UNIVERSITÀ DEGLI STUDI DI TORINO

Department of Life Sciences and Systems Biology

Ph.D. Program in Complex Systems for Life Sciences

Cycle: XXXIII

Transcriptional and epigenetic regulation in embryonic stem cell pluripotency and lineage differentiation

Thesis's author: Guohua Meng

Tutor: Prof. Salvatore Oliviero

Code of scientific discipline: BIO/11

Academic years of enrolment: AA 2017-2020

Contents

Preface

Mouse embryonic stem (mES) cells are pluripotent cells derived from the inner cell mass of blastocyst-stage embryos and have an indefinite expansion potential to produce progeny through self-renewal or differentiation processes. Their importance to modern biology and medicine derives from these two unique characteristics that distinguish them from all other organ-specific stem cells identified to date.

The transforming growth factor-beta (TGF-β) family of ligands plays significant roles in embryonic development, tissue homeostasis, adult immunity, and wound repair. Dysregulation of the TGF-β signaling pathway leads to severe diseases. Its key components have been revealed over the past two decades. This family of cytokines acts by activating receptor-activated Smad (R-Smad) transcription factors, which in turn modulate the expression of specific sets of target genes. Cells of a multicellular organism have the same genetic information, yet they show structural and functional differences due to differential expression of their genes. Studies have demonstrated that epigenetic regulation, an integral part of the TGF-β signaling, enables cells to sense and respond to TGF-β signaling in a cell context-dependent manner. As the central transcription factor of TGF-β signaling, R-SMAD can recruit various epigenetic regulators to shape the transcriptome.

About six decades ago, Conrad Hal Waddington proposed the epigenetic landscape model of development. He envisioned development starts with a marble representing a totipotent zygote cell at the top of a hill. After the marble rolls down the hill, it enters a series of furrows that represent increasing restrictive and committed cell fates from a totipotent state to a pluripotent state, and then to a multipotent state, finally to a terminal state, i.e., terminally differentiated cells of adult tissue. He speculated that the different cell populations through the trip in the valley are further associated with specific epigenetic status. Here, I focus on epigenetic regulatory mechanisms in the TGF-β signaling pathway during mammalian development. The crosstalk between TGF-β signaling and the epigenome could serve as a versatile fine-tuning mechanism for transcriptional regulation during embryonic development and the progression of diseases.

In the first part of my Ph.D., detailed in Part One of the present thesis, I have focused my attention on the transcriptional regulation of Smad7 in mouse embryonic stem cells. Transforming Growth Factor beta (TGF-β) pathway activation in embryonic stem cells (ESCs), derived from the pre-implantation blastocyst, shows an opposed role in maintaining pluripotency and inducing differentiation. The activation of the TGF-β branch leads to the recruitment of SMAD transcription factors and translocation into the nucleus, where Smad7 is

transcriptionally activated. Smad7 is a major inhibitory SMAD protein that has been reported to promote ESC self-renewal and induce pluripotency stem cell reprogramming. Consequently, the nuclear localization is consistent with the hypothesis that Smad7 acts as a transcription factor involved in regulating target gene expression, but very little is known about it. Here, for the first time, we mapped target gene promoter occupancy of Smad7 on a genome-wide scale in mouse ESCs by using a rigorous streptavidin-based genome-wide ChIP-Seq approach with biotin-tagged Smad7 (Bio-Smad7). The identified sites are mainly enriched for active histone modifications. Furthermore, gene ontology analysis reveals that the newly identified Smad7 target genes are enriched for genes associated with cell metabolism and gene expression regulation and genes with molecular functions involved in nucleic acid binding and transcription regulator activity. These results provide insights into the contribution of Smad7 in maintaining stem cell self-renewal and keeping these cells in an undifferentiated state.

In my second part of my Ph.D., detailed in Part Two of the present thesis, I have focused on exploring the mechanism of *de novo* DNA methylation in stem cell pluripotency and Meso-endoderm differentiation. The establishment of DNA methylation patterns requires *de novo* methylation that occurs predominantly during early development and gametogenesis in mice. *De novo* methylation during early embryogenesis and differentiation is catalyzed by DNA methyltransferase 3a (Dnmt3a) and 3b (Dnmt3b), and the absence of DNA methylation leads to ectopic gene activation in the embryo. Here we observe that Dnmt3b depletion does not affect the transition from the ESCs to the post-implantation epiblast. Interestingly, Dnmt3b in the epiblast is also targeted to the lineage-specific genes. Subsequently, we find defects in meso-endoderm progenitor differentiation in Dnmt3b knockouts (KO) that the silencing of Sox2 can rescue. Since Sox2 has been demonstrated to play antagonistic roles with Brachyury in the specification of neural and mesodermal fates. Notably, neuro-ectoderm-specific genes are upregulated in Dnmt3b KO during the meso-endoderm differentiation. These results indicate that Dnmt3b has defined the methylation lineage differentiation pattern in the epiblast, which is essential for terminal lineage differentiation in mouse embryonic stem cells.

有志者事竟成

Where there is a will, there is a way!

To my families

and all the group members

Part One

Smad7-mediated transcriptional regulation in embryonic stem cells

Chapter 1

INTRODUCTION

1.1 TGF-β family signaling in embryonic stem cells

1.11 TGF-β family signaling

The transforming growth factor-β (TGF-β) family of ligands, which include the TGF-βs, the activins, NODAL, bone morphogenetic proteins (BMPs), and growth and differentiation factors (GDFs), elicit their pleiotropic effects on cell behavior by signaling to the nucleus and initiating new programs of gene expression. As shown in Figure 1.1 (A, B), ligand binding activates pairs of type I and type II serine/threonine kinase receptors, with specific combinations of receptors recognizing different ligands (Feng and Derynck 2005; Wakefield and Hill 2013). The type II receptor phosphorylates and activates the type I receptor, which activates the pathway's primary signal transducers, the receptor-regulated SMADs (R-SMADs) (Massagué 2012). The R-SMADs are divided into those activated predominantly in response to TGF-β, activin, and NODAL, which are SMAD2 and SMAD3, and those primarily activated by BMPs and GDFs, which are SMAD1, SMAD5, and SMAD8 (Miyazawa et al. 2002). Once phosphorylated, the R-SMADs form homomeric complexes and heteromeric complexes with the common SMAD, SMAD4. These complexes are thought to be predominantly trimers, although some evidence for dimers also exists (Chacko et al. 2001; Wu et al. 2001; Inman and Hill 2002). Activated SMAD complexes accumulate in the nucleus, where they bind DNA directly or indirectly via other transcription factors and regulate gene expression, both positively and negatively (Massagué 2012).

TGF-β also activates, in a Smad-independent manner (Figure 1.1 (C)), signaling pathways that are generally considered as essential effector pathways for tyrosine kinase receptors (Derynck R, 2003; Moustakas A, 2005; Zhang YE, 2009). TGF-β induces activation of these non-Smad pathways through interactions of signaling mediators with the type I or type II receptors, either directly or through adaptor proteins. Depending on the cell system, these non-Smad pathways can also be activated indirectly due to Smad-mediated changes in gene expression. TGF-β has been shown to directly activate the Ras-Raf-MEK-Erk MAPK pathway through the association of ShcA with the TGF-β receptor complex and direct tyrosine phosphorylation of ShcA by TGF-β type I receptor in response to TGF-β, taking advantage of the dual-specificity of the TGF-β receptor kinases. The phosphorylated tyrosines on ShcA then provide a docking site for the recruitment of Grb2 and Sos, and this complex initiates Ras activation leading to Erk MAPK

signaling cascade (Lee MK, 2007). TGF-β also induces p38 and JNK MAPK signaling through activation of TAK1 by the ubiquitin ligase TRAF6 that interacts with the TGF-β receptor complex (Sorrentino A, 2008; Yamashita M, 2008). Furthermore, TGF-β also regulates the activities of the small GTPase proteins Rho, Rac, and Cdc42, which regulate the cytoskeletal organization and gene expression (Bhowmick NA, 2001; Edlund S, 2002; Ozdamar B, 2005), but how receptor activation leads to signaling by small GTPases remains to be better defined. TGF-β-activated RhoA induces activation of its downstream targets ROCK and LIM kinase (Vardouli L, 2005). Finally, TβRII phosphorylates the polarity protein PAR6, which regulates the local degradation of the RHOA small GTPase that controls the assembly of tight intercellular junctions in mammalian cells (Ozdamar B, 2005). The roles of TGF-β-induced, Smad-independent signaling in stem cells are still unclear and remain to be elucidated.

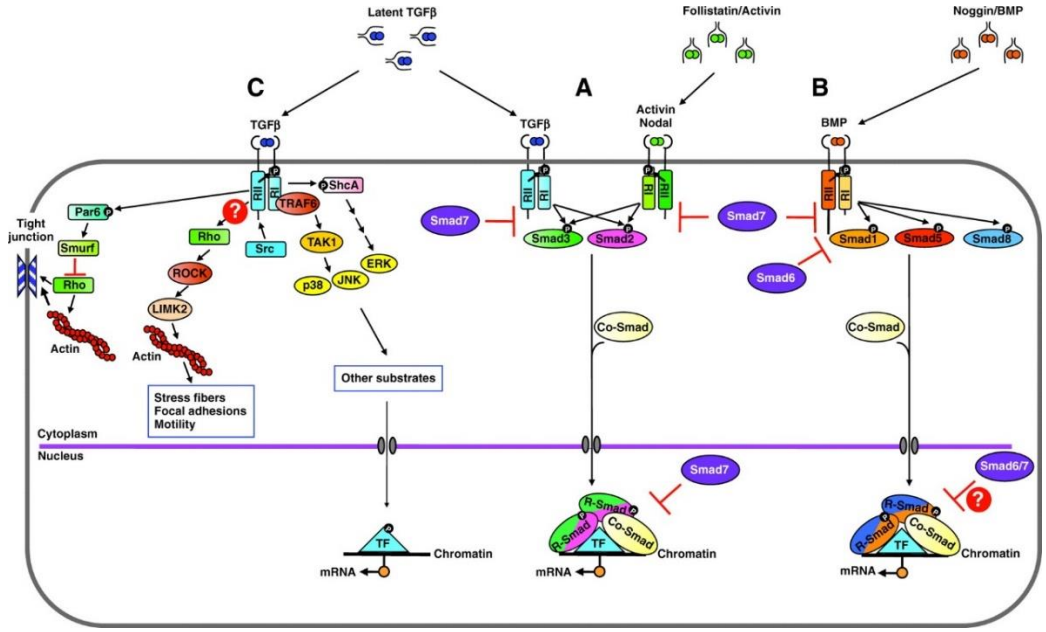


Figure 1.1 TGF-β and BMP signaling.

(A, B) The (A) TGF-β and activin/nodal and (B) BMP pathways, with their corresponding Smad proteins and mechanisms of inhibition by I-Smads (Smad6/7). The latent TGF-β complex and the extracellular antagonists, follistatin (bound to activin) and noggin (bound to BMP) are shown. (C) Non-Smad signaling pathways downstream of the TGF-β receptors [RI, TβRI (ALK5) and RII, TβRII (TGFβR2)]. The nuclear Smad complexes that lead to gene regulation are shown for each pathway. In these complexes, the Smad trimer most likely contains two R-Smad (identical or different) and one Co-Smad subunit. In addition to the major signaling pathways shown, TGF-β also activates BMP R-Smads in certain contexts. (*Moustakas A and Heldin CH, Development, 2009*)

1.12 Molecular mechanism of ESC pluripotency

The first embryonic stem cells (ESCs) were isolated 30 years ago from the inner cell mass of mouse blastocysts (Martin GR, 1981). ESCs are defined by their capacity for unlimited self-renewal, and upon differentiation, contribute to all three germ layers of the embryo proper. These cells can be derived from the blastocyst or epiblast stage of mammalian embryos and represent an actual pluripotent state (Smith AG, 2001; Hanna JH, 2010). Pluripotency of mouse embryonic stem cells (mESC) has been demonstrated *in vivo* by introducing these cells into pre-implantation embryos, where they contribute to all fetal lineages and form viable chimeric animals (Hanna JH, 2010).

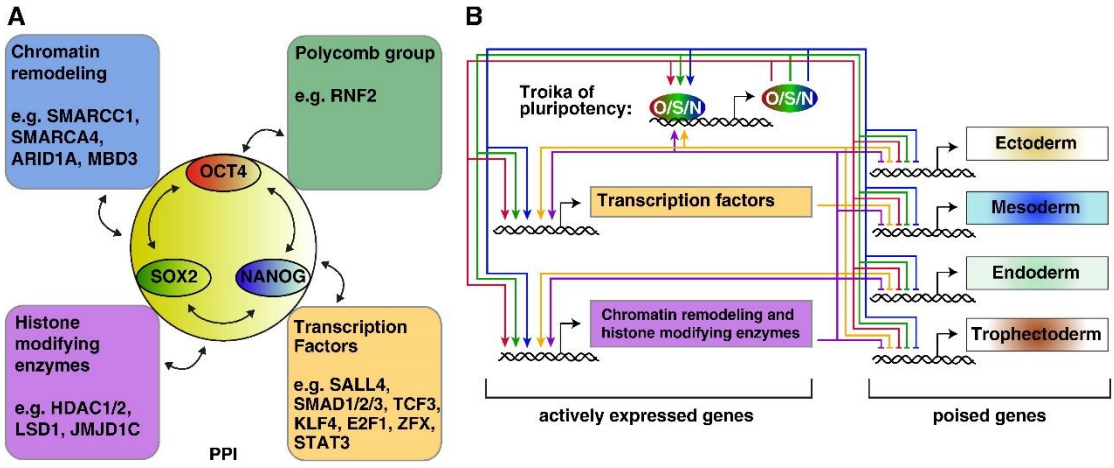


Figure 1.2 Regulation of pluripotency.

A: OCT4/SOX2/NANOG (O/S/N) interaction partners: O/S/N interact with each other and a wide variety of transcription factors, chromatin remodeling enzymes, histone modifiers, and proteins of the polycomb group. These interactions lay the base for a complex regulatory network to transcriptionally maintain pluripotency and suppress differentiation. B: Transcriptional targets of the pluripotency troika: O/S/N regulate their expression in complex feed-forward and feed-feed-back loops by directly binding their own and each others' regulatory regions (upper panel). In addition, O/S/N regulates a plethora of target genes that support the maintenance of pluripotency. Among them are transcription factors and suppressors of differentiation, such as LIN28A. In addition, O/S/N, together with other proteins, are bound to poised regulatory regions of factors important for differentiation, thereby repressing cell fate choices and stabilizing pluripotency. In summary, pluripotency is regulated by a complex network, which provides enough stability to maintain stemness and plasticity to allow cell fate specification. (*Beyer TA et al., Biochim Biophys Acta, 2013*)

Over recent years, the molecular circuitry of pluripotency has been extensively studied by employing various strategies, including genome-wide mRNA expression profiling, gene-knockdown/knockout, and protein-protein and protein-DNA interaction studies (Young RA, 2011; Jaenisch R, 2008). At the core of pluripotency lies a troika of transcription factors: OCT4 (also known as POU5F1), SOX2, and NANOG (Figure 1.2). These factors physically interact among themselves and with other factors to regulate the expression of many genes associated

with pluripotency, including their expression. This creates a complex network of feed-forward and feed-feed-back loops that ensure appropriate expression levels of key components of pluripotency. OCT4, SOX2, and NANOG are critical to driving the expression of other transcription factors involved in pluripotency, such as LIN28A, DPPA3/4, FOXD3, TERT, and UTF1. Furthermore, they suppress the expression of factors involved in differentiation such as BRACHYURY/T, FOXA2, EOMES, and GSC, and altering the balanced expression levels of OCT4 and SOX2 leads to commitment either towards neurectoderm (SOX2-positive, OCT4-negative) or mesendoderm (OCT4-positive, SOX2-negative) cell fates (Niwa H, 2000; Thomson M, 2011; Wang Z, 2012). In addition, OCT4 interacts with many regulators of the epigenetic state (e.g., histone-modifying enzymes, chromatin remodeling factors, and the polycomb repressive complex, Figure 1.2A). This indicates that OCT4 plays a crucial role in controlling the pluripotent state by controlling the epigenetic landscape (Ding J, 2012).

1.2 The structures of Smad proteins

Two highly similar amino-acid sequences compose the general structure of R-SMAD and Co-SMAD proteins at their N- and C-termini, which are termed Mad Homology 1 (MH1) and Mad Homology 2 (MH2) domains, respectively (Figure 1.3). A divergent proline-rich linker region separates the MH1 and MH2 domains with variable lengths. In the inactive state, the MH1 domain interacts with the MH2 domain, which represses the MH2 domain and MH1 domain function and binds to specific DNA sequences (TGF-β responsive sequence) in the nucleus. In contrast, the MH2 domain plays important roles in SMAD/receptor interactions, SMAD oligomerization, SMAD interaction with DNA binding partners, and transcriptional co-activators or co-repressors (Massagué et al. 2005; Miyazono et al. 2005).

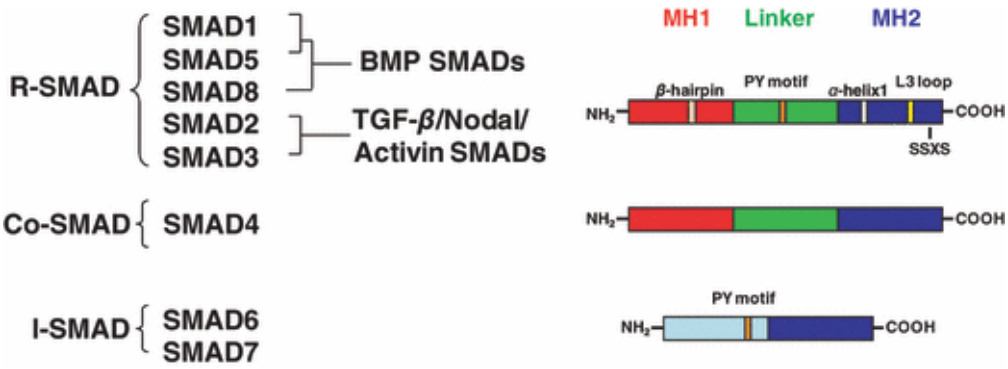


Figure 1.3. The structure of SMADs.

(*Tao S and Sampath K, Development, Growth & Differentiation, 2010*)

The Ser-Ser-X-Ser (SSXS) motif at the C-terminus is unique to R-SMADs, serving as the phosphorylation site by type I receptors (Massagué et al. 2005; Miyazono et al. 2005). The L45 loop within the type I receptor kinase domain and the L3 loop in the MH2 domains of R-SMADs determine intracellular signaling specificity induced by TGF-βs. The BMP-regulated SMADs, SMAD1, 5, and 8, are phosphorylated by BMP type I receptors (ALK-1, ALK-2, ALK-3, and ALK-6), whereas TGF-β/Activin-SMADs, SMAD2, and SMAD3 are activated by TGF-β and activin type I receptors (ALK-4 and ALK-5) and orphan type I receptor ALK-7. For the interaction of SMAD1 with ALK-1 and ALK-2, not only its L3 loop but also the α-helix 1 (H1) in the MH2 domain is required (Chen & Massagué 1999; Miyazono et al. 2005; Itoh & Ten Dijke 2007) (Figure 1.3).

The Co-SMAD, SMAD4 can form heteromeric complexes with all activated R-SMADs; it is a shared component in TGF-β, Activin, and BMP signal transduction. Gel-mobility shift assays, together with the crystal structure of the C-terminal domain of SMAD4, suggest that the SMAD complexes may be heterotrimers, composed of two R-SMADs and one SMAD4 or one R-SMAD and two SMAD4 molecules (Shi et al. 1997; Kawabata et al. 1998). Although phosphorylated R-SMADs can form oligomers and translocate into the nucleus even in the absence of Co-SMAD, the Co-SMAD stabilizes the SMAD structure oligomers and ensures the efficient transcriptional activity of the SMAD complexes (Shi & Massagué 2003).

The I-SMADs, SMAD6 and SMAD7, are distantly related members of the SMAD family. I-SMADs contain a conserved MH2 domain that interacts with type I receptors to compete with R-SMADs for receptor binding. The MH2 domain of SMAD6 binds to activated R-SMAD and prevents complex formation between R-SMAD and Co-SMAD (Heldin et al. 1997; Hata et al. 1998; Murakami et al. 2003). The expression of I-SMADs is induced by ligand stimulation. For example, the Drosophila homolog of SMAD6/7, Daughters against Dpp (Dad), is induced by Dpp (the Drosophila homolog of Bmp2/4) signal. This finding indicates that I-SMADs regulate TGF-β/BMP signaling by a negative feedback loop (Tsuneizumi et al. 1997; Christian & Nakayama 1999; Miyazono et al. 2005). Both SMAD6 and SMAD7 inhibit BMP signaling, while SMAD7 is more potent in inhibiting TGF-β/Activin signals than SMAD6. The N-terminal regions of I-SMADs show weak similarity to MH1 domains of R- and Co-SMADs and may have a role in determining signal specificity (Miyazono et al. 2005).

The proline-rich linker region of R-SMADs has consensus sequences that can be phosphorylated by MAPK (mitogen-activated protein kinase). MAPK phosphorylation inhibits R-SMADs' nuclear localization, thereby antagonizing TGF-β signaling (Kretzschmar et al. 1997). In Xenopus embryos, MAPK induced phosphorylation of the SMAD1 linker region by Insulin Growth Factor (IGF) and FGF (fibroblast growth factor) plays a vital role in neural

induction (Pera et al. 2003). Besides, the PPXY sequence (also called PY motif), which interacts with the WW domain-containing proteins such as the E3 ligase Smurf, is also present in the linker regions of most R-SMADs and I-SMADs (Kavsak et al. 2000) (Figure 1.3). The ubiquitin-proteasome pathway regulates R-SMAD proteins through association with the HECT domain ubiquitin ligases such as Smurf1 and Smurf2. Smurf1 interacts with SMAD1 and SMAD5, thereby affecting BMP responses, whereas Smurf2 interacts more broadly with different R-SMADs, allowing interference with BMP and TGF-β/activin signaling (Derynck & Zhang 2003; Miyazono et al. 2005).

1.3 Transcriptional regulation by Smads

The list of transcription factors to which Smads bind to regulate gene expression continues to grow. Nuclear Smad complexes bind with weak affinity to Smad-binding elements (SBEs) on DNA (Schmierer and Hill, 2007). Notably, the most common isoform of SMAD2 fails to bind to SBEs owing to insertion within its DNA-binding domain, which resides in the MH1 domain of all Smads. SMAD3 recognizes 5'-GTCTG-3' as its SBE. By contrast, the BMP Smads and SMAD4 recognize GC-rich sequences with less conserved motifs, which are sometimes close to an SBE. In general, recruitment of Smad complexes to chromatin is dependent on their direct interaction with transcription factors that bind to DNA with higher affinity.

Upon binding to DNA and their transcriptional partners, Smads recruit co-activators and histone acetyltransferases, such as p300, C/EBP-binding protein (CBP), and p300/CBP-associated factor (P/CAF), facilitating the initiation of transcription (Schmierer and Hill, 2007). Recent evidence has shown that p300/CBP also acetylates SMAD2/3, enhancing their DNA-binding activity in mammalian cells (Simonsson et al., 2006). Conversely, histone deacetylases inhibit SMAD1 transcriptional activity during neuronal differentiation in the mouse embryonic brain (Shakéd et al., 2008). However, direct acetylation or deacetylation of BMP-specific Smads has yet to be demonstrated.

Genome-wide screens have revealed an association between Smads and the SWI/SNF family chromatin remodeling protein Brahma-related gene 1 (BRG1; SMARCA4) and the DNA-binding proteins ETS1 and transcription factor activating enhancer-binding protein 2 (TFAP2) (Koinuma et al., 2009; Xi et al., 2008). A current model suggests that chromatin-bound Smads cannot perform transcriptional work without essential chromatin remodeling factors, such as BRG1 and the mediator component ARC105 (Schmierer and Hill, 2007). Interestingly, ARC105 localization in distinct chromatin domains is regulated by TAZ, the nuclear Smad-tethering factor (Varelas et al., 2008). These early reports open the door to future studies that might demonstrate how TGF-β alters chromatin's dynamic architecture, leading to gene-

specific transcriptional induction or repression. For the first time, such research might establish links between the epigenetic regulation of chromatin and the TGF-β pathways' function.

1.4 The role of Smad7 in embryonic stem cells

Smad7, induced by all ligands of the TGF-β superfamily, can act as a negative feedback product to inhibit TGF-β signaling (Yan X, 2011; Briones-Orta MA, 2011). Smad7 has been thought to function primarily through its inhibition of both TGF-β and BMP signaling. It has been reported that SMAD7 is sufficient to directly convert pluripotent hESCs to an anterior neural fate (Ozair MZ, 2013) (Figure 1.4). Time-course gene expression revealed down-regulation of MAPK components and combining MEK1/2 inhibition with SMAD7-mediated TGFβ inhibition promoted telencephalic conversion. Blank U, 2006 reported that the self-renewal capacity of murine hematopoietic stem cells is stimulated *in vivo* upon blocking of the entire Smad pathway by retroviral gene transfer of the inhibitory Smad7, as shown by both primary and secondary bone marrow (BM) transplantations. Besides, in mouse ESCs, an increased level of Smad7 due to loss of its E3 ligase RNF12 impairs both activin-induced anterior mesoderm formation and BMP-mediated repression of neural induction (Zhang L, 2012).

Members of the TGF-β family of proteins modulate the proliferation, differentiation, and survival of many different cell types. Neural stem and progenitor cells (NPCs) in the adult brain are inhibited in their proliferation by TGF-β and by bone morphogenetic proteins (BMPs). Surprisingly, Krampert M, 2010 demonstrated that endogenous Smad7 regulates neural stem/progenitor cell proliferation in a TGF-β and BMP-independent manner. The enhanced proliferation potential of Smad7 mutant cells was retained *in vitro* in neurosphere cultures, and a higher sphere-forming capacity and faster growth and cell cycle progression were also observed. Recently, Yu Y et al. (2017) describes an unexpected finding that Smad7 promotes self-renewal of embryonic stem cells (ESCs) in a manner independent of its inhibition on TGF-β signaling. Instead, Smad7 acts to induce activation of transcription factor signal transducers and activators of transcription 3 (STAT3) in ESCs. Smad7 activates STAT3 through its direct binding to the cytokine receptor upstream of STAT3 activation. In agreement with the role of STAT3 in maintaining ESC pluripotency, Smad7 promotes ESC self-renewal and induced pluripotent stem cell reprogramming (Figure 1.5). However, the transcriptional regulation mechanisms of Smad7 in terms of Smad7-DNA interactions in ESC are still not precise.

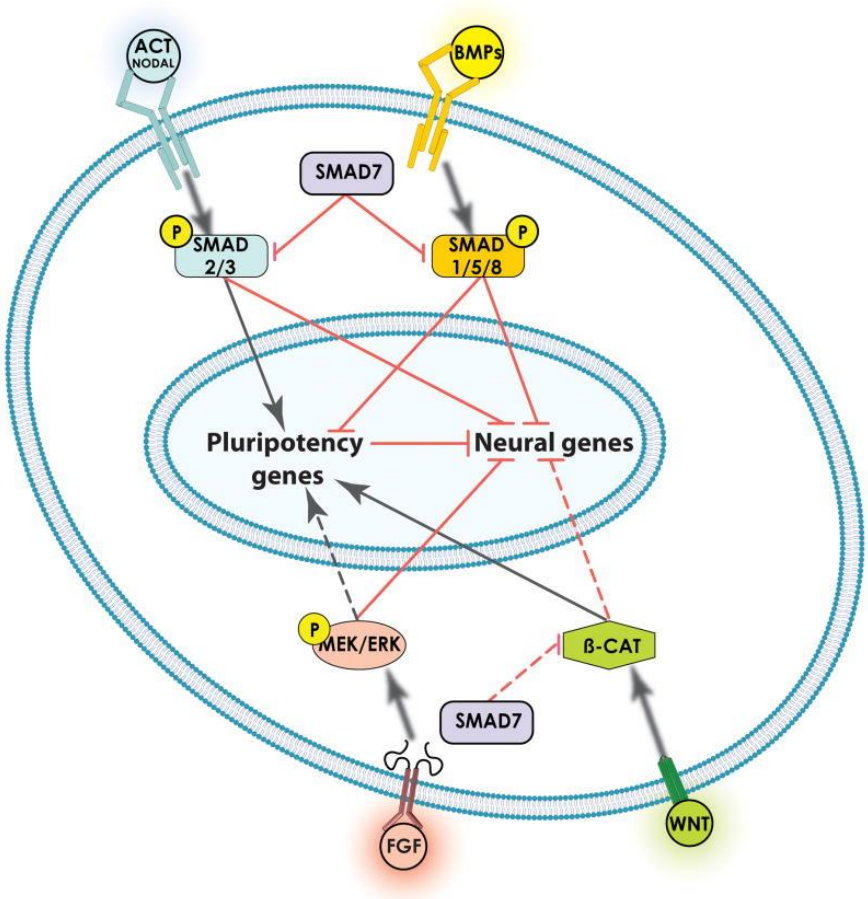


Figure 1.4. Proposed model showing that 'primed' pluripotent state requires inhibition of the 'default' neural state.

Schematic showing that the three pathways mediating pluripotency in hESCs: FGF-MEK, WNT-β CATENIN, and ACTIVIN/NODAL-SMAD2/3 may repress neural fate directly and indirectly via pluripotency genes like NANOG. This model proposes that the state of pluripotency requires inhibition of the default state of differentiation, i.e., the neural fate. Arrows represent activation, while hatches represent inhibition. Dotted lines denote postulated mechanisms from evidence in mouse ESCs.
(*Ozair MZ et al., Stem Cells, 2013*)

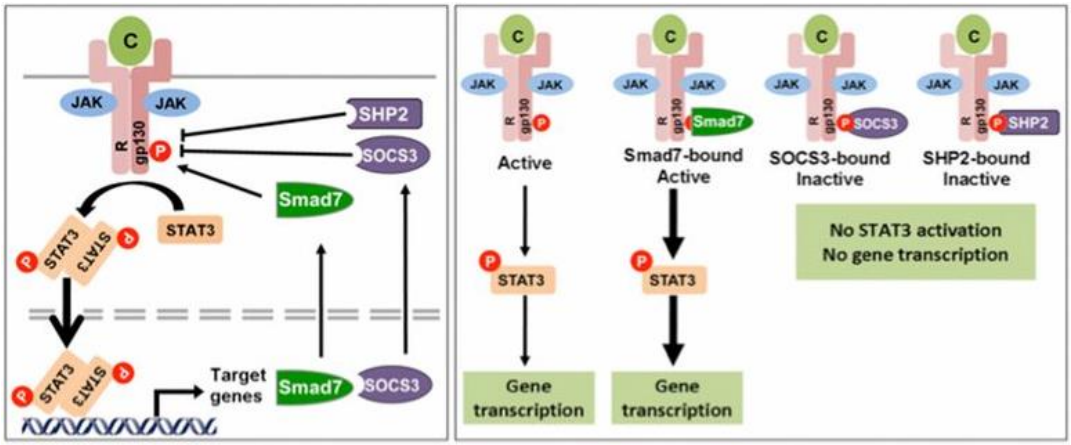


Figure 1.5. A working model for Smad7 potentiating STAT3 activation.

(Left) LIF and related cytokines (C) bind to the gp130 receptor complex. Receptor-associated JAK kinases phosphorylate STAT3 leading to STAT3 accumulation in the nucleus, where STAT3 controls the expression of target genes, including Smad7 and SOCS3. SOCS3 and SHP2 bind to gp130 to inhibit STAT3 activation. Smad7 can compete for the gp130 binding, maintaining STAT3 activation. (Right) Active and inactive forms of the cytokine-receptor–gp130 complex are shown.
(*Yu Y et al., Proc Natl Acad Sci U S A. 2017*)

1.5 Methods to study protein-DNA interactions

Traditionally, EMSA (Gaudreault M, 2009; Molloy PL, 2000) (electrophoretic mobility shift assay) and SELEX (Klug SJ, 1994) (systematic evolution of ligands through exponential enrichment) have been used *in vitro* to study protein-DNA interactions. On the contrary, ChIP is a method (Collas P, 2008; Turner FB, 2006) commonly used to determine the site-specific occupancy of genomic DNA by selected proteins *in vivo*. This technique gives a picture of protein-DNA interactions occurring in the nucleus of the living cell or tissue used. This assay's primary basis is that DNA-binding proteins (including transcription factors) in living cells can be cross-linked to their binding DNA. After shearing or cutting genomic DNA, the protein-DNA complex obtained by immunoprecipitation from cell lysates using antibodies specific to the target DNA binding protein can be used. Subsequently, the purified DNA can be hybridized with a tiled oligonucleotide microarray (ChIP-on-chip or ChIP-on-chip) (Boyer LA, 2006) to discover where the protein binds across the genome. Besides, ChIP-Sequencing (ChIP-Seq) (Valouev A, 2008) has recently emerged as a new technology that can locate protein binding sites in a high-throughput, cost-effective manner.

The critical reagent for chromatin immunoprecipitation is the immunoprecipitation antibody. However, antibodies suitable for chromatin immunoprecipitation are not available for many transcription factors. Besides, antibodies may have a nonspecific binding that is idiosyncratic to the particular antibody preparation, and the requirement for different antibodies for each

factor complicates the comparison between factors. Another method is to express a transcription factor fused to a short *bio* peptide tag specifically biotinylated by the *Escherichia coli* (*E. coli*) enzyme *BirA* (Beckett et al., 1999; de Boer et al., 2003).

In detail, *in vivo* biotinylation is based upon a short 'biotinylation peptide' (Schatz PJ, 1993) fused to a protein of interest (e.g., a transcription factor; bioTF). The bioTF serves as an *in vivo* substrate mimic for *E. coli* biotin holoenzyme synthetase (*BirA*), an enzyme that performs highly selective biotinylation of the fusion protein. In mammalian cells, plasmid expression vectors carrying the bioTF and *birA* genes can be used to obtain high-level production of soluble bioTF and *BirA* proteins. Under appropriate culture conditions, the bioTF protein produced by this system is entirely biotinylated (Figure 1.6).

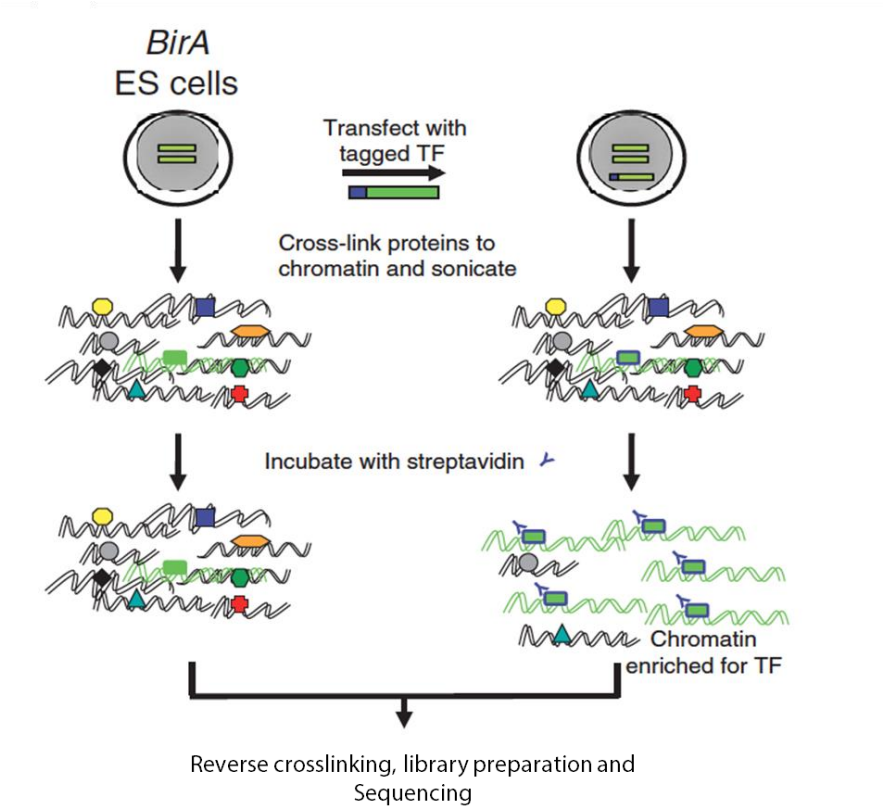


Figure 1.6. A diagram of biotin chromatin immunoprecipitation (BioChIP).

ES cells that express the biotin ligase gene BirA are transfected with an expression construct that encodes a tagged version (blue) of the protein of interest (green). Chromatin is then prepared from the BirA line and the derivative line expressing the protein of interest, except that sonicated chromatin is incubated with streptavidin coupled to a solid phase. Following the collection and washing of streptavidin-coupled chromatin, protein-DNA cross-links are reversed, and the purified DNA is submitted for ChIP-seq library construction and high-throughput sequencing. (*Modified from Chambers I & Tomlinson SR. Development. 2009*)

Biotinylation offers several advantages over traditional immunoaffinity approaches to protein complex purification. First, the high affinity of biotin for streptavidin (10^{-15} M kDa) allows efficient purification of the biotinylated protein. Second, for ChIP applications, the high biotin-streptavidin affinity allows very high stringency washing conditions (2% (vol/vol) sodium dodecyl sulfate (SDS)), thus reducing background binding that may be observed with other affinity tags or native antibodies. Third, there are few naturally biotinylated proteins, thus reducing the possibility for cross-reaction. Fourth, the approach avoids the need to generate protein-specific antibodies, which often may cross-react with other cellular proteins. In cases where investigated, biotinylation of a tagged transcription factor has not significantly altered its DNA-binding properties *in vivo* and subnuclear distribution (de Boer E et al., 2003). Therefore, it offers a unique methodology to study protein-DNA interactions simultaneously. However, potential limitations do exist: first, additional time and effort are required to establish cell lines for *in vivo* biotinylation compared with direct antibody-based immunoprecipitation; second, similar to any other overexpression system, ectopic expression of a protein significantly beyond endogenous levels may result in spurious protein complexes and increase nonspecific DNA binding. Therefore, it is crucial to select cell lines expressing the lowest level of the biotinylated proteins required for analysis (typically below the endogenous target protein level).

1.6 Aim of this study

Smad7 plays distinct roles in various cells and is essential for ESC self-renewal and pluripotency maintenance. As described in the introduction of this work, as a member of the TGF-β signaling pathway, Smad7 has the potential to regulate the target gene expression in a TGF-β dependent or independent manner. The hypothesis is that Smad7 may act as a transcription factor involved in regulating target gene expression in ESC, but the genome-wide binding pattern of Smad7 has not been thoroughly investigated. In this work, I used a rigorous streptavidin-based genome-wide chromatin immunoprecipitation approach with biotin-tagged Smad7 (Bio-Smad7 ChIP-seq) to systematically identify Smad7 binding sites in mouse ESCs. Following the integrative analyses of chromosomal targets and global gene expression profiles following knockdown (KD) of Smad7, this work will demonstrate the transcriptional regulation mechanisms of Smad7 in ESC.

Chapter 2

RESULTS

2.1 Smad7 regulates the genes involved in ESC self-renewal and pluripotency maintenance

Smad7 has been demonstrated to promote mouse ESC self-renewal and pluripotency maintenance and function to attenuate mouse ESC differentiation (Yu Y et al., 2017). We asked whether Smad7 is involved in transcriptional regulation in mouse ES cells. Here, we investigated whether Smad7 can regulate the gene expression by first performing knockdown (KD) of Smad7 using short hairpin RNAs (shRNAs) interference. We confirmed Smad7 knockdown by RT-qPCR, western blot and immunofluorescence assay (Figure 2.1A, D, E, F). Furthermore, compared with the typical round colony morphology of control ES cells, Smad7 KD in ESCs resulted in compressed cell morphology, monolayer growth, and a decreased alkaline phosphatase (AP) activity (Figure 2.1B and 2.1C), indicating exit from self-renewal. To investigate to what extent Smad7 KD affects the global transcriptome in ES cells, we then performed RNA-seq on Smad7 KD samples and used four biological replicates in this experiment. We found 1188 differentially expressed genes, among which 561 genes significantly downregulated, and 627 genes upregulated considerably (Figure 2.1G). Notably, gene ontology analysis showed that the upregulated genes involved in cell migration regulation, animal organ morphogenesis, and cell differentiation (Figure 2.1H top). In contrast, the downregulated genes were enriched in the cellular metabolic process, response to LIF, and cell cycle process (Figure 2.1H bottom).

In particular, we found that several genes (i.e., Celsr1, Fgf8, Id2, Smarca2, Wnt6, Wnt9a, Wnt11) involved in neural system development significantly upregulated, and the genes (i.e., Cxcr4, Gata6, Kdr, T, Wls) related to meso-endoderm differentiation significantly increased. The genes involved in the TGFβ/SMAD signaling pathway (i.e., Bmp1, Lefty1, Serpine1, Tgfbr2, Tgfbr3) likewise upregulated. Besides, we found that the epithelial-mesenchymal transition (EMT)-associated genes, such as Nfatc1, Serpine2, Lrp6, Col4a1, Eng, Ecm1, and Notch1, upregulated upon Smad7 KD. Furthermore, several downregulated genes (i.e., Pou5f1, Ccnd1, Psph, Ptma) were involved in self-renewal and pluripotency maintenance. Thus, our results demonstrate that Smad7 does regulate the expression of the genes involved in ESC self-renewal and pluripotency maintenance.

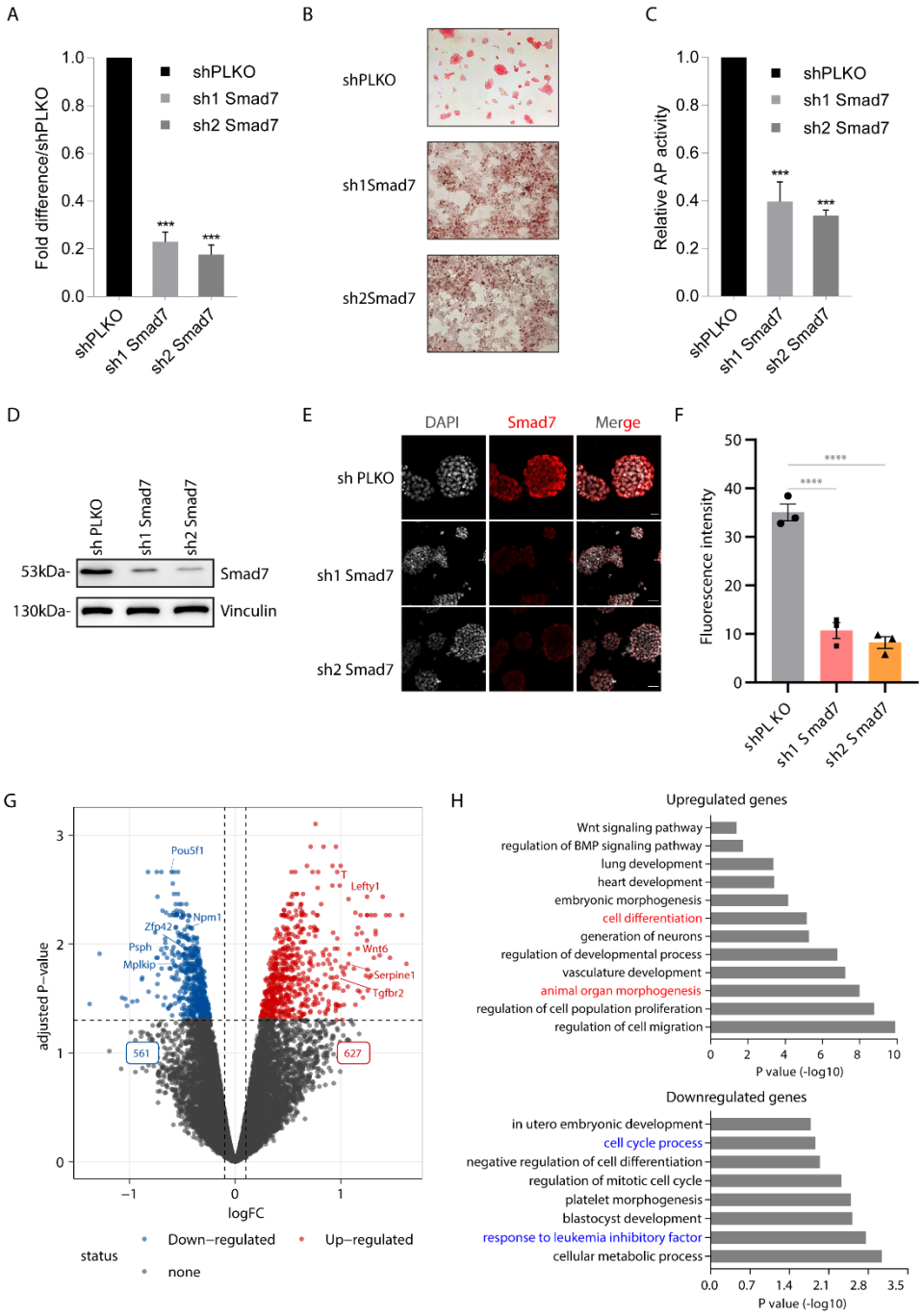


Figure 2.1. Smad7 regulates the expression of the genes involved in pluripotency maintenance in ESCs.

A. qRT-PCR analysis of Smad7 expression in ESCs. Data are shown as mean ± SEM; n = 3. The results are shown as fold difference normalized to shPLKO. ***P <0.001.
B. AP staining of ES cells expressing shSmad7.
C. Quantification of the AP positive colonies. Data are shown as mean ± SEM; n = 3. The results are shown as fold difference normalized to shPLKO. **P <0.01.
D. Western blot showing the knockdown of Smad7 after transfecting shRNA of Smad7. Cells were harvested 48 hours after the transfection.
E. Representative confocal fluorescence images of Smad7 (red) in the ESCs transfected with 2 different Smad7 shRNAs compared with the shPLKO control. The nuclei were counter-stained with DAPI (grey). Scale bar = 50 μm.

F. Quantification of fluorescence intensity of Smad7 expression in the ESCs transfected with 2 different Smad7 shRNAs compared with the shPLKO control. n = 3, ****, P < 0.0001.
G. Volcano plot showing the 1188 expressed genes. Significantly differentially expressed genes in response to Smad7 KD (FDR < 0.05 and log2 fold change > 0.2) are indicated in red and blue. While 627 genes are upregulated, 561 genes are down-regulated. The dashed line indicates the p-value significance threshold.
H. Bar graph representing the enriched GO terms of biological processes in upregulated and downregulated genes after the KD of Smad7.

2.2 In vivo biotinylation of Smad7 in mESCs

Mouse embryonic stem cells have been widely used to study the transcription factor's functions in transcription regulation, self-renewal, and pluripotency maintenance. First, by co-transfection with different combinations of the BirA and Avitag-tagged Smad7 constructs in ES cells, Smad7 with Avitag added at its N-terminus in ESCs, which were ~two kDa larger than endogenous Smad7 (Figure 2.2A), as exemplified by Biotin detected in the western blot. We observed an excellent biotinylated Smad7 expression by the co-transfection with the same amount of the two constructs (Figure 2.2B). We next generated the BirA-ES cells, which can stably express BirA, and BirA-ES cells were then used for transfection with the Avi-Smad7 construct to confirm the biotinylation (Figure 2.2C and 2.2D). By using alkaline phosphatase (AP) staining, we observed a similar AP activity in one of the BirA clones (BirA cl#1) compared to wild-type ES cells (Figure 2.2E). Then we used this BirA-ES clone for the transient transfection with Avi-Smad7 construct in further experiments. We chose transient transfection of a BirA-ES cell population to avoid the risk of isolating cell variants during cloning. This procedure took advantage of the high efficiency (75-95%) of the transfection protocol that we used (See EXPERIMENTAL PROCEDURES).

To determine the physiological level of exogenous Smad7 expression, we assessed Smad7 expression levels in individual cell nuclei of Avi-Smad7 transfected cells by Smad7 and Biotin immunofluorescence intensities (Figure 2.3A). The majority of transfected cells had a similar nuclear Smad7 expression compared to untransfected cells (Figure 2.3A top). In addition, a significant Biotin expression in the nuclei was noted in the Avi-Smad7 transfected cell population (Figure 2.3A bottom). To further delineate the subcellular localization of Smad7 in mESCs, BirA-ES cells were transfected with the Avi-Smad7 construct and harvested to analyze subcellular fractionation. As shown in Figure 2.3B, the biotinylated Smad7 highly enriched in the chromatin in BioSmad7-expressed samples compared to control samples (no BioSmad7).

Next, to confirm that exogenous Smad7 expression does not interfere with the ESC state, we analyzed the expression levels of ESC core pluripotency factors (Nanog, Pou5f1, and Sox2) and ESC naïve state markers (Zfp42 and Prdm14). RT-qPCR analysis showed that cells expressing BioSmad7 are functionally equivalent to wild-type ES cells in stem-cell marker gene

expression (Figure 2.3C). Altogether, this approach can facilitate the analysis of chromatin-templated events by characterizing the entire set of Smad7 binding sites in mESCs.

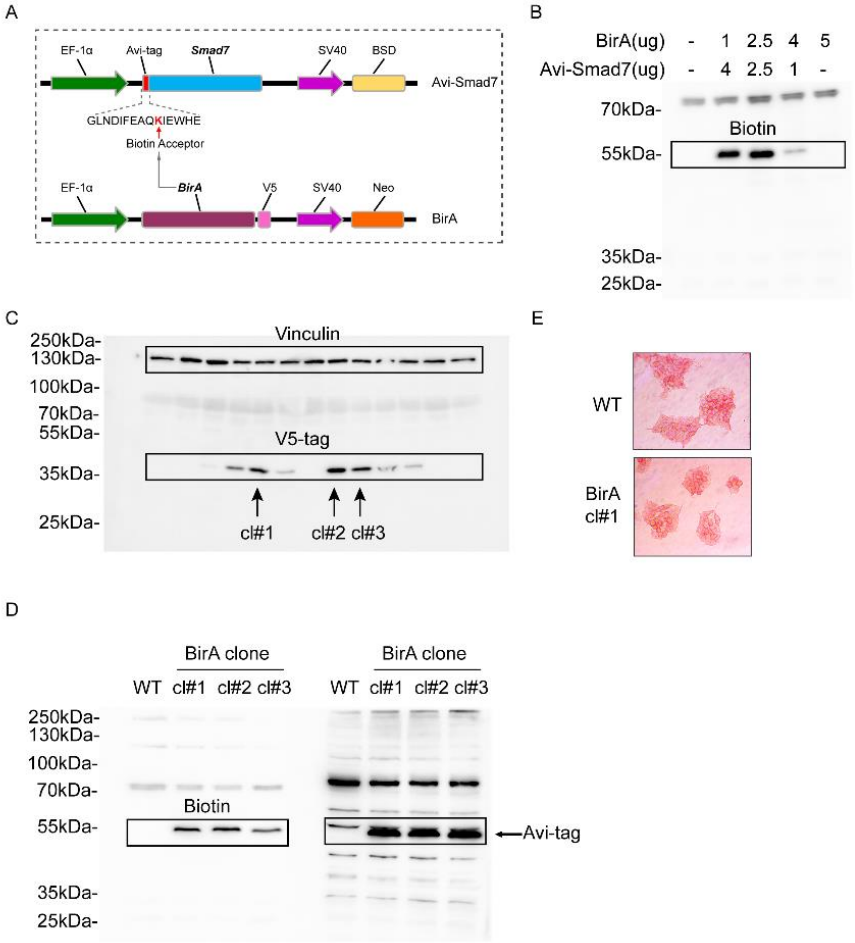


Figure 2.2. Characterization and functional validation of BirA-ES stable clone.

A. Schematic of the two components of the *in vivo* biotinylation system: an Avi-Smad7 vector and a biotin ligase BirA.
B. Western blot using an anti-Biotin antibody comparing the expression of biotinylated Smad7 with different combinations of the transfected expression vector. Five ug vector was used for each transfection.
C. Western blot using an anti-V5-tag antibody screened the expression of BirA in the stable clones. Vinculin was used as a loading control.
D. Western blot showing the overexpression of Smad7 after transfecting Avi-Smad7 vector in three different BirA-ES clones. Cells were harvested 30 hours after the transfection. Biotin and Avi-tag antibody were used to determine the transfection efficiency.
E. AP staining of WT and BirA ESCs.

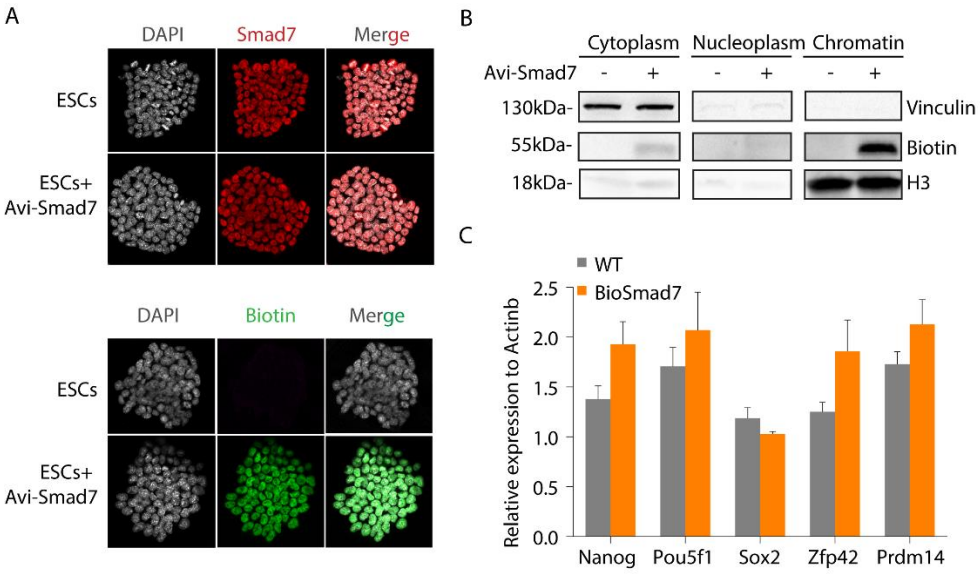


Figure 2.3. Characterization and functional validation of biotinylated Smad7 in ESCs.

A. Immunofluorescence assay of Smad7 (up) and Biotin (down) in the ES cells transfected with Avi-Smad7 compared with the untransfected control. Nuclei were stained with DAPI.
B. Western blot analysis using an anti-Biotin antibody compared the expression of exogenous (bio-tagged) Smad7 in three cells' fractions. BirA cells and Avi-Smad7 transfected cells were used in this experiment. Vinculin was used as a cytoplasm marker, and H3 was used as a chromatin marker.
C. qRT-PCR analysis of Nanog, Pou5f1, Sox2, Zfp42, and Prdm14 in Avi-Smad7 transfected cells (BioSmad7) in comparison to WT ES cells. The expression level was normalized to Actinb. Data are shown as mean ± SEM; n = 3.

2.3 Genome-wide identification and characterization of Smad7 binding sites in ESCs

To elucidate the transcriptional regulatory mechanisms of Smad7 in cell pluripotency maintenance and transcriptional regulation of mouse ESCs, we tried to identify the direct targets of Smad7 by BioChIP-seq in undifferentiated E14 ES cells. We performed the BioChIP-seq experiments in two independent biological replicates. One day after transfection with a biotinylated Smad7 expression vector for 6 hours, we processed the cells for chromatin fragmentation using the whole lysate of formaldehyde-crosslinked cells, yielding chromatin fragments with an average DNA length from 100 bp to 400 bp (Figure 2.4B). Meanwhile, we performed western blotting to confirm the equivalent biotinylated and endogenous Smad7 expression in the two replicates experiment (Figure 2.4A). Upon *in vivo* biotinylation of Smad7 by the biotin ligase BirA, we isolated the genomic locus-associated macromolecules by high-affinity streptavidin purification under stringent wash conditions (See EXPERIMENTAL PROCEDURES). Significant enrichment of the DNA recovered in the Bio-ChIP experiment was observed (Figure 2.4C). We then submitted the ChIP DNA and the Mock control DNA to ChIP-seq library preparation and high-throughput sequencing analyses. As shown in Figure 2.4D and 2.4E, we obtained the DNA length of around 300bp in our ChIP-seq libraries and the successful ChIP-seq signal enrichment in the two replicates.

The principal component analysis (PCA) exposed a close relationship in the binding patterns of Smad7 between the two replicates, whereas a diverse binding pattern in the Mock (Figure 2.5A). Following the ENCODE ChIP-seq pipeline, we mapped global binding targets of Smad7, and a total of 928 ChIP-seq peaks were identified in the pooled replicates. Overall, 48% of the BioSmad7-bound regions were located in promoter regions (Figure 2.5B), indicating that Smad7 regulates gene expression via binding to core promoters of its targets. Then, an additional 39% were located in intergenic regions (21%) or introns (18%), suggesting that Smad7 may regulate gene expression via binding to distal regulatory elements (Figure 2.5B). Moreover, the motif discovery analysis on the Smad7 ChIP-seq dataset showed that enriched Smad7 binding sites were for ETS transcription factor family motifs (Figure 2.5C).

Further gene ontology (GO) analysis revealed that the identified Smad7 target genes are involved in the stem cell population maintenance, cellular response to LIF, cellular metabolism, and molecular function involved in DNA binding, transcription coregulator activity, as well cellular component enriched in the nucleus localization, transcription regulator complex (Figure 2.5E). Altogether, our studies reveal the mode of target gene regulation of Smad7 in mESC and the putative involvement of Smad7 in stem cell pluripotency maintenance.

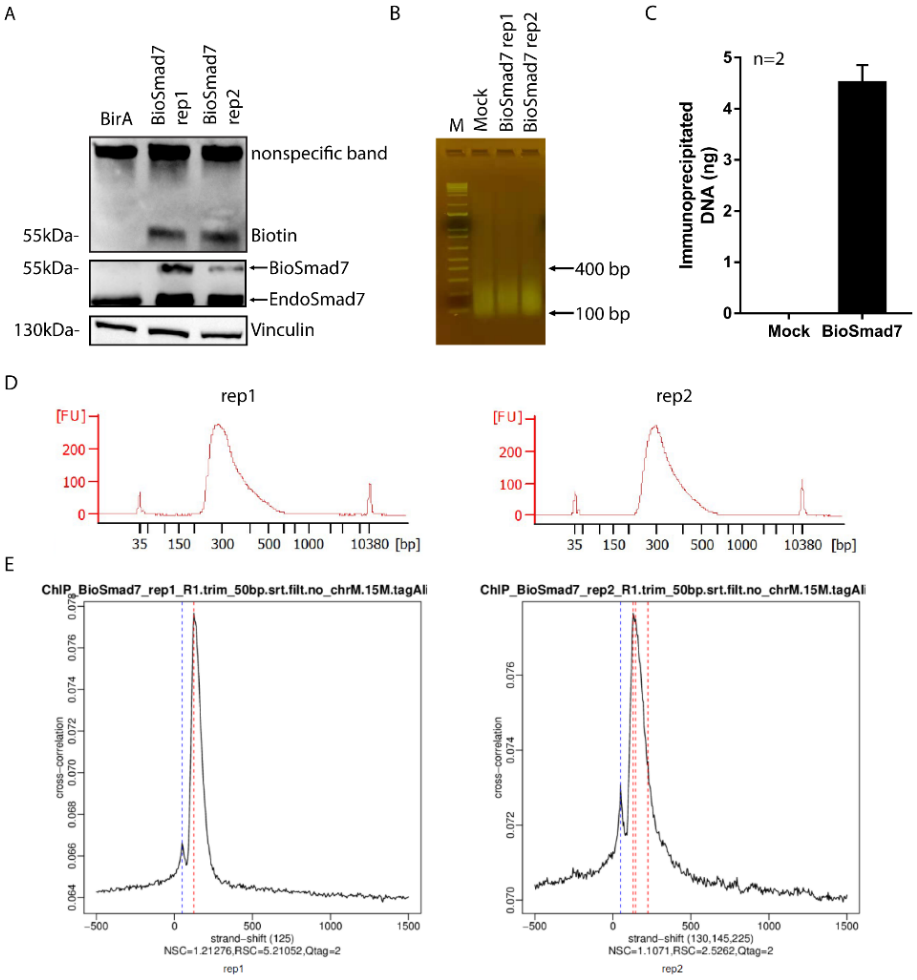


Figure 2.4. An overview of the experimental procedures for BioChIP of Smad7 in mESCs.

A. Western blot shows the *in vivo* biotinylation of Smad7. Biotin antibody was used to determine the expression of biotinylated Smad7. The native antibody was used against both endogenous and biotinylated proteins. The subendogenous level of the biotinylated Smad7 (BioSmad7) and the expression level of the endogenous Smad7 (EndoSmad7) are indicated. Vinculin was used as a loading control. Two replicates were used in this experiment.
B. Agarose gel electrophoresis shows the sonicated chromatin extracts with reversed crosslinking. 1 ml of nuclear extract was sonicated for 20 minutes using a Pico bioruptor with the "HIGH" setting, cycled 30 sec on and 30 sec off. The size of the sonicated chromatin ranges from 100 bp to 400 bp.
C. The bar graph shows the immunoprecipitated DNA after the BioChIP assay. Mock (BirA cells) was used as a negative control. Two replicates were used here.
D. Two replicates of the BioChIP libraries were assessed on a bioanalyzer.
E. Strand cross-correlation analysis shows the enrichment of the BioChIP peak signals and the fragment length.

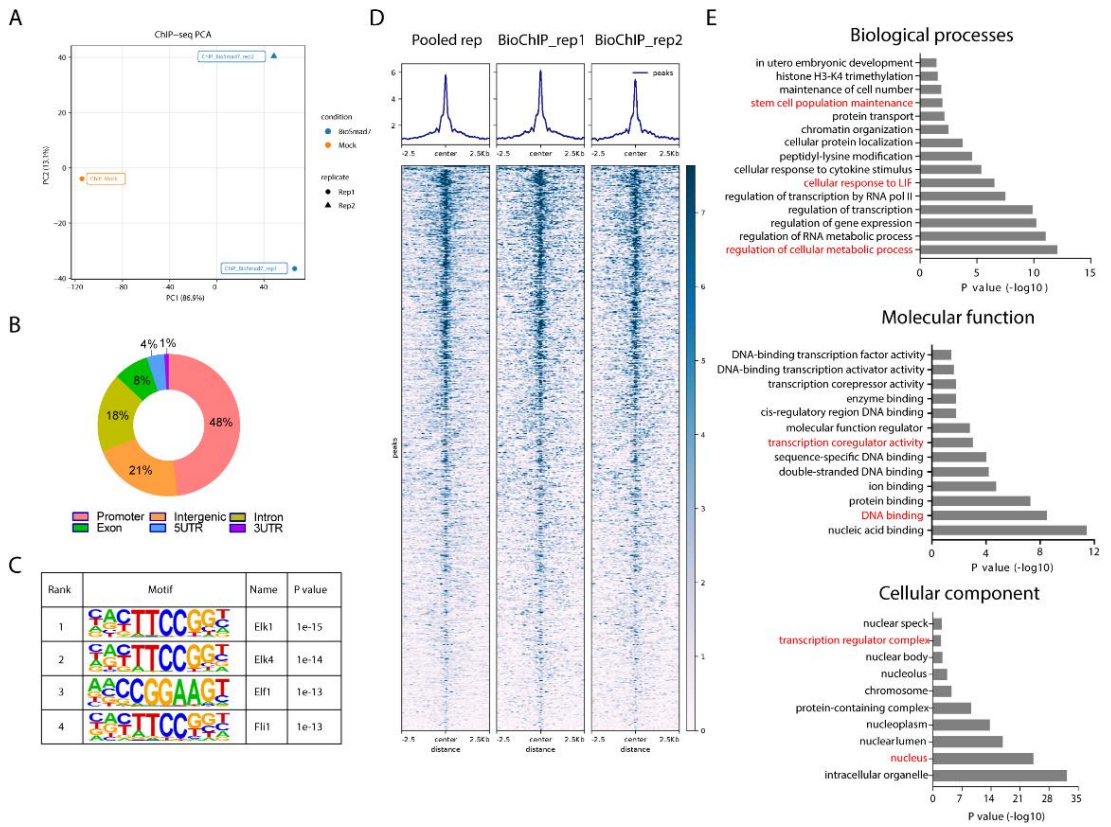


Figure 2.5. Statistics on the Smad7 ChIP-seq analysis.

A. Principal component analysis (PCA) analysis showing two replicates of BioChIP-seq of Smad7 and the Mock. Note that the first PCA component (PC1) explained > 86.9 % of the variation, and the PC1 values for the two replicates were similar.
B. A pie chart is presenting the distribution of Smad7 binding sites in candidate cis-regulatory elements (cCRE). This data is corresponding to Figure 3C.
C. The consensus sequences of Smad7 motif and sequences that are the most enriched in Smad7 binding sites.
D. Heatmaps show the colocalization of Smad7 ChIP-seq peaks in pooled replicate, replicate 1 and replicate 2. Occupancy signals within ± 2.5 Kb of the center of Smad7-binding sites are shown.
E. Bar graphs showing enriched GO terms of biological processes, molecular function, and cellular component of Smad7 target genes.

2.4 Smad7 binds to enhancers and promoters of TF and pluripotency genes on the genome

We next sought to understand the molecular mechanism by which Smad7 promotes pluripotency and self-renewal. We generated heatmaps of the identified binding sites within 2.5 kb up- or down-stream of the transcription start sites (TSSs), and the Smad7 binding sites were localized around the TSSs (Figure 2.6A and 2.5D). We then annotated the binding sites to the Candidate cis-Regulatory Elements (cCREs), representing DNase hypersensitivity sites (DHSs) supported by histone modifications. We found that approximately 40% of the binding sites were annotated with promoter-like signatures (PLS) (fall within 200 bp of the annotated TSSs) and 11% of targets had high DNase and H3K4me3 signals, and 13.9% of binding regions were with low DNase (Figure 2.6B). Moreover, around 23% of target genes were with enhancer-like signatures (ELS), including 13.4% proximal enhancer-like signatures (pELS) and 9.7% distal enhancer-like signatures (dELS) of the known TSSs. Besides, 7.5% binding targets were classified into DNase-only group, containing cCREs with high DNase Z-scores but low H3K4me3 and H3K27ac (Figure 2.6B).

Histone acetylation is associated with both active promoters and enhancers, so we performed clustering analysis on histone modifications-associated Smad7 binding regions with regions annotated by acetylation of histone 3 on lysine 9 or 27. Clustering results showed a strong colocalization of Smad7, H3K9ac, and H3K27ac (Figure 2.6A and 2.6C). We next analyzed the distribution of mono-methylation and tri-methylation of histone 3 on lysine 4, associated with active enhancers and promoters, respectively. We found that Smad7 also strongly colocalized with H3K4me1 and H3K4me3 (Figure 2.6A and 2.6C). These results are consistent with the target occupancy patterns of many ESC core transcriptional factors and regulators.

Besides, we performed a similar clustering analysis on histone modifications-associated Smad7 binding regions with regions annotated by another active histone modification (H3K36me3), repressed histone modification (H3K9me3), and “open” chromatin signature (DNAse) (Figure 2.6C). Clustering results showed strong colocalizations of Smad7, H3K36me3, and DNAse (Figure 2.6A and 2.6C) and a weak correlation of Smad7 and H3K9me3 (Figure 2.6C). Consequently, this result demonstrated that Smad7 binding targets preferentially localize at active enhancers and promoters in ESCs.

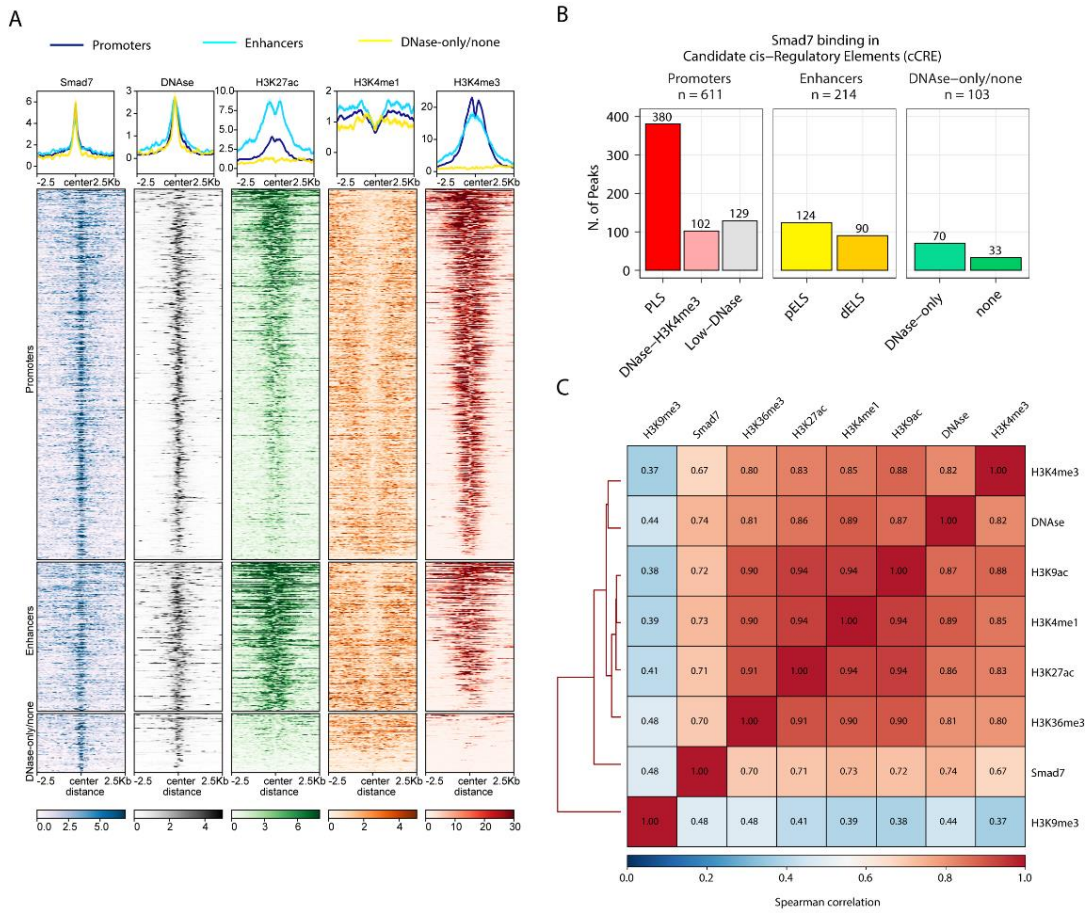


Figure 2.6. Mapping of global binding sites of Smad7 unveils that Smad7 activates target genes by occupying their promoters and enhancers.

A. Heatmaps show the colocalization of Smad7 and DNAse, H3K27ac, H3K4me1, H3K4me3. Occupancy signals within ± 2.5 Kb of the center of Smad7-binding sites are shown. Note the presence of Smad7 both at active enhancers (H3K4me1 positive) and active promoters (H3K4me3 positive). Aggregation plots illustrating the fraction of relative midpoint positions of the ChIP-seq peaks of Smad7 are shown above the heatmaps.
B. A bar graph presenting the distribution of Smad7 binding sites in candidate *cis*-regulatory elements (cCRE). cCREs with PLS fall within 200 bp (center to center) of an annotated GENCODE TSS and have high DNase and H3K4me3 signals; The subset of cCREs-ELS within 2 kb of a TSS is denoted proximal (cCRE-pELS), while the remaining subset is denoted distal (cCRE-dELS); DNase-H3K4me3 cCREs have high H3K4me3 max-Z scores but low H3K27ac max-Z scores and do not fall within 200 bp of a TSS; DNase-only group contains cCREs with high DNase Z-scores but low H3K4me3 and H3K27ac. PLS: promoter-like signatures; pELS: proximal enhancer-like signatures; dELS = distal enhancer-like signatures.
C. Hierarchical clustering of pairwise Spearman correlation of Smad7 and ChIP-seq datasets indicated. Colors indicate the level of correlation (red indicates perfect correlation, blue indicates ideal anticorrelation). Smad7 clusters together with active histone marks.

2.5 Smad7 shares common genomic targets with ESC core transcription factors

The core transcriptional regulatory network has been well discussed in mouse stem cell systems. Therefore, we asked whether Smad7 has a similar genome-wide binding pattern compared to other transcriptional regulators. We identified genomic targets of Smad7 by performing ChIP-seq, and we used published datasets of whole-genome binding sites of 14 sequence-specific transcription factors (TFs) ((Nanog, Oct4, Sox2, STAT3, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, and E2f1) in mouse ES cells (Chen X et al., 2008) and Smad2, Smad3, Smad4 (Aragón E et al., 2019)). Nanog, Oct4, and Sox2 are well-known pluripotency or self-renewal regulators, as well as Esrrb and Zfx. Stat3 is a critical component of the LIF signaling pathway. The expression of Tcfcp2l1 preferentially increased in ES cells (Ivanova N et al., 2006), and it acts as an Oct4-interacting ESC transcription factor. E2F1 is best known for its role in cell cycle progression regulation and has also been demonstrated to contribute to regulating a substantial percentage of the genome (Bieda M et al., 2006). Klf4 and Myc TFs are reprogramming factors that are also suggested in maintaining ES cells' undifferentiated state (Cartwright P et al., 2005; Jiang J et al., 2008). Smad complexes (Smad2, Smad3, and Smad4) are important factors in the TGFβ-SMAD pathway. To assess the genome-wide binding preferences of Smad7 versus other regulatory factors in ES cells, we calculated target co-occupancy scores for genomic regions identified as highly enriched in ChIP-seq experiments and subjected the matrix of scores to hierarchical clustering analysis. Clustering of the 15 factors based on their target co-occupancy (Figure 2.7A) showed that of the 15 tested factors, as expected, Nanog, Oct4, Sox2, and Stat3 exhibit overall similarity in their targets. We found it interesting that Smad7 consistently occupied sites within promoter regions that were only 50-100 base pairs (bp) upstream of TSSs (Figure 2.7B, 2.9B, and Supplementary Figure 1), in contrast to the more distant sites occupied by other core transcription factors (data not shown). Surprisingly, we found that targets of Smad7, Smad2, Smad3, Smad4, Klf4, Zfx, E2f1, c-Myc, and n-Myc segregate to a distinct cluster (Figure 2.7A). As examples, the binding profiles for all 15 factors at the Pou5f1 (Oct4) and Psph gene loci are shown in Figure 2.7B. These results indicate that Smad7 might interconnect with the ESC core regulatory circuitry.

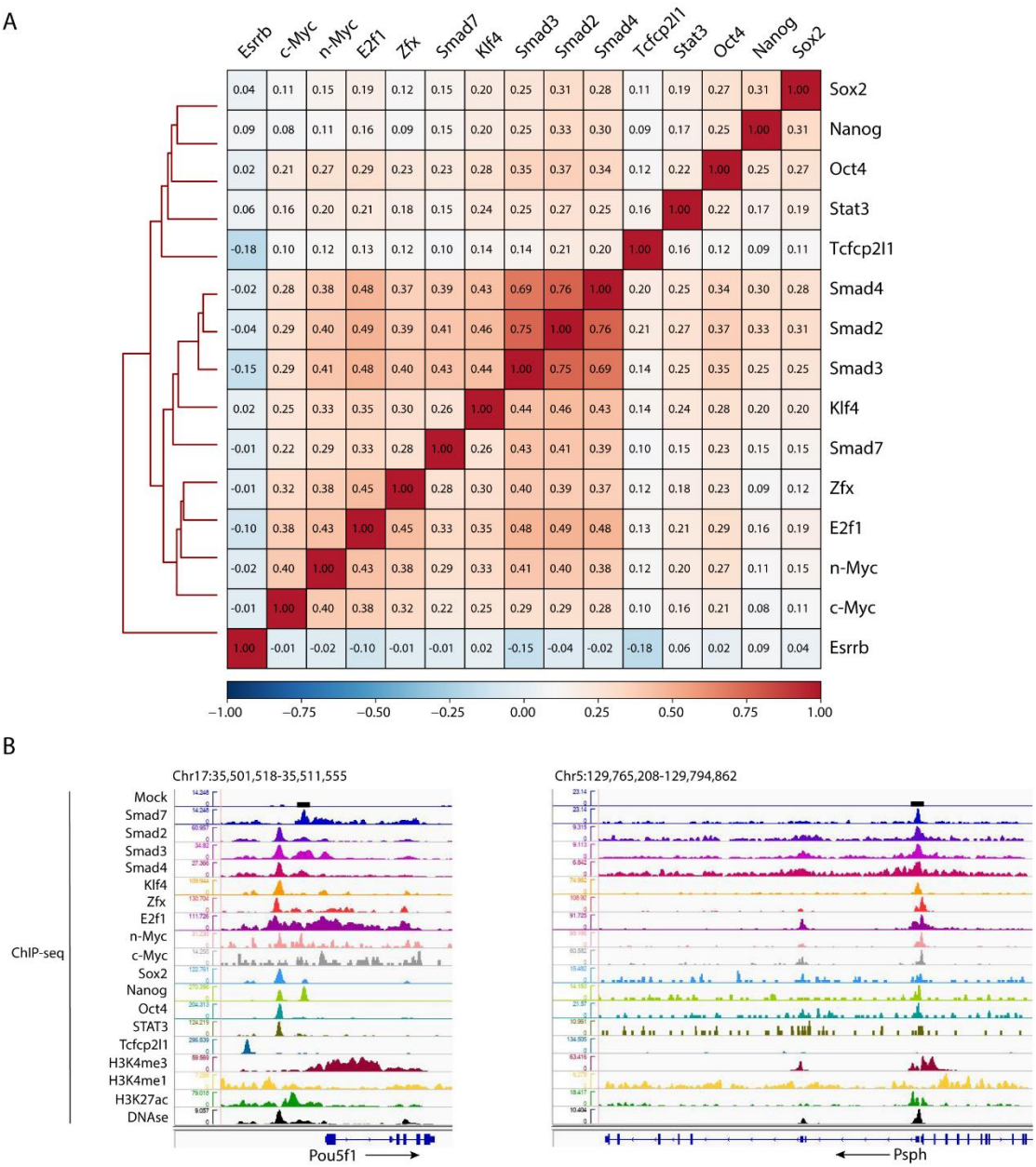


Figure 2.7. Smad7 shares common genomic targets with ESC core transcription factors.

A. Heatmap showing the similarity of their binding sites among diverse transcription factors and coregulators in mouse ES cells (Chen X et al., 2008), Smad2/3/4 (Aragón E et al., 2019) as well as Smad7 (current study) in mouse ES cells. Colors in the heatmap show the level of overlap for each pair of samples (red, all binding sites overlapped; orange, overlap expected by chance; blue, mutually exclusive binding).

B. Genome-wide mapping of 15 factors in ES cells. Transcription binding sites profiles for the sequence-specific transcription factors and mock ChIP control at the Pou5f1 and Psph gene loci are shown. Smad7 ChIP-seq data is from current study, and the ChIP-seq data of other 14 factors are from Chen X et al. (2008) and Aragón E et al. (2019).

2.6 Smad7 directly regulates the target genes expression in ESC

To verify Smad7 does bind at these genes, we performed the Bio-ChIP assay on Smad7 followed by RT-qPCR analysis to reveal the binding activity of Smad7. We noted significant enrichment of Smad7 by Bio-ChIP at the promoter regions of the target genes identified by Bio-Smad7 ChIP-Seq compared with Mock (Figure 2.8B). Then, to determine the structural features for the Smad7-DNA interaction, we generated the Smad7 mutant (Smad7Δ407), which lacks 19 amino acids at its C-terminus and contains an Avi-tag at its N-terminus (Figure 2.8A top). Surprisingly, subcellular fractionation results showed that the expression of Smad7Δ407 dramatically decreased in nucleoplasm and chromatin fractions (Figure 2.8A bottom). Consistently, we observed significantly reduced enrichment of Smad7Δ407 at the corresponding binding regions compared to Smad7 WT by Bio-ChIP followed by RT-qPCR analysis (Figure 2.8B). This result indicates that the intact C-terminus could be potentially critical for the Smad7 binding and target gene expression regulation.

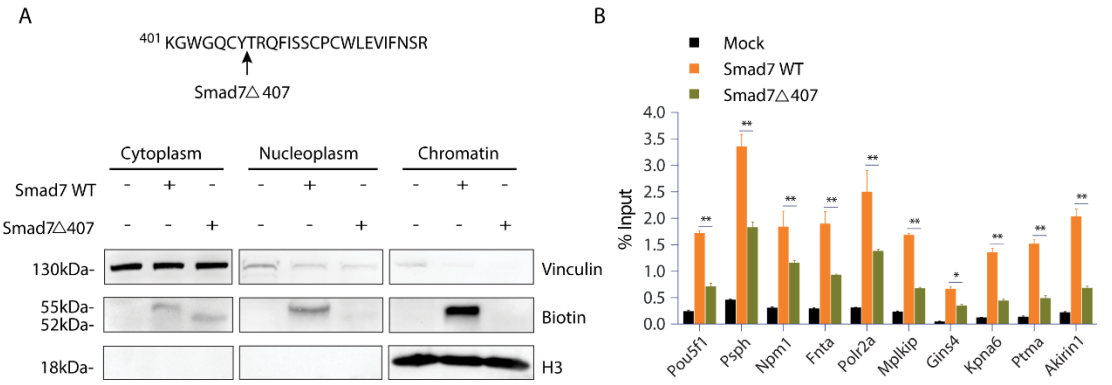


Figure 2.8. Smad7 C-terminus is vital for the Smad7 binding activity.

A. Top: Schematic of the mutant of Smad7 (Smad7Δ407), which lacks the last 19 amino acids at the C terminal. Smad7Δ407 construct also has an Avi-tag at its N terminal. Bottom: Western blot showing the subcellular fractionation of Smad7 WT and Smad7Δ407. Biotin antibody was used to compare the expression of exogenous (bio-tagged) Smad7 WT and Smad7Δ407 in three fractions of cells. BirA cells, Smad7 WT, and Smad7Δ407 transfected cells were used in this experiment. Vinculin was used as a cytoplasm marker, and H3 was used as a chromatin marker.
B. qRT-PCR analysis of BioSmad7 WT and Smad7Δ407 ChIP samples at the promoters of the indicated genes. The results are shown as the percentage (1/100) of the input. Data are shown as mean ± SEM; n = 3. $^{**}P < 0.01$. $^{*}P < 0.05$.

To determine the direct transcriptional influence of Smad7 on its binding targets, we performed RNA-seq analysis and ranked the expression values following KD of Smad7, and found the corresponding Smad7 target gene expression profile following Smad7 KD. The results showed that the Smad7 target genes were generally highly transcribed in ES cells. What's more, several Smad7 target genes significantly downregulated after the KD of Smad7 compared to the control (Supplementary Table 2). We then confirmed this result by RT-qPCR analysis showing that these target genes' expression significantly decreased after the Smad7 silencing (Figure 2.9A). For example, we observed that Smad7 bound and activated phosphoserine phosphatase Psph, the cell proliferation regulator Akirin1, cell cycle regulator Mplkip, and cell apoptosis inhibitor Ptma (Figure 2.9B, (b-e)), indicating the pluripotency and self-renewal activity of Smad7.

Interestingly, we observed that Pou5f1, a well-known pluripotency marker, was also bound and positively regulated by Smad7 (Figure 2.9B, (a)). Similarly, we also found other target genes (i.e., Fnta, Polr2a, Kpna6, Npm1, Gins4) involved in the cell cycle and cell metabolism process positively regulated by Smad7 (Supplementary Figure 1). Therefore, Smad7 up-regulates many of its target genes, and the knowledge of Smad7-mediated regulation could allow a better understanding of the biological role of Smad7 in ES cells.

Besides, several lineage-specific markers (i.e., T, Wls, Gata6, and Wnt6) significantly upregulated by KD of Smad7 were not the most vital targets of Smad7, suggesting that upregulation of lineage markers following KD of Smad7 was not due to a direct suppression by Smad7. Instead, the differentiation of ESCs upon downregulation of ESC core factors might trigger the upregulation of lineage markers.

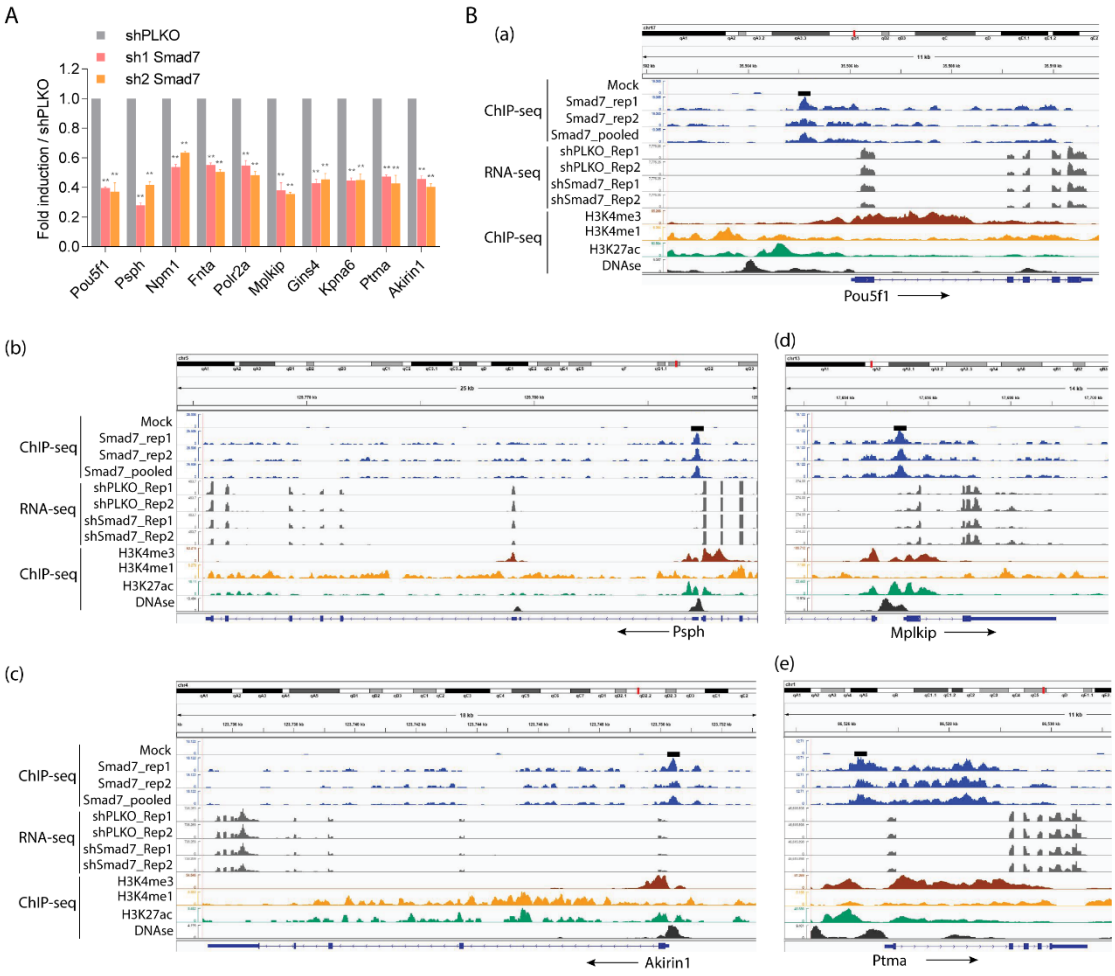


Figure 2.9. The newly identified genes by Bio-Smad7 ChIP-Seq are directly regulated by Smad7 in ES cells.

A. qRT-PCR analysis of the indicated transcripts upon a knockdown (KD) of Smad7 proteins. The results are shown as fold difference normalized to shPLKO. Data are shown as mean ± SEM; n = 3. $^{**}P < 0.01$.

B. Representative genomic occupancy profiles of genes (a, b, c, d, e) identified by Bio-Smad7 ChIP-Seq. IGV images showing the comparison of ChIP-seq peaks of Smad7 binding, the locations of histone H3 modifications, DNAse, and RNA-seq of Smad7 knockdown around the genes which were downregulated by the knockdown of Smad7. The genes and their direction of transcription are indicated by arrows. The binding loci of Smad7 are indicated by black boxes.

2.7 Smad7-mediated transcriptional regulatory networks in ESC

Through ChIP-seq and global gene expression profiling analysis, it is possible to identify the TFs whose expression was changed upon the KD of Smad7 and directly or indirectly regulated by Smad7. We classified the TFs into three groups by integrating the results. As shown in Figure 2.10, the first group contained direct Smad7 targets, including Pou5f1, Kpna6, Ctbp2, Akirin1, Zfp42, Actr2, and Myb, which were regulated (Smad7 KD versus control) and bound by Smad7 (Smad7 ChIP-seq). The second group contained indirect Smad7 targets, for example, Nanog, Lrrc34, Dnmt3b, Sox2, T, Id2, Wnt6, Jun, Ets1, Tbx3, Vegfa, and Gata6, whose expression was changed after the silencing of Smad7 and that were regulated, but not bound, by Smad7, indicating that Smad7 influences their expression through other mechanisms or factors. The third group comprised putative Smad7 targets that are bound, but not regulated, by Smad7. In particular, Several TFs involved in pluripotency maintenance as well as in TGF-β signaling were part of this group, including Sall4, Myc, Zfp398, Ctnnb1, Med6, Klf5, Dcp1a, and Brf2. These results suggest that the identified Smad7 targets are likely to play essential roles in regulating ESC pluripotency and coordinating the downstream processes by influencing other TFs. Collectively, our findings demonstrate that Smad7 binds to the genome and functions as a transcription factor in ESCs, promoting pluripotency and self-renewal (Figure 2.11).

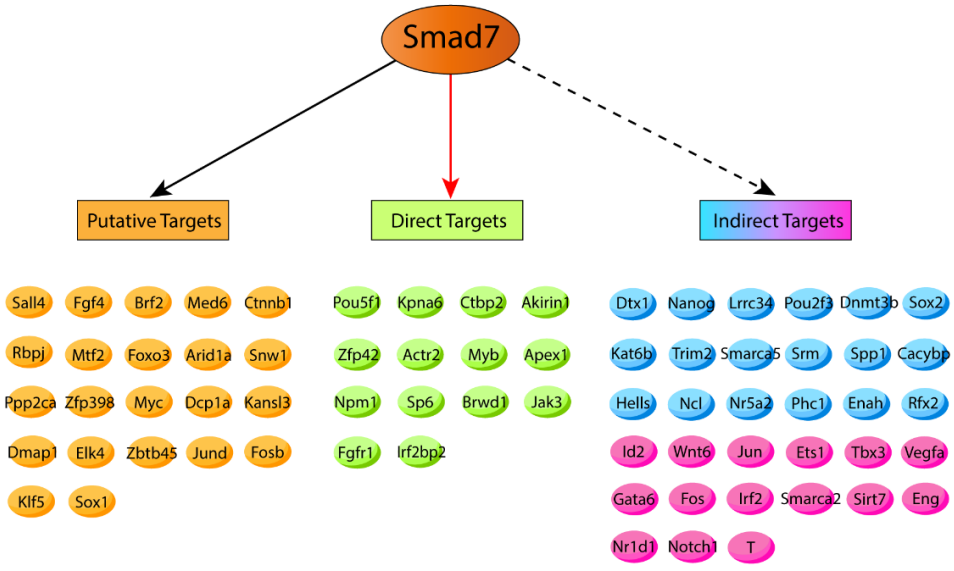


Figure 2.10. Smad7-mediated dynamic transcription network in ESC.

Smad7-dependent transcription circuitries. Smad7-dependent genes were chosen based on their regulation by Smad7 KD (FDR < 0.05), or by identified Smad7 binding. Smad7 direct target genes were regulated by direct DNA binding as well as in Smad7 KD experiments. In particular, Smad7 indirect targets (Blue: down-regulated genes, Pink: up-regulated genes) are regulated by the KD of Smad7, but not bound by Smad7. Putative Smad7 target genes were bound by Smad7 in ChIP-seq experiments but were not regulated by Smad7 in the KD experiments.

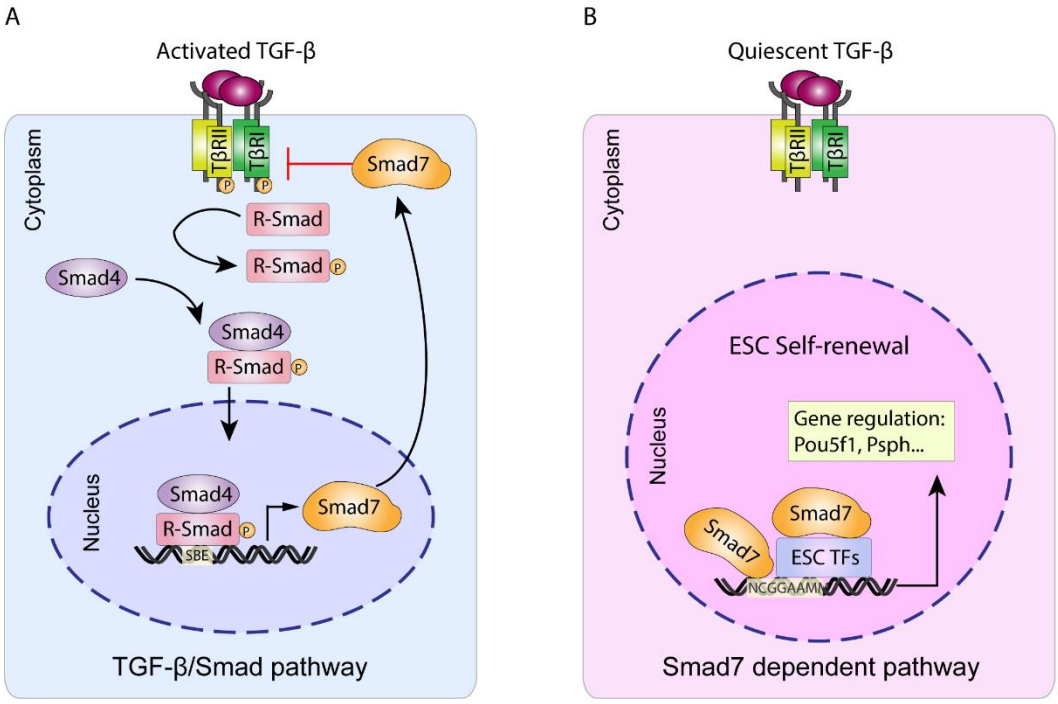


Figure 2.11. A working model for Smad7-mediated transcriptional regulation.

A. Upon TGFβ/Smad pathway activation, the Smad complex translocates from the cytoplasm into the nucleus, where the expression of Smad7 is activated. Then Smad7 can function as a TGFβ antagonist in the cytoplasm.
B. Our study illustrates Smad7 functions as a transcription factor and regulates the target gene expression independently or by interacting with other ESC factors in the nucleus, thus promoting ESC self-renewal and pluripotency maintenance.

DISCUSSION

Previous investigations have elucidated the function of Smad7 in promoting ESC self-renewal and maintaining pluripotency through various mechanisms. However, the molecular mechanisms by which Smad7 influences downstream genes in ES cells have rarely been explored on a genome-wide level. Therefore, it is essential to examine in ES cells the targets of Smad7 that may promote the self-renewal or function of ES cells. Using mouse ES cells, our data revealed that endogenous and exogenous Smad7 are predominantly located in the nucleus, despite the low presence in the cytoplasm. Indeed, the subcellular localization of Smad7 varies in different cells or various culture conditions. Exogenously expressed Smad7 has been reported to be located in the nucleus in COS1, COS7, and Mv1Lu cells, while it is predominantly localized in the cytoplasm in HepG2 cells and TβRI-deficient R mutant Mv1Lu cells (Itóh S et al., 1998; Hanyu A et al., 2001). Itóh S et al. (1998) and Zhang S et al. (2007) reported that Smad7 would transport from the nucleus into the cytoplasm in COS1, COS7, and Mv1Lu cells under TGF-β stimulation, and in COS1 cells when co-expressed with TβRI. Besides, endogenous Smad7 predominantly remains in the nucleus in Hep3B, and ectopically expressed Smad7 localizes in the nucleus in HeLa cells, although upon TGF-β1 treatment (Zhang S et al., 2007). Smad7 localization might be different because the interactions between Smad7 and various partners differ in various cells. To further understand the regions essential for its subcellular localization, we generated a Smad7 deletion mutant (Smad7Δ407), which lacks the last 19 amino acids. We found that the expression of Smad7Δ407 dramatically decreased in nucleoplasm and chromatin fractions; this may be due to the degradation of Smad7 gene expression in the nucleus. This result suggests that the intact C-terminus is critical for its nuclear localization and transcriptional regulation. These results are in accordance with the previous studies by Itoh et al. (1998) and Zhang S et al. (2007), who demonstrated that the integrity of this region is vital for Smad7 nuclear localization.

We report that Smad7, a classical TGF-β target gene so far recognized mainly as a negative feedback product of TGF-β superfamily signaling, can also function in the nucleus as a positive transcription factor in ESC. We identified the genome-wide binding sites of Smad7 that are likely to play essential roles in ES cells. The results show that over half of Smad7 binding sites are associated with promoter and enhancer regions. To evaluate the correlations between Smad7 binding and the combined histone H3 modification signatures, we performed cluster analysis indicating that Smad7 binding sites are strongly associated with active histone modification enzymes, including H3K9ac, H3K27ac, H3K4me1, H3K4me3, and H3K36me3. This result allows us to understand better that Smad7 can recruit these proteins to regulate chromatin structure and thus control gene expression epigenetically in ESC.

Moreover, the motif discovery analysis showed that Smad7 binding consensus was enriched for ETS transcription factor family motifs. Interestingly, Koinuma D et al. (2009) reported that ETS binding sites, including Elk-1 binding sites and other ETS family binding motifs, were significantly enriched in Smad2/3 binding sites. Our study showed that Smad7, Smad2, Smad3, and Smad4 exhibit partial similarity in their targets. This result may explain why ETS transcription factors are enriched in Smad7 binding sites. Altogether, the resultant effects of ETS1 on TGF-β-induced transcription may be critical for many of the Smad7-binding target genes. Furthermore, the predicted Smad7-bound sequences (5'-(C/T)TTCCG -3') is distinct from that of the other member of the Smad family as Smad2/3 preferentially bind to G(T/G)CT(G). This difference could be due to the fact that Smad2/3 bind the DNA via MH1 domain while Smad7 has been shown to bind DNA via its MH2 domain (Zhang S et al., 2007).

To explore whether the silencing of Smad7 correlates with the occupancy of Smad7 on its target genes, we first found that Smad7 KD influenced the expression of over 1,000 genes. In particular, the silencing of Smad7 significantly decreases the expression of the self-renewal and pluripotency markers and increases the expression of several genes involved in the neurectoderm and meso-endoderm differentiation. Previous studies demonstrated that Smad7 acts as a STAT3 target gene and contributes to the undifferentiated state's maintenance (Bourillot PY et al., 2009), promoting ESC self-renewal in a manner dependent on the LIF receptor (Yu Y et al., 2017). Nevertheless, our study showed that Smad7 knockdown enhances the expression of the genes relevant to the TGFβ/SMAD pathway, such as Lefty1, Serpine1, Tgfbr2. Since Smad7 is a well-established negative regulator in TGF-β/SMAD signaling, our results imply the critical function of Smad7 in inhibiting TGF-β/SMAD signaling and maintaining pluripotency in ESC. Significantly, a group of TFs, including Pou5f1, Kpna6, Akirin1, Myb, Nanog, Sox2, Dnmt3b, T, Id2, Tbx3, Gata6, and Ets1 were influenced directly or indirectly by Smad7. Notably, Chen X et al. (2008) demonstrated that the densest TF binding locus is the enhancer region of Pou5f1, and 11 TFs bound this region (Chew JL et al., 2005). And our study also showed that Smad7 directly binds to this region. Also, several TFs, such as Sall4, Myc, Zfp398, and Klf5, were the putative targets of Smad7, suggesting that Smad7 may interconnect with other transcriptional regulators and then regulate the expression of these genes.

Furthermore, Serpine1 (PAI-1), a target gene of TGF-β (Dennler S et al., 1998), was shown to be bound by Smad7 at the promoter in Hep3B cells (Zhang S et al., 2007). Moreover, Thakur N et al. (2020) reported that Smad7 binds to the regulatory regions of c-Jun and HDAC6 genes, thereby inducing their expression in response to TGF-β in prostate cancer cells. However, these genes were not in our Smad7 ChIP-seq target genes list, it might be due to the difference in

Smad7 localization in various cells, and the interaction between Smad7 and various partners differ contextually in response to TGF-β.

In summary, our data about Smad7-bound chromosomal regions provides a valuable resource for future studies aiming at a transcriptional-metabolic dissection of ESC self-renewal and pluripotency maintenance. Our study provides insight into the molecular mechanism of the transcriptional regulation of Smad7, and it gives a basis for understanding the epigenetic control of gene expression in ES cells.

Chapter 3

EXPERIMENTAL PROCEDURES

3.1 Embryonic stem cell culture

E14 mouse ES cells were grown on 0.1% gelatin-coated plates and maintained in high-glucose DMEM (Invitrogen) supplemented with 18% EmbryoMax® FBS (Millipore), 0.1 mM non-essential amino acids (Invitrogen), 1 mM sodium pyruvate (Invitrogen), 0.1 mM 2-mercaptoethanol, 1500 U/mL LIF (Millipore), 25 U/mL penicillin, and 25 μg/mL streptomycin. ES cells stably expressing *Escherichia coli* biotin ligase BirA enzyme was generated according to (Kim J et al., 2009). Positive clones were selected by growth in G418, and the level of ectopic expression of the protein was detected by Western blotting assay with anti-Biotin. Cells were routinely tested to be free of mycoplasma.

3.2 DNA construct and shRNA

pEF1α-BirA-V5 construct expresses C-terminal V5-tagged BirA biotin ligase. The full-length Smad7 cDNA and mutant (Smad7Δ407; it lacks 19 amino acids at its C-terminus) were obtained by PCR amplification and N-terminally tagged by Avitag. Then the expression constructs were cloned into the Not I and Xba I sites of the pEF6 vector (Invitrogen).

Smad7 shRNAs were constructed using the TRC hairpin design tool (http://www.broadinstitute.org/rnai/public/seq/search) targeting the following sequences:

5′- GCTTTCAGATTCCCAACTTCT-3′ (shRNA1 Smad7);

5′- GTCTTGTTCTTTGAGAAATTA-3′ (shRNA2 Smad7)

Annealed oligonucleotides were cloned into pLKO.1 lentiviral vector (Addgene:10879), and each construct was verified by sequencing.

3.3 Transient transfection

Transient transfection was performed using Lipofectamine 2000 Transfection Reagent (Invitrogen) for mouse ES cells according to the manufacturer's recommendations. For Smad7 knockdown, cells were transfected twice with 5 μg of the specific shRNA construct and maintained in the growth medium for 48 h.

3.4 In vivo biotinylation

E14 ES cells were transfected with a plasmid construct expressing BirAV5 and cultured for ten days in a growth medium with G418. Clones were picked and expanded for western blotting by using anti-V5–HRP (Invitrogen). A BirA-expressing ES clone was then used for transfection with a plasmid construct containing biotin-tagged Smad7 cDNA. ES cells expressing tagged Smad7 were identified by western blotting with anti-Biotin of the total lysates.

3.5 Subcellular fractionation

Cytoplasmic, nucleoplasmic, and chromatin fractions can be easily prepared from a pellet of cultured cells. In detail, 2×10^7 ES cells were harvested by centrifugation, washed in PBS, and directly resuspended in 300 µL of Fractionation buffer (PARIS, #AM1921) supplemented with anti-protease, and the cells were incubated for 5 min on ice. Nuclei were collected in the pellet (P1) by low-speed centrifugation (5 min, 1,500 rpm, 4°C). The supernatant (S1: Cytoplasmic fraction) was further clarified by high-speed centrifugation (10 min, 14,000 rpm, 4°C) to remove cell debris and insoluble aggregates. Nuclei (P1) were then lysed in 150 µL of Glycerol buffer (20 mM Tris-HCl, 75 mM NaCl, 0.5 mM EDTA, 0.85 mM DTT, 0.125 mM Phenylmethylsulfonyl fluoride, 50 % Glycerol, pH 8.0) and 150 µL of Nuclei Resuspension buffer (10 mM HEPES, 1 mM DTT, 7.5 mM $MgCl_2$, 0.2 mM EDTA, 0.3 M NaCl, 1 M Urea, 1 % NP-40, pH 7.5). The supernatant (S2: Nucleoplasmic fraction) was further collected by high-speed centrifugation (5 min, 14,000 rpm, 4°C). The Insoluble chromatin pellet (P2) was resuspended in 50 µL of F-buffer and sonicated for three pulses (30 sec “ON”, 30 sec “OFF”, High). The supernatant (S3: Chromatin fraction) was further clarified by high-speed centrifugation (10 min, 14,000 rpm, 4°C). The separated subcellular fractions were used for further assays and analysis.

3.6 Protein extraction and western blotting

For total cell extracts, cells were resuspended in cold F-buffer (10 mM Tris-HCl pH 7.0, 50 mM NaCl, 30 mM Na-pyrophosphate, 50 mM NaF, 1% Triton X-100, anti-proteases), sonicated for three pulses (30 sec "ON", 30 sec "OFF", High), and then stored on ice for 10 min. Cell extract was then centrifuged for 10 min at 13,000 rpm, and the pellet was discarded. Extracts were quantified using BCA assay (Pierce). Equivalent amounts of protein extract were separated by SDS-polyacrylamide gels (10% TGX Stain-Free Protein Gels, Bio-Rad) at different percentages, transferred to nitrocellulose membranes (iBlot™ 2 Transfer Stacks, Invitrogen). Membranes were blocked with 5% milk in TBS and incubated with specific primary antibodies overnight at 4°C. Membranes were rinsed and incubated with HRP-

conjugated secondary antibodies for one hour at room temperature, followed by chemiluminescent detection using Amersham ECL Western Blotting Detection Reagent (GE Healthcare) and Luminata™ Forte Western HRP Substrate (Millipore). Primary antibodies used were: mouse monoclonal IgG Smad7 (MAB2029; R&D systems, 1:1000), rabbit polyclonal IgG Smad7 (PA1-41506; Invitrogen, 1:1000), rabbit monoclonal IgG Biotin (5597s; Cell signaling, 1:1000), mouse monoclonal IgG Vinculin (SAB4200080; Sigma, 1:5000), rabbit polyclonal IgG Histone H3 (ab1791: Abcam, 1:2000), and mouse monoclonal β-Actin (A5441; Sigma, 1:5000).

3.7 Immunofluorescence

Cells were prepared for immunofluorescence by fixation in 4% paraformaldehyde (Electron Microscopy Sciences) for 10 mins and subsequent permeabilization for 15 mins with 0.1% Triton X-100 in PBS at room temperature. Cells were then blocked for nonspecific binding with 2% BSA in PBS for 2 h at room temperature. Primary antibodies were diluted in 0.1% BSA and incubated with the samples overnight at 4°C. Samples were rinsed with PBS and incubated with the appropriate fluorescent dye-labeled secondary antibodies (Alexa Fluor; Invitrogen, 1:1000) for 45 min at room temperature protected from light. Primary antibodies used were: mouse monoclonal IgG2a Smad7 (Z8-B; Sant Cruz, 1:100) and goat IgG Biotin (559286; BD Pharmingen, 1:100). Nuclei were visualized with DAPI (1 μg mL^{-1}).

3.8 AP staining

Cells were plated at a density of $5\times10^3/cm^2$, and after 2 days, were fixed with 4% paraformaldehyde for 2 min and then stained with alkaline phosphatase (AP) solution (Vector® Red alkaline phosphatase substrate kit; catalog no. SK-5100).

3.9 RNA extraction and qRT-PCR

RNA samples were extracted directly from cultured cells using Trizol reagent (Invitrogen) followed by isopropanol precipitation, and sample quality control was performed with Agilent Bioanalyzer 2100. All of the samples had an RNA integrity number ranging from 9.9 to 10. Briefly, mRNA quantitation was performed by qRT-PCR using the Superscript III platinum One-step qRT-PCR System kit (Invitrogen) and normalized on Actb mRNA. Oligonucleotide sequences are indicated in Supplementary Table 1.

3.10 RNA-seq library preparation

For RNA-seq library preparation, 1 μg of total RNA was depleted of ribosomal RNA using the

RiboMinus Eukaryote System v2 kit (Invitrogen), following manufacturer instructions. Ribo-RNA was resuspended in 17 µ l of EFP buffer (Illumina), heated to 94 °C for 8 min, and used as input for the first-strand synthesis, using the TruSeq RNA Sample Prep kit, following manufacturer instructions. Poly(A) RNA-seq library was performed by using the TruSeq RNA Sample Prep kit, following the manufacturer's instructions.

3.11 Bio-ChIP

4×10^7 ES cells expressed Bio-Smad7 protein were harvested, cross-linked with 1% formaldehyde for 10 min, quenched with 0.125 M glycine for 5 min, and then washed twice in cold 1×PBS. Cells were lysed in Isotonic Buffer (20 mM HEPES, 100 mM NaCl, 250 mM Sucrose, 5 mM $MgCl_2$, 5 µM $ZnCl_2$, pH 7.5) supplemented with 1% NP-40 and anti-protease. Cell lysates were centrifuged at 1,000 rpm for 5 min at 4°C to isolate the nuclei. Nuclei were suspended in 1 mL of 0.15 % SDS Lysis Buffer (20 mM Tris-HCl, 150 mM NaCl, 2 mM EDTA, 0.15 % SDS, 1 % Triton X-100, pH 8.0) supplemented with anti-protease and subjected for sonication to shear chromatin fragments to an average size between 100bp and 400bp on the Pico Bioruptor® (Diagenode) (2 runs of 10 cycles [30 sec "ON", 30 sec "OFF"]) at high power setting. Fragmented chromatin was centrifuged at 13,000 rpm for 15 min at 4°C.

50 µL MyOne™ Streptavidin T1 Dynabeads (Thermo-Fisher Scientific) were saturated with PBS/1% BSA at RT for 1 hour and then incubated with the sample at 4°C overnight on a rotator. After incubation, Dynabeads were washed twice with 1 mL of Washing Buffer I (2 % SDS), once with 1 mL of Washing Buffer II (50 mM HEPES, 500 mM NaCl, 0.1% Sodium Deoxycholate, 1% Triton X-100, 1 mM EDTA, pH 7.5), once with 1 mL Washing Buffer III (10 mM Tris-HCl, 250 mM LiCl, 0.5% NP-40, 0.5% Sodium Deoxycholate, 1 mM EDTA, pH 8.0), and once with 1 mL of TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7.5). All the washing steps were performed at RT for 8 mins on a rotator. The chromatin was eluted in 250 µL SDS Elution Buffer (1% SDS, 50 mM Tris-HCl, 10 mM EDTA, pH 8.0) followed by reverse cross-linking at 65°C overnight. After decrosslinking, the ChIP DNA was purified using QIAQuick PCR Purification Kit (Qiagen) according to the manufacturer's instructions.

For the reference sample, BirA-ES cells without tagged protein were used.

3.12 Bio-ChIP-seq library preparation

4 ng of purified ChIP DNA was processed for library generation using the NEBNext ChIP-seq Library Prep Master Mix (New England Biolabs or NEB) following the manufacturer's protocol. Libraries were pooled and sequenced on an Illumina NextSeq 500 Platform using the 75bp high

output sequencing kit.

3.13 Bio-ChIP assay

For ChIP-qPCR analysis, 1×10^7 ES cells transduced with BirA and linearized Smad7 constructs were used. The immunoprecipitated DNA was isolated using the protocol described for Bio-ChIP was analyzed by quantitative PCR (qPCR) using the SYBR GreenER kit (Invitrogen). For input samples, 250 µL of SDS Elution Buffer was added into 20 µL of the sheared chromatin. The samples were incubated at 65°C overnight to reverse cross-linking. DNA fragments were purified with the QIAquick PCR Purification Kit and eluted with 40 µL of H_2O. Quantitative PCR was performed with approximately 2% of the ChIP sample. The amount of each amplification product was determined relative to a standard curve, and fold enrichment was calculated by comparison of amplified product from bioChIP sample and ChIP samples from BirA containing ES cells. The amount of immunoprecipitated DNA was calculated relative to input. Primer pairs for quantitative ChIP-PCR were designed using ±150bp genomic sequence information specific to the predicted target loci to generate 100bp to 125bp amplified products. Primers targeting mouse --- sequences or --- as a control were used for qPCR analysis. Primer sequences are listed in Supplementary Table 1.

3.14 RNA-seq analysis

Total RNA was isolated using TRIzol reagent (Invitrogen), according to the manufacturer's protocol. Quantity and quality of the starting RNA were checked by Qubit and Bioanalyzer (Agilent). 1 µg of total RNA was subjected to poly(A) selection, and libraries were prepared using the TruSeq RNA Sample Prep Kit (Illumina) following the manufacturer's instructions. Sequencing was performed on the Illumina NextSeq 500 platform. After quality controls with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), raw reads were aligned to the NCBIm37 mouse genome reference (mm9) using HiSat2 v2.2.0 (Kim D et al., 2019) with options: -N 1 -L 20 -i S,1,0.5 -D 25 -R 5 --pen-noncansplice 20 --mp 1,0 --sp 3,0. Pre-built indexes based on the Ensembl transcript annotation (release 84) for guided alignment to transcriptome were retrieved from the HiSat2 web site (https://cloud.biohpc.swmed.edu/index.php/s/grcm38_tran/download).

Gene expression levels were quantified with featureCounts v1.6.1 (Liao Y et al., 2014) (options: -t exon -g gene_name) using GENCODE comprehensive gene annotation (release M23 – GRCm38.p6 https://www.gencodegenes.org/mouse/release_M23.html). Multi-mapped reads were excluded from quantification. Gene expression counts were next analysed using the edgeR package (Robinson MD et al., 2010) . After filtering lowly expressed genes (1 count per million

(CPM) in less than 3 samples), normalization factors were calculated using the trimmed-mean of M-values (TMM) method (implemented in the calcNormFactors function) and CPM were obtained using normalized library sizes. Differential expression analysis was carried out by fitting a Generalized Linear Model (GLM) to all groups and performing Quasi-Likelihood F-test while pairing each Smad7-shRNA with its respective scramble control (design matrix formula = "~treatment+batch"). Genes with False Discovery Rate (FDR) less than or equal to 0.05 were considered as differentially expressed. Hierarchical clustering of gene expression profiles was performed on differentially expressed genes only, using Euclidean distances and the complete linkage method. CPM values were scaled as Z-scores across samples before computing distances. Gene expression heatmaps were generated using the ComplexHeatmap R package.

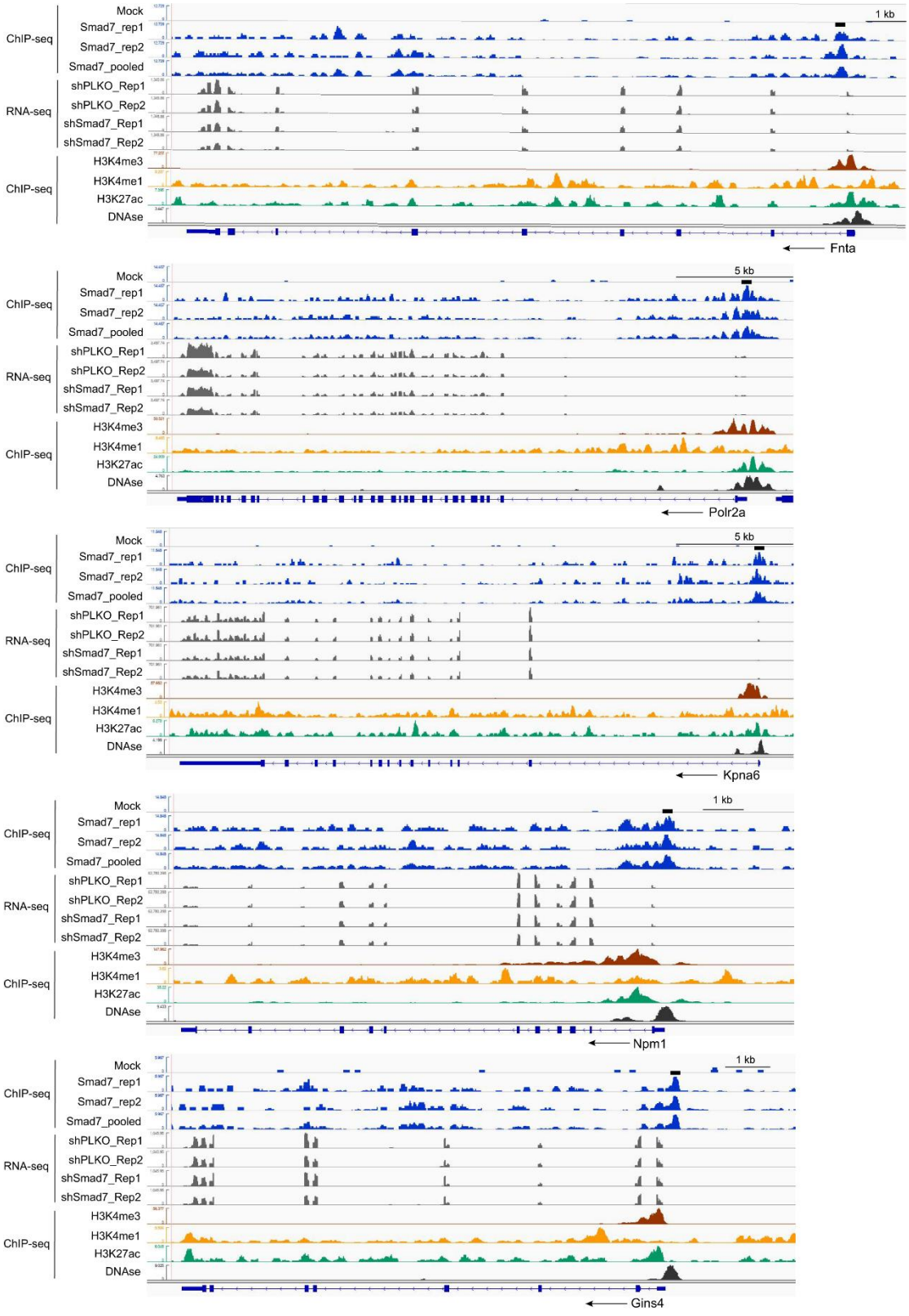
3.15 Bio-ChIP-seq data analysis

Following quality controls (performed with FastQC v0.11.2), sequencing reads were processed with Trim Galore! v0.5.0 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) to perform quality and adapter trimming (parameters: --stringency 3 –q 20). Trimmed reads were next analysed with the ENCODE Transcription Factors and Histone Modifications ChIP-seq pipeline 2 (v1.6.1, available from https://github.com/ENCODE-DCC/chip-seq-pipeline2), using default software and parameter settings for the 'Transcription Factors' processing mode. Briefly, reads were aligned to the mouse reference genome (UCSC mm10) using Bowtie2 (Langmead B and Salzberg SL, 2012). Duplicated, multi-mapping and poor-quality alignments were discarded, and peak calling was performed using MACS2 (Zhang Y et al., 2008) , using the input DNA as control library. Signal tracks were generated as fold enrichment over control for both individual and pooled replicates using MACS2 (Zhang Y et al., 2008) . To determine a consensus set of Smad7 putative genomic binding sites conserved across biological replicates, the Irreproducible Discovery Rate (IDR) procedure was carried out, as implemented in the ENCODE pipeline (Landt SG et al., 2012) . Starting from overlapped peak calls in individual libraries (with SPP (Kharchenko PV et al., 2008) peak caller at FDR < 0.01 and using the mock bioChIP-seq as control), the conservative set (i.e. from the comparison of true replicates) of IDR thresholded (IDR < 0.05) peaks was retained as the final set of Smad7 binding sites. Genomic annotation of Smad7 peaks was carried out using the HOMER (Heinz S et al., 2010) suite (annotatePeaks.pl utility). Association of Smad7 peaks to regulatory elements was performed using the annotated list of Candidate cis-Regulatory Elements (cCRE) for mESC cell line E14 available from the SCREEN (ENCODE Project Consortium, 2010) database. Heatmaps of ChIP-seq signals over regulatory regions were generated using the deepTools (Ramírez F et al., 2014) computeMatrix and plotHeatmap utilities. ChIP-seq datasets for

comparative analysis in mESC E14 were retrieved from ENCODE for histone marks/chromatin accessibility, and from GEO for transcription factors (accession code GSE11431, GSE125116). Correlation heatmaps were generated on a merged set of binding sites (defined by the ENCODE pipeline or by MACS2 (Zhang Y et al., 2008) at Q-value < 0.05 for the GEO-retrieved datasets) using the deepTools multiBamSummary and plotCorrelation utilities, using Spearman correlation.

SUPPLEMENTAL INFORMATION

Supplementary Figure 1



Supplementary Table 1: Oligonucleotide sequences used as primers with the indication of the corresponding genes.

Gene	Sequence	Strand	Application
Actb	TCTTTGCAGCTCCTTCGTTG	Fw	RT-qPCR
Actb	ACGATGGAGGGGAATACAGC	Rev	RT-qPCR
Nanog	AAGTACCTCAGCCTCCAGCA	Fw	RT-qPCR
Nanog	GTGCTGAGCCCTTCTGAATC	Rev	RT-qPCR
Pou5f1	CTGAGGGCCAGGCAGGAGCACGAG	Fw	RT-qPCR
Pou5f1	CTGTAGGGAGGGCTTCGGGCACTT	Rev	RT-qPCR
Sox2	ATGATGGAGACGGAGCTGAA	Fw	RT-qPCR
Sox2	TTGCTGATCTCCGAGTTGTG	Rev	RT-qPCR
Zfp42	GGACTAAGAGCTGGGACACG	Fw	RT-qPCR
Zfp42	CCTGCTTTTTGGTCAGTGGT	Rev	RT-qPCR
Prdm14	CTCTACAATCTGCCCTGGTA	Fw	RT-qPCR
Prdm14	GGGACAGATATTGACTGGGA	Rev	RT-qPCR
Psph	GAGTTCTGGCTGCTCCAGTT	Fw	RT-qPCR
Psph	AACTTTCCGGTAGCAGTGGG	Rev	RT-qPCR
Npm1	TGAAGATGAAGATGAGGAGGACG	Fw	RT-qPCR
Npm1	TTACCACCTCCAGGAGCAGA	Rev	RT-qPCR
Fnta	GTGTATAGCAGCGAGGGCAT	Fw	RT-qPCR
Fnta	TTTAGCGCACAGGCAAGAGA	Rev	RT-qPCR
Polr2a	CCAACCTCCCCGACATACTC	Fw	RT-qPCR
Polr2a	GCTCAGTTCTCCTCATCGCT	Rev	RT-qPCR
Mplkip	GGAACATCCTGGACAGAACTT	Fw	RT-qPCR
Mplkip	ATTGTCCTATCATTTCCGCTGTT	Rev	RT-qPCR
Gins4	CATCGCACCTCTTGTTGCTT	Fw	RT-qPCR
Gins4	GCGCCTTTCTACGTCCAAGT	Rev	RT-qPCR
Kpna6	TGAACCCTGTTGGCATCTCC	Fw	RT-qPCR
Kpna6	TAGAGAGCGGCTAAGGGTGT	Rev	RT-qPCR
Ptma	AGCAGAAGACCGAGGAGGAT	Fw	RT-qPCR
Ptma	GGTGGAGAGCGCATGTCATA	Rev	RT-qPCR
Akirin1	GGACTGTTTCCAGGCACAGAA	Fw	RT-qPCR
Akirin1	AGAAGTTCAGCGTTCAGGGA	Rev	RT-qPCR
Pou5f1	TTAGGGTTAGAGCTGCCCCC	Fw	ChIP-qPCR
Pou5f1	CTGGGTGGAGAAACCCAACG	Rev	ChIP-qPCR
Psph	GAAGCATCCCTCACACACGA	Fw	ChIP-qPCR
Psph	CGAAAAGTCTTACGCACCCG	Rev	ChIP-qPCR
Npm1	CTTTCATCCTCACGCCTAGGTA	Fw	ChIP-qPCR
Npm1	GAGGCCACAAAGAGAAAACGC	Rev	ChIP-qPCR
Fnta	TCCCATAGCGAGCTGTCAAAA	Fw	ChIP-qPCR
Fnta	TCTGAAGCTAGGGTGCGGT	Rev	ChIP-qPCR
Polr2a	GGCTGAGTGACGCCCTTTAT	Fw	ChIP-qPCR
Polr2a	AAACCCGAAAGGAGGACGAC	Rev	ChIP-qPCR
Mplkip	AGTGACCTGTTTACCTGCCC	Fw	ChIP-qPCR
Mplkip	CGGACTACCGCACCTTAGTT	Rev	ChIP-qPCR
Gins4	GGCGTTCGGCTCACAGTAA	Fw	ChIP-qPCR
Gins4	TGGCAATGGAACCAACCCAC	Rev	ChIP-qPCR

Kpna6	GCAGACAATATGGCGGACCT	Fw	ChIP-qPCR
Kpna6	CTTTAAGCCCGCCCCTATCC	Rev	ChIP-qPCR
Ptma	AGTCACAGTCACCCCGTTGT	Fw	ChIP-qPCR
Ptma	CTGACGCAGAAAGCTAAAGCG	Rev	ChIP-qPCR
Akirin1	CCCCCTGCGGAAGAAGTAGG	Fw	ChIP-qPCR
Akirin1	AGGTGCGAAAGACCTCCCAG	Rev	ChIP-qPCR
Smad7	CCGGGCTTTCAGATTCCCAACTTCTCTCG AGAGAAGTTGGGAATCTGAAAGCTTTTTG	Fw	shRNA1 Smad7
Smad7	AATTCAAAAAGCTTTCAGATTCCCAACTTC TCTCGAGAGAAGTTGGGAATCTGAAAGC	Rev	shRNA1 Smad7
Smad7	CCGGGTCTTGTTCTTTGAGAAATTACTCGA GTAATTTCTCAAAGAACAAGACTTTTTG	Fw	shRNA2 Smad7
Smad7	AATTCAAAAAGTCTTGTTCTTTGAGAAATTA CTCGAGTAATTTCTCAAAGAACAAGAC	Rev	shRNA2 Smad7

Supplementary Table 2: Representative Smad7 directly regulated target genes.

Gene Name	Annotation	Nearest Promoter ID	Nearest Refseq	Gene Alias	Gene Description	Gene Type	peak ID	logFC	FDR
Actr2	Intergenic	NM_146243	NM_146243	4921510D23Rik|AA409782|Arp2|D6Ertd746e	ARP2 actin-related protein 2	protein-coding	chr11.20139035.20139418	-0.27686	0.040157
Akirin1	promoter-TSS (NM_023423)	NM_023423	NM_023423	6330407G11Rik	akirin 1	protein-coding	chr4.123750240.123750623	-0.34102	0.035702
Apela	promoter-TSS (NM_001297554)	NM_001297554	NM_001297554	Ela|Elabela|Ende|Gm10664	apelin receptor early endogenous ligand	protein-coding	chr8.65037150.65037533	-0.38592	0.012718
Apex1	TTS (NM_001310506)	NM_009687	NM_009687	APE|Apex|HAP1|Ref-1	apurinic/apyrimidinic endonuclease 1	protein-coding	chr14.50927013.50927396	-0.31358	0.031104
Ctbp2	intron (NM_009980, intron 1 of 10)	NM_009980	NM_009980	AA407280|D7Ertd45e|Gtrgeo6|Ribeye	C-terminal binding protein 2	protein-coding	chr7.133114889.133115272	-0.25703	0.043516
Cyb5b	Intergenic	NM_025558	NM_025558	1810044O22Rik|AU015618|Cyb5m	cytochrome b5 type B	protein-coding	chr8.107133289.107133672	-0.31511	0.024064
Cyb5b	promoter-TSS (NM_025558)	NM_025558	NM_025558	1810044O22Rik|AU015618|Cyb5m	cytochrome b5 type B	protein-coding	chr8.107150505.107150888	-0.31511	0.024064
Dhodh	5' UTR (NM_020046, exon 1 of 9)	NM_020046	NM_020046	2810417D19Rik|AI834883	dihydroorotate dehydrogenase	protein-coding	chr8.109608207.109608590	-0.30004	0.031453
Eloc	promoter-TSS (NR_037998)	NM_026456	NM_026456	2610043E24Rik|2610301I15Rik|AA407206|AI987979|AW049146|Tceb1	elongin C	protein-coding	chr1.16656579.16656962	-0.34559	0.026769
Fnta	promoter-TSS (NM_008033)	NM_008033	NM_008033	FTA	farnesyltransferase, CAAX box, alpha	protein-coding	chr8.26015447.26015830	-0.41584	0.028666
Gcsh	promoter-TSS (NM_026572)	NM_026572	NM_026572	1100001L02Rik|5730591C18Rik	glycine cleavage system protein H (aminomethyl carrier)	protein-coding	chr8.116994097.116994480	-0.42996	0.015179
Gins4	promoter-TSS (NM_024240)	NM_024240	NM_024240	2810037C03Rik|4933405K01Rik|Sld5	GINS complex subunit 4 (Sld5 homolog)	protein-coding	chr8.23237666.23238049	-0.40532	0.017316
Itgb7	promoter-TSS (NM_013566)	NM_013566	NM_013566	Ly69	integrin beta 7	protein-coding	chr15.102231692.102232075	-0.36307	0.035049
Kpna6	promoter-TSS (NM_008468)	NM_008468	NM_008468	IPOA7|Kpna5|NPI-2	karyopherin (importin) alpha 6	protein-coding	chr4.129672536.129672919	-0.34604	0.018835
Lss	promoter-TSS (NM_146006)	NM_146006	NM_146006	2810025N20Rik|D10Ertd116e|Osc	lanosterol synthase	protein-coding	chr10.76531479.76531862	-0.33055	0.030206
Mplkip	promoter-TSS (NM_138654)	NM_025479	NM_025479	2810021B07Rik|C330007M08Rik|Ttdn1	M-phase specific PLK1 intereacting protein	protein-coding	chr13.17695112.17695495	-0.51042	0.017286
Myb	5' UTR (NM_010848, exon 1 of 15)	NM_010848	NM_010848	AI550390|M16449|c-myb	myeloblastosis oncogene	protein-coding	chr10.21160644.21161027	-0.40712	0.028666

Npm1	promoter-TSS (NM_001252260)	NM_001252260	NM_008722	B23|NO38|Npm	nucleophosmin 1	protein-coding	chr11.33163099.33163482	-0.45084	0.006906
Platr10	intron (NR_040539, intron 1 of 2)	NR_040539	NR_040539	2410007B07Rik	pluripotency associated transcript 10	ncRNA	chr3.75651719.75652102	-0.47611	0.012216
Polr2a	exon (NM_001291068, exon 1 of 29)	NM_001291068	NM_009089	220kDa|Rpb1|Rpo2-1	polymerase (RNA) II (DNA directed) polypeptide A	protein-coding	chr11.69758013.69758396	-0.42052	0.007112
Pou5f1	promoter-TSS (NM_013633)	NM_013633	NM_013633	NF-A3|Oct-3|Oct-3/4|Oct-4|Oct3|Oct3/4|Oct4|Otf-3|Otf-4|Otf3|Otf3-rs7|Otf3g|Otf4	POU domain, class 5, transcription factor 1	protein-coding	chr17.35504913.35505296	-0.60182	0.002177
Psph	promoter-TSS (NM_133900)	NM_133900	NM_133900	AI480570|PSP|PSPase	phosphoserine phosphatase	protein-coding	chr5.129787014.129787397	-0.58336	0.010694
Ptma	Intergenic	NM_001360830	NM_008972	Thym	prothymosin alpha	protein-coding	chr1.86533824.86534207	-0.32186	0.027408
Ptma	Intergenic	NR_153833	NM_008972	Thym	prothymosin alpha	protein-coding	chr1.86487399.86487990	-0.32186	0.027408
Ptma	Intergenic	NR_153833	NM_008972	Thym	prothymosin alpha	protein-coding	chr1.86525140.86525723	-0.32186	0.027408
Ptma	promoter-TSS (NM_001360830)	NR_153833	NM_008972	Thym	prothymosin alpha	protein-coding	chr1.86526041.86526424	-0.32186	0.027408
Rpl18a	intron (NM_029751, intron 1 of 4)	NM_029751	NM_029751	2510019J09Rik	ribosomal protein L18A	protein-coding	chr8.70897041.70897424	-0.27496	0.030997
Scd1	Intergenic	NM_009127	NM_009127	AA589638|AI265570|Scd|Scd-1|ab	stearoyl-Coenzyme A desaturase 1	protein-coding	chr19.44388230.44388815	-0.40187	0.011754
Selenot	exon (NM_001040396, exon 1 of 6)	NM_001040396	NM_001040396	2810407C02Rik|5730408P04Rik|Selt	selenoprotein T	protein-coding	chr3.58576479.58577100	-0.34823	0.039618
Sgk1	promoter-TSS (NM_011361)	NM_001161850	NM_011361	Sgk	serum/glucocorticoid regulated kinase 1	protein-coding	chr10.21994001.21994384	-0.4247	0.012003
Snhg3	Intergenic	NR_003270	NR_002904	Rnu17d|U17HG	small nucleolar RNA host gene 3	ncRNA	chr4.132349828.132350211	-0.35947	0.039208
Uck2	5' UTR (NM_030724, exon 1 of 7)	NM_030724	NM_030724	AA407809|AI481316|AU018180|AU020720|TSA903|UK|UMK|Umpk	uridine-cytidine kinase 2	protein-coding	chr1.167284821.167285204	-0.28473	0.042171
Usp10	5' UTR (NM_001310630, exon 1 of 14)	NM_009462	NM_009462	2610014N07Rik|UBPO|Uchrp|mKIAA0190	ubiquitin specific peptidase 10	protein-coding	chr8.119910492.119910875	-0.25536	0.038707
Zc3h18	promoter-TSS (NM_001310650)	NM_001310650	NM_001029993	1190001B23Rik|5830416A07Rik|Nhn1	zinc finger CCCH-type containing 18	protein-coding	chr8.122376265.122376648	-0.25323	0.043591
Zfp42	intron (NM_009556,	NM_009556	NM_009556	Rex-1|Rex1|Zfp-42	zinc finger protein 42	protein-coding	chr8.43300235.43300618	-0.38375	0.011017

	intron 3 of 3)								
Zfp42	Intergenic	NM_009556	NM_009556	Rex-1|Rex1|Zfp-42	zinc finger protein 42	protein-coding	chr8.43290319.43290702	-0.38375	0.011017
Zfp981	Intergenic	NM_001243138	NM_001243138	Gm13247	zinc finger protein 981	protein-coding	chr4.146492487.146492870	-0.40935	0.032798
Zswim1	intron (NM_028028, intron 1 of 1)	NM_028028	NM_028028	2410003H12Rik|AI850991|BB046916	zinc finger SWIM-type containing 1	protein-coding	chr2.164823904.164824492	-0.42083	0.023737

REFERENCES (I)

Annes JP, Munger JS, Rifkin DB. Making sense of latent TGFbeta activation. J Cell Sci. 2003;116(Pt 2):217-224. doi:10.1242/jcs.00229.

Aragón E, Wang Q, Zou Y, et al. Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF-β signaling. Genes Dev. 2019;33(21-22):1506-1524. doi:10.1101/gad.330837.119

Bhowmick NA, Ghiassi M, Bakin A, et al. Transforming growth factor-beta1 mediates epithelial to mesenchymal transdifferentiation through a RhoA-dependent mechanism. Mol Biol Cell. 2001;12(1):27-36. doi:10.1091/mbc.12.1.27.

Bieda M, Xu X, Singer MA, Green R, Farnham PJ. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res. 2006;16(5):595-605. doi:10.1101/gr.4887606

Blank U, Karlsson G, Moody JL, et al. Smad7 promotes self-renewal of hematopoietic stem cells. Blood. 2006;108(13):4246-4254. doi:10.1182/blood-2006-02-005611.

Bourillot PY, Aksoy I, Schreiber V, et al. Novel STAT3 target genes exert distinct roles in the inhibition of mesoderm and endoderm differentiation in cooperation with Nanog. Stem Cells. 2009;27(8):1760-1771. doi:10.1002/stem.110

Boyer LA, Plath K, Zeitlinger J, et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature. 2006; 441(7091):349-353. doi:10.1038/nature04733.

Briones-Orta MA, Tecalco-Cruz AC, Sosa-Garrocho M, Caligaris C, Macías-Silva M. Inhibitory Smad7: emerging roles in health and disease. Curr Mol Pharmacol. 2011;4(2):141-153.

Brons IG, Smithers LE, Trotter MW, et al. Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature. 2007;448(7150):191-195. doi:10.1038/nature05950.

Cartwright P, McLean C, Sheppard A, Rivett D, Jones K, Dalton S. LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. Development. 2005;132(5):885-896. doi:10.1242/dev.01670

Chacko BM, Qin B, Correia JJ, Lam SS, de Caestecker MP, Lin K. The L3 loop and C-terminal phosphorylation jointly define Smad protein trimerization. Nat Struct Biol. 2001;8(3):248-253. doi:10.1038/84995.

Chambers I, Tomlinson SR. The transcriptional foundation of pluripotency. Development. 2009;136(14):2311-2322. doi:10.1242/dev.024398

Chen X, Xu H, Yuan P, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell. 2008;133(6):1106-1117. doi:10.1016/j.cell.2008.04.043

Chen YG, Massagué J. Smad1 recognition and activation by the ALK1 group of transforming growth factor-beta family receptors. J Biol Chem. 1999;274(6):3672-3677.

doi:10.1074/jbc.274.6.3672.

Chew JL, Loh YH, Zhang W, et al. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. Mol Cell Biol. 2005;25(14):6031-6046. doi:10.1128/MCB.25.14.6031-6046.2005

Christian JL, Nakayama T. Can't get no SMADisfaction: Smad proteins as positive and negative regulators of TGF-beta family signals. Bioessays. 1999;21(5):382-390. doi:10.1002/(SICI)1521-1878(199905)21:5<382::AID-BIES5>3.0.CO;2-V.

Collas P, Dahl JA. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. Front Biosci. 2008; 13:929-943. Published 2008 Jan 1. doi:10.2741/2733.

de Boer E, Rodriguez P, Bonte E, et al. Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. Proc Natl Acad Sci U S A. 2003;100(13):7480-7485. doi:10.1073/pnas.1332608100.

Dennler S, Itoh S, Vivien D, ten Dijke P, Huet S, Gauthier JM. Direct binding of Smad3 and Smad4 to critical TGF beta-inducible elements in the promoter of human plasminogen activator inhibitor-type 1 gene. EMBO J. 1998;17(11):3091-3100. doi:10.1093/emboj/17.11.3091

Derynck R, Zhang YE. Smad-dependent and Smad-independent pathways in TGF-beta family signalling. Nature. 2003; 425(6958):577-584. doi:10.1038/nature02006.

Ding J, Xu H, Faiola F, Ma'ayan A, Wang J. Oct4 links multiple epigenetic pathways to the pluripotency network. Cell Res. 2012;22(1):155-167. doi:10.1038/cr.2011.179.

Edlund S, Landström M, Heldin CH, Aspenström P. Transforming growth factor-beta-induced mobilization of actin cytoskeleton requires signaling by small GTPases Cdc42 and RhoA. Mol Biol Cell. 2002;13(3):902-914. doi:10.1091/mbc.01-08-0398.

ENCODE Project Consortium, Moore JE, Purcaro MJ, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583(7818):699-710. doi:10.1038/s41586-020-2493-4

Feng XH, Derynck R. Specificity and versatility in tgf-beta signaling through Smads. Annu Rev Cell Dev Biol. 2005; 21:659-693. doi:10.1146/annurev.cellbio.21.022404.142018.

Gaudreault M, Gingras ME, Lessard M, Leclerc S, Guérin SL. Electrophoretic mobility shift assays for the analysis of DNA-protein interactions. Methods Mol Biol. 2009; 543:15-35. doi:10.1007/978-1-60327-015-1_2.

Hanna JH, Saha K, Jaenisch R. Pluripotency and cellular reprogramming: facts, hypotheses, unresolved issues. Cell. 2010;143(4):508-525. doi:10.1016/j.cell.2010.10.008.

Hanyu A, Ishidou Y, Ebisawa T, Shimanuki T, Imamura T, Miyazono K. The N domain of Smad7 is essential for specific inhibition of transforming growth factor-beta signaling. J Cell Biol. 2001;155(6):1017-1027. doi:10.1083/jcb.200106023

Hata A, Lagna G, Massagué J, Hemmati-Brivanlou A. Smad6 inhibits BMP/Smad1 signaling

by specifically competing with the Smad4 tumor suppressor. Genes Dev. 1998;12(2):186-197. doi:10.1101/gad.12.2.186.

Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576-589. doi:10.1016/j.molcel.2010.05.004

Heldin CH, Miyazono K, ten Dijke P. TGF-beta signalling from cell membrane to nucleus through SMAD proteins. Nature. 1997;390(6659):465-471. doi:10.1038/37284.

Inman GJ, Hill CS. Stoichiometry of active smad-transcription factor complexes on DNA [published correction appears in J Biol Chem. 2003 May 9;278(19):17580]. J Biol Chem. 2002;277(52):51008-51016. doi:10.1074/jbc.M208532200.

Itóh S, Landström M, Hermansson A, et al. Transforming growth factor beta1 induces nuclear export of inhibitory Smad7. J Biol Chem. 1998;273(44):29195-29201. doi:10.1074/jbc.273.44.29195

Itoh S, ten Dijke P. Negative regulation of TGF-beta receptor/Smad signal transduction. Curr Opin Cell Biol. 2007;19(2):176-184. doi:10.1016/j.ceb.2007.02.015.

Ivanova N, Dobrin R, Lu R, et al. Dissecting self-renewal in stem cells with RNA interference. Nature. 2006;442(7102):533-538. doi:10.1038/nature04915

Jaenisch R, Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell. 2008;132(4):567-582. doi:10.1016/j.cell.2008.01.015.

Jiang J, Chan YS, Loh YH, et al. A core Klf circuitry regulates self-renewal of embryonic stem cells. Nat Cell Biol. 2008;10(3):353-360. doi:10.1038/ncb1698

Kavsak P, Rasmussen RK, Causing CG, et al. Smad7 binds to Smurf2 to form an E3 ubiquitin ligase that targets the TGF beta receptor for degradation. Mol Cell. 2000;6(6):1365-1375. doi:10.1016/s1097-2765(00)00134-9.

Kawabata M, Inoue H, Hanyu A, Imamura T, Miyazono K. Smad proteins exist as monomers in vivo and undergo homo- and hetero-oligomerization upon activation by serine/threonine kinase receptors. EMBO J. 1998;17(14):4056-4065. doi:10.1093/emboj/17.14.4056.

Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008;26(12):1351-1359. doi:10.1038/nbt.1508

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907-915. doi:10.1038/s41587-019-0201-4

Kim J, Cantor AB, Orkin SH, Wang J. Use of in vivo biotinylation to study protein-protein and protein-DNA interactions in mouse embryonic stem cells. Nat Protoc. 2009;4(4):506-517. doi:10.1038/nprot.2009.23

Klug SJ, Famulok M. All you wanted to know about SELEX. Mol Biol Rep. 1994;20(2):97-107. doi:10.1007/BF00996358.

Koinuma D, Tsutsumi S, Kamimura N, et al. Chromatin immunoprecipitation on microarray analysis of Smad2/3 binding sites reveals roles of ETS1 and TFAP2A in transforming growth factor beta signaling. Mol Cell Biol. 2009;29(1):172-186. doi:10.1128/MCB.01038-08.

Krampert M, Chirasani SR, Wachs FP, et al. Smad7 regulates the adult neural stem/progenitor cell pool in a transforming growth factor beta- and bone morphogenetic protein-independent manner. Mol Cell Biol. 2010;30(14):3685-3694. doi:10.1128/MCB.00434-09.

Kretzschmar M, Doody J, Massagué J. Opposing BMP and EGF signalling pathways converge on the TGF-beta family mediator Smad1. Nature. 1997;389(6651):618-622. doi:10.1038/39348.

Landt SG, Marinov GK, Kundaje A, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22(9):1813-1831. doi:10.1101/gr.136184.111

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-359. Published 2012 Mar 4. doi:10.1038/nmeth.1923

Lee MK, Pardoux C, Hall MC, et al. TGF-beta activates Erk MAP kinase signalling through direct phosphorylation of ShcA. EMBO J. 2007;26(17):3957-3967. doi:10.1038/sj.emboj.7601818.

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656

Martin GR. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc Natl Acad Sci U S A. 1981;78(12):7634-7638. doi:10.1073/pnas.78.12.7634.

Massagué J. TGFβ signalling in context. Nat Rev Mol Cell Biol. 2012;13(10):616-630. doi:10.1038/nrm3434.

Massagué J, Seoane J, Wotton D. Smad transcription factors. Genes Dev. 2005;19(23):2783-2810. doi:10.1101/gad.1350705.

Mesnard D, Guzman-Ayala M, Constam DB. Nodal specifies embryonic visceral endoderm and sustains pluripotent cells in the epiblast before overt axial patterning. Development. 2006;133(13):2497-2505. doi:10.1242/dev.02413.

Miyazawa K, Shinozaki M, Hara T, Furuya T, Miyazono K. Two major Smad pathways in TGF-beta superfamily signalling. Genes Cells. 2002;7(12):1191-1204. doi:10.1046/j.1365-2443.2002.00599.x.

Miyazono K, Maeda S, Imamura T. BMP receptor signaling: transcriptional targets, regulation of signals, and signaling cross-talk. Cytokine Growth Factor Rev. 2005;16(3):251-263. doi:10.1016/j.cytogfr.2005.01.009.

Molloy PL. Electrophoretic mobility shift assays. Methods Mol Biol. 2000; 130:235-246.

doi:10.1385/1-59259-686-x:235.

Moustakas A, Heldin CH. Non-Smad TGF-beta signals. J Cell Sci. 2005;118(Pt 16):3573-3584. doi:10.1242/jcs.02554.

Murakami G, Watabe T, Takaoka K, Miyazono K, Imamura T. Cooperative inhibition of bone morphogenetic protein signaling by Smurf1 and inhibitory Smads. Mol Biol Cell. 2003;14(7):2809-2817. doi:10.1091/mbc.e02-07-0441.

Najm FJ, Chenoweth JG, Anderson PD, et al. Isolation of epiblast stem cells from preimplantation mouse embryos. Cell Stem Cell. 2011;8(3):318-325. doi:10.1016/j.stem.2011.01.016.

Niwa H, Miyazaki J, Smith AG. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat Genet. 2000;24(4):372-376. doi:10.1038/74199.

Ozair MZ, Noggle S, Warmflash A, Krzyspiak JE, Brivanlou AH. SMAD7 directly converts human embryonic stem cells to telencephalic fate by a default mechanism. Stem Cells. 2013;31(1):35-47. doi:10.1002/stem.1246.

Ozdamar B, Bose R, Barrios-Rodiles M, Wang HR, Zhang Y, Wrana JL. Regulation of the polarity protein Par6 by TGFbeta receptors controls epithelial cell plasticity. Science. 2005;307(5715):1603-1609. doi:10.1126/science.1105718.

Pera EM, Ikeda A, Eivers E, De Robertis EM. Integration of IGF, FGF, and anti-BMP signals via Smad1 phosphorylation in neural induction. Genes Dev. 2003;17(24):3023-3028. doi:10.1101/gad.1153603.

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42(Web Server issue):W187-W191. doi:10.1093/nar/gku365

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-140. doi:10.1093/bioinformatics/btp616

Schatz PJ. Use of peptide libraries to map the substrate specificity of a peptide-modifying enzyme: a 13 residue consensus peptide specifies biotinylation in Escherichia coli. Biotechnology (N Y). 1993; 11(10):1138-1143. doi:10.1038/nbt1093-1138.

Schmierer B, Hill CS. TGFbeta-SMAD signal transduction: molecular specificity and functional flexibility. Nat Rev Mol Cell Biol. 2007;8(12):970-982. doi:10.1038/nrm2297.

Shakèd M, Weissmüller K, Svoboda H, et al. Histone deacetylases control neurogenesis in embryonic brain by inhibition of BMP2/4 signaling. PLoS One. 2008;3(7):e2668. Published 2008 Jul 16. doi:10.1371/journal.pone.0002668.

Shi Y, Hata A, Lo RS, Massagué J, Pavletich NP. A structural basis for mutational inactivation of the tumour suppressor Smad4. Nature. 1997;388(6637):87-93. doi:10.1038/40431.

Shi Y, Massagué J. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. Cell. 2003;113(6):685-700. doi:10.1016/s0092-8674(03)00432-x.

Simonsson M, Kanduri M, Grönroos E, Heldin CH, Ericsson J. The DNA binding activities of Smad2 and Smad3 are regulated by coactivator-mediated acetylation. J Biol Chem. 2006;281(52):39870-39880. doi:10.1074/jbc.M607868200.

Smith AG. Embryo-derived stem cells: of mice and men. Annu Rev Cell Dev Biol. 2001;17:435-462. doi:10.1146/annurev.cellbio.17.1.435.

Sorrentino A, Thakur N, Grimsby S, et al. The type I TGF-beta receptor engages TRAF6 to activate TAK1 in a receptor kinase-independent manner. Nat Cell Biol. 2008;10(10):1199-1207. doi:10.1038/ncb1780.

Tesar PJ, Chenoweth JG, Brook FA, et al. New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature. 2007;448(7150):196-199. doi:10.1038/nature05972.

Thakur N, Hamidi A, Song J, et al. Smad7 Enhances TGF-β-Induced Transcription of c-Jun and HDAC6 Promoting Invasion of Prostate Cancer Cells [published online ahead of print, 2020 Sep 3]. iScience. 2020;23(9):101470. doi:10.1016/j.isci.2020.101470

Thomson M, Liu SJ, Zou LN, Smith Z, Meissner A, Ramanathan S. Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell. 2011;145(6):875-889. doi:10.1016/j.cell.2011.05.017.

Tsuneizumi K, Nakayama T, Kamoshida Y, Kornberg TB, Christian JL, Tabata T. Daughters against dpp modulates dpp organizing activity in Drosophila wing development. Nature. 1997;389(6651):627-631. doi:10.1038/39362.

Turner FB, Cheung WL, Cheung P. Chromatin immunoprecipitation assay for mammalian tissues. Methods Mol Biol. 2006; 325:261-272. doi:10.1385/1-59745-005-7:261.

Valouev A, Johnson DS, Sundquist A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods. 2008; 5(9):829-834. doi:10.1038/nmeth.1246.

Vardouli L, Moustakas A, Stournaras C. LIM-kinase 2 and cofilin phosphorylation mediate actin cytoskeleton reorganization induced by transforming growth factor-beta. J Biol Chem. 2005;280(12):11448-11457. doi:10.1074/jbc.M402651200.

Varelas X, Sakuma R, Samavarchi-Tehrani P, et al. TAZ controls Smad nucleocytoplasmic shuttling and regulates human embryonic stem-cell self-renewal. Nat Cell Biol. 2008;10(7):837-848. doi:10.1038/ncb1748.

Wakefield LM, Hill CS. Beyond TGFβ: roles of other TGFβ superfamily members in cancer. Nat Rev Cancer. 2013;13(5):328-341. doi:10.1038/nrc3500.

Wang Z, Oron E, Nelson B, Razis S, Ivanova N. Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. Cell Stem Cell. 2012;10(4):440-454. doi:10.1016/j.stem.2012.02.016.

Wu JW, Fairman R, Penry J, Shi Y. Formation of a stable heterodimer between Smad2 and Smad4. J Biol Chem. 2001;276(23):20688-20694. doi:10.1074/jbc.M100174200.

Xi Q, He W, Zhang XH, Le HV, Massagué J. Genome-wide impact of the BRG1 SWI/SNF chromatin remodeler on the transforming growth factor beta transcriptional program. J Biol Chem. 2008;283(2):1146-1155. doi:10.1074/jbc.M707479200.

Yamashita M, Fatyol K, Jin C, Wang X, Liu Z, Zhang YE. TRAF6 mediates Smad-independent activation of JNK and p38 by TGF-beta. Mol Cell. 2008;31(6):918-924. doi:10.1016/j.molcel.2008.09.002.

Yan X, Chen YG. Smad7: not only a regulator, but also a cross-talk mediator of TGF-β signalling. Biochem J. 2011;434(1):1-10. doi:10.1042/BJ20101827.

Young RA. Control of the embryonic stem cell state. Cell. 2011;144(6):940-954. doi:10.1016/j.cell.2011.01.032.

Yu Y, Gu S, Li W, et al. Smad7 enables STAT3 activation and promotes pluripotency independent of TGF-β signaling. Proc Natl Acad Sci U S A. 2017;114(38):10113-10118. doi:10.1073/pnas.1705755114.

Zhang L, Huang H, Zhou F, et al. RNF12 controls embryonic stem cell fate and morphogenesis in zebrafish embryos by targeting Smad7 for degradation [published correction appears in Mol Cell. 2012 Jul 27;47(2):330]. Mol Cell. 2012;46(5):650-661. doi:10.1016/j.molcel.2012.04.003.

Zhang S, Fei T, Zhang L, et al. Smad7 antagonizes transforming growth factor beta signaling in the nucleus by interfering with functional Smad-DNA complex formation. Mol Cell Biol. 2007;27(12):4488-4499. doi:10.1128/MCB.01636-06

Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137. doi:10.1186/gb-2008-9-9-r137

Zhang YE. Non-Smad pathways in TGF-beta signaling. Cell Res. 2009;19(1):128-139. doi:10.1038/cr.2008.328.

Part Two

Exploring the mechanism of Dnmt3b in stem cell pluripotency and lineage-specific differentiation

Chapter 4

INTRODUCTION

The establishment of DNA methylation patterns requires de novo methylation that occurs predominantly during early development and gametogenesis in mice. The two DNA methyltransferases, Dnmt3a and Dnmt3b, are essential for de novo methylation and for mouse development. Inactivation of both genes by gene targeting blocks de novo methylation in ES cells and early embryos, but it does not affect the maintenance of imprinted methylation patterns. Dnmt3a and Dnmt3b also exhibit non-overlapping functions in development, with Dnmt3b required explicitly for methylation of centromeric minor satellite repeats. In this part of the thesis, I focused my attention on the mechanism of Dnmt3b in mouse stem cell pluripotency and lineage-specific differentiation.

4.1 CpG Islands and DNA methylation

DNA methylation mainly occurs by covalent addition of a methyl group position C5 of cytosine (5 methylcytosine, 5mC) and constitutes the most conserved epigenetic modifications in animals and plants (Figure 4.1). In mammals, DNA methylation is essential for normal embryonic development, and it plays crucial roles in gene expression regulation, X chromosome inactivation, genomic imprinting, chromatin modification, and endogenous retroviruses silencing. Cytosine methylation in mammalian cells happens predominantly in CpG dinucleotides, resulting in a cell-type or developmental stage-specific pattern of DNA methylation. There is about 4% of methylated C residues in adult tissue, while DNA across the genome of mammalian somatic tissues is methylated at 70%-80% of all CpG sites.

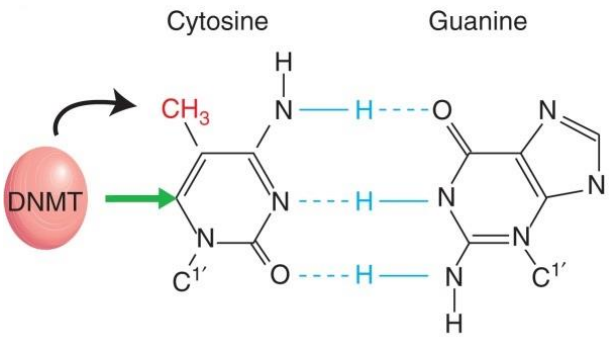


Figure 4.1 Cytosine methylation in DNA.

Addition of a methyl group, CH3 (red), at the five position of the cytosine pyrimidine ring (black arrow) does not sterically interfere with GC base pairing (blue lines). DNA methyltransferases associate covalently with the carbon 6 position (straight green arrow) during methyl group transfer.
(*Li, E., & Zhang, Y. Cold Spring Harbor perspectives in biology. 2014*)

CpG islands (CGIs) are ~1 kb GC-rich sequences nonmethylated in germ cells, the early embryo, and most somatic tissues (Bird A et al., 1985). These GC-rich regions were identified in early mapping studies of individual gene promoters (McKeon C et al., 1982) (Figure 4.2). It is now evident that most (if not all) CGIs mark the promoters and 5′ upstream regions of the genes. In fact, approximately 60% of human genes have CGI promoters in the proximity of the transcription start site (TSS), while other CpGs are mostly located within gene bodies.

Genes with CGI promoters expressed in a tissue-specific manner are usually expressed in early embryos and then in the somatic cells. In contrast to the norm, a distinct DNA methylation pattern is found on the inactive X chromosome in females, where CGIs become *de novo* methylated in large numbers during the embryonic process of X-chromosome inactivation in female placental mammals (Wolf SF et al., 1984). This process is essential for the leak-proof silencing of genes on the inactivated chromosome necessary for dosage compensation, as DNA methylation-deficient mice or cells show frequent transcriptional reactivation of X-linked genes (Brockdorff N and Turner BM, 2015).

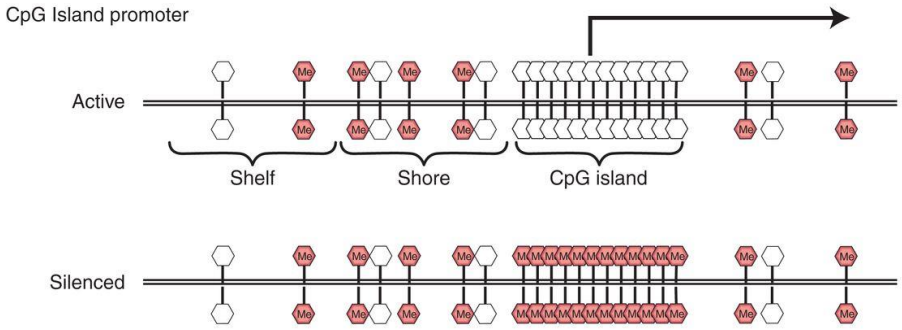


Figure 4.2 CpG island promoter.

CGIs are regions of high CpG density (>50%), usually 200 bp–2 kb in length that lack CpG methylation, found at promoters of most human genes. Long-term silencing of the gene can be insured by methylation of the CGI region. For example, genes on the inactive X chromosome and certain imprinted genes are silenced in this way. Also, in cancer cells certain genes are aberrantly silenced by CGI methylation. Shores are regions of the genome that reside up to 2 kb from CGIs, whereas shelves are found 2–4 kb away from CGIs. (*Li, E., & Zhang, Y. Cold Spring Harbor perspectives in biology. 2014*)

4.2 Epigenetic reprogramming during mammalian development

The variety of cellular states in multicellular organisms reflects the diversity of the transcriptional program of cells, even though nearly all cells in any given organism bear an identical genome sequence. The transcriptional state of a cell is governed by a specific set of transcriptional regulators and also by chemical modifications of the genome, including cytosine methylation (5-methylcytosine; 5mC) and post-translational modifications of histone tails, which regulate the accessibility of transcriptional regulators to, and the on or off states or expression levels of, all genes in the genome (Bonasio R et al., 2010). The stability of the phenotype of a cell upon mitosis/meiosis is considered to be underpinned by the stability of these modifications. Modifications that regulate the identity of a cell without changing the DNA sequence are referred to as epigenetic modifications (Bird A., 2007; Bonasio R et al., 2010), and the genome-wide state of the epigenetic status of a cell is known as the epigenome (Bernstein BE et al., 2007). Epigenetic modifications are crucial for the identity and stability of cells, and when aberrant, they can lead to disease.

The genome-wide epigenetic states of pre-implantation embryos and primordial germ cells (PGCs) undergo extensive reprogramming during mouse development (Figure 4.3). In detail, global DNA demethylation occurs in pre-implantation embryos and PGCs. DNA methylation in the paternal genome reaches the level in the maternal genome after fertilization, and then both genomes are further demethylated until a zygote develops into a blastocyst. After blastocyst development, post-implantation epiblast gains *de novo* global methylation during gastrulation and differentiates into three germ layers. PGCs emerged from the epiblast also undergo global demethylation. Germ cells (sperm or oocyte) gain *de novo* methylation during their maturation in a sex-dependent manner to establish genomic imprints. An improved understanding of the epigenetic reprogramming mechanisms that occur in these cells should provide important new information about the regulation of the epigenetic state and the mechanisms of induced pluripotency.

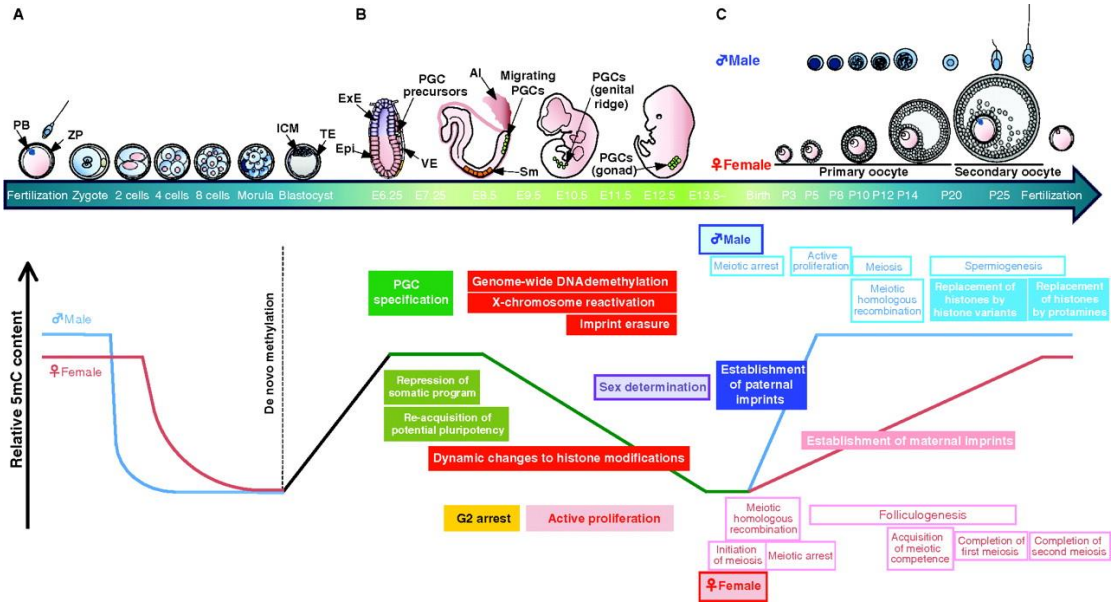


Figure 4.3 A schematic of mouse pre-implantation and germ cell development.

(Top) A schematic of pre-implantation and germ-cell development in mice. (A) Pre-implantation development stages; (B) post-implantation embryonic development, following blastocyst implantation at around E4.5; and (C) postnatal germ cell development and maturation. Primordial germ cell (PGC) precursors (E6.25) and PGCs are shown as green circles in embryos from E6.25 to E12.5. (Bottom) Key genetic and epigenetic events are shown that are associated with pre-implantation and germ cell development, together with relative levels of 5-methylcytosine (5mC) at different developmental stages. Al, allantois; Epi, epiblast; ExE, extraembryonic ectoderm; ICM, inner cell mass; PB, polar body; PGCs, primordial germ cells; Sm, somite; TE, trophectoderm; VE, visceral endoderm; ZP, zona pelucida. E indicates the embryonic day. (*Saitou M et al. Development. 2012*)

4.2.1 de novo DNA methyltransferases

De novo methyltransferases were discovered by searching for sequence homology with prokaryotic cytosine DNA methyltransferases using expressed sequence tag databases. Prokaryotic cytosine DNA methyltransferases share a set of conserved protein motifs (Pósfai J et al., 1989), and these features were also discovered in the mammalian maintenance DNA methyltransferase, Dnmt1. The homology search identified three genes that could potentially encode novel DNA methyltransferases (Figure 4.4). The other two genes, Dnmt3a and Dnmt3b, encoded related catalytically active polypeptides that showed no preference for methylating hemimethylated DNA *in vitro*, unlike Dnmt1 (Okano M et al., 1998). Inactivation of both Dnmt3a and Dnmt3b by gene targeting in ES cells confirmed that these genes constituted the missing *de novo* DNA methyltransferases — ES cells and embryos lacking both proteins were unable to *de novo* methylate proviral genomes and repetitive elements (Okano M et al., 1999). Moreover, Dnmt3a and an associated regulatory factor Dnmt3L were shown to be required for the establishment of distinct DNA methylation patterns found at imprinted genes (Hata K et al., 2002).

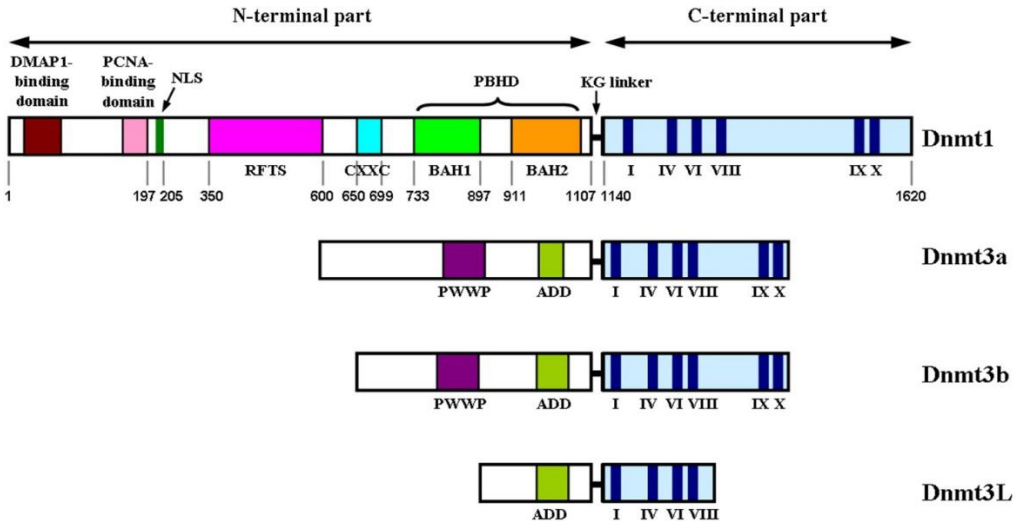


Figure 4.4 Structure and motif organization of mammalian DNA cytosine methyltransferases.

The C-terminal domains are marked by light blue filling. The other domains are signed. The conservative motifs are numbered using Roman numerals. (*Ryazanova, A. Yu. 2012*)

4.2.2 DNA methylation and histone modifications

It is crucial to indicate that the relationship between promoter DNA methylation and promoter activity depends on the CpG content of the promoters: high CpG promoters (HCPs), intermediate CpG promoters, or low CpG promoters (LCPs). In ES cells, HCP promoters are characterized by low DNA methylation levels, whereas LCP promoters are enriched in DNA methylation (Fouse SD et al., 2008; Mohn F et al., 2008; Meissner A et al., 2008). Moreover, specific histone modifications (H3K4me3 and H3K27me3) in HCPs appear to be more crucial for the expression of the corresponding genes and suggest a degree of protection from DNA methylation (Sørensen AL et al., 2010). On the contrary, methylated LCP promoters are depleted of bivalent histone marks and are mostly repressed in ES cells. It is suggested that silencing of pluripotency-related genes occurs through CpG promoter hypermethylation, while the gain of differentiation features is defined by gene regulation of Polycomb complex targets.

Specific epigenetic features at a global level also underpin the pluripotency of ES cells. Previous studies have demonstrated that ES cell chromatin is in a highly dynamic state with global DNA hypomethylation and a general abundance of transcriptionally active chromatin marks such as H3K4me3 and acetylation of histone H4, which is reflected in the relatively decondensed chromatin of ES cells (Reik W, 2007; Azuara V et al., 2006). This global lack of DNA methylation in stem cells could be associated with the ability of such cells to activate a wide range of cell-type-specific genes during the differentiation programs. It is true that DNA methylation and histone modifications work together and that the epigenetic inactivation of differentiation-specific genes in stem cells is usually repressed by alternative chromatin

remodeling factors, such as Polycomb proteins (Azuara V et al., 2006; Boyer LA et al., 2006). Consequently, further study of the interplay of all of the chromatin regulators is essential for understanding the dynamism of transcriptional control during stem cell self-renewal and differentiation.

4.2.3 Naïve and primed mouse pluripotent stem cells

After maternal predetermination gives way to zygotic regulation, a ground state is established within the mammalian embryo. This tabula rasa for embryogenesis is present only transiently in the preimplantation epiblast. Nichols J and Smith A (2009) proposed that two phases of pluripotency can be defined: naive and primed (Figure 4.5). This distinction extends to pluripotent stem cells derived from embryos or by molecular reprogramming ex vivo.

Since ICM cells can contribute to all cell lineages of the body, they are functionally described as having 'naive pluripotency'. Mouse naive ES cells derived from ICM of blastocysts have historically been maintained in serum and leukemia inhibitory factor (LIF) on feeder cells. And they indefinitely proliferate on a culture dish and differentiate into all three germ layers, including germ cell lines, when they are injected into blastocysts (Martin GR and Evans MJ, 1975; Smith AG et al., 1988; Williams RL et al., 1988). After post-implantation, mouse epiblast (EpiSC or EpiS) cells form an egg cylinder structure around E5.5-E6.5. EpiSC cells are derived from post-implantation epiblast under media containing basic fibroblast growth factor and activin and defined as having 'primed pluripotency' (Tesar PJ et al., 2007; Brons IG et al., 2007; Nichols J and Smith A, 2009) (Figure 4.6). EpiSC cells can differentiate into various cell types *in vitro* and form teratomas, but they cannot contribute to chimeras when injected into blastocysts (Tesar PJ et al., 2007). While the maintenance of mouse ES cells is dependent on the LIF/Stat3 signaling pathway, the maintenance of EpiS cells is dependent on the FGF/ERK pathway.

Consistent with the *in vivo* epigenetic property of female developing embryos, mouse ES cells have two active X chromosomes (XaXa), whereas one copy of X chromosomes is inactive (XaXi) in EpiSC cells (Figure 4.6). Transcriptionally, EpiSC cells express core pluripotent marker genes, including Nanog and Oct3/4. However, Klf4, Zfp42 (Rex1), and Dppa3 (Stella) are downregulated in primed EpiSC cells; Lefty1, Otx2, and Cer1 are upregulated in EpiSC cells compared to naive ES cells (Festuccia N et al., 2012; Tesar PJ et al., 2007). Thus, there are marked functional and molecular differences between the naive and primed states.

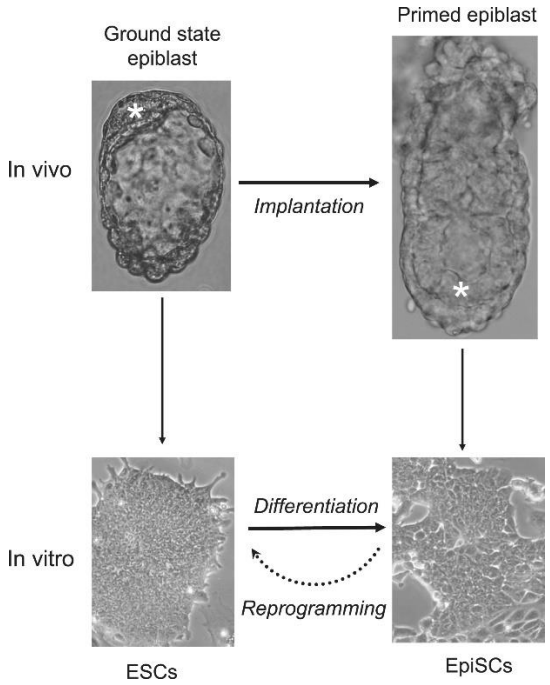


Figure 4.5 Two Phases of Pluripotency.

Upper images show mouse embryos at E4.5 and E5.5, or shortly before and shortly after implantation. The white asterisks indicate the epiblast. Note the layer of hypoblast underlying the epiblast in the blastocyst and the proamniotic cavity surrounded by epiblast in the post-implantation embryo. The epiblast is displaced downward after implantation due to proliferation of the trophectoderm-derived extraexembryonic ectoderm and the constraint of the uterine wall. Lower images show representative colonies of ESCs and EpiSCs.
(*Nichols J, Smith A. Cell Stem Cell. 2009*)

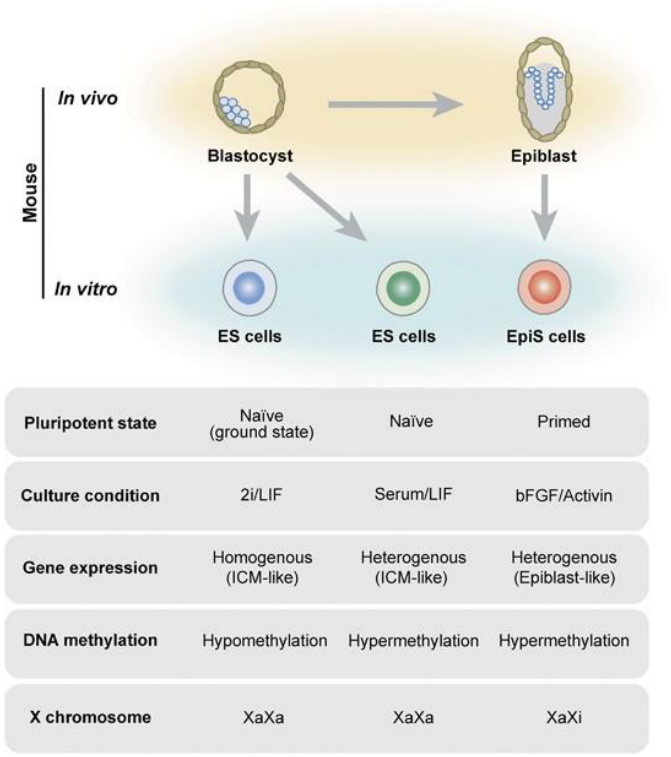


Figure 4.6 Transcriptional and epigenetic signatures in naive and primed mouse pluripotent stem cells.

In mice, ES cells derived from ICM of blastocysts are defined as having naive pluripotency, while EpiS cells derived from post-implantation epiblasts are defined as having primed pluripotency. The X chromosome status is XaXa in naive state and XaXi in primed state (Xa; active X chromosome and Xi; inactive X chromosome). Serum/LIF-cultured ES cells display global DNA hypermethylation and heterogeneous expression patterns of naive pluripotent marker genes. 2i/LIF media enables ES cells to maintain homogenous ground-state pluripotency and global DNA hypomethylation.
(*Yagi M et al. Lab Invest. 2017*)

4.2.4 de novo Methylation in ESC pluripotency and differentiation

Mouse ESCs are one of the most extensively studied systems for dissecting epigenetic mechanisms. The transcriptional circuitry associated with pluripotency is rapidly silenced during differentiation — in part through *de novo* methylation — as embryonic programs are resolved towards specific lineage differentiation. Furthermore, pluripotency represents a unique developmental window in which repetitive elements can be silenced *de novo*.

DNA methylation has a crucial role in ESC commitment but not in pluripotency maintenance or establishment. The complete erasure of methylation does not affect either the molecular signature of pluripotency or self-renewal. ESCs that lack all three DNA methyltransferases remain viable and do not show notable aneuploidy (Tsumura A et al., 2006). ESCs that are depleted explicitly of either maintenance or *de novo* methylation machinery lose nearly all

global methylation, albeit at markedly disparate rates and steady-state global values: loss of Dnmt1 induces rapid demethylation that stabilizes to ~20% of the average value, whereas the loss of Dnmt3a and Dnmt3b ESCs lose nearly all methylation over progressive divisions, indicating that the DNMT3 enzymes provide additional robustness to the inheritance of DNA methylation (Chen T et al., 2003; Jackson M et al., 2004). However, reintroduction of DNMT1 into Dnmt1-knockout ESCs restores previous methylation patterns, except imprints (Tucker KL et al., 1996). Although stem cell molecular identity is not impaired in the absence of DNA methylation, differentiation is almost completely inhibited. Methylation-free ESCs cannot upregulate germ-layer-associated markers and do not efficiently silence pluripotency factors. Acute deletion of the DNMT3 enzymes does not completely inhibit differentiation, suggesting that DNA methylation levels themselves, and not necessarily *de novo* silencing, may be responsible for this phenotype (Chen T et al., 2003; Jackson M et al., 2004).

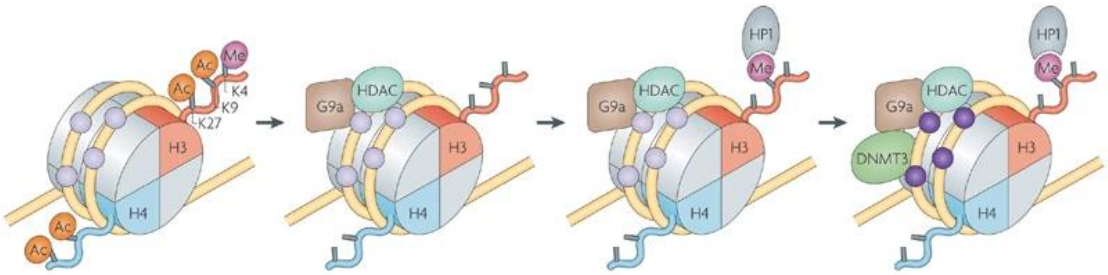


Figure 4.7 Turning off pluripotency genes.

In embryonic stem cells, pluripotency genes such as Oct3/4 and Nanog have unmethylated CpG islands (light purple circles) and are packaged with acetylated (Ac) histone H3 and H4 and methylated (Me) lysine 4 of histone H3 (H3K4). With the onset of differentiation, the SET domain-containing histone methyltransferase G9a is recruited, together with a histone deacetylase (HDAC), and this causes deacetylation of local histones. In addition, H3K4 is demethylated, but the enzymatic machinery responsible for this has not yet been identified. In the next step, G9a catalyzes the methylation of H3K9, and this modification serves as a binding site for the chromodomain protein heterochromatin protein 1 (HP1), thus generating a form of local heterochromatin. Finally, G9a recruits the methylases DNMT3A and DNMT3B, which mediate de novo methylation (dark purple circles) of the underlying DNA.
(*Cedar H, Bergman Y. Nat Rev Genet. 2009*)

It has been widely reported that maintenance of the pluripotency state is conferred by a set of development-associated transcription factors, such as OCT4, NANOG, and SOX2, that occupy promoters of active genes associated with self-renewal (Loh YH et al., 2006; Boyer LA et al., 2005). Expression of the transcription regulators mentioned above is usually controlled by CpG promoter methylation, and differentiation of ES cells is accomplished by partial or complete methylation of pluripotency-associated genes, resulting in their downregulation (Fouse SD et al., 2008; Farthing CR et al., 2008; Hawkins RD et al., 2010). As shown in Figure 4.7, the binding on repressors initiates silencing, followed by G9A-mediated H3K9 methylation, heterochromatin protein 1 (HP1) recruitment, and finally, *de novo* DNA methylation (Feldman N et al., 2006). The regulatory regions of Oct4 during silencing have been studied closely and,

although DNMT3A and DNMT3B show equal potential to initiate methylation at the proximal enhancer, DNMT3A more robustly triggers stable inheritance (Athanasiadou R et al., 2010). On the genome-scale, nucleosome-depleted regions associated with cell-type-specific regulation show pluripotency factor binding and DNA hypomethylation in ESCs (You JS et al., 2011). During differentiation, DNA methylation co-occurs with nucleosome assembly, thus inhibiting the binding of transcription factors. Oct4 silencing can be ectopically induced through artificial targeting of HP1α, which instructs H3K9 methylation followed by DNA methylation (Hathaway NA et al., 2012). In this system, *de novo* silencing outside pluripotent cells remains heritable after removing the targeting initiator, but in ESCs, removal leads to simultaneous Oct4 reactivation and demethylation.

4.2.5 Key regulators during cell fate determination

The trunk of murine embryos forms by continuous recruitment of cells generated in the primitive streak (PS), node-streak border (NSB), and caudal lateral ectoderm (CLE), located at the caudal end of the embryo, into the neural or mesodermal lineage thereby elongating the body anlage (Wilson et al., 2009). The source of cells giving rise to the spinal cord and mesodermal tissues, comprising the vertebral column, skeletal musculature, ventral body wall, kidneys, gonads, limbs, and others, is a resident progenitor cell type with self-renewing capability, the neuro-mesodermal progenitor (NMP). T, Tbx6, Fgfr1 and Wnt3a are expressed in the PS region and required for correct mesoderm production, and loss of each of them leads both to shortened axes, and the ectopic production of neural tissue at the expense of somitic mesoderm (Chapman DL and Papaioannou VE, 1998; Yamaguchi TP et al., 1999; Yoshikawa Y et al., 1997; Ciruna BG et al., 1997). This suggests that NMP maintenance is intimately linked with preserving a balance between neurectoderm and mesoderm production. Tbx6 expression in the midline PS represses Sox2 in mesoderm-fated cells, ensuring suppression of the neural transcription program (Takemoto T et al., 2011). Furthermore, in zebrafish, Wnt/β-catenin activation influences the decision of cells in both gastrula- and somite-stage embryos to enter neural or mesodermal lineages (Martin BL and Kimelman D, 2012). More recently, lineage-tracing experiments showed that conditional deletion of Wnt3a or β-catenin in the T+ cell compartment leads to a switch of primitive streak progenitors towards a neural fate (Garriock RJ et al., 2015). However, constitutive Wnt/β-catenin activity in the T+ cell compartment is not sufficient to divert all neural progenitors to mesoderm fates: providing cells in the caudal progenitor region with a stabilized form of β-catenin results in an enlarged PSM domain, but does not lead to loss of neural cell production (Aulehla A et al., 2008; Jurberg AD et al., 2014). Moreover, enhanced β-catenin activity does not necessarily compromise the presence of NMPs in the CLE (Garriock RJ et al., 2015). While these experiments point to an important role of

Wnt signalling in axial progenitors, the promoters used do not specifically target NMPs. Grafting of precise NMP areas can provide a complementary approach that allows a direct assessment of the currently unresolved roles of Wnt signalling in NMPs and the caudal-most CLE.

In particular, the Sox2 gene is one of the earliest transcription factors expressed in the neurodevelopment. It plays a critical role in maintaining the pluripotency of stem cells and is also involved in neural differentiation. In pluripotent embryonic stem (ES) cells, Sox2 is intimately involved in the control of Nanog and Oct-4, two key determinants of self-renewal. Several regulatory regions have been identified for the Sox2 gene, including the core promoter (Lengler J et al., 2005) as well as upstream [∼4 kb from the transcription start site (TSS)] and downstream (∼4 kb from the TSS) enhancers (Uchikawa M et al., 2003; Miyagi S et al., 2004, 2006; Catena R et al., 2004). In mouse, these regions are designated SRR1 and SRR2, respectively, and are conserved in human DNA (Tomioka M et al., 2002). SRR1 contains POU transcription factor motifs and can direct neural-specific Sox2 expression. SRR2 contains a bipartite Oct4-Sox2 motif and is essential for the expression of Sox2 in both ES cells and neural stem cells; thus, the SRR2 enhancer participates in the tight coregulation of Sox2, Oct4, and Nanog expression to maintain pluripotency in stem cells.

4.3 Methods for genome-wide quantitation of DNA methylation

To understand the functions of DNA methylation, it is first necessary to map methyl CpG in the genome and its dynamic changes during cell proliferation and differentiation or in development and disease. Several methods have been developed for quantitative analysis of DNA methylation at the genome scale and in a gene locus-specific manner.

Bisulfite-sequencing (Frommer M et al., 1992) This is the most reliable method for testing all cytosines within a region of the genome. It involves the “bisulfite modification” of single-stranded DNA, which leads to the deamination of unmodified cytosines, whereas 5-methylcytosine is protected. As a result, cytosines that survive bisulfite treatment are identified as methylated. Because of its high resolution and identification of methylated cytosine, this is the method of analyzing DNA methylation patterns. In combination with next-generation sequencing, the technique is widely used for methylome analysis at the whole-genome scale. Several polymerase chain reaction (PCR)-based methods that depend on prior bisulfite treatment of DNA have also been developed for rapid analysis of methylation of genes of interest (Herman JG et al., 1996).

MeDIP (Weber M et al., 2005) Methylated DNA immunoprecipitation (MeDIP) is a versatile

approach for unbiased detection of methylated DNA and can be applied to generate comprehensive DNA methylation profiles on a genome-wide scale. This method uses a monoclonal antibody that explicitly recognizes 5-methylcytidine to enrich methylated genomic DNA fragments by immunoprecipitation. The methylation status is then determined by PCR for specific regions or DNA microarrays for the whole genome.

Pyrosequencing (Tost J and Gut IG, 2007) Analysis of DNA methylation patterns by pyrosequencing yields reproducible and accurate measures of the degree of methylation at several CpGs near high resolution. After bisulfite treatment and PCR, the degree of each methylation at each CpG position in a sequence is determined from the ratio of T and C. The process of purification and sequencing can be repeated for the same template to analyze other CpGs in the same amplification product. The method is susceptible and quantitative, and it is often applied to methylation analysis of specific regions.

CHARM DNA methylation analysis (Irizarry RA et al., 2009) Comprehensive high-throughput array-based relative methylation (CHARM) analysis is a microarray-based method. It can be applied to custom-designed microarray covering the whole genome (usually nonrepetitive sequences) or specific regions (e.g., all CGIs). The technique is quantitative, and data analysis is straightforward. It has an advantage over other methods when comparing DNA methylation patterns from a large number of samples.

Immunofluorescence analysis 5mC/5hmC is marked by specific antibodies and fluorophore-conjugated secondary antibodies in situ, followed by fluorescent microscopic analysis. This method cannot distinguish the methylation state of specific sequences but detects genome-wide methylation levels in single cells. 5mC/5hmC antibodies detect densely methylated sequences efficiently, but single CpG methylation less efficiently (Pastor WA et al., 2011; Suzuki MM and Bird A, 2008). As transposon-related elements occupy ~40% of the genome and genes only ~2-3%, most 5mC/5hmC signals should be derived from the methylation of transposon-related elements.

4.4 Aim of this study

The establishment of DNA methylation patterns requires *de novo* methylation that occurs predominantly during early development and gametogenesis in mice. *De novo* methylation during early embryogenesis is catalyzed by DNA methyltransferase 3b (Dnmt3b), and the absence of DNA methylation leads to ectopic gene activation in the embryo. Therefore, it will be interesting to define the role of Dnmt3b-mediated methylation in early embryo development and afterward in lineage specification. In this work, we will take advantage of the loss of

function and multi-omics analysis and develop the EpiSC model *in vitro* directly from ESC to decipher the role of Dnmt3b in lineage-specific determination.

Chapter 5

RESULTS

5.1 Comparison between two states of pluripotency: mESCs and mEpiSCs

The establishment of DNA methylation patterns requires de novo methylation that occurs predominantly during early development and gametogenesis in mice. De novo methylation during early embryogenesis is catalyzed by DNA methyltransferase 3b (Dnmt3b), and the absence of DNA methylation leads to ectopic gene activation in the embryo. Intriguingly, at implantation, the mouse embryo is principally composed of the epiblast, a pluripotent derivative of the inner cell mass (ICM), which was referred to as EpiSCs (post-implantation epiblast derived stem cells) (Tesar PJ et al., 2007). The EpiSC lines are distinct from mouse ES cells in their epigenetic state and the signals controlling their differentiation (Tesar PJ et al., 2007; ten Berge D et al., 2011). To investigate the transcriptional and epigenetic regulation mechanism in mouse ESC and EpiSC, we tried to obtain the EpiSC through *in vitro* induction starting from E14 ES cells by taking advantage of ActivinA/Nodal signaling pathway.

5.1.1 In vitro mouse epiblast (mEpiSCs) induction

To obtain EpiSC *in vitro*, we induced ES cells by adding ActivinA in the medium followed by 14 days of culture, and we observed significant morphology changes. Mouse ES cells grew in small, compact, domed colonies, while EpiSC colonies were larger and grew as a monolayer (Figure 5.1A). Then, we performed real-time PCR analysis on the cells from different passages, the results showed that Nanog and Pou5f1 maintained a relatively stable expression level during the EpiSC induction, and the expression level of Pou5f1 was also confirmed by western blot (Figure 5.1B and C). The expression levels of two EpiSC markers Fgf5 and Lefty1, strikingly increased with passages (Figure 5.1B). Western blot analysis also showed that Dnmt3b and two isoforms of Dnmt3a increased the expression in EpiSC compared with ESC, while Dnmt3l was expressed in ES cells and significantly decreased or not detected in EpiSC (Figure 5.1C). Genes associated with the epiblast and early germ layers such as Lefty1 and Otx2 were expressed at higher levels in EpiSC (Figure 5.1C). This data was also confirmed by immunofluorescence assay, which showed the increased expression of Dnmt3b and Fgf5 in EpiSC compared with ESC (Figure 5.1D and E). Furthermore, we performed methylation analysis at the promoter region for the genes associated with ICM (Zfp42 and Dppa3) in ESC and EpiSC through bisulfite conversion. As shown in Figure 5.1F, the promoter regions were hypermethylated in EpiSC for both two genes compared with ESC. These data suggest that distinct transcriptional networks operate to maintain pluripotency in epiblast cells.

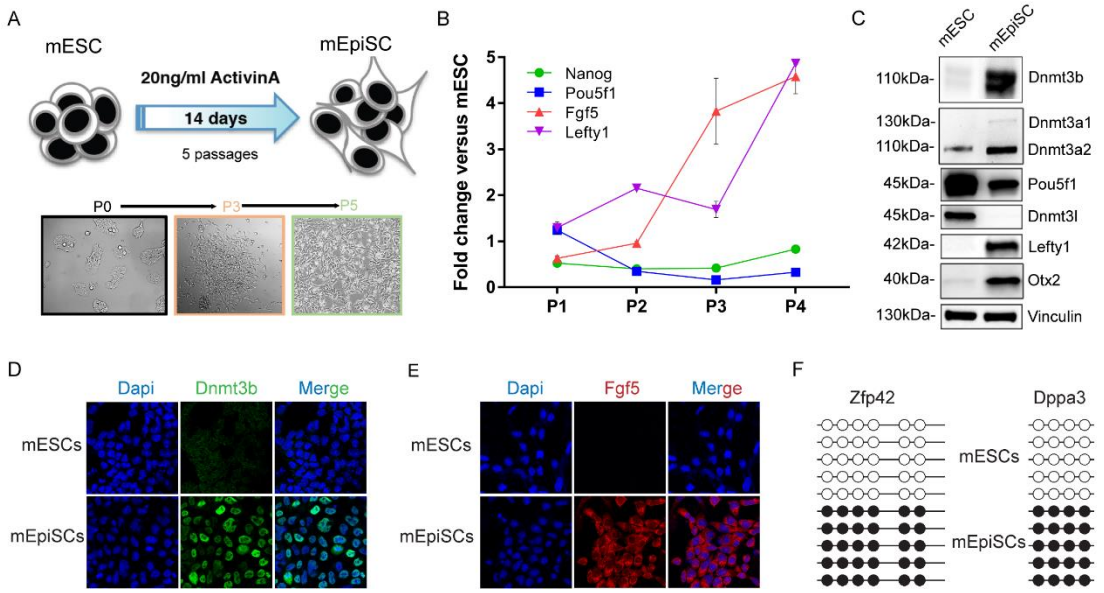


Figure 5.1. Characterization of *in vitro* mEpiSCs induction.

(A) Schematic of the experimental set-up of *in vitro* mEpiSCs induction starting from mESCs. (Up) mEpiSCs were obtained by using 20 ng/ml ActivinA in the medium following by five passages during the 14-days culture. (Down) Bright-field images of the cells at passage 0, 3, 5 (P0, P3, P5).

(B) Line graph showing the relative expression of naïve markers (Nanog and Pou5f1) and primed markers (Fgf5 and Lefty1) normalized to mESCs. P1, P2, P3, P4 represent passage 1, 2, 3, 4. Data were represented as mean ± SEM.

(C) Western blotting showing the protein expression of DNA methyltransferase (Dnmt3b, Dnmt3a (Dnmt3a1, Dnmt3a2), Dnmt3l), pluripotency marker (Pou5f1), and EpiSC markers (Lefty1, Otx2) in mESC and mEpiSC. Vinculin was used as a loading control.

(D) Immunofluorescence assay showing the expression of Dnmt3b in mESC and mEpiSC. Nuclei were stained with DAPI.

(E) Immunofluorescence assay showing the expression of EpiSC marker (Fgf5) in mESC and mEpiSC. Nuclei were stained with DAPI.

(F) Bisulfite sequencing of Zfp42 and Dppa3 in mESC and mEpiSC at the promoter region (n = 5 per group). Black dot, methylated CpG site; white dot, unmethylated CpG site.

5.1.2 Lack of Dnmt3b does not affect the transition from mESC to mEpiSC

After the successful induction of EpiSC *in vitro*, we asked ourselves whether the lack of Dnmt3b affect the formation of EpiSC. As shown in Figure 5.2A, AP staining showed that the EpiSC colonies in the 3BKO cells were similar compared with WT cells. To better understand the transcriptome regulation mechanism in EpiSC, we performed RNA-seq analysis in ESC and EpiSC cells for both WT and 3BKO cells. The data showed that 3BKO cells had similar transcription regulation pattern in comparison to WT cells. In detail, the core pluripotency markers (Tdgf1, Gdf3, Fgf4 and Sox2) significantly decreased in EpiSC and were expressed at equivalent levels in 3BKO cells compared to WT cells (Figure 5.2B). Transcripts from genes related to naïve pluripotency such as Dppa3, Prdm14, Zfp42, Esrrb and Dnmt3l were expressed by mouse ES cells and significantly decreased or not detected in EpiSCs (Figure 5.2B). Genes associated with the epiblast and early germ layers such as Dnmt3a, Fgf5, Lefty1, Foxa2, Cer1 and Fgf15 were expressed at higher levels in EpiSCs (Figure 5.2B). This data was also confirmed by western blot and immunofluorescence assay (Figure 5.2C and 5.2D). Taken together, these data emphasize that Dnmt3b depletion does not impact on the EpiSC induction *in vitro*.

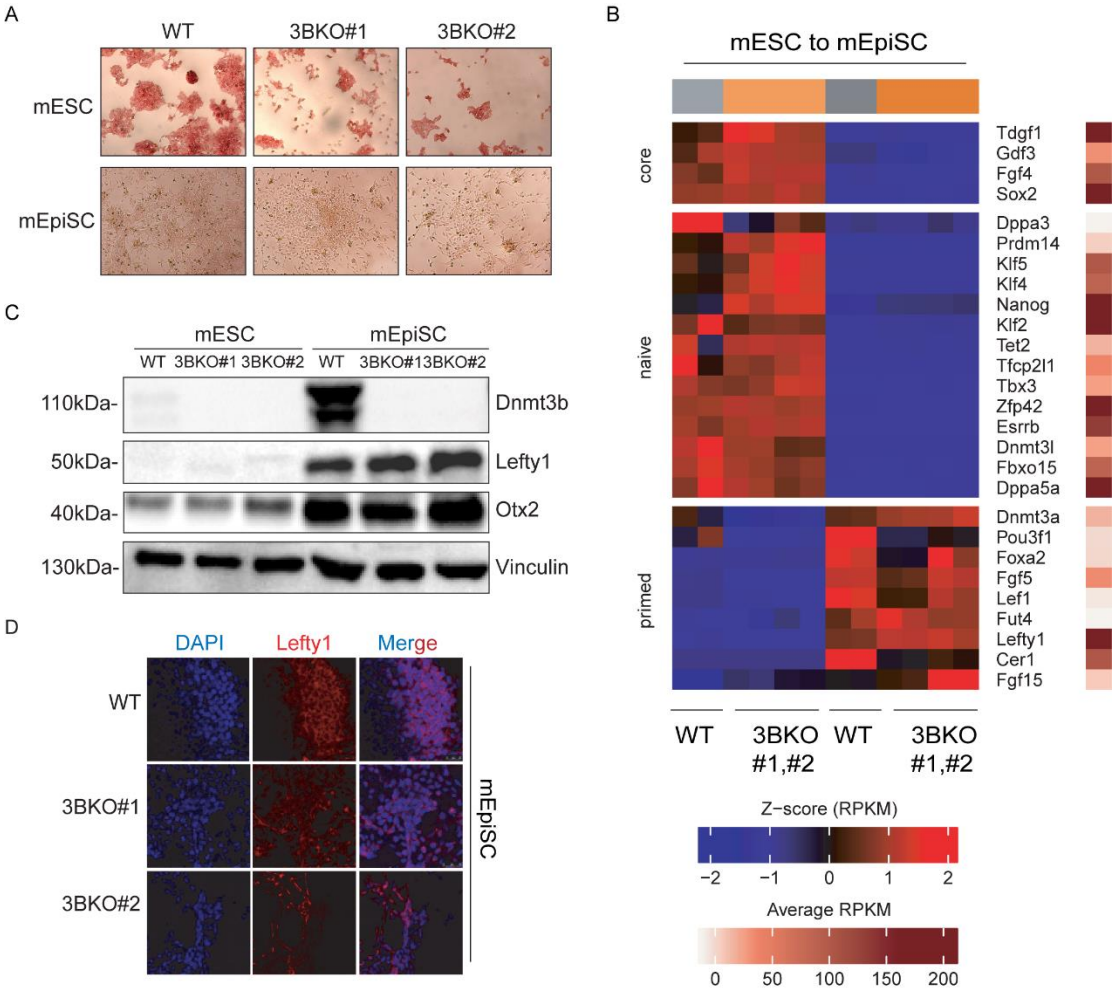


Figure 5.2. Characterization of *in vitro* mEpiSCs induction in Dnmt3b knockout cells.

(A) Photos of Alkaline Phosphatase (AP) staining in WT and Dnmt3b knockout mESC and mEpiSC cells.

(B) RNA-seq heatmap displaying the significantly dysregulated genes related to core, naïve and primed pluripotency when comparing 3BKO with WT expression levels during the transition from mESC to mEpiSC. Two replicates for each were used in this experiment. The significance threshold is fold change ≥ 1.5 and FDR ≤ 0.01. Relative expressions were represented by Z-scores to generate the heatmap.

(C) Western blotting showing the protein expression of Dnmt3b and EpiSC markers (Lefty1, Otx2) in WT and two Dnmt3b knockout cells (3BKO#1, 3BKO#2) during the transition from mESC to mEpiSC. Vinculin was used as a loading control.

(D) Immunofluorescence assay showing the expression of Lefty1 in WT and two Dnmt3b knockout cells (3BKO#1, 3BKO#2) in mESC and mEpiSC. Nuclei was stained with DAPI.

5.2 In vitro lineage-specific differentiation protocol set-up in mESCs

Cell differentiation during embryogenesis is a delicate process in which transcription and repression of genes must be scrupulously coordinated; pluripotency genes are switched off, and lineage-specific genes start to be transcribed. Embryonic stem cells (ESCs) are pluripotent cells derived from the inner cell mass (ICM) of the blastocyst, and they can self-renew indefinitely. Under proper stimuli, ESCs can differentiate into three germ layers: ectoderm, mesoderm, and endoderm, and develop potentially into different organs. To explore the transcriptional regulation of the embryonic developmental process, we set the protocol to induce the differentiation of ESCs into the progenitors of meso-endoderm (ME) and neuro-ectoderm (NE).

5.2.1 In vitro meso-endoderm differentiation in mESCs.

ME progenitors were obtained through the WNT pathway activation by using a Wnt3a agonist (iGsk3-β) (Figure 5.3A). We collected the RNA and protein at different time points (0 hr, 24 hrs, 48 hrs, 72 hrs, and 96 hrs) to characterize the differentiation process. It is already known that WNT/β-catenin signaling promotes differentiation in ESCs, by the induction of Brachyury (T) expression. Gene expression was analyzed by Real-time PCR (RT-PCR), detecting stemness marker (Nanog and Oct4) and lineage-specific markers for ME (T, Gata4).

As expected, the induction of ME differentiation *in vitro* was characterized by down-regulation of stemness markers (Nanog and Oct4); and the concurrent up-regulation of specific ME genes (T and Gata4) (Figure 5.3B). The result was also confirmed by western blotting detecting the increased expression of Gata4 (Figure 5.3C).

To better define the gene transcriptional regulation in our differentiation model, we performed paired-end, high throughput RNA sequencing (RNA-seq) on five different differentiation time points (0 hr, 24 hrs, 48 hrs, 72 hrs, and 96 hrs). In particular, we determined 1101 genes upregulated (Log (fold change) > 2) and 2543 genes downregulated (Log (fold change) < -2) in ME differentiation; Gene Ontology (GO) analysis showed that mesoderm development and nervous system development were significantly overrepresented (P-value < 0.05) in up-regulation categories and down-regulation categories, respectively (Figure 5.3D and E).

5.2.2 In vitro neuro-ectoderm differentiation in mESCs.

NE progenitors were induced by taking the advantage of all-trans retinoic acid (atRA), a metabolic product of vitamin A (retinol) (Figure 5.4A). Several papers demonstrated that atRA induced the expression of neural-specific genes, such as Sox1, and a consequent down-regulation of pluripotency markers. As shown in Figure 5.4B and 5.4C, the expression of ESC pluripotency markers (Nanog and Pou5f1) was decreased, and Sox1 achieved a peak expression at 48 hours; the expression of Nestin increased gradually during the time course differentiation.

Meanwhile, we performed RNA-seq analysis on five different differentiation time points (0 hr, 24 hrs, 48 hrs, 72 hrs, and 96 hrs). We identified 3135 upregulated genes (Log (fold change) > 2) and 791 downregulated genes (Log (fold change) < -2) in NE differentiation. Gene ontology (GO) analysis showed that ectoderm development and mesoderm development were significantly overrepresented (P-value < 0.05) in up-regulation categories and down-regulation categories, respectively (Figure 5.4D and E).

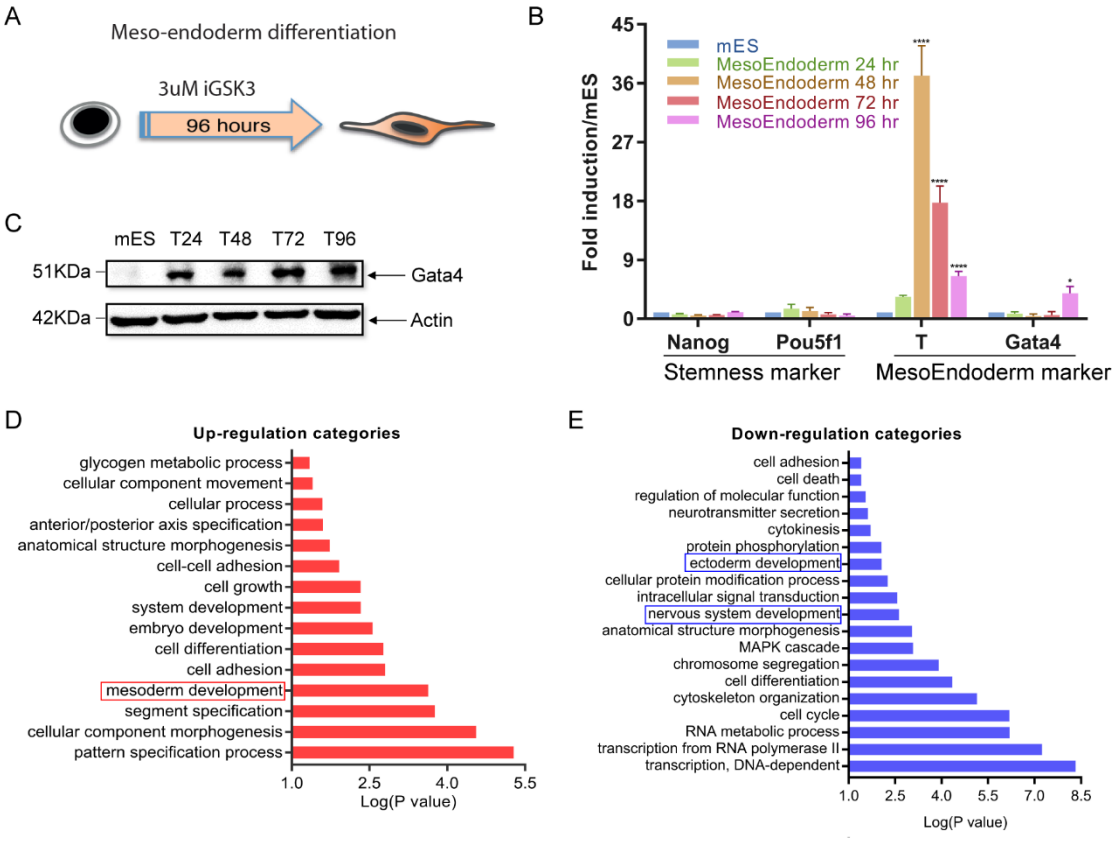


Figure 5.3. *In vitro* meso-endoderm differentiation protocol.

(A) Diagram of the meso-endoderm differentiation. mESCs were grown in N2B27 medium supplied with 3mM of iGsk3 for 96 hours followed by RNA and protein samples collection.
(B) Bar graphs showing relative mRNA expression of stemness markers (Nanog, Oct4) and meso-endoderm markers (T, Gata4) in mESCs during time course differentiation. Each time point was normalized to day 0. Error bars indicate SEM (n = 3). P values were calculated using Two-way ANOVA test. *P < 0.05; ****P < 0.001.
(C) Protein levels of Gata4 during time course differentiation. β-Actin was used as a loading control.
(D) A bar graph showing enriched GO terms of strong upregulated genes (Log (fold change) > 2).
(E) A bar graph showing enriched GO terms of strong downregulated genes (Log (fold change) < -2).

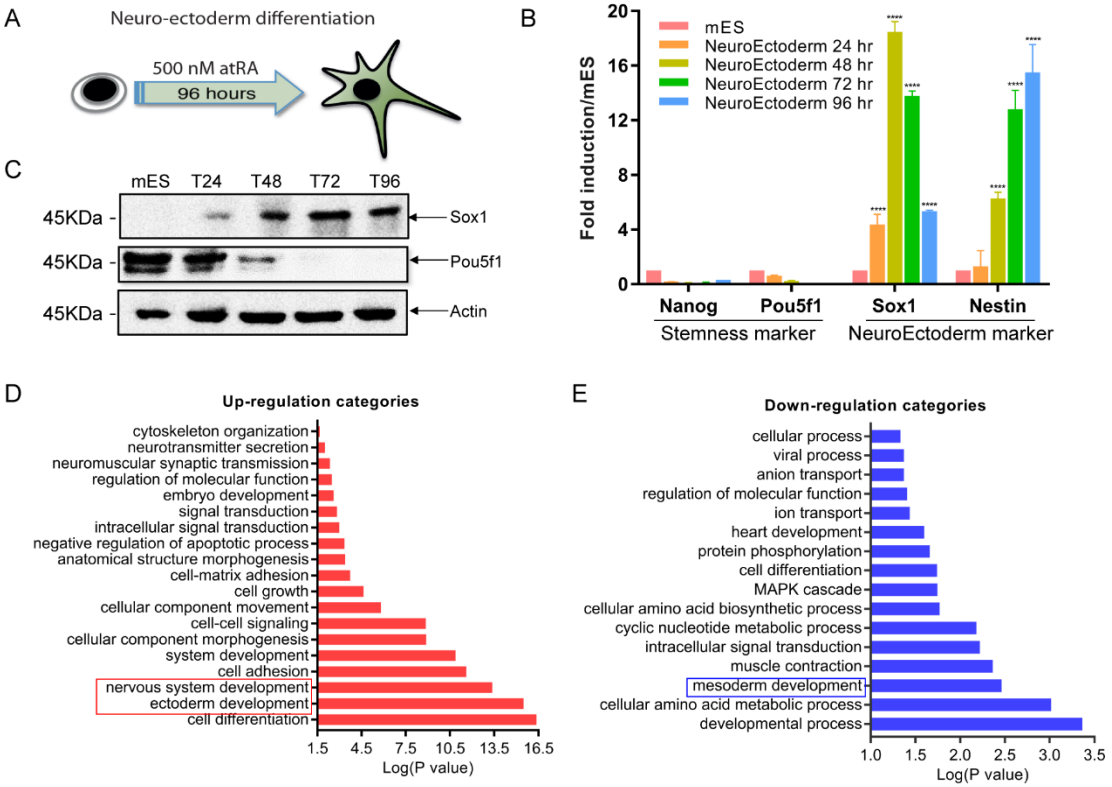


Figure 5.4. *In vitro* neuro-ectoderm differentiation protocol.

(A) Diagram of the neuro-ectoderm differentiation. mESCs were grown in N2B27 medium supplied with 500 nM of atRA for 96 hours followed by RNA and protein samples collection.

(B) Bar graphs showing relative mRNA expression of stemness markers (Nanog, Oct4) and neuro-ectoderm markers (Sox1, Nestin) in mESCs during time course differentiation. Each time point was normalized to day 0. Error bars indicate SEM (n = 3). P values were calculated using Two-way ANOVA test. ****P < 0.001.

(C) Protein levels of Sox1 and Pou5f1 during time course differentiation. β-Actin was used as a loading control.

(D) A bar graph showing enriched GO terms of strong upregulated genes (Log (fold change) > 2).

(E) A bar graph showing enriched GO terms of strong downregulated genes (Log (fold change) < -2).

5.3 Dnmt3b depletion impairs the meso-endoderm differentiation

5.3.1 Dnmt3b lack impairs the expression of meso-endoderm markers.

To further investigate whether the lack of Dnmt3b during cell differentiation can directly impair the physiological properties of the resulting cellular populations, in addition to modulating lineage-fate decisions or inducing proliferation or apoptosis. We performed *in vitro* meso-endoderm lineage differentiation in EpiSC through the activation of the WNT pathway by using the Gsk3β inhibitor, and the epiblast was committed towards a meso-endodermal fate (Figure 5.5A). As shown in Figure 5.5B, both *de novo* DNMTs (Dnmt3a and Dnmt3b) were significantly upregulated in the EpiSC stage and downregulated during meso-endoderm fate commitment; Intriguingly, the expression of Dnmt3b was much higher than Dnmt3a during this developmental process which suggests a vital role of Dnmt3b in meso-endoderm lineage specification.

To better understand the role of Dnmt3b during meso-endoderm specification, we first performed gene expression profiling using bulk RNA sequencing on EpiSC and meso-endoderm stages for WT and 3BKO cells. In particular, as shown in Figure 5.5C, the genes that show a similar expression pattern in both WT and 3BKO cells are characterized by pluripotent genes that are stably downregulated after the meso-endodermal differentiation, for example, Nanog and Pou5f1 for naïve pluripotency, Fgf5, Lefty1 and Otx2 for primed pluripotency. On the other hand, several genes show a divergent expression pattern between WT and 3BKO cells. Specifically, the genes that are upregulated in WT cells after meso-endoderm differentiation and downregulated in 3BKO cells compared to WT cells show the mesoderm formation characteristics (i.e., Kdr, T, Tbx6, Wnt3a); the genes that are upregulated in 3BKO cells with respect to WT after meso-endodermal commitment are associated with neural development (i.e., Sox1, Sox2, Pax6). The result is also confirmed by western blot which showed that T, a mesoderm marker, is highly expressed in WT cells compared to 3BKO cells at meso-endoderm T24 (Figure 5.5B); while the expression level of Sox1 significantly increased in 3BKO cells during the differentiation (Figure 5.5B). We evaluated the efficiency of meso-endoderm through immunofluorescence assay at meso-endoderm T24 and T48 (Figure 5.5E). We found a significantly reduced percentage of T^+ cells in the 3BKO cells (Figure 5.5F; from 70% in WT to 15% in 3BKO cells at T24; and from 24% in WT to 0.35% in 3BKO cells at T48, $p < 0.0001$). In contrast, 3BKO cultures expressed more $Sox1^+$ cells (Figure 5.5F; from 0.5% in WT to 55% in 3BKO cells at T48, $p < 0.0001$). Furthermore, FACS analysis (Figure 5.5D) showed that the meso-endodermal surface marker CD31 is significantly decreased in 3BKO cells after the meso-endoderm commitment compared to WT cells.

Taken together, we observe a robust impairment during the differentiation towards meso-endoderm from 3BKO ESCs. These differences were consistent for both 3BKO clones and indicate that Dnmt3b and/or at least a subset of its targets are involved in the meso-endoderm differentiation process.

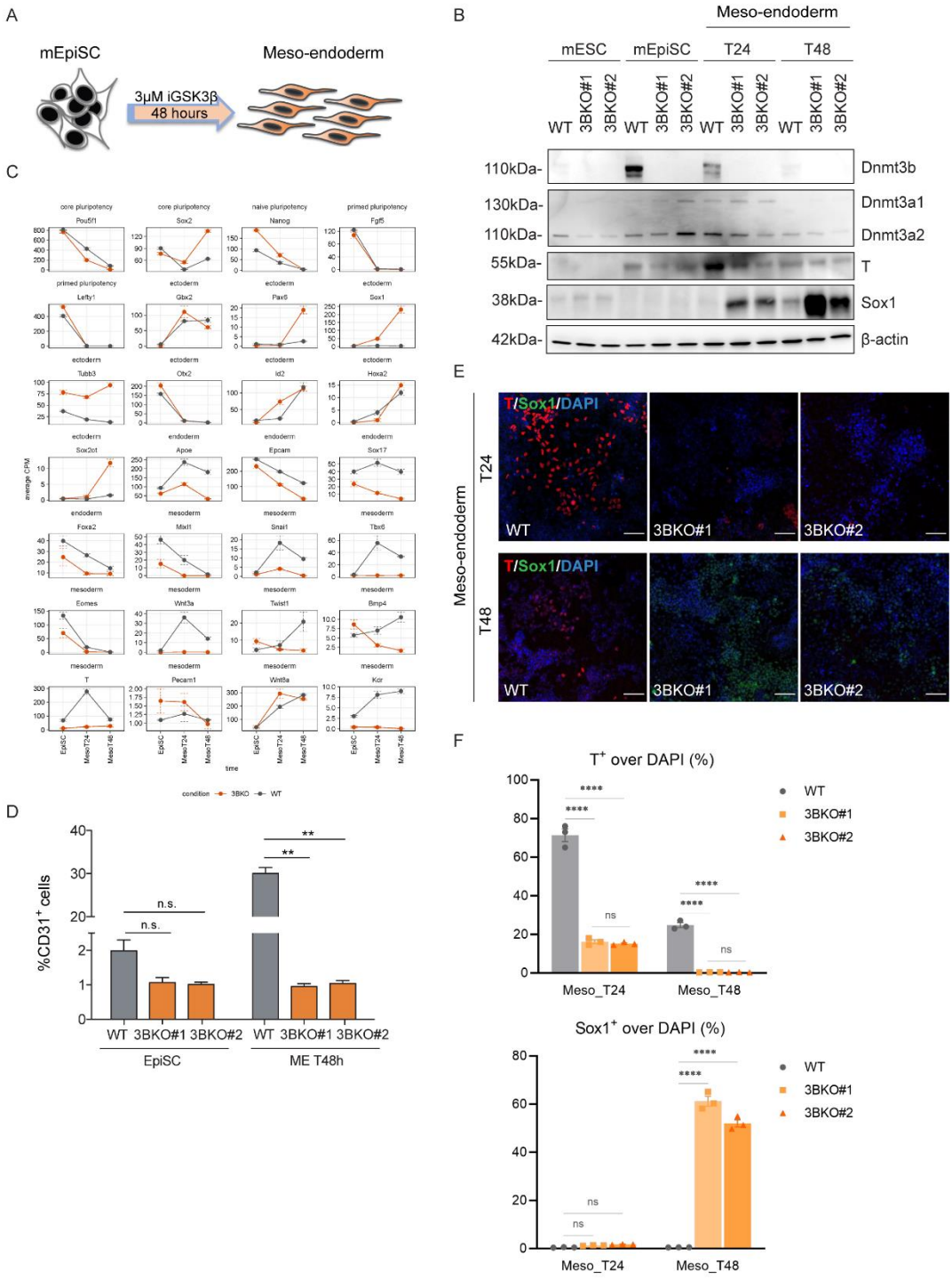


Figure 5.5. Dnmt3b KO cells had defects towards meso-endoderm progenitor differentiation.

(A) Schematic of experimental set-up for *in vitro* meso-endoderm progenitor differentiation starting from mEpiSCs. Meso-endoderm were obtained by using 3 μM iGSK3β in the medium during the 48-hour culture.
(B) Western blotting showing the expression of Dnmt3b and Dnmt3a (Dnmt3a1 and Dnmt3a2) in mESC, mEpiSC and meso-endoderm differentiation. β-Actin was used as a loading control.
(C) RNA-seq heatmap displaying the significantly dysregulated genes related to ectoderm, endoderm, mesoderm and primed pluripotency when comparing 3BKO with WT expression levels during the meso-endoderm differentiation starting from mEpiSC. Three time points (mEpiSC, meso-endoderm T24 and T48) were represented here. Two replicates for each time point were used in this experiment. The significance threshold is fold change ≥ 1.5 and FDR ≤ 0.01. Relative expressions were represented by average CPM to generate the heatmap.
(D) Flow cytometry analysis showing the percentage of CD31 positive cells in EpiSC and 48-hour meso-endoderm differentiation for WT and two Dnmt3b KO cells. Error bars indicate SEM (n = 3). P values were calculated using Two-way ANOVA test. **P < 0.01.
(E) Immunofluorescence assay showing the expression level of T and Sox1 at meso-endoderm differentiation T24 and T48 for WT and two Dnmt3b KO cells. Nuclei was stained with DAPI. Scale bar, 50 μm.
(F) Quantification of T, and Sox1+ cell percentages across cell lines and the parental isogenic control at Meso_T24 and Meso_T48 presented as means ± SEM; ANOVA (****p < 0.0001), n.s: not significant.

5.3.2 Defects in meso-endoderm progenitor differentiation in Dnmt3b KO cells can be rescued by the silencing of Sox2.

It has been reported that Sox2 and Brachyury play antagonistic roles in the specification of neural and mesodermal fates in mESCs (Koch F et al., 2017), which arouses our interests in the reason why Dnmt3b KO cells were not able to differentiate into meso-endoderm progenitors. We were asking ourselves whether Sox2 blocks the meso-endoderm differentiation in Dnmt3b KO cells, and whether the silencing of Sox2 could rescue the meso-endoderm lineage markers expression.

First, we looked at the CpG methylation levels at the promoter of Sox2 in mEpiSCs for WT and two Dnmt3b KO cells by using bisulfite sequencing analysis, the result showed that the Sox2 promoter were hypomethylated in both WT and Dnmt3b KO cells (Figure 5.6A). Then we performed methylation analysis at the enhancer of Sox2, surprisingly, the SRR2 element of Sox2 enhancer were hypermethylated in WT cells while hypomethylated in Dnmt3b KO cells (Figure 5.6B). The RT-qPCR and immunofluorescence assay showed that Sox2 dramatically upregulated in Dnmt3b KO cells compared to WT cells at the meso-endoderm progenitor differentiation 48 hours (Figure 5.6C and D).

Second, as shown in Figure 5.6E, we tried to perform the silencing of Sox2 through shRNA interference in Dnmt3b KO cells during the meso-endoderm differentiation (See EXPERIMENTAL PROCEDURES). Finally, we obtained about 50% silencing of Sox2 in both two Dnmt3b KO cells (Figure 5.6F and G), then qRT-PCR analysis showed that the meso-endoderm markers (T and Gata4) were significantly increased after the Sox2 silencing in comparison with the control (Figure 5.6G). This suggests that Dnmt3b may predetermine the meso-endoderm progenitor lineage commitment in epiblast stage, where it blocks the role of Sox2 acting as a neural development marker.

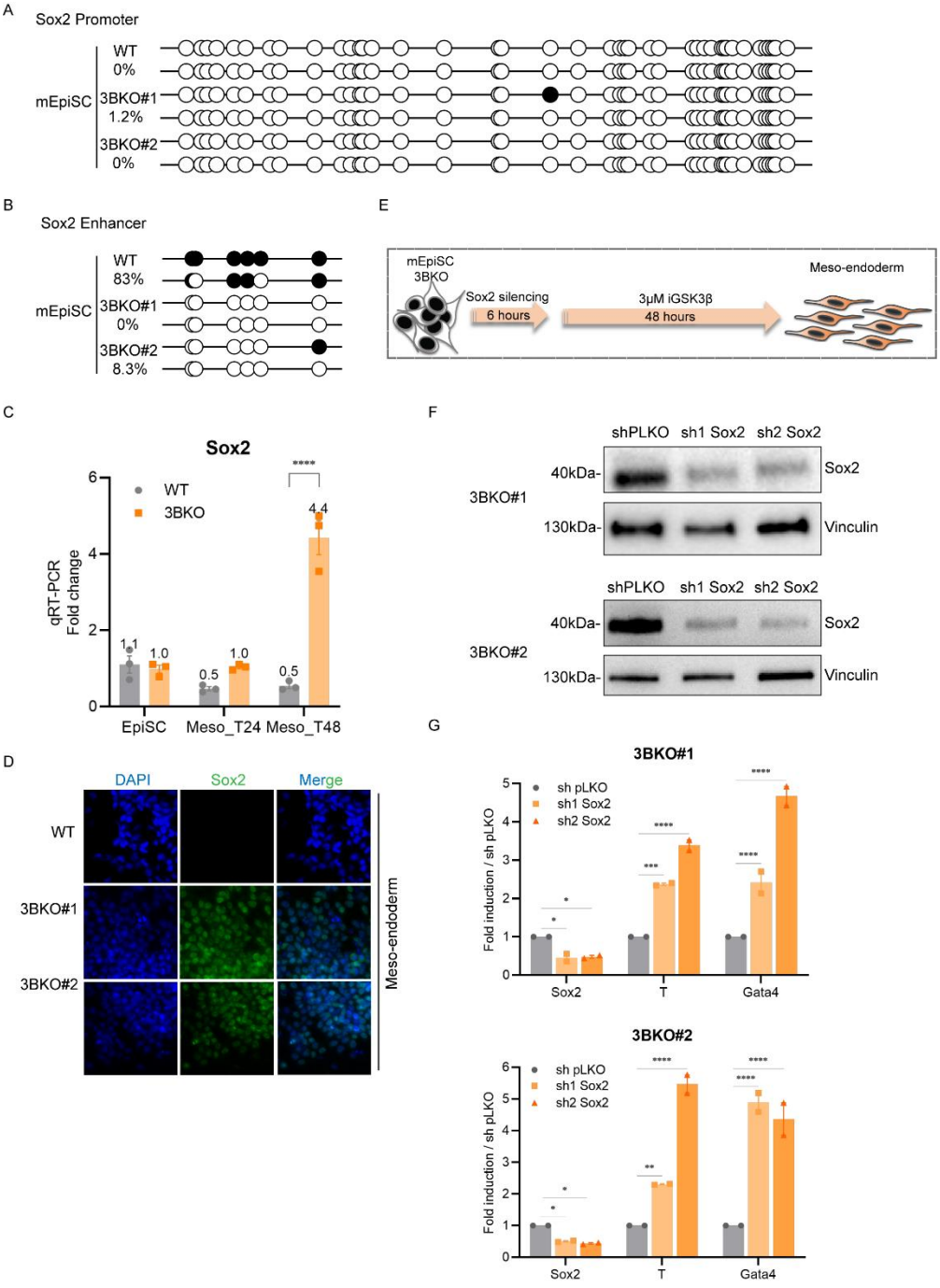


Figure 5.6. Silencing of Sox2 rescued the meso-endoderm markers expression in Dnmt3b knockout cells.

A. The lollipop-style representation of methylation analysis on Sox2 promoter in EpiSC for both WT and two Dnmt3b KO cells. Each horizontal line represents one clone (in each sample 2 clones were sequenced). The number of columns corresponds to the number of CpG dinucleotides in each fragment. Each CpG position is indicated by white circles for unmethylated CpGs and black circles for methylated CpGs.

B. The lollipop-style representation of methylation analysis on Sox2 enhancer (SRR2 element) in mEpiSC for both WT and two Dnmt3b KO cells. Each horizontal line represents one clone (in each sample 2 clones were sequenced). The number of columns corresponds to the number of CpG dinucleotides in each fragment. Each CpG position is indicated by white circles for unmethylated CpGs and black circles for methylated CpGs.

C. RT-qPCR analysis showing the expression level of Sox2 at mEpiSC stage and meso-endoderm

differentiation T48 for WT and Dnmt3b KO cells. n=3. ****P < 0.0001.

D. Immunofluorescence assay showing the expression level of Sox2 at mEpiSC stage and meso-endoderm differentiation T48 for WT and two Dnmt3b KO cells. Nuclei was stained with DAPI.
E. Schematic of experimental set-up for Sox2 silencing in mEpiSCs for Dnmt3b KO cells following *in vitro* meso-endoderm progenitor differentiation. Cells were transfected with shRNAs of Sox2 and meso-endoderm were obtained by using 3 μM iGSK3β in the medium during the 48-hour culture.
F. Western blot analysis showing the expression of Sox2 in two Dnmt3b KO cells after the silencing with 2 different shRNAs of Sox2. Vinculin was used as a loading control.
G. Bar graphs showing relative mRNA expression of Sox2 and two meso-endoderm markers (T, Gata4) in two Dnmt3b KO cells after the silencing with two different shRNAs of Sox2. The results were represented as fold induction normalized to shPLKO. Error bars indicate SEM (n = 2). P values were calculated using Two-way ANOVA test. *P < 0.05, **P < 0.01.

DISCUSSION

The present thesis has to be considered part of a major study in which we aim to deeply describe the role of *de novo* methylation in the meso-endoderm progenitor differentiation in mouse epiblast cells.

The *de novo* DNA methylases establish DNA methylation in the early stages of mouse embryo development at the exit from pluripotency and differentiation into somatic lineages. It has been reported that DNMT3A plays a role at later stages in differentiation, while DNMT3B-dependent methylation is involved in the regulation of early development and explains the early embryonic lethality observed *in vivo* in Dnmt3b knockout embryos, while Dnmt3a knockout mice only die after birth (Okano M et al., 1999; Li E et al., 1992; Nguyen S et al., 2007). In the first part of this study, we took advantage of the study from Tesar PJ et al. (2007) and developed the *in vitro* EpiSC induction protocol directly starting from mouse ES cells. Our data suggest that distinct transcriptional networks operate to maintain pluripotency in epiblast cells, which is consistent with the findings of Tesar PJ et al. (2007), Nichols J and Smith A (2009), and ten Berge D et al. (2011). To better understand the mechanism in which Dnmt3b is involved in early embryo development, we characterized the gene expression pattern of the EpiSC model by RNA-seq. The results showed that the lack of Dnmt3b does not prevent the establishment of EpiSC, as we observed similar expression of core, naïve, and primed markers in both WT and Dnmt3b KO cells.

Consequently, the induction of EpiSC towards the meso-endodermal fate demonstrated that the lack of Dnmt3b in EpiSC impairs the differentiation towards the meso-endoderm lineage. This data agrees with the idea that DNA methylation at the epiblast stage is responsible for cell priming to ensure the correct cell specification at later stages. Interestingly, we found that the expression of meso-endoderm markers (i.e., T, Sox17, Kdr) significantly increased in WT cells rather than in Dnmt3b KO cells after the meso-endoderm differentiation, whereas the neuroectoderm markers (i.e., Sox2, Sox1, Pax6) significantly upregulated in the Dnmt3b KO cells. Then, we focused our attention on Sox2, a well-known master regulator of neural development (Koch F et al., 2017) and a direct target of Dnmt3b. Indeed, we found that the silencing of Sox2 in Dnmt3b KO epiblast cells rescues the expression of the meso-endoderm markers upon their further differentiation. It has been shown that the differentially methylated CpGs in the Sox2 SRR2 element are immediately downstream of the SOX2-POU motif, and their methylation may silence the gene against all stimuli acting via SRR2 (Sikorska M et al., 2008). This can explain why we observed the hypomethylation of this element in 3B KO epiblast rather than WT epiblast. We also performed WGBS and found that the enhancers associated with neuroectodermal development genes (i.e., Sox2, Sox1, Tubb3) are

demethylated in Dnmt3b KO cells, which fail to be silenced at the later stages resulting in their differentiation towards to neuro-ectoderm (Data not shown). This data is consistent with the finding of Argelaguet R et al. (2019) and suggests that epiblast cells should repress several neuroectoderm gene enhancers that are open and demethylated at this stage to differentiate towards meso-endoderm.

Therefore, Dnmt3b-dependent DNA methylation establishes the epigenetic inheritance by suppressing the expression of critical neuro-ectoderm markers to specify the meso-endodermal lineage differentiation. This work provides the functional characterization of the *de novo* Dnmt3b during lineage determination and suggests that Dnmt3b-dependent methylation is essential to prime EpiSCs for their further differentiation into the meso-endodermal lineages.

Chapter 6

EXPERIMENTAL PROCEDURES

6.1 Cell culture

Embryonic stem cells were generated and cultured as described previously (Neri F et al., 2007). E14 mouse WT, DNMT3B $^{-/-}$ (cl.B77/cl.2) and DNMT3B $^{-/-}$ (cl.B126/cl.1) ES cells were cultivated in high-glucose DMEM (Euroclone) supplemented with 15% FBS (Millipore Corp., Billerica, MA, USA), 0.1 mmol/l nonessential amino acids (Invitrogen), 1 mmol/l sodium pyruvate (Invitrogen), 0.1 mmol/l β-mercaptoethanol, 1500 U/ml Leukemia Inhibitory Factor (LIF; Millipore), 25 U/ml penicillin, and 25 μg/ml streptomycin. All the cells were mycoplasma-free.

6.2 EpiSCs induction from ESCs

Epiblast (EpiSCs) induction was modified from ten Berge D et al. (2011). Briefly, a single-cell suspension was seeded onto Geltrex (A1413202 GIBCO)-coated plates at a density of 10,000 cells cm^{-2} in N2B27 medium supplemented with 20ng/ml ActivinA (PHC9564 GIBCO) and 12 ng/ml bFGF (PHG0026 GIBCO). The cells were passaged 1:3 as small clumps using Collagenase IV (17104019 GIBCO). EpiSCs were collected for DNA and RNA analyses after 14 days of induction followed by daily medium changes.

N2B27 medium is composed by 50% advanced DMEM/F12 (12634028 GIBCO) and 50% Neurobasal medium (21103049 GIBCO), supplemented with 0.5% N2 Supplement (17502048 GIBCO), 1% B27 Supplement (17504044 GIBCO), 0.033% BSA solution (A9647 SIGMA), 50 uM β-mercaptoethanol (M3148 Sigma), 2mM Glutamax (35050038 GIBCO), 100U/ml penicillin and 100 ug/ml streptomycin (DE17-602E LONZA).

6.3 In vitro lineage differentiation

Mesoendoderm (ME) differentiation was modified from ten Berge D et al. (2008). For ME lineage-specific differentiation, cells were plated in the growth medium for 24 hours. The day after, the medium was replaced with N2B27 medium consisted of 50% advanced DMEM/F12 (12634028 GIBCO) and 50% Neurobasal medium (21103049 GIBCO), supplemented with 0.5% N2 Supplement (17502048 GIBCO), 1% B27 supplement minus Vitamin A (12587010 GIBCO) and 3 uM iGSK3β (CHIR99021 SIGMA).

Neuroectoderm (NE) differentiation was modified from Ying QL et al. (2003). For NE lineage-

specific differentiation, cells were plated in the growth medium for 24 hours. The day after, the medium was replaced with N2B27 medium consisted of 50% advanced DMEM/F12 (12634028 GIBCO) and 50% Neurobasal medium (21103049 GIBCO), supplemented with 0.5% N2 Supplement (17502048 GIBCO), 1% B27 supplement (17504044 GIBCO) and 500 nM atRA (sc-200898 SANTA CRUZ).

Cells were fed daily until the end of differentiation.

6.4 FACS analysis

FACS analysis was performed by using Annexin V-FITC kit (Miltenyi Biotec) following the manufacturer's instruction. Approximately 10^6 cells were dissociated after 48 hours of differentiation toward ME lineage. 2 ug antibody was incubated with the cells on ice for 30 mins in the dark. Shortly cells were washed twice in the Annexin binding buffer and then incubated with Annexin V-FITC for 15 mins in the dark at room temperature. After that, cells were washed once in Annexin binding buffer, and then, PI solution was added immediately before analysis by flow cytometry.

6.5 Protein extraction and Western blotting

For total cell extracts, cells were resuspended in F-buffer (10mM TRIS-HCl pH 7.0, 50mM NaCl, 30mM Na-pyrophosphate, 50mM NaF, 1% Triton X-100, anti-proteases) and sonicated for 3 pulses. Extracts were quantified using bicinchoninic acid (BCA) assay (BCA protein assay kit; catalog no. 23225; Pierce) and were run on SDS-polyacrylamide gels at different percentages, transferred to nitrocellulose membranes, and incubated with specific primary antibodies overnight.

6.6 shRNA Constructs

Custom shRNAs against Sox2 were constructed using the TRC hairpin design tool (http://www.broadinstitute.org/rnai/public/seq/search), and designed to target the following sequences:

5′-ACCAATCCCATCCAAATTAAC-3′ (shRNA1)

5′-GCACAGTTTGAGATAAATAAA-3′ (shRNA2)

Hairpins were cloned into pLKO.1 vector (Addgene: 10878), and each construct was verified by sequencing.

6.7 Transfections

Transfections of mouse EpiSCs were performed using Lipofectamine 2000 Transfection Reagent (INVITROGEN) following the manufacturer's protocol using equal amounts of each plasmid in multiple transfections. For Sox2 knockdown, cells were transfected twice with 5 μg of the specific shRNA construct and maintained in the growth medium for 48 h.

6.8 Alkaline phosphatase (AP) staining and Immunostaining

According to the manufacturer's protocol, ES cells and EpiSCs were fixed with 4% paraformaldehyde for 2 min and then stained with Vector® Red alkaline phosphatase substrate kit (SK-5100) 's protocol.

For immunostaining, cells were fixed with 4% paraformaldehyde for 10 min at room temperature. Permeabilization was performed in 0.1% Triton X-100 in PBS for 15 min, and then the cells were blocked in 2% BSA in PBS at room temperature for 2 hours. Cells were stained with primary antibodies at 4 °C overnight. The secondary antibody was applied for 1 hour at room temperature. Nuclei were stained with DAPI (D21490 INVITROGEN). Images were acquired using a Leica TCS SP5 Confocal microscope and LAS AF Lite software.

6.9 Antibodies

The following antibodies were used for western blotting: Dnmt3a (NB120-13888, Novus Biologicals), Dnmt3b (ab122932, Abcam), Dnmt3l (provided by Dr. S. Yamanaka, Kyoto University, Japan), Pou5f1 (sc5279, Santa Cruz), Lefty1 (ab22569, Abcam), Otx2 (ab21990, Abcam), Sox2 (sc-365823, Santa Cruz), T (Brachyury, AF2085, R&D systems), Sox1 (AF3369, R&D systems), Vinculin (SAB4200080, Sigma), β-Actin (A5441, Sigma).

The following antibodies were used for immunostaining: Dnmt3b (ab122932, Abcam), Lefty1 (ab22569, Abcam), Fgf5 (ab88118, Abcam), T (Brachyury, AF2085, R&D systems), Sox1 (07-1673, Millipore), Sox2 (sc-365823, Santa Cruz).

The following antibody was used for FACS analysis: CD31 (553370, BD Pharmingen).

6.10 DNA extraction

Genomic DNA was extracted from cells using the DNeasy Blood and Tissue kit (QIAGEN, 69506) following the manufacturer's instructions.

6.11 DNA methylation analysis

For DNA methylation analysis, 1 µg of genomic DNA was used for bisulfite conversion by using the EpiTect Conversion Kit (QIAGEN, 59104) according to the manufacturer's protocol. Converted DNA was eluted in 20 µl, and 3 µl of converted DNA was used in a 50 µL PCR reaction (INVITROGEN, 12346). Primer sequences are in Supplementary Table 2. PCR products were purified and cloned into TOPO-TA vector (INVITROGEN 450030), and positive clones were verified by sequencing. The bisulfite sequencing analysis of CpG methylation was performed as described in Kumaki Y et al. (2008).

6.12 RNA extraction and RT-PCR analysis

Total RNA was extracted by using TRIzol reagent (Invitrogen). Real-time PCR was performed using the SuperScript III Platinum One-Step Quantitative RT-PCR System (Invitrogen, cat.11732-020) following the manufacturer's instructions. Oligonucleotide sequences are reported in Supplementary Table 2.

6.13 RNA-seq library preparation

mRNA-seq libraries were generated from 1.5 ug of total RNA using TruSeq RNA Sample Preparation v2 according to the manufacturer's protocol.

6.14 RNA-seq analysis

Following quality controls (performed with FastQC v0.11.2), sequencing reads were aligned to the mouse reference genome (mm10/GRCm38 Ensembl release 84). Gene expression levels were quantified with featureCounts v1.6.1. Multi-mapped reads were excluded from quantification.

Gene expression counts were next analyzed using the edgeR package. Lowly expressed/detected genes (i.e., 1 RPKM in less than two samples) were filtered out, obtaining a total of 16,755 expressed genes for downstream analysis. Normalization factors were calculated using the trimmed mean of M-values (TMM) method, and RPKM was computed using normalized library sizes and gene lengths from the Ensembl release 84 annotation (rpkm function). Following dispersion estimation, an ANOVA-like test was implemented by fitting a Generalized Linear Model (GLM) to all sample groups and performing Quasi-Likelihood F-test in order to identify the genes that were significantly varying during the differentiation time course (i.e., differentially expressed genes in any of the sample groups during the time course, using the ESC-WT condition as baseline in the design matrix formula). The resulting 4,624

genes (|logFC|≥1 and FDR≤0.001) were used for clustering of gene expression profiles with K-means followed by hierarchical clustering. RPKM values were scaled as Z-scores across samples before computing distances. The optimal number of K-means clusters (n=4) was estimated using the within-cluster sum of squares methodology. Gene expression heatmaps were generated using the ComplexHeatmap R package. Gene set over-representation analysis was performed for each cluster with the gProfileR package, using all the expressed genes as background.

Differentially expressed genes between WT and 3BKO cells at each time point were obtained from the same GLM, comparing each contrast with the Quasi-Likelihood F-test (|logFC|≥1 and FDR≤0.05).

SUPPLEMENTAL INFORMATION

Supplementary Table 2: Oligonucleotide sequences used as primers with indication of the corresponding gene.

Gene	Sequence	Strand	Application
Actb	TCTTTGCAGCTCCTTCGTTG	Fw	RT-qPCR
Actb	ACGATGGAGGGGAATACAGC	Rev	RT-qPCR
Nanog	AAGTACCTCAGCCTCCAGCA	Fw	RT-qPCR
Nanog	GTGCTGAGCCCTTCTGAATC	Rev	RT-qPCR
Pou5f1	CTGAGGGCCAGGCAGGAGCACGAG	Fw	RT-qPCR
Pou5f1	CTGTAGGGAGGGCTTCGGGCACTT	Rev	RT-qPCR
Sox2	ATGATGGAGACGGAGCTGAA	Fw	RT-qPCR
Sox2	TTGCTGATCTCCGAGTTGTG	Rev	RT-qPCR
Fgf5	CTCCCACGAAGCCAGTGTGT	Fw	RT-qPCR
Fgf5	CAGGGCCACGTACCACTCTC	Rev	RT-qPCR
Lefty1	CTGCAGCTCGATCAACCGCC	Fw	RT-qPCR
Lefty1	GCACCAGCTCGCTGTTAGG	Rev	RT-qPCR
Sox1	GCGATGCCAACTTTTGTATG	Fw	RT-qPCR
Sox1	AGAGGGGATTGCGGTATAAA	Rev	RT-qPCR
Nestin	TGGGCAGCAACTGGCACACC	Fw	RT-qPCR
Nestin	TGGGCTGAGGACAGGGAGCA	Rev	RT-qPCR
T	CTGTGACTGCCTACCAGAATGAGGAG	Fw	RT-qPCR
T	GGTCGTTTCTTTCTTTGGCATCAAG	Rev	RT-qPCR
Gata4	CTCTATCACAAGATGAACGGCATCAAC	Fw	RT-qPCR
Gata4	TCTGGCAGTTGGCACAGGAGAG	Rev	RT-qPCR
Zfp42	GGGGATGATAGGAGGTTTATTTTAT	Fw	Methylation analysis_promoter
Zfp42	AAACAACACAACTCACTTTAAAAAC	Rev	Methylation analysis_promoter
Dppa3	TTTTTTTATTTTGTGATTAGGGTTG	Fw	Methylation analysis_promoter
Dppa3	CTTCACCTAAACTACACCTTTAAAC	Rev	Methylation analysis_promoter
Sox2	GGAAAAGGTTGGGAATAAGGTT	Fw	Methylation analysis_promoter
Sox2	CCCAACCCTAATCTTAAAAAAACA	Rev	Methylation analysis_promoter
Sox2	AAAAGGTTTAAATTGTAAGATTAGGT	Fw	Methylation analysis_enhancer
Sox2	TATTTTCTAAAAACCACAAAAAAAA	Rev	Methylation analysis_enhancer

Sox2	CCGGCACCAATCCCATCCAAATTAACTCGAGTTAATTTGGATGGGATTGGTGTTTTTG	Fw	shRNA1 Sox2
Sox2	AATTCAAAAACACCAATCCCATCCAAATTAACTCGAGTTAATTTGGATGGGATTGGTG	Rev	shRNA1 Sox2
Sox2	CCGGGCACAGTTTGAGATAAATAAACTCGAGTTTATTTATCTCAAACTGTGCTTTTTG	Fw	shRNA2 Sox2
Sox2	AATTCAAAAAGCACAGTTTGAGATAAATAAACTCGAGTTTATTTATCTCAAACTGTGC	Rev	shRNA2 Sox2

REFERENCES (II)

Argelaguet R, Clark SJ, Mohammed H, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature. 2019;576(7787):487-491. doi:10.1038/s41586-019-1825-8

Athanasiadou R, de Sousa D, Myant K, Merusi C, Stancheva I, Bird A. Targeting of de novo DNA methylation throughout the Oct-4 gene regulatory region in differentiating embryonic stem cells. PLoS One. 2010;5(4):e9937. Published 2010 Apr 1. doi:10.1371/journal.pone.0009937

Aulehla A, Wiegraebe W, Baubet V, et al. A beta-catenin gradient links the clock and wavefront systems in mouse embryo segmentation. Nat Cell Biol. 2008;10(2):186-193. doi:10.1038/ncb1679

Azuara V, Perry P, Sauer S, et al. Chromatin signatures of pluripotent cell lines. Nat Cell Biol. 2006;8(5):532-538. doi:10.1038/ncb1403

Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007;128(4):669-681. doi:10.1016/j.cell.2007.01.033

Bird A. Perceptions of epigenetics. Nature. 2007;447(7143):396-398. doi:10.1038/nature05913

Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell. 1985;40(1):91-99. doi:10.1016/0092-8674(85)90312-5

Bonasio R, Tu S, Reinberg D. Molecular signals of epigenetic states. Science. 2010;330(6004):612-616. doi:10.1126/science.1191078

Boyer LA, Lee TI, Cole MF, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005;122(6):947-956. doi:10.1016/j.cell.2005.08.020

Boyer LA, Plath K, Zeitlinger J, et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature. 2006;441(7091):349-353. doi:10.1038/nature04733

Brenet F, Moh M, Funk P, et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. PLoS One. 2011;6(1):e14524. Published 2011 Jan 18. doi:10.1371/journal.pone.0014524

Brockdorff N, Turner BM. Dosage compensation in mammals. Cold Spring Harb Perspect Biol. 2015;7(3):a019406. Published 2015 Mar 2. doi:10.1101/cshperspect.a019406

Brons IG, Smithers LE, Trotter MW, et al. Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature. 2007;448(7150):191-195. doi:10.1038/nature05950

Catena R, Tiveron C, Ronchi A, et al. Conserved POU binding DNA sites in the Sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells. J Biol Chem. 2004;279(40):41846-41857. doi:10.1074/jbc.M405514200

Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and

paradigms. Nat Rev Genet. 2009;10(5):295-304. doi:10.1038/nrg2540

Chapman DL, Papaioannou VE. Three neural tubes in mouse embryos with mutations in the T-box gene Tbx6. Nature. 1998;391(6668):695-697. doi:10.1038/35624

Chen T, Ueda Y, Dodge JE, Wang Z, Li E. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. Mol Cell Biol. 2003;23(16):5594-5605. doi:10.1128/mcb.23.16.5594-5605.2003

Ciruna BG, Schwartz L, Harpal K, Yamaguchi TP, Rossant J. Chimeric analysis of fibroblast growth factor receptor-1 (Fgfr1) function: a role for FGFR1 in morphogenetic movement through the primitive streak. Development. 1997;124(14):2829-2841.

Farthing CR, Ficz G, Ng RK, et al. Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. PLoS Genet. 2008;4(6):e1000116. Published 2008 Jun 27. doi:10.1371/journal.pgen.1000116

Feldman N, Gerson A, Fang J, et al. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. Nat Cell Biol. 2006;8(2):188-194. doi:10.1038/ncb1353

Festuccia N, Osorno R, Halbritter F, et al. Esrrb is a direct Nanog target gene that can substitute for Nanog function in pluripotent cells. Cell Stem Cell. 2012;11(4):477-490. doi:10.1016/j.stem.2012.08.002

Fouse SD, Shen Y, Pellegrini M, et al. Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. Cell Stem Cell. 2008;2(2):160-169. doi:10.1016/j.stem.2007.12.011

Frommer M, McDonald LE, Millar DS, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992;89(5):1827-1831. doi:10.1073/pnas.89.5.1827

Garriock RJ, Chalamalasetty RB, Kennedy MW, Canizales LC, Lewandoski M, Yamaguchi TP. Lineage tracing of neuromesodermal progenitors reveals novel Wnt-dependent roles in trunk progenitor cell maintenance and differentiation. Development. 2015;142(9):1628-1638. doi:10.1242/dev.111922

Hata K, Okano M, Lei H, Li E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development. 2002;129(8):1983-1993.

Hathaway NA, Bell O, Hodges C, Miller EL, Neel DS, Crabtree GR. Dynamics and memory of heterochromatin in living cells. Cell. 2012;149(7):1447-1460. doi:10.1016/j.cell.2012.03.052

Hawkins RD, Hon GC, Lee LK, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell. 2010;6(5):479-491. doi:10.1016/j.stem.2010.03.018

Herman JG, Graff JR, Myöhänen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A.

1996;93(18):9821-9826. doi:10.1073/pnas.93.18.9821

Irizarry RA, Ladd-Acosta C, Wen B, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178-186. doi:10.1038/ng.298

Jackson M, Krassowska A, Gilbert N, et al. Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. Mol Cell Biol. 2004;24(20):8862-8871. doi:10.1128/MCB.24.20.8862-8871.2004

Jurberg AD, Aires R, Nóvoa A, Rowland JE, Mallo M. Compartment-dependent activities of Wnt3a/β-catenin signaling during vertebrate axial extension. Dev Biol. 2014;394(2):253-263. doi:10.1016/j.ydbio.2014.08.012

Koch F, Scholze M, Wittler L, et al. Antagonistic Activities of Sox2 and Brachyury Control the Fate Choice of Neuro-Mesodermal Progenitors. Dev Cell. 2017;42(5):514-526.e7. doi:10.1016/j.devcel.2017.07.021

Kumaki Y, Ukai-Tadenuma M, Uno KD, et al. Analysis and synthesis of high-amplitude Cis-elements in the mammalian circadian clock. Proc Natl Acad Sci U S A. 2008;105(39):14946-14951. doi:10.1073/pnas.0802636105

Lengler J, Bittner T, Münster D, Gawad Ael-D, Graw J. Agonistic and antagonistic action of AP2, Msx2, Pax6, Prox1 AND Six3 in the regulation of Sox2 expression. Ophthalmic Res. 2005;37(6):301-309. doi:10.1159/000087774

Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992;69(6):915-926. doi:10.1016/0092-8674(92)90611-f

Li E, Zhang Y. DNA methylation in mammals. Cold Spring Harb Perspect Biol. 2014;6(5):a019133. Published 2014 May 1. doi:10.1101/cshperspect.a019133

Loh YH, Wu Q, Chew JL, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet. 2006;38(4):431-440. doi:10.1038/ng1760

Martin BL, Kimelman D. Canonical Wnt signaling dynamically controls multiple stem cell fate decisions during vertebrate body formation. Dev Cell. 2012;22(1):223-232. doi:10.1016/j.devcel.2011.11.001

Martin GR, Evans MJ. Differentiation of clonal lines of teratocarcinoma cells: formation of embryoid bodies in vitro. Proc Natl Acad Sci U S A. 1975;72(4):1441-1445. doi:10.1073/pnas.72.4.1441

McKeon C, Ohkubo H, Pastan I, de Crombrugghe B. Unusual methylation pattern of the alpha 2 (l) collagen gene. Cell. 1982;29(1):203-210. doi:10.1016/0092-8674(82)90104-0

Meissner A, Mikkelsen TS, Gu H, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008;454(7205):766-770. doi:10.1038/nature07107

Miyagi S, Saito T, Mizutani K, et al. The Sox-2 regulatory regions display their activities in two distinct types of multipotent stem cells. Mol Cell Biol. 2004;24(10):4207-4220.

doi:10.1128/MCB.24.10.4207-4220.2004

Miyagi S, Nishimoto M, Saito T, et al. The Sox2 regulatory region 2 functions as a neural stem cell-specific enhancer in the telencephalon. J Biol Chem. 2006;281(19):13374-13381. doi:10.1074/jbc.M512669200

Mohn F, Weber M, Rebhan M, et al. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell. 2008;30(6):755-766. doi:10.1016/j.molcel.2008.05.007

Neri F, Rapelli S, Krepelova A, et al. Intragenic DNA methylation prevents spurious transcription initiation. Nature. 2017;543(7643):72-77. doi:10.1038/nature21373

Nguyen S, Meletis K, Fu D, Jhaveri S, Jaenisch R. Ablation of de novo DNA methyltransferase Dnmt3a in the nervous system leads to neuromuscular defects and shortened lifespan. Dev Dyn. 2007;236(6):1663-1676. doi:10.1002/dvdy.21176

Nichols J, Smith A. Naive and primed pluripotent states. Cell Stem Cell. 2009;4(6):487-492. doi:10.1016/j.stem.2009.05.015

Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247-257. doi:10.1016/s0092-8674(00)81656-6

Okano M, Xie S, Li E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat Genet. 1998;19(3):219-220. doi:10.1038/890

Pastor WA, Pape UJ, Huang Y, et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature. 2011;473(7347):394-397. doi:10.1038/nature10102

Pósfai J, Bhagwat AS, Pósfai G, Roberts RJ. Predictive motifs derived from cytosine methyltransferases. Nucleic Acids Res. 1989;17(7):2421-2435. doi:10.1093/nar/17.7.2421

Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature. 2007;447(7143):425-432. doi:10.1038/nature05918

Saitou M, Kagiwada S, Kurimoto K. Epigenetic reprogramming in mouse pre-implantation development and primordial germ cells. Development. 2012;139(1):15-31. doi:10.1242/dev.050849

Sikorska M, Sandhu JK, Deb-Rinker P, et al. Epigenetic modifications of SOX2 enhancers, SRR1 and SRR2, correlate with in vitro neural differentiation. J Neurosci Res. 2008;86(8):1680-1693. doi:10.1002/jnr.21635

Smith AG, Heath JK, Donaldson DD, et al. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature. 1988;336(6200):688-690. doi:10.1038/336688a0

Sørensen AL, Jacobsen BM, Reiner AH, Andersen IS, Collas P. Promoter DNA methylation patterns of differentiated cells are largely programmed at the progenitor stage. Mol Biol Cell. 2010;21(12):2066-2077. doi:10.1091/mbc.e10-01-0018

Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet. 2008;9(6):465-476. doi:10.1038/nrg2341

Takemoto T, Uchikawa M, Yoshida M, et al. Tbx6-dependent Sox2 regulation determines neural or mesodermal fate in axial stem cells. Nature. 2011;470(7334):394-398. doi:10.1038/nature09729

ten Berge D, Koole W, Fuerer C, Fish M, Eroglu E, Nusse R. Wnt signaling mediates self-organization and axis formation in embryoid bodies. Cell Stem Cell. 2008;3(5):508-518. doi:10.1016/j.stem.2008.09.013

ten Berge D, Kurek D, Blauwkamp T, et al. Embryonic stem cells require Wnt proteins to prevent differentiation to epiblast stem cells. Nat Cell Biol. 2011;13(9):1070-1075. Published 2011 Aug 14. doi:10.1038/ncb2314

Tesar PJ, Chenoweth JG, Brook FA, et al. New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature. 2007;448(7150):196-199. doi:10.1038/nature05972

Tomioka M, Nishimoto M, Miyagi S, et al. Identification of Sox-2 regulatory region which is under the control of Oct-3/4-Sox-2 complex. Nucleic Acids Res. 2002;30(14):3202-3213. doi:10.1093/nar/gkf435

Tost J, Gut IG. DNA methylation analysis by pyrosequencing. Nat Protoc. 2007;2(9):2265-2275. doi:10.1038/nprot.2007.314

Tsumura A, Hayakawa T, Kumaki Y, et al. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes Cells. 2006;11(7):805-814. doi:10.1111/j.1365-2443.2006.00984.x

Tucker KL, Beard C, Dausmann J, et al. Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. Genes Dev. 1996;10(8):1008-1020. doi:10.1101/gad.10.8.1008

Uchikawa M, Ishida Y, Takemoto T, Kamachi Y, Kondoh H. Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. Dev Cell. 2003;4(4):509-519. doi:10.1016/s1534-5807(03)00088-1

Weber M, Davies JJ, Wittig D, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet. 2005;37(8):853-862. doi:10.1038/ng1598

Williams RL, Hilton DJ, Pease S, et al. Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. Nature. 1988;336(6200):684-687. doi:10.1038/336684a0

Wilson V, Olivera-Martinez I, Storey KG. Stem cells, signals and vertebrate body axis extension [published correction appears in Development. 2009 Jun;136(12):2133]. Development. 2009;136(10):1591-1604. doi:10.1242/dev.021246

Wolf SF, Jolly DJ, Lunnen KD, Friedmann T, Migeon BR. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-

chromosome inactivation. Proc Natl Acad Sci U S A. 1984;81(9):2806-2810. doi:10.1073/pnas.81.9.2806

Yagi M, Yamanaka S, Yamada Y. Epigenetic foundations of pluripotent stem cells that recapitulate in vivo pluripotency. Lab Invest. 2017;97(10):1133-1141. doi:10.1038/labinvest.2017.87

Yamaguchi TP, Takada S, Yoshikawa Y, Wu N, McMahon AP. T (Brachyury) is a direct target of Wnt3a during paraxial mesoderm specification. Genes Dev. 1999;13(24):3185-3190. doi:10.1101/gad.13.24.3185

Ying QL, Stavridis M, Griffiths D, Li M, Smith A. Conversion of embryonic stem cells into neuroectodermal precursors in adherent monoculture. Nat Biotechnol. 2003;21(2):183-186. doi:10.1038/nbt780

Yoshikawa Y, Fujimori T, McMahon AP, Takada S. Evidence that absence of Wnt-3a signaling promotes neuralization instead of paraxial mesoderm development in the mouse. Dev Biol. 1997;183(2):234-242. doi:10.1006/dbio.1997.8502

You JS, Kelly TK, De Carvalho DD, Taberlay PC, Liang G, Jones PA. OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. Proc Natl Acad Sci U S A. 2011;108(35):14497-14502. doi:10.1073/pnas.1111309108

ACKNOWLEDGEMENT

Throughout the writing of this thesis, I realize that time flies, and I have learned a lot during my Ph.D. journey. I have also received a great deal of support and assistance.

I would first like to thank my tutor, Professor Salvatore Oliviero, who provides me the opportunity to pursue my Ph.D. and work at this lab. Your invaluable expertise formulates my research hypothesis and methodology, and your insightful feedbacks pushed me to sharpen my thinking and brought my work to a higher level. Meanwhile, I would also like to thank China Scholarship Council for providing me the fellowship.

I want to acknowledge my group colleagues for their excellent collaboration. I would like to thank Stefania Rapelli and Mara Maldotti for your patient support when I just arrived in Torino, and all of the solutions to the experiment issues to further my research. I want to thank Andrea Lauria for your helpful data analysis to finish the project. Besides, I want to thank other colleagues, Isabelle, Francesca, Annalaura, Valentina, Mirko, Ivan, Fatemeh, Daniela, Chiara, Hassan, Lisa, Caterina, Danny, Edoardo, Giulia, for creating a pleasant lab atmosphere.

In addition, I would like to thank my family for their wise counsel and sympathetic ear. You are always there for me. I also want to thank my Chinese friends in Torino; our adventures in the mountains are my happiest time and unforgettable experience. Finally, I could not have completed this thesis without the support of Lei, who provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

Infine, vorrei ringraziare tutti i miei coinquilini italiani durante i miei 4 anni di vita a Torino. Ragazzi, mi avete aiutato a imparare l'italiano, la cultura italiana e la cucina italiana. Mi sono divertita molto per questo periodo felice.

Thanks also to those who are reading my thesis, and I hope you enjoy our work.