonic stem cells. Stem Cells.

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.

Ernst, M., Dawud, R.A., Kurtz, A., Schotta, G., Taher, L., and Fuellen, G. (2015). Comparative computational analysis of pluripotency in human and mouse stem cells. Sci. Rep.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589.

Jauch, R., Ng, C.K.L., Saikatendu, K.S., Stevens, R.C., and Kolatkar, P.R. (2008). Crystal Structure and DNA Binding of the Homeodomain of the Stem Cell Transcription Factor Nanog. J. Mol. Biol. *376*, 758–770.

Jerabek, S., Merino, F., Schöler, H.R., and Cojocaru, V. (2014). OCT4: Dynamic DNA binding pioneers stem cell pluripotency. Biochim. Biophys. Acta - Gene Regul. Mech.

Ji, X., Dadon, D.B., Abraham, B.J., Lee, T.I., Jaenisch, R., Bradner, J.E., and Young, R.A. (2015). Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. Proc. Natl. Acad. Sci.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I., et al. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell.

Kidder, B.L., and Palmer, S. (2012). HDAC1 regulates pluripotency and lineage specific transcriptional networks in embryonic and trophoblast stem cells. Nucleic Acids Res.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: A fast spliced aligner with low memory requirements. Nat. Methods.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S.H. (2008). An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. Cell *132*, 1049–1061.

King, H.W., and Klose, R.J. (2017a). The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. Elife.

King, H.W., and Klose, R.J. (2017b). The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. Elife.

Koche, R.P., Smith, Z.D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B.E., and Meissner, A. (2011). Reprogramming factor expression initiates widespread targeted

chromatin remodeling. Cell Stem Cell *8*, 96–105.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., et al. (2007). Directly Reprogrammed Fibroblasts Show Global Epigenetic Remodeling and Widespread Tissue Contribution. Cell Stem Cell *1*, 55–70.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766–770.

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

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

Moorthy, S.D., Davidson, S., Shchuka, V.M., Singh, G., Malek-Gilani, N., Langroudi, L., Martchenko, A., So, V., Macpherson, N.N., and Mitchell, J.A. (2017). Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. Genome Res.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods.

Nichols, J., and Smith, A. (2009). Naive and Primed Pluripotent States. Cell Stem Cell.

Nichols, J., and Smith, A. (2010). The origin and identity of embryonic stem cells. Development.

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

Orkin, S.H. (2005). Chipping away at the embryonic stem cell network. Cell *122*, 828–830.

Park, S., Park, S.H., Kook, M.-C., Kim, E., Park, S., and Lim, J.H. (2004). Ultrastructure of human embryonic stem cells and spontaneous and retinoic acid-induced differentiating cells. Ultrastruct. Pathol. *28*, 229–238.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics.

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

Raisner, R., Kharbanda, S., Jin, L., Jeng, E., Chan, E., Merchant, M., Haverty, P.M., Bainer, R., Cheung, T., Arnott, D., et al. (2018). Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. Cell Rep.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling. Cell.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.D.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res.

Simandi, Z., Horvath, A., Wright, L.C., Cuaranta-Monroy, I., De Luca, I., Karolyi, K., Sauer, S., Deleuze, J.F., Gudas, L.J., Cowley, S.M., et al. (2016). OCT4 Acts as an Integrator of Pluripotency and Signal-Induced Differentiation. Mol. Cell.

Singhal, N., Esch, D., Stehling, M., and Schöler, H.R. (2014). BRG1 Is Required to Maintain Pluripotency of Murine Embryonic Stem Cells. Biores. Open Access.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2014). Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. Cell *161*, 555–568.

Di Stefano, B., Collombet, S., Jakobsen, J.S., Wierer, M., Sardina, J.L., Lackner, A., Stadhouders, R., Segura-Morales, C., Francesconi, M., Limone, F., et al. (2016). C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. Nat. Cell Biol.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663–676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell.

Tazi, J., and Bird, A. (1990). Alternative chromatin structure at CpG islands. Cell.

Warrier, S., Van Der Jeught, M., Duggal, G., Tilleman, L., Sutherland, E., Taelman, J., Popovic, M., Lierman, S., Chuva De Sousa Lopes, S., Van Soom, A., et al. (2017).

Direct comparison of distinct naive pluripotent states in human embryonic stem cells. Nat. Commun.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

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

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: Establishing competence for gene expression. Genes Dev.

Zhang, X., Zhang, J., Wang, T., Esteban, M.A., and Pei, D. (2008). Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. J. Biol. Chem.

Zhou, C., Yang, X., Sun, Y., Yu, H., Zhang, Y., and Jin, Y. (2016). Comprehensive profiling reveals mechanisms of SOX2-mediated cell fate specification in human ESCs and NPCs. Cell Res.

CHAPTER 4

Concluding Remarks: Conclusions and Future Directions

The preceding chapters have attempted to investigate mechanisms of Oct4/Sox2 binding at the composite enhancer motif responsible for coordinating the maintenance and the establishment of pluripotency. The in-depth genetic and genomic analyses of transcriptional regulation by Oct4 and Sox2 in embryonic stem cells elaborate mechanistic insights by gene-centric system approaches. We took a combination of bioinformatics analysis and CRISPR/Cas9 experiments approach to reveal the distinct functions of the Oct4/Sox2 binding associated with well-defined gene clusters. Several findings have been highlighted and discussed at the end of the previous chapters. In this concluding chapter, we provide a brief discussion on the future prospects of studies focused on understanding the mechanisms required for maintaining and establishing pluripotency.

One exciting result we have discovered, after the deletion of Oct4/Sox2 binding, is the residual transcription of the ESC-specific gene *Pla2g1b* and the Dynamic genes *Hnrnpr* and *Pds5a* in ESCs and iPSCs. Since Oct4 and Sox2 are considered to be critical for pluripotency, the residual transcription leads to an intriguing question of why the expression of these genes are not fully-dependent on Oct4/Sox2. To answer this question, there are several possibilities we can address further. First, Oct4/Sox2 binding at enhancer composite motif may be critical for establishing pluripotency, but it may not be required for maintaining the gene transcription in established ESC lines. In our study, we have tested this hypothesis by building the model of mouse tet-on iPSC secondary reprogramming assay. The functional validations employed CRISPR/Cas9 and HDR mutation in this cell model have proven the hypothesis correct. Second, the Oct4/Sox2

composite binding site may play a redundant role with other regulatory elements further away for the genes. Chromatin in nucleus exhibits long-range interaction to form a three-dimensional architecture for gene regulatory functions. Another distal regulatory element may contribute to a portion of the gene expression. While the synergistic regulation is retained, solely deleting the Oct4/Sox2 motifs do not completely eliminate the gene transcription. This hypothesis also implicates potential co-factor regulation synergize on the pluripotency gene regulation. To elucidate potential co-factor regulation and distal regulatory elements interaction, a combination approach integrating co-factor ChIP-seq and Hi-C analysis will be valuable to identify the functional distal regulatory elements (Belton et al., 2012; van Berkum et al., 2010). Third, the residual transcription may be attributed to the transcription heterogeneity of the cell populations. Since we measured the transcript level in a bulk condition, the initiation and activation of transcription may not be homogenous in the cell populations. To assess this possibility, single-molecule technology could offer exquisite sensitivity in space and time along with the ability to observe transcriptional heterogeneity, which is difficult to distinguish in a bulk condition (Chen and Larson, 2016). Moreover, single-cell RNA-sequencing focusing will be another high-throughput method to compare the differences between individual cells and their transcriptional output (Kolodziejczyk et al., 2015).

Another exciting discovery in our study is the distinct function Oct4/Sox2 is playing between different gene clusters. Particularly, these binding sites are not functional at the Non-Dynamic genes Hnrnpr and Pds5a; and mediate transcriptional repression at the Silent genes *Oxgr1*, *Gnrhr*, *Uba7*, and *Lax1*. Understanding the selective mechanisms

that determine the Oct4/Sox2 functions will be necessary to uncover more mechanistic insights into their role in pluripotency. The comparative genetic and genomic analyses act as the initial step to investigate possible mechanisms distinguishing the functions of Oct4/Sox2 binding. A few studies have described the association of Oct4 with repression complex or the cooperation of Oct4/Sox2 with tissue-specific transcription factors at the somatic enhancers (Chronis et al., 2017; Liang et al., 2008). However, we did not obtain any impressive findings at this stage. An essential effort for the future direction should focus on the effort to decipher the selective mechanisms for Oct4/Sox2-mediated gene silencing will shed insight for pluripotency regulation.

REFERENCE

Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. Methods.

van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: A method to study the three-dimensional architecture of genomes. J. Vis. Exp.

Chen, H., and Larson, D.R. (2016). What have single-molecule studies taught us about gene expression? Genes Dev.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. Cell.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. Mol. Cell.

Liang, J., Wan, M., Zhang, Y., Gu, P., Xin, H., Jung, S.Y., Qin, J., Wong, J., Cooney, A.J., Liu, D., et al. (2008). Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. Nat. Cell Biol.