CORE PLURIPOTENCY FACTORS AND SMAD SIGNALING ORCHESTRATE HUMAN EMBRYONIC STEM CELL DIFFERENTIATION

by

TIANMING WU

(Under the Direction of STEPHEN DALTON)

ABSTRACT

OCT4, NANOG, and SOX2 (abbreviated OSN) are the core transcription factors (TFs) regulating pluripotency human embryonic stem cells (hESCs). Although their basic importance in maintaining self-renewal and reprogramming has been demonstrated, their mechanistic functions in differentiation are not well studied. To address this question, we report the integrative genome-wide chromatin accessibility and TF binding data with extensive transcriptomic and epigenomic data across the differentiation of hESC to the three germ layers. Using ATAC-seq (the assay for transposase-accessible chromatin using sequencing) technique, we identify 12,3483 transposase-accessible DNA elements. Integration of ATAC-seq with multi-omics data identifies a substantial number of putative distal enhancers that distinguish cell fate commitments. These data reveal regulatory TF binding motifs, validated by the DNA occupation by respective TF from in-house and published ChIP-seq dataset. For example, SMAD1, HAND1 for mesoderm (Meso), SMAD2/3, GATA6 for definitive endoderm (DE), SOX2 for neural ectoderm (Ecto). To our surprise, Active enhancers are densely occupied by OSN not only in hESCs, but also in early differentiated cells, in a motif independent manner.

The co-binding of OSN with extracellular signaling effectors, SMAD1 and SMAD2/3, are in a cell fate specific manner. In hESCs, OSN individually interact with each other and with SMAD1, SMAD2/3. Upon differentiation, interactions within OSN decrease, whereas Activin-A activated SMAD2/3 specifically interacts with NANOG in DE, but not in Meso or Ecto Similarly, BMP4 activated SMAD1 specifically interacts with OCT4 in Meso, but not in DE or Ecto. During Ecto differentiation, dual SMAD signaling inhibition enables SOX2 to release from the OSN-SMAD complex and to re-distribute onto Ecto-specific enhancers. In addition, OSN mediate higher-order chromatin remodeling in respective cell fate. Our studies provide a new model of how OSN function as cell fate specifiers in collaboration with SMAD signaling.

INDEX WORDS: human embryonic stem cells (hESCs), pluripotency factors, SMAD, chromatins, promoter-enhancer interactions, next generation sequencing, stem cell differentiation

CORE PLURIPOTENCY FACTORS AND SMAD SIGNALING ORCHESTRATE HUMAN EMBRYONIC STEM CELL DIFFERENTIATION

by

TIANMING WU

BS, SICHUAN UNIVERSITY, CHINA, 2011

MS, SICHUAN UNIVERSITY, CHINA, 2014

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

CORE PLURIPOTENCY FACTORS AND SMAD SIGNALING ORCHESTRATE HUMAN EMBRYONIC STEM CELL DIFFERENTIATION

by

TIANMING WU

Major Professor:	Stephen Dalton
Committee:	Hang Yin
	Shaying Zhao
	Takahiro Ito

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2021

DEDICATION

This work is dedicated to my grandparents, Qingcheng Jin and Fulan Wang. Thank you for raising me up.

ACKNOWLEDGEMENTS

First and foremost, I must thank my family. Without their support, I wouldn't have gone this far. Next, I want to thank my wife, Wenting Zhao, for her unconditioned love without reservation.

As for my academic path, I am very grateful to Dr. Stephen Dalton as my advisor. His passion in research has inspired me to dive deeper and deeper into my project. He always asked questions central to basic science, guiding me to think critically. Besides the rigorous scientific training, he also encouraged me to expand my interests into data science and bioinformatics, which allowed me to address biological and real-world questions at a greater scale.

My fellow Dalton lab members have been helpful to me. They offered me the chance to work interactively and efficiently. Dr. Amar Singh and Dr. Timothy Cliff are very knowledge persons in our lab. They always have an answer to my quest. Dr. Liang Zhang contributed to the preliminary analysis in my research.

I also want to thank Julie Nelson for the use of sorting machine, her expertise in flow cytometry and consistent service are indispensable for completing my work.

Finally, thank you my committee members, Dr. Hang Yin, Dr. Shaying Zhao, Dr. Takahiro Ito for your thoughtful guidance.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

PLURIPOTENT STEM CELLS

Pluripotent stem cells (PSCs) are characterized by self-renewal and the unlimited capacity to differentiate towards every cell type in the body (Niwa, 2007). During self-renewal, PSCs can be passaged continuously for years without losing pluripotency. While under conditioned signaling induction, PSCs can be programmed into lineage-specific cell types belonging to the three germ layers, namely ectoderm, mesoderm and endoderm (Cliff et al., 2017) (Figure 1.1). Given these characteristics, PSCs not only hold great values as models to study basic signaling mechanisms *in vitro* (Silva and Smith, 2008), but also open the gate for regenerative medicine applications *in vivo* (Tabar and Studer, 2014). How the PSCs exit pluripotency and commit into a termed cell fate are fundamental questions. Therefore, any insight into the regulatory network underlying the fundamental question would benefit the biomedical usage of PSCs in the future. PSCs exist in primarily two states, naïve or primed. Due to the physiological resemblance and convenient *in vitro* differentiation, we used a human embryonic stem cell (hESC) line representing a primed embryonic state in this research.

CORE PLURIPOTENCY FACTORS

The conventional criteria to define pluripotency transcription factors are based on their genetic functions in relation to distinct features of PSCs, restricted expression profile within the inner cell mass (ICM), proliferation capability, sustained global hypomethylation, inhibition of Erk

pathways, and maintenance of Lif/Stat3 signals. Utilizing the stringent genetically modified mouse models, an extended list of candidates were tested, including c-myc (Davis et al., 1993), Oct4 (Nichols et al., 1998), Sox2 (Avilion et al., 2003), Nanog (Niwa et al., 1998), Dnmt3a (Kaneda et al., 2004), etc. Disruption of those candidate genes result in early embryonic death or severe differentiation deficiencies. However, it is difficult to elucidate whether the transcription factor is the causation or the consequence of pluripotency.

In 2006, Takahashi and Yamanaka first revert mouse fibroblast cells back to PSCs. The reprogrammed PSCs were named induced pluripotent stem cells (iPSCs) (Takahashi and Yamanaka, 2006). The core pluripotency factor cohort in Yamanaka's work is Oct4, Sox2, c-myc, and Klf4. However, endogenous Oct4 and Nanog genes are expressed at a lower level from those mouse iPSCs than in mouse embryonic stem cells (mESCs), and their promoter regions are not fully erased from DNA methylation. This problem is quickly solved by replacing the Fbx15 reporter (Takahashi and Yamanaka, 2006; Tokuzawa et al., 2003) with Oct4 and/or Nanog reporter(s) in mouse model (Meissner et al., 2007; Wernig et al., 2007). These results suggest the importance of elevated expression level of Oct4 and the requirement of Nanog during pluripotency initiation.

Shortly after the generation of mouse iPSC, the first human iPSC was generated using a different four factor cocktail, Oct4, Sox2, Nanog, and Lin28, albeit at a lower efficiency (Yu et al., 2007). This result challenges the requirement of the two oncogenes, Klf4 and c-myc, in initiating pluripotency. c-Myc is shown to be independent of transcriptional regulation, it instead associates with pre-replication complexes and promotes DNA synthesis (Dominguez-Sola et al., 2007). ChIP-seq data also demonstrated that c-Myc binding sites are highly correlated with E2f1 (a transcription factor participates in DNA synthesis), instead of Oct4, Sox2 or Nanog in mESCs

(Chen et al., 2008). These data could explain why c-Myc is important for proliferation during self-renewal, rather than be necessary for reprogramming. As for Klf4, its co-binding to Oct4 and Sox2 gradually shifts to Klf4 specific binding pattern at the onset of reprogramming (Chen et al., 2008; Chronis et al., 2017). In addition, Klf4 is involved in unique higher-order chromatin reorganization during reprogramming (Di Giammartino et al., 2019). These data imply that Klf4 only functions to facilitate Oct4 and/or Sox2 binding to their targets at the early state of reprogramming by providing certain epigenetic environment, but it is not necessary for pluripotency establishment. Another explanation could be that mESCs and hESCs utilize divergent pool of key genes regulating pluripotency (Wei et al., 2005).

Lin28 is an RNA binding factor, thus it is excluded from this research.

After considerable efforts have been made to elucidate the core transcriptional organizers of pluripotency in the past decades, per the prevailing model, OCT4, SOX2, and NANOG (abbreviated as OSN) are the most important pluripotency factors in hESCs (Jaenisch and Young, 2008; Silva and Smith, 2008).

OCT4

OCT4, encoded by POU5F1 gene, is a homeodomain transcription factor belonging to POU (Pit, Oct, Unc) family. It regulates its target genes by binding to the octamer motif of ATGCAAAT at promoter or enhancer regions (Saijoh et al., 1996; Schöler et al., 1989a). OCT4 protein exists as two isoforms, OCT4A and OCT4B, which differ by the truncated N-terminal part of OCT4A. OCT4A has the full length of 134aa of N-terminus and is responsible for stemness in ESCs, because OCT4A is present in nucleus as a transcription factor, in contrast, OCT4B is mainly

cytoplasmic (Cauffman et al., 2004). Hereafter, the OCT4 studied in this research refers to OCT4A.

OCT4 protein consists of 3 domains, N-terminal domain, POU domain, and C-terminal domain. The POU domain is flanked by a POU specific domain (POUs) and POU homeodomain (POU_{HD}) by a α -helix linker. The two POU sub-domains both have DNA binding ability through the linker region, forming a helix-turn-helix structure (Reményi et al., 2003). The POU domain is highly conserved during evolution, while the N- and C-terminal exhibit a high sequence diversity in the OCT4 family members (Radzisheuskaya and Silva, 2014), suggesting the unique role of OCT4A in embryo development. There are various post-translational modification sites alongside the POU domain and C-terminal domain, potentially regulating OCT4's interaction with DNA, partner proteins, and its degradation (Abulaiti et al., 2017; Jin et al., 2016).

Oct4 protein can be detected in mouse oocyte (Schöler et al., 1989b), but Oct4 transcript is not present before the first cleavage. The first wave of Oct4 transcription is activated prior to the 8-cell stage of zygotic, hence the zygotic activation is mediated by maternal Oct4 protein (Gao et al., 2018; Ram and Schultz, 1993). Oct4 knockout embryos failed to develop ICM and resulted in a complete embryo failure (Nichols et al., 1998). Therefore, Oct4 is thought to be the first and most upstream gene in the molecular circuitry of pluripotency.

During mouse embryo development, Oct4 expression is restricted within the ICM in the blastocyst. After implantation, Oct4 protein level remains through gastrulation until ~E7.5 (DeVeale et al., 2013; Mulas et al., 2018; Radzisheuskaya et al., 2013), and it is co-localizing with chromatin remodeler Sall4, early mesoderm mark T-Brachyury, and cardiac lineage gene Mesp1 in the region of mesendoderm and lateral plate mesoderm at E7.5 (Abboud et al., 2015; Downs, 2008; Mulas et al., 2018; Radzisheuskaya et al., 2013). Knockout of Oct4 further

abrogated mesoderm lineage differentiation into paraxial mesoderm (Mulas et al., 2018). Interestingly, ectopic expression of Oct4 by less than two-fold led to differentiation into primitive endoderm and mesoderm phenotype in mESCs (Niwa et al., 2000; Radzisheuskaya et al., 2013; Thomson et al., 2011), despite that no obvious effects were observed in hESCs (Wang et al., 2012). Conversely, shRNA knockdown of OCT4 impaired expression of BMP4 mediated mesoderm target genes (Wang et al., 2012) and caused a leakage of endoderm gene expression (Teo et al., 2011) in hESCs. These *in vivo* and *in vitro* data imply that OCT4 plays a diverging role in pluripotency and differentiation potentiation.

NANOG

NANOG is named after 'land of ever young (Tir Na Nog)', due to its capability to maintain self-renewal and Oct4 expression in absence of Lif/Stat3 signals in mESCs (Chambers et al., 2003; Mitsui et al., 2003). NANOG is a member of homeobox family and binds to DNA in a sequence-specific manner. It consist of 3 domains, N- and C-terminal regions are transactivation domains (Pan and Pei, 2003), the central homeobox domain is responsible for DNA binding (Chambers et al., 2003; Mitsui et al., 2003). Of note, the C-terminal regions encompasses a tryptophan repeat (WR) domain which mediates NANOG dimerization and is required to confer Lif-independent self-renewal in mESCs (Pan and Pei, 2003).

Nanog is not present in the unfertilized egg. Rather, during embryogenesis, Nanog is present since the late morula. After the expansion of ICM, Nanog specifically demarcates the epiblast between E3.5 and E4.5. Nanog-null ICM is unable to progress into a viable epiblast (Silva et al., 2009). Notably, studies performing knockout of Oct4 throughout E7.0, Nanog positive cells expand out of primitive streak and compensate loss of Oct4 positive cells in

mesoderm domain. This progress is accompanied by significantly reduced mesoderm marker T-Brachyury and dramatically increased endoderm marker Foxa2 (Mulas et al., 2018; Suzuki et al., 2006b). These results indicate that besides the synergistic function with Oct4 to govern pluripotency (Festuccia et al., 2013; Loh et al., 2006; Rafiee et al., 2016; Yu et al., 2007), Nanog seems to play a distinct role in endoderm cell fate commitment while suppressing other lineages *in vivo*. In agreement with this speculation, shRNA knockdown (KD) experiments in hESCs showed that both OCT4 KD and SOX2 KD cause spontaneous definitive endoderm marker gene expression (Teo et al., 2011), whereas NANOG KD only induces neural ectoderm markers (Wang et al., 2012). On the other hand, overexpression of NANOG results in elevated level of endoderm marker genes, such as SOX17, FOXA2, and EOMES, under self-renewal condition (Mendjan et al., 2014; Teo et al., 2011; Wang et al., 2012). Given these results, endoderm fate commitment seems permissive under the pluripotency gatekeeper, NANOG.

SOX2

SOX (SRY homology box) family of proteins are defined by their characteristic conserved DNA binding region, the high mobility group (HMG) domain (Schepers et al., 2002). While the N-terminus containing the HMG domain is well-resolved (Michael et al., 2020), the C-terminus thought to engage co-factor interactions and transactivation remains structurally enigmatic. In ESCs, SOX2 is always found to be adjacent to OCT4 at a nearby motif of CTTTGTT (Chen et al., 2008; Hou et al., 2017; Michael et al., 2020), emphasizing its function with OCT4 to maintain pluripotency, collaboratively.

Sox2 is present in oocyte. Deletion of Sox2 resulted in early mouse embryonic lethality (Avilion et al., 2003; Campolo et al., 2013). While Sox2 is homogenously expressed in the ICM along with Oct4 and Nanog, it becomes exclusive to progenitor cells mostly of neuronal or epithelial fate by mid-late-streak stages E7.0 – E7.5 (Avilion et al., 2003; Schaefer and Lengerke, 2020). Like Oct4, the expression of Sox2 needs to be fine-tuned to maintain a subtle balance between pluripotency and differentiation, albeit there are controversial results between mouse and human. In mESCs, a less than 2-fold increase of Sox2 represses Oct4, Nanog, and a Oct4-Sox2 target gene Lefty1, causing differentiation of neuroectoderm and trophectoderm (Kopp et al., 2008). In contrast, in hESCs, overexpression of SOX2 has no apparent effect on inducing differentiation (Wang et al., 2012), while knockdown of SOX2 blocks ectoderm differentiation and leads to spontaneous endoderm gene expression (Teo et al., 2011; Wang et al., 2012). Taken together, SOX2 is important for self-renewal, as well as ectoderm differentiation. Both *in vivo* and *in vitro* data suggest that OSN functions are mutually exclusive at the onset of cell fate commitment (Mulas et al., 2018; Thomson et al., 2011), despite those core factors work synergistically to govern pluripotency (Boyer et al., 2005; Chen et al., 2008; Chronis et al., 2017).

Interactome studies reveal the core pluripotency factors OSN have physically contact with hundreds of important binding partners, including stem cell maintenance factors, chromatin remodelers, DNA methyltransferases, histone modifiers, chromatin looping anchor proteins (Chronis et al., 2017; Rafiee et al., 2016). The details will be reviewed in chapter 2.

SMAD SIGNALING IN STEM CELLS

The transforming growth factor β (TGFβ) superfamily are important morphogens in metazoans. There are 42 members in humans, of which Nodal, Activin and BMP families are considered as the evolutionarily most ancient family members (Pang et al., 2011). Nodal and Activin ligands are mutually replaceable at signaling the same receptors and effectors, SMAD2/3, in terms of gene regulation, whereas BMP families take effects through SMAD1/5/8. During gastrulation, combined gradients of Nodal/Activin and BMP signaling within primitive streak control endoderm and mesoderm germ layer specification and also their subsequent patterning whilst blocking neuroectoderm formation (Camus et al., 2006; Mesnard et al., 2006).

SMADs and gene regulation

The Activin/Nodal/BMP pathways are mediated by three classes of SMAD proteins: the receptor-regulated SMADs (R-SMADs, SMAD1/2/3/5/8), the common-mediator SMAD (Co-SMAD, SMAD4), and the inhibitory SMADs (I-SMADs, SMAD6/7) (Shi et al., 1997). SMAD1/5/8 signaling is mostly activated by BMP4, whereas SMAD2/3 signaling is activated by Nodal/Activin-A in humans. Upon receptor activation, the phosphorylated R-SMADs form a complex with the SMAD4 in the cytoplasm and then translocate into the nucleus to co-regulate target gene transcription. SMAD6/7 function as intracellular antagonist of R-SMADs complexes. R-SMADs and SMAD4 contain a highly conserved N-terminal DNA-binding MH1 domain, a weakly conserved linker region and a C-terminal transactivation MH2 domain. Both MH1 and MH2 domains can mediate protein-protein interactions (Pauklin and Vallier, 2015). Phosphorylation of the linker region of SMADs affects their translocation into nucleus and

mediates proteasome-dependent degradation through interaction with Smurf proteins (Zhang et al., 2001).

R-SMADs and SMAD4 bind directly to DNA, but with low affinity and low specificity (Aragón et al., 2019). This nature renders SMADs high tolerance in terms of motif recognition. Recent studies reveal that R-SMADs and SMAD4 can recognize a 5-bp consensus sequence GGC(GC)|(CG), in addition to the canonical SMAD binding element (SBE) of CAGAC (Kusanagi et al., 2000; Martin-Malpartida et al., 2017). Particularly, R-SMADS often require other co-factors to facilitate their binding to DNA, such as FOXH1 and C/EBPβ. The R-SMAD complex can also recruit active or repressive factors, such as histone acetyltransferase p300, histone deacetylases HDAC1-6, respectively. The requirement for these co-factors depends on molecular and cellular context. The resulting complexes regulate target gene expression in a cell type specific manner (Ross and Hill, 2008).

SMAD signaling and self-renewal

As introduced previously, both mEpiSCs and hESCs require extrinsic Activin-A signaling to maintain self-renewal state. Upon activation and dimerization, SMAD2/3 binds to promoters of key pluripotency genes, OCT4 and NANOG, and maintains pluripotent state through regulation of their transcription (Mullen et al., 2011; Vallier et al., 2009). It is also shown that SMAD2/3 cooperates with OSN as master regulators (Mullen et al., 2011; Teo et al., 2011) (Figure 1.2). Chemical inhibition of ALK4/7 (Nodal/Activin receptors) drives hESCs differentiation towards neuroectoderm lineage (Vallier et al., 2009). Nevertheless, in hESCs, the Activin-A/SMAD2/3 activity is suppressed at a lower level by PI3K/Akt signaling than that in differentiated cells (Singh et al., 2012).

Unlike in naïve mESCs, BMP4-SMAD1 signals seem unnecessary to block spontaneous differentiation in hESCs and mEpiSCs (Dahéron et al., 2004). The role of SMAD1 in self-renewal remains poorly understood due to limited studies in hESCs, despite that intrinsic phosphorylated form of SMAD1 is detectable in hESCs (Singh et al., 2012). Genome wide mapping of SMAD1 in both mESCs and hESCs confirm that SMAD1 cooperates with OSN at substantial portion of binding sites (Chen et al., 2008; Tsankov et al., 2015). However, it is shown that SMAD1 and NANOG reciprocally bind to each other's promoter by repressing their respective expression (Suzuki et al., 2006a; Xu et al., 2008). Combing with that OSN pre-marks differentiation genes (Boyer et al., 2005), it is assumed that BMP4-SMAD1 signaling may not be responsible for self-renewal, but rather for lineage priming predisposition in collaboration with OSN (Figure 1.2). Intriguingly, OCT4 knockdown by siRNA greatly disrupts SMAD1 binding to OCT4-SMAD1 co-occupied sites (Chen et al., 2008), emphasizing the important role of OCT4 in stabilizing SMAD1 complex loading onto DNA.

SMAD signaling and differentiation

Despite that extrinsically driven Activin-A signaling governs self-renewal state, high doses of Activin-A induce differentiation of hESCs into definitive endoderm (Singh et al., 2012; Vallier et al., 2009) (Figure 1.1). As major effector of Activin-A, SMAD2/3 has been found to relocate from OSN co-occupied sites to endoderm marker genes, including EOMES, GSC, SOX17 and LEFTY1, hereby activating their expression (Bertero et al., 2015). The underlying mechanism of gene regulation is by SMAD2/3 binding to distal enhancers (Kim et al., 2011). It is also shown the induced endoderm TF, EOMES, in turn interacts with SMAD2/3 to cooperate gene regulation (Faial et al., 2015; Teo et al., 2011).

Accordingly, inhibition of Activin-A blocks SMAD2/3 target gene expression and promotes the expression of mesoderm markers in the presence of BMP4 *in vitro* (Kubo et al., 2004) (Figure 1.1). This phenomenon is well studied at the promoter of NANOG in hESCs. Inhibitor of Activin or BMP4 treatment concordantly reduces SMAD2/3 binding, whereas increases SMAD1 binding (Xu et al., 2008). Furthermore, upon BMP4 induction, SMAD1 interacts with important mesoderm regulators such as T-Brachyury (Faial et al., 2015). Genome-wide analysis also propose that SMAD1, together with GATA4, recruits P300 to induce acetylation of H3K27 at mesoderm enhancers (Tsankov et al., 2015). WNT signaling is also present in both *in vitro* endoderm and mesoderm differentiation in this study, the WNT effectors TCF3/4 and β-catenin are thought to be co-factors of R-SMADs, interacting and stabilizing R-SMADs complexes (Pauklin and Vallier, 2015).

Lastly, for neuroectoderm differentiation, dual SMAD inhibition is widely used *in vitro* (Chambers et al., 2009; Chambers et al., 2012) (Figure 1.1). Two small molecular inhibitors, SB431542 and LDN193189, are capable of inhibiting SMAD signaling at high efficiency. The dual SMAD inhibition recapitulates mutually exclusive expression pattern of EOMES/T-Brachyury and SOX1/SOX2 between primitive streak and neuroectodermal epiblast in embryo development (Costello et al., 2011; Mulas et al., 2018). Although SMADs are reported to co-occupy with SOX2 in ESCs and believed to be part of the master regulatory network in self-renewal state, SMADs and SOX2 seem to be reciprocally inhibitory in regulating respective cell fate commitment.

CHROMATIN DYNAMICS IN STEM CELLS

The genome in PSCs is in a non-random and highly plastic state to meet the bifurcated needs in pluripotency (Parada and Misteli, 2002). Self-renewal requires that the genome maintains a cellular memory sustaining the pluripotency regulatory circuitry. On the other hand, once differentiation signaling cues, the genome undergoes widespread conformational and epigenetic changes, allowing a pluripotency-to-lineage transcriptional profile switch (Dixon et al., 2015; Dixon et al., 2012). One exception is that gene promoters are often consecutively accessible and largely invariant across cell types (Corces et al., 2016; Klemm et al., 2019). There are a few aspects of genome regulation, including nuclear lamina, chromosome positioning, the nucleolus, heterochromatin, topologically activation domain, and chromatins. In this chapter, the focus will be on linear and topological organizations of chromatins and their roles in gene regulation in pluripotency and differentiation.

Chromatin accessibility: a window to uncover the regulatory transcription factors

Eukaryotic chromatin is a complex of DNA and histone octamer, tightly packed into an array of nucleosomes, each of which is wrapped around by 147 bp of DNA and separated by linker DNA (Luger et al., 1997). Chromatin accessibility is the degree to which transcription factors (TFs) or other macromolecules physically contact DNA and is determined by the level of nucleosome occupancy. The distribution of nucleosomes alongside the chromatinized DNA is non-uniform, high degree of nucleosome eviction or destabilization often occur at cis-regulatory elements, such as promoters, enhancers, insulators, and actively transcribed gene bodies (Klemm et al., 2019). Such accessibility can be quantified using next-generation sequencing techniques, such as DNase-seq, MNase-seq, ATAC-seq, and FAIRE-seq (Tsompana and Buck, 2014).

It is reported that over 90% of profiled TFs fall within accessible chromatins, except for a small subset of heterochromatin associated proteins. Most of TFs exhibit high DNA binding affinity to certain representative sequence pattern, named motif (Thurman et al., 2012). Coordinatively, TF-bound distal accessible regions are usually flanked by histone modification marks, such as histone H3 lysine 4 mono-methylation (H3K4me1) and histone H3 lysine 27 acetylation (H3K27ac), which are well-documented marks for enhancers (Ong and Corces, 2011). A well-known example of this biophysical restriction is the core pluripotency factors OSN, which synergistically bind to their respective motif in close proximity during reprogramming (Chronis et al., 2017) and self-renewal (Chen et al., 2008). Currently there are four working models that explain how chromatin accessibility is established and maintained by motif-oriented TFs and active nucleosome remodelers (Klemm et al., 2019) (Figure 1.3).

The first model is passive competition between TFs and core histones, which is arguably the most elementary model given the fact that DNA is almost universally bound by either histones or other DNA-binding factors (Thurman et al., 2012). In this model, chromatin accessibility changes primarily in response to ratio of TF-histone turn-over rate. In support of this model, it has been shown that upregulation of C/EBPα competes histones at distal regulatory regions to induce pluripotency-related genes and facilitates reprogramming (Di Stefano et al., 2016; Svaren et al., 1994).

A second model is based on *cis* regulation through proximal linker histone displacement. This is a multistep process, including TFs initially binds to internucleosomal DNA and subsequently destabilizing core histone particle. At the initial step, the TF binding to internucleosomal region often relies on displacement of either linker histone H1 or other architectural proteins. Subsequently, stabilizing factors and/or chromatin remodelers are

recruited to maintain the chromatin accessibility. There is a line of evidence supporting this model. FOXA1 can access H1-compact chromatin and (Cirillo et al., 2002) and pre-marks inactive enhancers at early stages through pancreatic lineage differentiation by establishing a transcriptionally permissive enhancer chromatin state (Wang et al., 2015).

In the third model, chromatin remodeling is in *trans* regulation. Simply speaking, a TF bind to an accessible regulatory element for an induced TF that in turn recruits other co-factors to evict nucleosomes. Direct evidence for this mechanism is provided extensively in stem cell research. Endogenous OCT4 and NANOG reactivation and chromatin opening at respective promoter are employed as reporters for successful reprogramming (Meissner et al., 2007; Wernig et al., 2007). Induced OCT4 and NANOG, together with various of cofactors, bind to one another promoters to augment chromatin accessibility (Chen et al., 2008; Chronis et al., 2017).

The last model is the direct binding of pioneer TFs to nucleosomal DNA. Conventionally, a pioneer factor is arguably believed to be capable of binding to nucleosomal DNA and independently resulting in chromatin opening. However, the term is now used to describe TFs that are first to bind to DNA and then evict nucleosomes via one of the mechanisms outlined above (Klemm et al., 2019). Recent studies addressing how OCT4 functions as a pioneer factors favors the latter definition (King and Klose, 2017; Soufi et al., 2015). Soufi et al. demonstrated that OCT4 can recognize partial motif embedded within closed chromatin. Notably, 48 hours post reprogramming initiation, ~85% of OCT4 bound regions are also nucleosome occupied. King and Klose prove that depletion of BRG1, an ATP-dependent chromatin remodeler, cause dramatically increased nucleosome occupancy at OCT4 target sites. OCT4 and BRG1 are also inter-dependent to stabilize each other's binding to target sites. Therefore, it is reasonable to

deduce that OCT4 first binds to partial DNA motif wrapped around nucleosome and subsequently evict core histones via the cofactor BRG1.

Taken together, as chromatin accessibility is dynamically regulated through cell fate commitment, TFs play a critical organizing role in this process due to their DNA sequence specificity. In a reverse point of view, capturing differential chromatin accessibility across cell types can reveal the footprints of key TFs which determine the cell fate decision.

Higher-order chromatin structure and gene regulation

In a simplified model, the linear structure of chromatin, in a unit of nucleosome, determines accessibility. However, higher-order nucleosome organization impact access to DNA in a dimension-restricted manner. For example, higher-order chromatin organizations are often linked to long-range promoter-enhancer interactions that in turn control gene expression during ell fate commitment (Gorkin et al., 2014). Such non-linear chromatin interaction in transcriptional regulation is possible given two key fundamental properties: (1) most promoters are affected by more than one distal enhancer, and (2) one enhancer can influence gene expression regardless of linear distance (Corces et al., 2018; Gorkin et al., 2014; Thurman et al., 2012). This physical proximity allows TFs or protein complexes bound at enhancers to interact with those bound at promoters, hence regulating transcriptional machinery. The conformation of the higher-order chromatin structure is highly dynamic during early embryonic development and lineage commitment (Dixon et al., 2015), unravelling the molecular basis underlying this process gains critical knowledge of a multilayered gene regulatory network.

With the rapid development of technique based on the concept of chromatin conformation capture (3C) (Miele et al., 2006), high resolution of sub-topologically-associated-

domain (sub-TAD) promoter-enhancer interactions, at scale of 1kb, become observable (Rao et al., 2014). TADs are regions of high local chromatin contact frequency that are segregated by sharp borders, across which chromatin contacts are at low frequency. TADs, at scale of Mb, are largely invariant across cell types, whereas sub-TAD chromatin features are often cell type specific (Dixon et al., 2012). During ESC differentiation towards the three germ layers, although TADs remain stable across cell types, drastic intra-TAD interaction frequency changes in accordance with gene expression. Especially, the poised enhancer mark H3K4me1 is the most correlated histone mark in response to cell type specific intra-TAD changes, suggesting that enhancer dynamics may play a role in regulating local interaction changes during cell determination (Dixon et al., 2015).

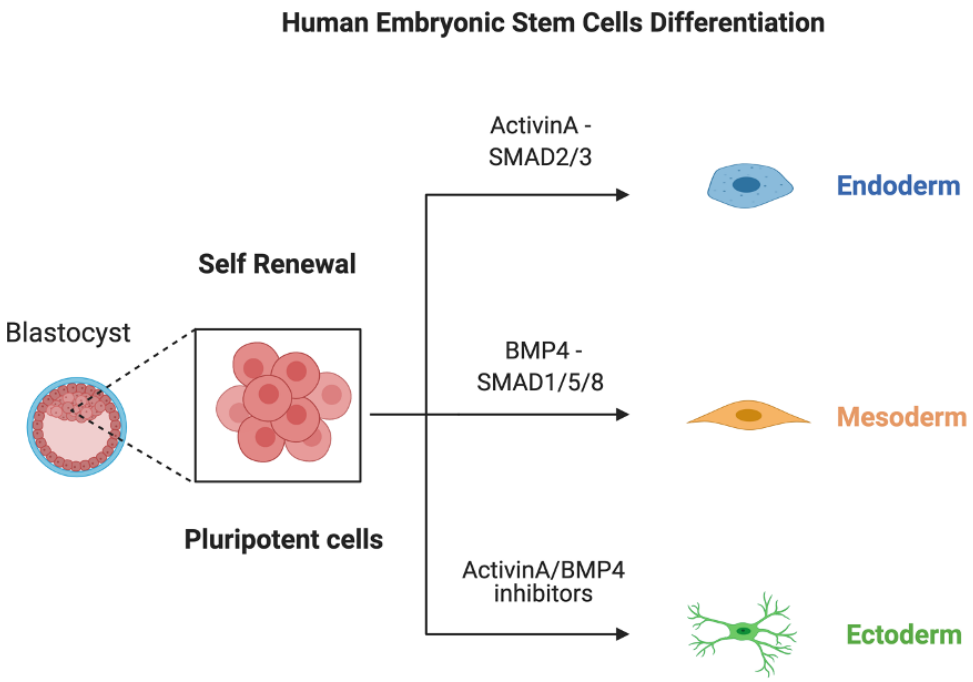


Figure 1.1 – Scheme of pluripotent stem cell differentiation *in vitro*. Cell fate commitment is under spatial and temporal control of extrinsic signaling cues. In this research, under chemically defined conditions, hESCs are induced towards three well-characterized germ layers, definitive endoderm (DE), splanchnic mesoderm (Meso), and ectoderm (Ecto). Briefly, hESCs are maintained in low Activin-A basic medium for self-renewal. For differentiation, high concentration of Activin-A treatment for 4 days results in DE. BMP4 treatment for 4 days results in Meso. Dual inhibition of Activin and BMP receptors for 6 days leads to Ecto. See Methods for details.

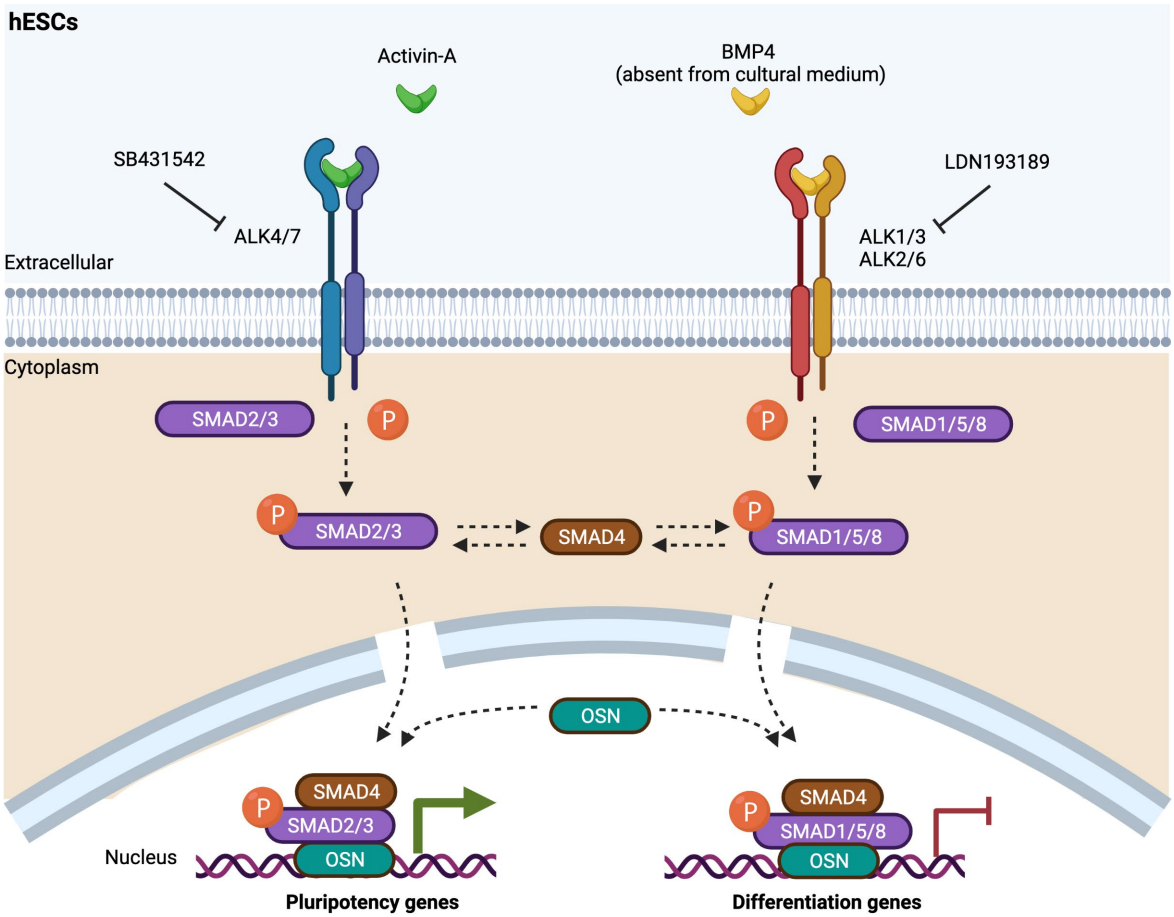


Figure 1.2 – SMAD signaling in hESCs. Low dose of Activin is required to maintain SMAD2/3 activation in hESCs. Phosphorylated SMAD2/3 is integrated into OSN core network to regulate pluripotency genes during self-renewal. Although external BMP ligands are absent in basic medium, intrinsic SMAD1 is found to be translocated into nucleus. The exact role of SMAD1 is poorly understood in hESC self-renewal. It is suggested that SMAD1 and SMAD2/3 signals are reciprocally inhibitory at regulating pluripotency gene expression, such as NANOG. Moreover, SMAD1 and OSN are found to be binding tomesendoderm genes at self-renewal state, leading to the assumption that SMAD1 is responsible for differentiation priming.

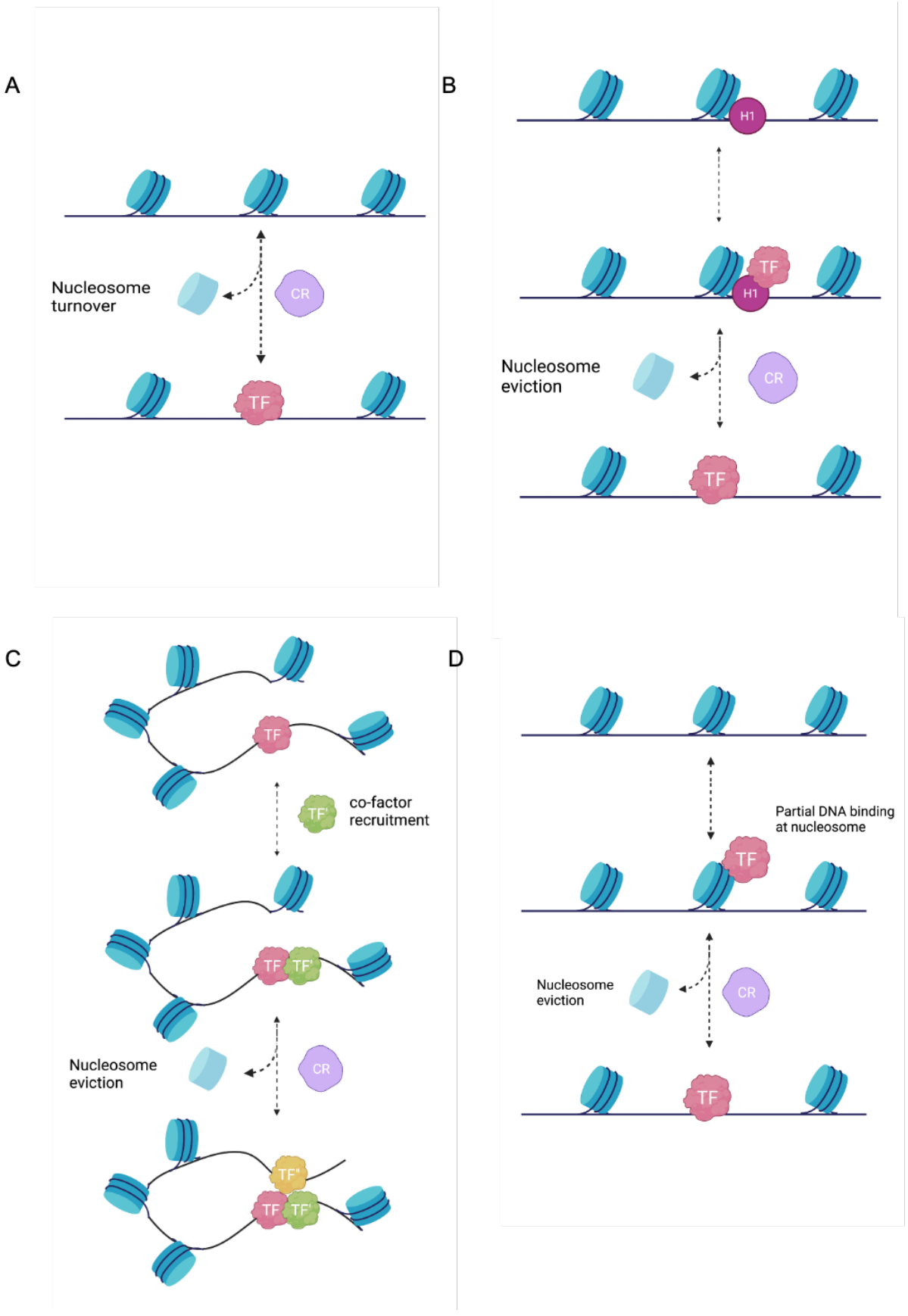

A
Nucleosome
turnover
CR
TF
B
H1
TF
H1
Nucleosome
eviction
CR
TF
C
TF
co-factor
recruitment
TF'
TF
TF'
Nucleosome
eviction
CR
TF'
TF
TF'
D
Partial DNA binding
at nucleosome
TF
Nucleosome
eviction
CR
TF

Figure 1.3 – Models of chromatin accessibility remodeling. (A)Transcription factors (TFs) compete with dynamic nucleosomes to initiate chromatin accessibility. The turnover of nucleosomes is regulated by specific chromatin remodeler (CR). (B) *cis*-regulation of chromatin accessibility is triggered by TF binding at internucleosomal regions like H1-histone cap or other architectural proteins. The displacement by TF causes destabilization of nucleosome. CR subsequently help TF evict the nucleosome (C) The primary TF binds to consecutively accessible regions and recruits other synergistic TF in *trans*. Active TF complex mediates nucleosome eviction by recruiting more TFs and CR. As a result, the newly formed protein complex stabilizes the nucleosome free state and mediates chromatin looping. (D) Pioneer TFs bind to nucleosomal DNA and nucleosome directly open chromatin either independently or collaboratively with CR.

REFERENCES

Abboud, N., Moore-Morris, T., Hiriart, E., Yang, H., Bezerra, H., Gualazzi, M.-G., Stefanovic, S., Guénantin, A.-C., Evans, S.M., and Pucéat, M. (2015). A cohesin–OCT4 complex mediates Sox enhancers to prime an early embryonic lineage. Nature communications *6*, 1-14.

Abulaiti, X., Zhang, H., Wang, A., Li, N., Li, Y., Wang, C., Du, X., and Li, L. (2017). Phosphorylation of Threonine343 Is Crucial for OCT4 Interaction with SOX2 in the Maintenance of Mouse Embryonic Stem Cell Pluripotency. Stem cell reports *9*, 1630-1641.

Aragón, E., Wang, Q., Zou, Y., Morgani, S.M., Ruiz, L., Kaczmarska, Z., Su, J., Torner, C., Tian, L., and Hu, J. (2019). Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF-β signaling. Genes & development *33*, 1506-1524.

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

Bertero, A., Madrigal, P., Galli, A., Hubner, N.C., Moreno, I., Burks, D., Brown, S., Pedersen, R.A., Gaffney, D., and Mendjan, S. (2015). Activin/Nodal signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. Genes & development *29*, 702-717.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., and Jenner, R.G. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. cell *122*, 947-956.

Campolo, F., Gori, M., Favaro, R., Nicolis, S., Pellegrini, M., Botti, F., Rossi, P., Jannini, E.A., and Dolci, S. (2013). Essential role of Sox2 for the establishment and maintenance of the germ cell line. Stem cells *31*, 1408-1421.

Camus, A., Perea-Gomez, A., Moreau, A., and Collignon, J. (2006). Absence of Nodal signaling promotes precocious neural differentiation in the mouse embryo. Developmental biology *295*, 743-755.

Cauffman, G., Van de Velde, H., Liebaers, I., and Van Steirteghem, A. (2004). Oct-4 mRNA and protein expression during human preimplantation development. Molecular human reproduction *11*, 173-181.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell *113*, 643-655.

Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nature biotechnology *27*, 275-280.

Chambers, S.M., Qi, Y., Mica, Y., Lee, G., Zhang, X.-J., Niu, L., Bilsland, J., Cao, L., Stevens, E., and Whiting, P. (2012). Combined small-molecule inhibition accelerates developmental timing and converts human pluripotent stem cells into nociceptors. Nature biotechnology *30*, 715.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J.*, et al.* (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell *133*, 1106-1117.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative binding of transcription factors orchestrates reprogramming. Cell *168*, 442-459. e420.

Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Molecular cell *9*, 279-289.

Cliff, T.S., Wu, T., Boward, B.R., Yin, A., Yin, H., Glushka, J.N., Prestegaard, J.H., and Dalton, S. (2017). MYC Controls Human Pluripotent Stem Cell Fate Decisions through Regulation of Metabolic Flux. Cell Stem Cell *21*, 502-516.e509.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., and Greenleaf, W.J. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nature genetics *48*, 1193-1203.

Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., and Cho, S.W. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*.

Costello, I., Pimeisl, I.-M., Dräger, S., Bikoff, E.K., Robertson, E.J., and Arnold, S.J. (2011). The T-box transcription factor Eomesodermin acts upstream of Mesp1 to specify cardiac mesoderm during mouse gastrulation. Nature cell biology *13*, 1084-1091.

Dahéron, L., Opitz, S.L., Zaehres, H., Lensch, W.M., Andrews, P.W., Itskovitz‐Eldor, J., and Daley, G.Q. (2004). LIF/STAT3 signaling fails to maintain self‐renewal of human embryonic stem cells. Stem cells *22*, 770-778.

Davis, A.C., Wims, M., Spotts, G.D., Hann, S.R., and Bradley, A. (1993). A null c-myc mutation causes lethality before 10.5 days of gestation in homozygotes and reduced fertility in heterozygous female mice. Genes & development *7*, 671-682.

DeVeale, B., Brokhman, I., Mohseni, P., Babak, T., Yoon, C., Lin, A., Onishi, K., Tomilin, A., Pevny, L., and Zandstra, P.W. (2013). Oct4 is required~ E7. 5 for proliferation in the primitive streak. PLoS Genet *9*, e1003957.

Di Giammartino, D.C., Kloetgen, A., Polyzos, A., Liu, Y., Kim, D., Murphy, D., Abuhashem, A., Cavaliere, P., Aronson, B., and Shah, V. (2019). KLF4 is involved in the organization and

regulation of pluripotency-associated three-dimensional enhancer networks. Nature cell biology *21*, 1179-1190.

Di Stefano, B., Collombet, S., Jakobsen, J.S., Wierer, M., Sardina, J.L., Lackner, A., Stadhouders, R., Segura-Morales, C., Francesconi, M., and Limone, F. (2016). C/EBPα creates elite cells for iPSC reprogramming by upregulating Klf4 and increasing the levels of Lsd1 and Brd4. Nature cell biology *18*, 371-381.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., and Xie, W. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331-336.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376-380.

Dominguez-Sola, D., Ying, C.Y., Grandori, C., Ruggiero, L., Chen, B., Li, M., Galloway, D.A., Gu, W., Gautier, J., and Dalla-Favera, R. (2007). Non-transcriptional control of DNA replication by c-Myc. Nature *448*, 445-451.

Downs, K.M. (2008). Systematic localization of Oct‐3/4 to the gastrulating mouse conceptus suggests manifold roles in mammalian development. Developmental Dynamics *237*, 464-475.

Faial, T., Bernardo, A.S., Mendjan, S., Diamanti, E., Ortmann, D., Gentsch, G.E., Mascetti, V.L., Trotter, M.W., Smith, J.C., and Pedersen, R.A. (2015). Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. Development *142*, 2121-2135.

Festuccia, N., Osorno, R., Wilson, V., and Chambers, I. (2013). The role of pluripotency gene regulatory network components in mediating transitions between pluripotent cell states. Current Opinion in Genetics & Development *23*, 504-511.

Gao, L., Wu, K., Liu, Z., Yao, X., Yuan, S., Tao, W., Yi, L., Yu, G., Hou, Z., and Fan, D. (2018). Chromatin accessibility landscape in human early embryos and its association with evolution. Cell *173*, 248-259. e215.

Gorkin, David U., Leung, D., and Ren, B. (2014). The 3D Genome in Transcriptional Regulation and Pluripotency. Cell Stem Cell *14*, 762-775.

Hou, L., Srivastava, Y., and Jauch, R. (2017). Molecular basis for the genome engagement by Sox proteins. Paper presented at: Seminars in cell & developmental biology (Elsevier).

Jaenisch, R., and Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell *132*, 567-582.

Jin, W., Wang, L., Zhu, F., Tan, W., Lin, W., Chen, D., Sun, Q., and Xia, Z. (2016). Critical POU domain residues confer Oct4 uniqueness in somatic cell reprogramming. Scientific reports *6*, 20818.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900-903.

Kim, S.W., Yoon, S.-J., Chuong, E., Oyolu, C., Wills, A.E., Gupta, R., and Baker, J. (2011). Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. Developmental biology *357*, 492-504.

King, H.W., and Klose, R.J.J.E. (2017). The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. *6*, e22631.

Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics *20*, 207-220.

Kopp, J.L., Ormsbee, B.D., Desler, M., and Rizzino, A. (2008). Small increases in the level of Sox2 trigger the differentiation of mouse embryonic stem cells. Stem cells *26*, 903-911.

Kubo, A., Shinozaki, K., Shannon, J.M., Kouskoff, V., Kennedy, M., Woo, S., Fehling, H.J., and Keller, G. (2004). Development of definitive endoderm from embryonic stem cells in culture. Development *131*, 1651-1662.

Kusanagi, K., Inoue, H., Ishidou, Y., Mishima, H.K., Kawabata, M., and Miyazono, K. (2000). Characterization of a bone morphogenetic protein-responsive Smad-binding element. Molecular Biology of the Cell *11*, 555-565.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., and Liu, J. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nature genetics *38*, 431-440.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature *389*, 251-260.

Martin-Malpartida, P., Batet, M., Kaczmarska, Z., Freier, R., Gomes, T., Aragón, E., Zou, Y., Wang, Q., Xi, Q., and Ruiz, L. (2017). Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors. Nature communications *8*, 1-15.

Meissner, A., Wernig, M., and Jaenisch, R. (2007). Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. Nature biotechnology *25*, 1177-1181.

Mendjan, S., Mascetti, V.L., Ortmann, D., Ortiz, M., Karjosukarso, D.W., Ng, Y., Moreau, T., and Pedersen, R.A. (2014). NANOG and CDX2 pattern distinct subtypes of human mesoderm during exit from pluripotency. Cell stem cell *15*, 310-325.

Mesnard, D., Guzman-Ayala, M., and Constam, D.B. (2006). Nodal specifies embryonic visceral endoderm and sustains pluripotent cells in the epiblast before overt axial patterning. Development *133*, 2497-2505.

Michael, A.K., Grand, R.S., Isbel, L., Cavadini, S., Kozicka, Z., Kempf, G., Bunker, R.D., Schenk, A.D., Graff-Meyer, A., and Pathare, G.R. (2020). Mechanisms of OCT4-SOX2 motif readout on nucleosomes. Science.

Miele, A., Gheldof, N., Tabuchi, T.M., Dostie, J., and Dekker, J. (2006). Mapping chromatin interactions by chromosome conformation capture. Current protocols in molecular biology, 21.11. 21-21.11. 20.

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

Mulas, C., Chia, G., Jones, K.A., Hodgson, A.C., Stirparo, G.G., and Nichols, J. (2018). Oct4 regulates the embryonic axis and coordinates exit from pluripotency and germ layer specification in the mouse embryo. Development *145*.

Mullen, A.C., Orlando, D.A., Newman, J.J., Lovén, J., Kumar, R.M., Bilodeau, S., Reddy, J., Guenther, M.G., DeKoter, R.P., and Young, R.A. (2011). Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell *147*, 565-576.

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

Niwa, H. (2007). How is pluripotency determined and maintained? Development *134*, 635-646.

Niwa, H., Burdon, T., Chambers, I., and Smith, A. (1998). Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. Genes & development *12*, 2048-2060.

Niwa, H., Miyazaki, J.-i., and Smith, A.G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nature genetics *24*, 372-376.

Ong, C.-T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Reviews Genetics *12*, 283-293.

Pan, G.J., and Pei, D.Q. (2003). Identification of two distinct transactivation domains in the pluripotency sustaining factor nanog. Cell research *13*, 499-502.

Pang, K., Ryan, J.F., Baxevanis, A.D., and Martindale, M.Q. (2011). Evolution of the TGF-β signaling pathway and its potential role in the ctenophore, Mnemiopsis leidyi. PloS one *6*, e24152.

Parada, L.A., and Misteli, T. (2002). Chromosome positioning in the interphase nucleus. Trends in cell biology *12*, 425-432.

Pauklin, S., and Vallier, L. (2015). Activin/Nodal signalling in stem cells. Development *142*, 607-619.

Radzisheuskaya, A., Chia, G.L.B., Dos Santos, R.L., Theunissen, T.W., Castro, L.F.C., Nichols, J., and Silva, J.C. (2013). A defined Oct4 level governs cell state transitions of pluripotency entry and differentiation into all embryonic lineages. Nature cell biology *15*, 579.

Radzisheuskaya, A., and Silva, J.C. (2014). Do all roads lead to Oct4? The emerging concepts of induced pluripotency. Trends in cell biology *24*, 275-284.

Rafiee, M.-R., Girardot, C., Sigismondo, G., and Krijgsveld, J. (2016). Expanding the Circuitry of Pluripotency by Selective Isolation of Chromatin-Associated Proteins. Molecular Cell *64*, 624-635.

Ram, P.T., and Schultz, R.M. (1993). Reporter gene expression in G2 of the 1-cell mouse embryo. Developmental biology *156*, 552-556.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., and Lander, E.S. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

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

Ross, S., and Hill, C.S. (2008). How the Smads regulate transcription. The international journal of biochemistry & cell biology *40*, 383-408.

Saijoh, Y., Fujii, H., Meno, C., Sato, M., Hirota, Y., Nagamatsu, S., lkeda, M., and Hamada, H. (1996). Identification of putative downstream genes of Oct‐3, a pluripotent cell‐specific transcription factor. Genes to Cells *1*, 239-252.

Schaefer, T., and Lengerke, C. (2020). SOX2 protein biochemistry in stemness, reprogramming, and cancer: the PI3K/AKT/SOX2 axis and beyond. Oncogene *39*, 278-292.

Schepers, G.E., Teasdale, R.D., and Koopman, P. (2002). Twenty pairs of sox: extent, homology, and nomenclature of the mouse and human sox transcription factor gene families. Developmental cell *3*, 167-170.

Schöler, H., Balling, R., Hatzopoulos, A.K., Suzuki, N., and Gruss, P. (1989a). Octamer binding proteins confer transcriptional activity in early mouse embryogenesis. The EMBO journal *8*, 2551-2557.

Schöler, H., Hatzopoulos, A.K., Balling, R., Suzuki, N., and Gruss, P. (1989b). A family of octamer‐specific proteins present during mouse embryogenesis: evidence for germline‐specific expression of an Oct factor. The EMBO journal *8*, 2543-2550.

Shi, Y., Hata, A., Lo, R.S., Massague, J., and Pavletich, N.P. (1997). A structural basis for mutational inactivation of the tumour suppressor Smad4. Nature *388*, 87-93.

Silva, J., Nichols, J., Theunissen, T.W., Guo, G., van Oosten, A.L., Barrandon, O., Wray, J., Yamanaka, S., Chambers, I., and Smith, A. (2009). Nanog is the gateway to the pluripotent ground state. Cell *138*, 722-737.

Silva, J., and Smith, A. (2008). Capturing Pluripotency. Cell *132*, 532-536.

Singh, A.M., Reynolds, D., Cliff, T., Ohtsuka, S., Mattheyses, A.L., Sun, Y., Menendez, L., Kulik, M., and Dalton, S. (2012). Signaling Network Crosstalk in Human Pluripotent Cells: A Smad2/3-Regulated Switch that Controls the Balance between Self-Renewal and Differentiation. Cell Stem Cell *10*, 312-326.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555-568.

Suzuki, A., Raya, A., Kawakami, Y., Morita, M., Matsui, T., Nakashima, K., Gage, F.H., Rodríguez-Esteban, C., and Belmonte, J.C.I. (2006a). Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells. Proceedings of the National Academy of Sciences *103*, 10294-10299.

Suzuki, A., Raya, Á., Kawakami, Y., Morita, M., Matsui, T., Nakashima, K., Gage, F.H., Rodríguez-Esteban, C., and Belmonte, J.C.I. (2006b). Maintenance of embryonic stem cell pluripotency by Nanog-mediated reversal of mesoderm specification. Nature clinical practice Cardiovascular medicine *3*, S114-S122.

Svaren, J., Klebanow, E., Sealy, L., and Chalkley, R. (1994). Analysis of the competition between nucleosome formation and transcription factor binding. Journal of Biological Chemistry *269*, 9335-9344.

Tabar, V., and Studer, L. (2014). Pluripotent stem cells in regenerative medicine: challenges and recent progress. Nature Reviews Genetics *15*, 82-92.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. cell *126*, 663-676.

Teo, A.K.K., Arnold, S.J., Trotter, M.W., Brown, S., Ang, L.T., Chng, Z., Robertson, E.J., Dunn, N.R., and Vallier, L. (2011). Pluripotency factors regulate definitive endoderm specification through eomesodermin. Genes & development *25*, 238-250.

Thomson, M., Liu, S.J., Zou, L.-N., Smith, Z., Meissner, A., and Ramanathan, S. (2011). Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. Cell *145*, 875-889.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., and Vernot, B. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Tokuzawa, Y., Kaiho, E., Maruyama, M., Takahashi, K., Mitsui, K., Maeda, M., Niwa, H., and Yamanaka, S. (2003). Fbx15 is a novel target of Oct3/4 but is dispensable for embryonic stem cell self-renewal and mouse development. Molecular and cellular biology *23*, 2699.

Tsankov, A.M., Gu, H., Akopian, V., Ziller, M.J., Donaghey, J., Amit, I., Gnirke, A., and Meissner, A. (2015). Transcription factor binding dynamics during human ES cell differentiation. Nature *518*, 344-349.

Tsompana, M., and Buck, M.J. (2014). Chromatin accessibility: a window into the genome. Epigenetics & chromatin *7*, 1.

Vallier, L., Mendjan, S., Brown, S., Chng, Z., Teo, A., Smithers, L.E., Trotter, M.W., Cho, C.H.-H., Martinez, A., and Rugg-Gunn, P. (2009). Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. Development *136*, 1339-1349.

Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N.A., Muth, K., Palmer, J., Qiu, Y., and Wang, J. (2015). Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. Cell stem cell *16*, 386-399.

Wang, Z., Oron, E., Nelson, B., Razis, S., and Ivanova, N. (2012). Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells. Cell stem cell *10*, 440-454.

Wei, C.L., Miura, T., Robson, P., Lim, S.K., Xu, X.Q., Lee, M.Y.C., Gupta, S., Stanton, L., Luo, Y., and Schmitt, J. (2005). Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state. Stem cells *23*, 166-185.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., and Jaenisch, R. (2007). In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. nature *448*, 318-324.

Xu, R.-H., Sampsell-Barron, T.L., Gu, F., Root, S., Peck, R.M., Pan, G., Yu, J., Antosiewicz-Bourget, J., Tian, S., and Stewart, R. (2008). NANOG is a direct target of TGFβ/activin-mediated SMAD signaling in human ESCs. Cell stem cell *3*, 196-206.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., and Stewart, R. (2007). Induced pluripotent stem cell lines derived from human somatic cells. science *318*, 1917-1920.

Zhang, Y., Chang, C., Gehling, D.J., Hemmati-Brivanlou, A., and Derynck, R. (2001). Regulation of Smad degradation and activity by Smurf2, an E3 ubiquitin ligase. Proceedings of the National Academy of Sciences *98*, 974-979.

CHAPTER 2

CORE PLURIPOTENCY FACTORS AND CELL FATE COMMIMENT: IMPLICATIONS FOR CHROMATIN REMODELING, DIFFERENTIATION GENE ACTIVATION AND HIGHER-ORDER CHROMATIN INTERACTION

INTRODUCTION

The maintenance of ESC self-renewal requires a regulatory circuity of TFs, of which OSN are regarded as the core. Surprisingly, in addition to pluripotency related targets, OSN promiscuously bind to promoters of a substantial portion of development related genes (Guenther et al., 2007). It is also reported that OSN bound promoters tend to coincide with bivalent domains (H3K4me3 and H3K27me3 modifications of histone) (Bernstein et al., 2006). Previously, OSN are thought to play a repressive role at these targets to prevent unwanted differentiation (Jaenisch and Young, 2008; Young, 2011). Recent hESC studies show that bivalent domain marked differentiation genes are temporally activated in a cell cycle dependent manner (Pauklin and Vallier, 2013; Singh et al., 2015), whereas the DNA binding profiles of OSN are relatively consecutive across cell cycle stages (Friman et al., 2019; Osnato and Vallier, 2020). These results indicate the OSN may be permissive, rather than repressive, for differentiation gene priming, albeit their drastic down-regulated expression during gastrulation.

It has been introduced in Chapter 1 that the precise levels of OSN govern distinct fates of PSCs. Briefly, a less than two-fold overexpression of OCT4 causes differentiation into mesoderm in hESCs and mESCs; overexpression of NANOG specifies definitive endoderm in hESCs and mESCs; overexpression of SOX2 prompts ectoderm differentiation in mESCs only.

Conversely, repression of individual of OSN results in spontaneous differentiation towards the lineage other than the one induced by overexpression.

Indirect evidence suggests BMP4/ACTIVIN-A activated SMAD signals are involved in these observations. Faial et al. and Mendjan et al. established a method whereby inhibition of PI3K pathway for 24-48 hr in addition of ACTIVIN-A or BMP4 to induce early differentiation. In BMP4 induced mesoderm cells, the dominant effector T-Brachyury binds to a group of target sites which enrich for partial motif of OCT4 (Faial et al., 2015). Although direct interaction between T-Brachyury and OCT4 has not been detected, it is interesting to know that T-Brachyury physically interact with SMAD1, a major BMP4 effector. Additionally, OCT4 has been shown to bind to a subset of SMAD1 sites in in vitro differentiated mesoderm cells (Tsankov et al., 2015). These are supportive of a long-standing hypothesis back in 2001, that OCT4 could be involved in cell commitment during gastrulation by inductive signals either of OCT4 itself or of OCT4 partners (Smith, 2001).

On the other hand, under ACTIVIN-A induction condition, NANOG protein colocalizes with EOMES. Overexpression of NANOG only results in a endoderm fate, not mesoderm (Mendjan et al., 2014). EOMES has been demonstrated to cooperate with SMAD2/3 to activate endoderm target genes (Faial et al., 2015; Teo et al., 2011). The direct physical interaction between SMAD2/3 and NANOG, and SMAD2/3 and EOMES, have been validated in hESCs as well (Faial et al., 2015; Xu et al., 2008).

Out of the three core pluripotency factors, SOX2 is the most understood due to its exclusive expression pattern in neural ectoderm development. Although SOX2 co-binds to SMAD1 and SMAD2/3 to some extent in ESCs (Chen et al., 2008; Mullen et al., 2011). The differentiation of ectoderm requires dual SMAD inhibition. Neural progenitor cells (NPCs) are

well-defined and induced by two inhibitors antagonizing BMP4/ACTIVIN-A *in vitro* (Chambers et al., 2009). SOX2 binding re-distribution take places at the onset of NPCs and activates a myriad of neural fate specification genes (Zhou et al., 2016).

In summary, these results could possibly be integrated into the classic Waddington's landscape model, in which PSCs would resemble the rolling balls down a hill, deciding their lineage path at key branching points in choice of OSN and SMAD signals.

BEYOND PLURIPOTENCY

Intuitively, if OSN function as lineage specifiers, they must bind to and activate transcriptional machinery of target loci encoding differentiation genes. In agreement with this idea, OSN are known to bind to differentiation genes under self-renewal state (Boyer et al., 2005), and bind to a subset of distal elements during early differentiation (Tsankov et al., 2015). Combining the fact that pluripotency factors are recognized as pioneer factors during reprogramming, it is reasonable to speculate that OSN may function as pioneer factors to potentiate chromatin remodeling and facilitate the binding of first responders of extrinsic signaling cues to regulatory elements. In here, I will address the possibility of OSN as pioneer factors in differentiation.

In PSCs, OSN are known to recognize target loci based on their respective DNA motif in undifferentiated state. Surprisingly, at the beginning of reprogramming process in mouse fibroblasts, Oct4 and Sox2 (OS) promiscuously bind to genetic loci unique to fibroblast cell type (Chronis et al., 2017; Li et al., 2017). These loci do not have canonical OS motifs, but are instead enriched with motifs for differentiation genes, such as TEAD, RUNX, and AP-1 family TFs. The relocation of OS from developmental loci to their OS motif enriched pluripotency loci happens in the latter stage of reprogramming, which takes a few days. These observations indicate that

OS do not strictly follow the rule of DNA sequence specific binding. On the contrary, they may be recruited by developmental TFs, possibly in a way of physical protein interaction, at distal elements specific to differentiation genes. Their role in co-binding with developmental TFs could be explained in one of the models of chromatin remodeling (see Chapter 1 and Figure 1.3). It will be illuminating to test whether ectopically expression and tethering of pluripotency factors at those loci is enough to initiate/maintain nucleosome-free status and to activate gene expression.

Moreover, during the switch of OS binding events from differentiation loci to pluripotency loci, OS are found to bind partial motifs embedded within intact nucleosome (Soufi et al., 2015). It is reported that BRG1 is required for OCT4 to shape chromatin accessibility landscape (King and Klose, 2017). These results implicate the pioneer activity of OS, although the detailed mechanism of how OS individually or cooperatively evict nucleosome remains unclear. Intriguingly, the ability of OCT4 binding to partial core motif is coinciding with the observation that partial OCT4 motif is enriched in SMAD1/T-Brachyury target sites in BMP4-induced mesoderm cells (Faial et al., 2015). It is tempting to assume that OCT4 might cooperate with SMAD1 and T-Brachyury to potentiate mesoderm fate, given that SMAD1 itself has low affinity to DNA (Hill, 2016) and T-Brachyury is not expressed instantly in response to BMP4 cues (Singh et al., 2012). Let alone OCT4 knockdown impairs SMAD1 binding to OCT4-SMAD1 co-binding regions (Chen et al., 2008). In this hypothesis, the ability of OCT4 directly binding to intact nucleosome can aid SMAD1 contact with nucleosome wrapped SMAD motif at the first step of mesoderm differentiation. Should my hypothesis hold true, OCT4 dependent eviction of nucleosome will fully expose SMAD motif to SMAD1. Subsequent expressed T-

Brachyury will later participate the complex, whereby stabilizes SMAD1 binding and recruits other cofactors to activate target genes.

As for NANOG, although direct evidence pointing to a pioneer activity is missing, the utility of NANOG as a reprogramming factor predicts that NANOG might bind to closed chromatin just like OCT4 and SOX2. It is also hinted that NANOG is involved in establishing a repertoire for SMAD2/3 binding to DNA. Bertero et al. and Xu et al. demonstrate that NANOG physically interacts with SMAD2/3 to recruit H3K4 methyltransferase, DPY30, on ACTIVIN-A responsive genes in hESCs (Xu et al., 2008). Importantly, NANOG knockdown recapitulates ACTIVIN-A pathway inhibitor caused loss of SMAD2/3 binding events (Bertero et al., 2015). These data suggest that NANOG is necessary for depositing active histone marks nearby SMAD2/3 binding sites. Moreover, NANOG is necessary to initiate the expression of EOMES, which has been validated to interact with SMAD2/3 and cooperatively activates the transcriptional network directing definitive endoderm formation (Faial et al., 2015; Teo et al., 2011). Like the hypothesis proposed above, one may readily deduce that NANOG could act as a pioneer factor, thus, that NANOG will facilitate SMAD2/3 binding to SMAD motif at closed chromatin before EOMES is activated in endoderm fate.

In sum, I propose that OSN function as pioneer factors and that OSN modulate selective SMAD, in response to BMP4/ACTIVIN-A manipulation, to trigger lineage-restricted fate commitment (Figure 2).

CORE PLURIPOTENCY FACTORS ARE IMPLICATED IN CHROMATIN DYNAMICS DURING EARLY DIFFERENTIATION

As discussed in Chapter 1, as differentiation proceeds, PSCs gains high-order chromatin conformation changes which permit novel promoter-enhancer interactions for genes important at later stages while losing interactions required at earlier stages (Dixon et al., 2015; Gorkin et al., 2014). These promoter-enhancer interactions are described as a looping structure, mostly at a scale of hundreds of kb within TADs. The loop anchors are usually secured by architectural proteins and active co-factors, including TFs, Cohesin, Mediator, and RNA polymerase II (Gorkin et al., 2014). In both human and mouse ESCs, OSN are highly enriched at those sub-TAD promoter-enhancer loop anchors (De Wit et al., 2013; Dowen et al., 2014; Ji et al., 2015).

In particular, OSN have been found to physically interact with each other and with key anchor proteins, histones, and chromatin remodelers, such as Smc1a, Smc3 (component of Cohesin), histone variants (H1, H2a, H2b, H3, H4), and Brg1 (Rafiee et al., 2016). Loss-of-function experiments further confirm the necessity of OSN. Knockdown of either of Oct4 or Nanog in mouse ESCs disrupt loop formation at interactive loci by these TFs (Apostolou et al., 2013; De Wit et al., 2013). The OSN mediated short-range promoter-enhancer interactions are also Cohesin dependent (Phillips-Cremins et al., 2013; Whyte et al., 2013). The presence of homodimer and hetero-dimer forms of OSN (Michael et al., 2020; Mistri et al., 2020; Pan and Pei, 2003) suggest that OSN may also directly form loops in a context dependent manner. Put together, above observations suggest that OSN are directly involved in higher-order chromatin organization in ESCs.

In a case study, Oct4 is reported to shift from Sox2 to Sox17 loci between in ESCs and in mesendoderm (Abboud et al., 2015). Oct4 binding at Sox17 is dose dependent and requires

Cohesin loading. 3C experiments also confirm that overexpression of Oct4 results in loss of loops at Sox2 loci while gaining of loops at Sox17 loci.

In short, OSN are required for proper promoter-enhancer interactions to maintain self-renewal, more importantly, they are also implicated in reshaping higher-order chromatin landscape during differentiation.

CONCLUSIONS

The core pluripotency factor OSN are inherently unstable. They are interconnected to main self-renewal, meanwhile they are reciprocally inhibited at the onset of differentiation. In classic point of view, PSCs are protected by a shield of pluripotency factors that interactively repress differentiation. However, such model cannot explain how PSCs exit pluripotency and establish a chromatin landscape without the emergence of differentiation TFs, given that BMP4/ACTIVIN-A responsive SMADs only have weak affinity to DNA. To this point, I propose a new model in this research (Figure 2). That the plausible pioneer activity of OSN confers a fundamental mechanism to potentiate chromatin remodeling, allowing selected SMAD to find its niche. This process not only re-shapes cell identity in 2D, but also in 3D organization of chromatin. The resultant expression of differentiation TFs in turn can take place as OSN are down regulated. This model will be tested in the following chapters.

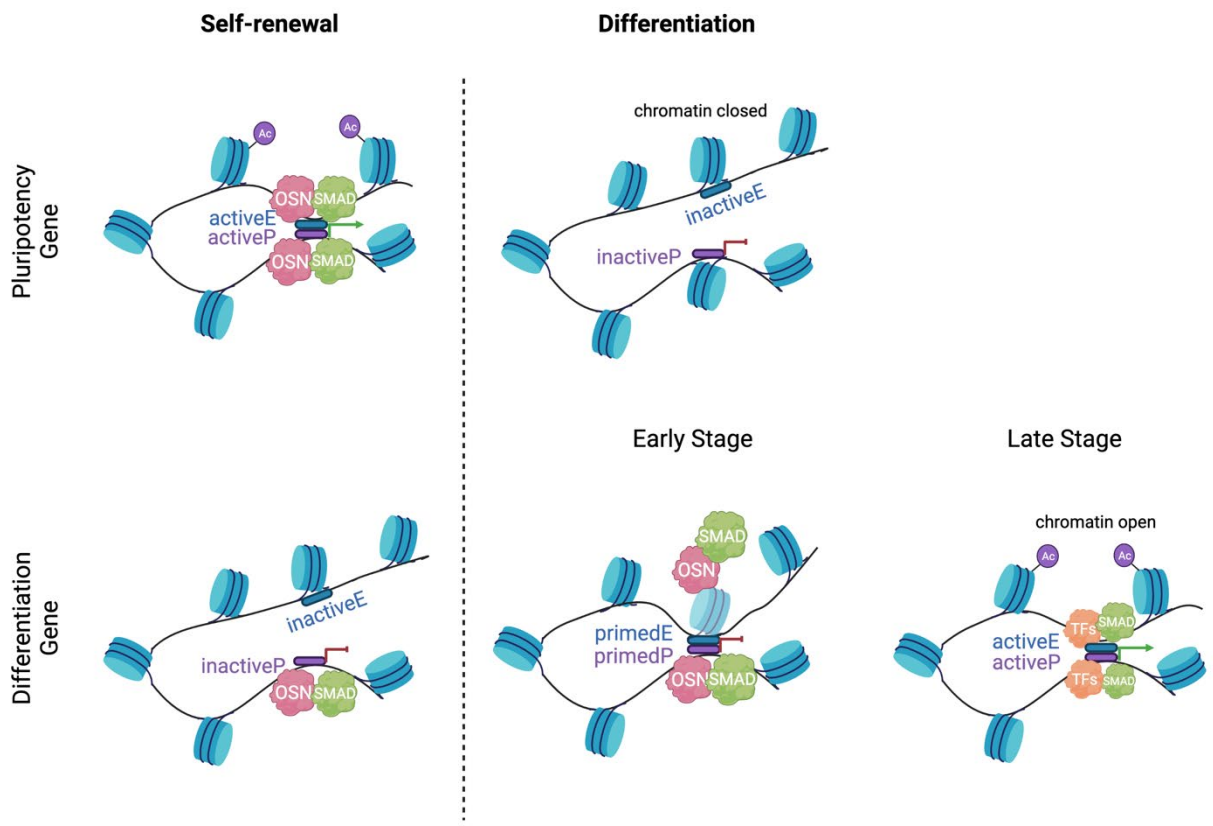


Figure 2 – Hypothesis: OSN factors and BMP4/ACTIVIN-A regulated SMADs orchestrate as cell fate specifiers during hESC differentiation. (Top panel**)** It has been extensively investigated that OSN (to clarify, OSN represents either individual one or collaborative ones of the three factors in this hypothesis) are key TFs regulating pluripotency genes under self-renewal state through binding to proximal and distal regulatory elements, such as promoters (P) and enhancers (E). Genome-wide analysis also show that SMADs often integrate with OSN regulatory complex. The regulatory effect within OSN-SMADs could be both synergistic and antagonistic (details discussed previously). At the onset of differentiation, external signaling cues can cause down-regulation of OSN expression and dynamic SMADs phosphorylation, resulting in chromatin closing at pluripotency gene loci and pluripotency exit (Bottom panel). Conversely, regulatory elements for differentiation genes undergo dynamic chromatin remodeling (Early

Stage) and consequently prompt transcription activation (Late Stage). OSN-SMADs pre-mark inactive promoters of differentiation genes under self-renewal state. At the initiation step of differentiation, BMP4/ACTIVIN-A activated SMADs recognize their motif at distal enhancers in Meso and DE, respectively. However, they cannot directly bind to DNA due to presence of intact nucleosomes. At this step, enhancers are at primed stage due to lack of active histone marks (H3K27ac, denoted as Ac in purple arrow). The eviction of these nucleosomes relies on the interacted OSN, individually or collaboratively. Additionally, the OSN facilitate the proximity interactions between promoters and enhancers via looping formation. The pioneer activity of OSN evicts nucleosomes and thus, making DNA motif accessible to SMAD proteins. Consequently, the active promoter-enhancer interactions promote expression of target differentiation TFs, such as GATA6 in DE and HAND1 in Meso. These differentiation TFs are in turn recruited to displace OSN, whereby stabilizing SMADs binding and maintaining chromatin configuration. In the scenario of ectoderm differentiation, since dual inhibition of SMAD is required, the SMAD module will be missing in the model. The block of all Nodal/Activin/BMP pathways results in quick loss of OCT4 and NANOG. SOX2 will be released from the OSN-SMADs complex and re-distributed to regulatory elements containing SOX and other related motif.

REFERENCES

Abboud, N., Moore-Morris, T., Hiriart, E., Yang, H., Bezerra, H., Gualazzi, M.-G., Stefanovic, S., Guénantin, A.-C., Evans, S.M., and Pucéat, M. (2015). A cohesin–OCT4 complex mediates Sox enhancers to prime an early embryonic lineage. Nature communications *6*, 1-14.

Apostolou, E., Ferrari, F., Walsh, R.M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S., Stuart, H.T., Polo, J.M., Ohsumi, T.K., and Borowsky, M.L. (2013). Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. Cell stem cell *12*, 699-712.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., and Plath, K. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bertero, A., Madrigal, P., Galli, A., Hubner, N.C., Moreno, I., Burks, D., Brown, S., Pedersen, R.A., Gaffney, D., and Mendjan, S. (2015). Activin/Nodal signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. Genes & development *29*, 702-717.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., and Jenner, R.G. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. cell *122*, 947-956.

Chambers, S.M., Fasano, C.A., Papapetrou, E.P., Tomishima, M., Sadelain, M., and Studer, L. (2009). Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nature biotechnology *27*, 275-280.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J.*, et al.* (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell *133*, 1106-1117.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative binding of transcription factors orchestrates reprogramming. Cell *168*, 442-459. e420.

De Wit, E., Bouwman, B.A., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J., Krijger, P.H., Festuccia, N., Nora, E.P., and Welling, M. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature *501*, 227-231.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., and Xie, W. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331-336.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K.*, et al.* (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374-387.

Faial, T., Bernardo, A.S., Mendjan, S., Diamanti, E., Ortmann, D., Gentsch, G.E., Mascetti, V.L., Trotter, M.W., Smith, J.C., and Pedersen, R.A. (2015). Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. Development *142*, 2121-2135.

Friman, E.T., Deluz, C., Meireles-Filho, A.C., Govindan, S., Gardeux, V., Deplancke, B., and Suter, D.M. (2019). Dynamic regulation of chromatin accessibility by pluripotency transcription factors across the cell cycle. Elife *8*, e50087.

Gorkin, David U., Leung, D., and Ren, B. (2014). The 3D Genome in Transcriptional Regulation and Pluripotency. Cell Stem Cell *14*, 762-775.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Hill, C.S. (2016). Transcriptional control by the SMADs. Cold Spring Harbor perspectives in biology *8*, a022079.

Jaenisch, R., and Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell *132*, 567-582.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., and Lee, T.I. (2015). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell stem cell.

King, H.W., and Klose, R.J.J.E. (2017). The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. *6*, e22631.

Li, D., Liu, J., Yang, X., Zhou, C., Guo, J., Wu, C., Qin, Y., Guo, L., He, J., and Yu, S. (2017). Chromatin accessibility dynamics during iPSC reprogramming. Cell stem cell *21*, 819-833. e816.

Mendjan, S., Mascetti, V.L., Ortmann, D., Ortiz, M., Karjosukarso, D.W., Ng, Y., Moreau, T., and Pedersen, R.A. (2014). NANOG and CDX2 pattern distinct subtypes of human mesoderm during exit from pluripotency. Cell stem cell *15*, 310-325.

Michael, A.K., Grand, R.S., Isbel, L., Cavadini, S., Kozicka, Z., Kempf, G., Bunker, R.D., Schenk, A.D., Graff-Meyer, A., and Pathare, G.R. (2020). Mechanisms of OCT4-SOX2 motif readout on nucleosomes. Science.

Mistri, T.K., Kelly, D., Mak, J., Colby, D., Mullin, N., and Chambers, I. (2020). Characterisation of interactions between the pluripotency transcription factors Nanog, Oct4 and Sox2. bioRxiv.

Mullen, A.C., Orlando, D.A., Newman, J.J., Lovén, J., Kumar, R.M., Bilodeau, S., Reddy, J., Guenther, M.G., DeKoter, R.P., and Young, R.A. (2011). Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell *147*, 565-576.

Osnato, A., and Vallier, L. (2020). Transcriptional networks are dynamically regulated during cell cycle progression in human Pluripotent Stem Cells. bioRxiv.

Pan, G.J., and Pei, D.Q. (2003). Identification of two distinct transactivation domains in the pluripotency sustaining factor nanog. Cell research *13*, 499-502.

Pauklin, S., and Vallier, L. (2013). The cell-cycle state of stem cells determines cell fate propensity. Cell *155*, 135-147.

Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.-T., Hookway, T.A., Guo, C., and Sun, Y. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281-1295.

Rafiee, M.-R., Girardot, C., Sigismondo, G., and Krijgsveld, J. (2016). Expanding the Circuitry of Pluripotency by Selective Isolation of Chromatin-Associated Proteins. Molecular Cell *64*, 624-635.

Singh, A.M., Reynolds, D., Cliff, T., Ohtsuka, S., Mattheyses, A.L., Sun, Y., Menendez, L., Kulik, M., and Dalton, S. (2012). Signaling Network Crosstalk in Human Pluripotent Cells: A Smad2/3-Regulated Switch that Controls the Balance between Self-Renewal and Differentiation. Cell Stem Cell *10*, 312-326.

Singh, A.M., Sun, Y., Li, L., Zhang, W., Wu, T., Zhao, S., Qin, Z., and Dalton, S. (2015). Cell-Cycle Control of Bivalent Epigenetic Domains Regulates the Exit from Pluripotency. Stem cell reports *5*, 323-336.

Smith, A.G. (2001). Embryo-derived stem cells: of mice and men. Annual review of cell and developmental biology *17*, 435-462.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555-568.

Teo, A.K.K., Arnold, S.J., Trotter, M.W., Brown, S., Ang, L.T., Chng, Z., Robertson, E.J., Dunn, N.R., and Vallier, L. (2011). Pluripotency factors regulate definitive endoderm specification through eomesodermin. Genes & development *25*, 238-250.

Tsankov, A.M., Gu, H., Akopian, V., Ziller, M.J., Donaghey, J., Amit, I., Gnirke, A., and Meissner, A. (2015). Transcription factor binding dynamics during human ES cell differentiation. Nature *518*, 344-349.

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

Xu, R.-H., Sampsell-Barron, T.L., Gu, F., Root, S., Peck, R.M., Pan, G., Yu, J., Antosiewicz-Bourget, J., Tian, S., and Stewart, R. (2008). NANOG is a direct target of TGFβ/activin-mediated SMAD signaling in human ESCs. Cell stem cell *3*, 196-206.

Young, R.A. (2011). Control of the embryonic stem cell state. Cell *144*, 940-954.

Zhou, C., Yang, X., Sun, Y., Yu, H., Zhang, Y., and Jin, Y. (2016). Comprehensive profiling reveals mechanisms of SOX2-mediated cell fate specification in human ESCs and NPCs. Cell research *26*, 171.

CHAPTER 3

CORE PLURIPOTENCY FACTORS COORDINATE WITH SMAD TO SPECIFY CELL FATE COMMITMENT

INTRODUCTION

As hESCs exit pluripotency they undergo dramatic changes in transcriptional machinery, gene expression profile, epigenomic modification and chromatin conformation (see Chapter 1&2). Using RNA-seq technique, we have provided a molecular roadmap for differentiation of hESCs toward DE, Meso, and Ecto (Cliff et al., 2017). The next step to elucidate the mechanistic network regulating differential gene needs information of causative TF profiles and regulatory DNA elements. Exhausted profiling of each candidate TF is however near impossible. It is widely accepted that active DNA elements are chromatin accessible and are often bound by regulatory TFs (Klemm et al., 2019), dissecting chromatin accessibility landscape across cell types confers a flexible approach to narrow down the candidate list of TFs responsible for divergent cell fate specification.

We carried out ATAC-seq (assay for transposase-accessible chromatin using sequencing) (Buenrostro et al., 2013) and mapped DNA accessibility at chromatin sites from hESC and derived DE, Meso, and Ecto. We show that during differentiation, cell fates are defined by unique open chromatin patterns. Promoter regions of differentiation genes are largely invariant across cell types and are pre-marked by core pluripotency factors, OCT4-SOX2-NANOG (abbreviated OSN), in combination of SMAD2/3. Whereas distal open chromatin loci are clustered in a cell type specific manner. A small fraction of the cell type specific open chromatin loci are putative active enhancers and are occupied by coordinated OSN-SMAD in respective

cell type. NANOG-SMAD2/3 bind to DE-specific active enhancers depending on SMAD phosphorylation, whereas OCT4-SMAD1 bind to Meso-specific active enhancers. NANOG and OCT4 also mediate unique higher-order chromatin re-organizations in DE and Meso, respectively. Direct tethering of NANOG or OCT4 to unopened enhancers is sufficient to activate developmental gene expression. Conditional knockout of NANOG or OCT4 impairs SMAD2/3 or SMAD1 binding to chromatinized DNA at enhancer sites, respectively, accompanied by decreased H3K27ac deposition. SOX2 is the sole pluripotency factor binding to Ecto-specific enhancers. Together, we expand the understanding of the core pluripotency as cell fate specifiers.

RESULTS

DISTAL OPEN CHROMATIN LOCI DEFINE CELL IDENTITY

hESCs are induced for differentiation depending on SMAD signaling (Figure 1.1 and Supplemental Fig4.1A). I generated high quality ATAC-seq data in two biological replicates from the four cell types, hESC and its early derivatives, DE, Meso, and Ecto. I quantified enrichment over whole genome background for each sample and obtained reproducible ATAC peaks (R ranges 0.73 to 0.91, Supplemental Figure 4.1B) with each cell type having a median of 60,809 peaks (range 55,692 to 64,352). Similarly, the accessibility data between hESC ATAC-seq and Roadmap Epigenomics Project DNase I hypersensitive sites sequencing (DNase-seq) from hESC (Table 1) are highly overlapped (Supplemental Figure 4.1C), highlighting the reproducibility of our results. By comparing enrichment scores based on *K*-means clustering of merged reproducible peaks, ATAC-seq data revealed 3 basic clusters, uniquely open in one cell

type (Cell-Type-Specific, CTS), consecutively open in all cell types (Stable), dynamically open in more than one cell type (Dynamic) (Figure 3.1A).

Of the 12,3483 combined ATAC-seq peaks, the three major categories display divergent genomic features: The Stable peaks are highly enriched for promoters and within 1 kb flanking the transcription start sites (TSS) of protein-coding genes; CTS peaks account for 59% of all combined peaks and are widely distributed along the chromosomes, but much less enriched for promoters, suggesting unique distal elements are determinants of cell type specification; Dynamic peaks are promiscuous in terms of genomic feature annotation, reflecting shared regulatory mechanism among different cells (Supplemental Figures 4.1D, E). These observations are consistent with published reports that chromatin accessibilities at promoters are largely invariant among different cell types, whereas distal accessibilities exhibited greater cell type specificity (Corces et al., 2018b; Klemm et al., 2019). To address whether the chromatin accessibility correlates with changes in gene expression during differentiation, I linked individual peak to nearest gene and found that one protein-coding gene could be linked by more than one distal peak in each cluster (Figure 3.1B and Supplemental Figure 4.1F). However, this peak-to-gene linkage is arbitrary, solely based on shortest distance. It is known that a *cis*-element may act regardless of distance and may regulate different genes, and one cis-element may regulate more than one gene (Gorkin et al., 2014; Thurman et al., 2012). Additionally, the prediction power of peak-to-gene linkage decays drastically as distance increases (Corces et al., 2018a). Therefore, we restricted the peak-to-gene linkage to 1 kb for Stable peaks (promoter) and within 1 kb to 10 kb for CTS peaks (distal) in gene expression analysis. Interestingly, we found a significant enrichment of distal-peak-linked genes having concordant promoter peaks with an average 49.4% (ranges from 47.1% to 52.4%, $p < 0.001$, overlapping of gene lists are permutated 1000

times over that from random genomic regions). In contrast, overlapping of the genes among distal clusters is usually low, on average of 12.5% (ranges from 2.1% to 26.2%, p = 0.79) (Figure 3.1C), resembling the clustering purity of ATAC-seq peaks in CTS category (Figure 3.1A). In total, we linked at least one peak-to-gene pair for 29,624 peaks and 11,923 protein-coding genes which accounts for around half of all protein-coding genes in the human genome. As a result, we found that the gene expression patterns are in association with ATAC-seq peak clustering (Figure 3.1D) and are enriched for biological pathways in respective cell fate (Supplemental Figure 4.1G), indicating that distal peaks are regions of functional *cis*-elements involved in transcription activation during differentiation (Figure 3.1E).

ENHANCER HIERACHY AT OPEN CHROMATIN LOCI

H3K4me1 and H3K27ac are known markers to predict functional enhancers in vicinity of open chromatin regions (Ong and Corces, 2011). Hence, not all open chromatin loci are believed to be functional enhancers shaping cell identity (Klemm et al., 2019). To further dissect the prediction power of chromatin accessibility in relation to gene expression, I mapped global H3K27ac ChIP-seq data, in combination of public H3K4me1 ChIP-seq, and distinguished enhancer hierarchy by assigning the CTS open chromatins into three enhancer sub-types: enhancers undergo full deposition of H3K4me1 and H3K27ac in close vicinity as Active enhancers; those only marked by H3K4me1 as Primed enhancers; the ones absent from either H3K4me1 or H3K27ac marker as Poised enhancers (Supplemental Figure 4.2A). In summary, the number of Active enhancers from hESC, DE, Meso, Ecto are 1925, 3211, 5624, 1847, accounting for 17.2%, 19%, 24.9%, 13.4% of all putative enhances, respectively. I then re-performed the peak-to-gene expression analysis and found that genes linked by Active enhancers are highest in transcription profile

(Figure 3.2B), indicating the functionality of Active sub-type of defined enhancers. We noted that number of genes linked by more than one enhancer group is substantially smaller than that by an individual enhancer group (Supplemental Figure 4.2B). An average of ~13% of genes linked by Active enhancers are also linked by Primed enhancers, suggesting the complexity of enhancer regulatory network. We reasoned the differential activity between enhancer sub-types by that transitioning from inactive state to active state requires enhancer activation, which requires recruitment of pioneer factors, active TFs, histone acylation transferase, chromatin remodelers (Ong and Corces, 2011), and that enhancer (in)activation is also a reversible progress (Zaret and Yamamoto, 1984). The ATAC-seq technique indistinguishably snapshot varied stages of enhancer activation at given time point. Together, we report that H3K4me1 and H3K27ac marked CTS open chromatins are indicative of functionally active enhancers.

OSN AND SMAD PREDOMINANTLY OCCUPY ACTIVE ENHANCERS DURING DIFFERENTIATION

To further gain mechanistic insight into drivers establishing transcriptional machinery at Active enhancers, we wished to see whether dynamic TF binding patterns are the reason for enhancer hierarchy. Given that TF bind to DNA elements depending on specific DNA motif, motif discovery analysis should provide insights in inferring TF binding pattern. As quality control, ChIP-seq data of OSN, SMAD1, and SMAD2/3 were generated across hESC and the three germ layer cells and *de novo* motif discovery were carried out. Similar to published results (Tsankov et al., 2015), our ChIP-seq data generated canonical motifs for these OSN, SMAD1, and SMAD2/3, individually (Figure 3C). Next, we applied a stringent threshold for motif analysis on categorized enhancer loci and found that motifs for OSN, SMAD1, and SMAD2/3 are preferably

enriched in Active enhancers (Figure 3.2D, Supplemental Figure 3.3). Motifs for key developmental TFs are highly enriched accordingly, such as SOX17, and GATA family TFs in DE, CDX2, TEAD, GATA, and JUN-AP1 family TFs in Meso, and OTX2, LHX and SOX family in Ecto. The evident enrichment of motif for canonical OCT4:SOX2: NANOG in Active and Primed enhancers suggests the accuracy of our enhancer classification. Similarly, in DE, since ACTIVIN-SMAD2/3 signaling triggers differentiation, *de novo* motif for SMAD2/3 is expectedly enriched in Active enhancers. Whereas *de novo* motif for BMP4-responsive SMAD1 is enriched in both Meso-Active and Primed enhancers. As for Ecto, SOX family motif is consensus and ubiquitously present in all sub-types. A clear observation from motif analysis is that motifs for TFs that most closely related to pluripotency and differentiation cues tend to be enriched at Active enhancers in all four cell types, while motifs for differentiation induced TFs are either ubiquitously distributed or prone to bind to Primed/Poised loci (Supplemental Figure 4.3). TFs binding to primed enhancers are believed to be involved in enhancer priming for later developmental stages (Wang et al., 2015). Motif analysis suggests diverse TF binding patterns at different stages of enhancer activation. It is reasonable to assume that pre-existing TFs binding to Active enhancers at the beginning of differentiation are the first responders to external signaling and drivers of cell fate specification

To see whether this logic holds true, I integrated a comprehensive analysis of genome-wide binding data for 21 TFs from in-house and publish resource across hESC, DE, Meso, and Ecto (Tsankov et al., 2015) (Table 1). All TF ChIP-seq signals were normalized and were clustered based on the high similarity in binding profile over CTS enhancers (Supplemental Figure 4.4). List of TFs was narrowed down if a TF was clustered with CTS open chromatin signals collected from the same cell type. In consistent with previous understanding of

differentiation, important developmental TF binding pattern are restricted to cell fate. For example, SMAD2/3, EOMES, GATA6 are clustered with DE-Active enhancers, SMAD1, HAND1, TCF4 are clustered with Meso-Active enhancers (Figure 3.3A, Supplemental Figure 4.4) (Faial et al., 2015; Tsankov et al., 2015). Strikingly, OSN binding events at enhancer loci are not only observed in hESC, but also in the three germ layer cells, even though canonical OSN motif is not present at these loci (Figure 3.3B, Figure 3.2D and Supplemental Figure 4.5). OSN at enhancer loci is not an artifact due to unspecific co-immunoprecipitation, evident by obvious lineage specific binding pattern. Based on the binary binding of a certain factor, we then categorized Active enhancers into the simplest tier of O, S, or N, in combination with differentiation signaling dependent SMAD that defines a cell type: OSN bind to Active enhancers in hESC; OSN-SMAD2/3 bind to Active enhancers in DE; OS-SMAD1 bind to Active enhancers in Meso; S alone occupies Active enhancers in Ecto (Figure 3.3B). These observations are apparently conflicting with the conventional model that OSN are reciprocally inhibitory in cell fate decision. Instead, redistributed OSN may work cooperatively as drivers of chromatin opening in early differentiated cells in a similar fashion as they take action during iPSC reprogramming (Li et al., 2017). Therefore, coordinated OSN-SMAD co-binding in differentiated cells favors our hypothesis that OSN function as pioneer factors to initiate chromatin opening at enhancers (Soufi et al., 2015) (Discussed in Chapter 2).

BINDING PARTNER SWITCH BETWEEN OSN AND SMAD

To understand OSN action, we asked why OSN are redistributed independent of binding motif in DE and Meso. We began by hypothesizing that OSN, individually or collaboratively, are recruited by SMAD, which in turn facilitate SMAD to bind SBE motif at Active enhancers (Figure 3.2D and Figure 3.3C). This reasoning is based on the following facts: (1) SMAD binds to chromatinized DNA weakly and often requires co-binding factors (Gaarenstroom and Hill, 2014; Hill, 2016). (2) SMAD co-factors and cell fate specifying TFs like EOMES (DE) and T-Brachyury (Meso) are not expressed yet at the beginning of differentiation (Faial et al., 2015). (3) OSN are potential pioneer factors to initiate nucleosome depletion (Soufi et al., 2015). To test our hypothesis, we first need to confirm a physical interaction between OSN and SMAD during differentiation.

I performed proximity ligation assay (PLA) for OSN, SMAD1, and SMAD2/3 in a pairwise and time dependent manner. Experiments were conducted through D0, D1, D4/6 of differentiation from hESC to DE, Meso, and Ecto (Figure 3.4A). In agreement with previous understanding of core pluripotency circuits, O, S, and N mutually interact with each other in hESC as core pluripotency circuitry (Festuccia et al., 2013). At day1 of differentiation, mutual interactions between O, S, N drop dramatically (~70%), in accordance with decreased total proteins (Supplemental Figure 4.6). By D4 of differentiation, mutual interactions among OSN eventually drop to a near zero level, except for the pronounced O-S interactions in DE at D4 (Supplemental Figures 4.7A-C). Endogenous SMAD1 and SMAD2/3 are maintained at low phosphorylated state and are partners of OSN complex in hESC as expected (Supplemental Figure 4.6B, Supplemental Figure 4.4D-I). Interestingly, O preferably interacts with SMAD1 in Meso and significantly loses contact with SMAD1 in DE or Ecto. (Supplemental Figures 4.7D).

N specifically interacts with SMAD2/3 in DE, even at a higher frequency than in hESC. In contrast, NANOG-SMAD2/3 interactions decrease to basal detection levels in Meso and Ecto (Supplemental Figures 4.7H). Furthermore, O and N exhibit BMP4/Activin-A signaling dependent aversion to interaction with SMAD2/3 and SMAD1, respectively (Supplemental Figure 4.6B, Supplemental Figures 4.7E, G). As for S, dual SMAD inhibition completely blocked its interaction with either SMAD1 or SMAD2/3 in Ecto since D1 (Supplemental Figures 4.7F, I). S seems to prefer the under-activated SMAD1 in DE, and under-activated SMAD2/3 in Meso by D1 of differentiation. These physical contact between S and SMADs are finally lost by D4 (Supplemental Figures 4.7G, I). Combining the DNA binding data, the results suggest that OSN-SMAD derail pluripotency regulatory complex upon external signaling cues and reform phosphorylation dependent complex to at putative enhancers in differentiation.

To further test this hypothesis, I used SB431542 (SB), a selective Activin receptor inhibitor, and LDN193189 (LDN), a selective BMP receptor inhibitor. Addition of SB for 8 hours blocked NANOG-SMAD2/3 interaction under DE differentiation condition, while addition of LDN blocked OCT4-SMAD1 interaction under Meso differentiation condition (Figure 3.4B). Next, we wished to test whether redistribution of OSN from hESC open chromatin loci to developmental open chromatin locus is a result of SMAD signaling activation. To minimize background noise caused by signaling pathway cross-talk, hESC was allowed to spontaneously differentiate toward DE for 24 hours, using a previously defined minimal Activin-A activation approach (Singh et al., 2012). Removal of Heregulinβ -1 and IGF from self-renewal maintaining medium permits activation of SMAD2/3 pathway. I selected a putative enhancer for GATA6 based on DE-specific binding profile of NANOG, SMAD2/3, H3K27ac. This enhancer locus contains canonical binding motif for SMAD, but not NANOG (Figure 3.4C). As a result, we

observed NANOG binding to tested enhancer of GATA6, accompanied by GATA6 transcription activation (Figure 3.4D). Notably, addition of SB blocked NANOG binding and shutdown GATA6 transcription. In conclusion, we re-categorized core OSN-SMAD interactome in differentiated cells based on direct protein-protein interaction and concurrent SMAD signaling activation: NANOG-SMAD2/3 for DE, OCT4-SMAD1 for Meso, SOX2 alone for Ecto.

NANOG AND OCT4 BIND TO UNOPENED CHROMATIN PRECEDING SMAD DIRECT BINDING TO DNA

It is known that phosphorylation of R-SMAD rapidly and reversibly changes R-SMAD function. Previously I have demonstrated that, in DE, activated SMAD2/3 signaling can initiate NANOG redistribution and cause GATA6 expression within 24 hours (Figure 3.4D). Given that the tested enhancer is enriched for SMAD motif, and that SMAD2/3 physically interact with NANOG, it is reasonable to deduce that SMAD2/3 is the driver for NANOG redistribution. To our surprise, we failed detect SMAD2/3 binding at GATA6 enhancer by day 1 of DE (data not shown). We reasoned that chromatin opening is a time dependent progress, thus the interrogated GATA6 locus is not open yet at that time. We then hypothesized that NANOG binds to unopened chromatin whereby SMAD2/3 cannot or weakly bind to DNA. Indeed, normalized DE D1 ATAC-seq signals over DE-Active enhancers are significantly lower than that in DE D4, but slightly higher than in hESC (Figure 3.5A), indicating incomplete nucleosome depletion within the first 24 hours of differentiation. To validate our hypothesis, I used a dual crosslinking approach including EGS (Ethylene glycol bis(succinimidylsuccinate)) and formaldehyde to detect indirect DNA binding profile (Tian et al., 2012). For SMAD1 and SMAD2/3, direct binding (formaldehyde alone) and indirect binding (EGS + formaldehyde) ChIP-seq experiments

were carried out. Here, we show that NANOG already occupies substantial amount of DE-Active enhancers in DE D1 (Figure 3.5F). Furthermore, indirect SMAD2/3 ChIP-seq detected stronger binding events of SMAD2/3 than direct ChIP-seq (Figure 3.5F), indicating the indirect binding of SMAD2/3 to chromatinzed DNA at DE-Active enhancers. In a more obvious fashion, OCT4 binding events can be detected at Meso-Active enhancers by the time direct SMAD1 binding cannot be observed (Figures 3.5B, F). The evident binding of O and N at unopened chromatins suggest their pioneer activity for nucleosome eviction (Soufi et al., 2015). Of note, the substantial overlap between direct and indirect SMAD2/3 binding profiles suggests that chromatin remodeling is faster in DE than Meso.

To better understand the stepwise activation of developmental enhancers, I conducted time-course ChIP-qPCR experiments to evaluate TF binding dynamic for the first two days of DE differentiation. NANOG, SMAD2/3, GATA6, and H3K27ac were tested at three previously investigated enhancers, GATA6 E2, GATA6 E3, and SOX17 E2, all of which are not full chromatin open by day 1 of DE differentiation. Notably, GATA6 E2 locus has been extensively studied whereby NANOG binding is SMAD2/3 phosphorylation dependent (Figures 3.4C, D). In consistent with NANOG ChIP-seq pattern, ChIP-qPCR data confirmed that NANOG binding to individual enhancer is pronounced in DE D1 (Figure 3.5G). Interestingly, the active enhancer mark H3K27ac follows the trend of NANOG (Figure 3.5H), suggesting that H3K27ac deposition is a process preceding enhancer activation. GATA6 binding signal is hardly detected by day1 (Figure 3.5I), which supports our original hypothesis that key developmental TFs are absent at the beginning of differentiation, hence SMAD requires pre-existing TFs to act. Direct SAMD2/3 bindings were merely detectable in D1 samples, reflecting a weak affinity to chromatinized DNA. In contrast, indirect SMAD signal is stronger by 1-2 orders (Figure 3.5J). This binary

binding pattern of SMAD2/3 between hESC and DE D1 is in concert with NANOG and H3K27ac deposition, but not with traditionally key DE marker GATA6 (Figure 3.5K). In conclusion, the time course global and local binding profile of NANOG-SMAD2/3 and OCT4-SMAD1, and ATAC-seq data together lead to a conclusion that NANOG and OCT4 are capable of binding to putative developmental enhancer loci which are occupied with nucleosomes and enriched for SMAD motif.

OSN AND SMAD PRE-MARK PROMOTER-PROXIMAL ELEMENTS OF DIFFERENTIATION GENES

It is known that promoters of differentiation genes are pre-marked by OSN-SMAD in self-renewal state (Chen et al., 2008; Guenther et al., 2007; Teo et al., 2011), but the biological relevance regarding gene regulation remains debating. We noted that promoter region of GATA6 (DE) and HAND1 (Meso) are broadly marked by NANOG-SMAD23 and OCT4-SMAD1, respectively (Figure 3.5B). To test whether OSN-SMAD display any complex variability at promoters of genes associated with previously categorized Active enhancers, I re-mapped individual O, S, N and SMAD ChIP-seq data along the gene bodies linked by Active enhancers. OSN bind promiscuously to promoter-proximal elements (100-200 bp upstream of TSS) of differentiation genes regardless of cell fate commitment (Figures 3.6A-C). However, these proximal regions are not nucleosome free in hESC (Supplemental Figure 4.8A) and are not enriched for canonical motifs for OSN, but for GC box elements (Supplemental Figure 4.8B). Promoter-proximal elements are extensively studied for transcriptional machinery, consisting of RNA polymerase II, Mediator complex, and P300 (Griffiths et al., 2000; Whyte et al., 2013). Indeed, MED1, a subunit of Mediator, is observed at the same promoter proximal regions in

early G1 fraction of hESC (Supplemental Figure 4.8C). Interestingly, MED1 signal in hESC seems to be indicative of higher gene expression, evident by its moderately higher occupancy at promoter-proximal regions of hESC-Active enhancer linked genes and unique binding to hESC-Active enhancers. (Figure 3.2B, Supplemental Figure 4.8C, D). Moreover, recent studies demonstrate that OSN can form phase-separated condensates with Mediator and such unique protein-protein interaction is important for transcriptional regulation (Boija et al., 2018). These results could provide a possible explanation regarding OSN pre-deposition at at promoter regions of unexpressed differentiation genes.

Direct SMAD2/3 binding to promoter proximal regions are generally low except for DE D4 (Figure 3.6D), because SMAD2/3 full activation and DNA binding is time dependent (Figure 3.5E, Supplemental Figure 4.6B). Whereas indirect SMAD2/3 binding are more profound, which could be due to physical interactions with OSN in ES state (Figure 3.4A, Figure 3.6D, Supplemental Figures 4.7D, E). As for SMAD1, both direct and indirect SMAD1 binding to promoter proximal elements cannot be observed until extracellular BMP4 addition (Figure 3.6E). Like OSN, SMAD binding to promoter-proximal regions are not lineage specific. Given that OSN-SMAD occupied enhancers are indicative of transcription of differentiation genes (Figure 3.2B, Figure 3.3B), the promiscuous binding of OSN-SMAD at promoter regions of both expressed and unexpressed genes point to the fact that OSN may not be repressors of differentiation as previously thought (Boyer et al., 2005; Jaenisch and Young, 2008).

OCT4 AND NANOG CONNECT ACTIVE ENHANCERS TO PROMOTERS OF DIFFERENTIATION GENES

OCT4 and NANOG are known to mediate Cohesin dependent enhancer-promoter interactions for pluripotency related genes, a process relatively well studied in ES state (Phillips-Cremins et al., 2013; Whyte et al., 2013). Indeed, intersection of binding sites with chromatin states revealed that SMC1, subunit of Cohesin, displays a cell fate specific binding pattern at Active enhancers (Figure 3.3A). Further analysis show that SMC1 binding concordantly increases with NANOG, SMAD2/3, and H3K27ac levels at DE-Active enhancers (Figure 3.7A). Similar phenomenon was observed for OCT4-SMAD1 bound Meso-Active enhancers as well (Figure 3.7E). Presence of cohesion at enhancer sites are indicative of looping formation. To see whether pluripotency factors play a role in reorganizing chromatin architecture for differentiation genes, we performed HiChIP (Mumbach et al., 2016) for OCT4 and NANOG and called statistically significant interactions at a resolution of 10 kb, in range of 1 kb to 1 Mb, with at least one anchor overlapping with respective ChIP-seq peaks. Differential loops between hESC and DE or Meso were filtered based on P values ($P < 0.05$ for NANOG HiChIP, $P < 0.1$ for OCT4, and fold change > 1.5). We then found both NANOG and OCT4 mediated loops are mostly specific to cell fate specification, while the stable loops only account for less than 3% (Figures 3.7B, G). The means of NANOG looping distance is 97 kb in hESC and 12.9 kb in DE (Figures 3.7C), while the means of OCT4 looping distance is 14.7 kb in hESC and 13 kb in Meso (Figures 3.7F). The frequency of NANOG or OCT4 mediated looping distance is in agreement with normal short-range enhancer-promoter interaction distance (Corces et al., 2018b). The intersection between cell fate specific loop anchor and promoter and Active enhancer enriched binding event

indicate topological reorganization around these regions (Figure 3.3B, Figure 3.6). GATA6 for DE and HAND1 for Meso are exemplified (Figures 3.7D, H).

ARTIFICIAL TETHERING OF PLURIPOTENCY FACTOR AT CLOSED CHROMATIN IS SUFFICIENT TO ACTIVATE SMAD TARGET GENE

The ability of NANOG and OCT4 protein to directly activate SMAD target genes, independent of SMAD recruitment, would be a strong test of model that NANOG and OCT4 act to initiate enhancer activation for differentiation. To experimentally validate the role of NANOG and OCT4, we carried out artificial tethering test with dCas9-NANOG and dCas-OCT4 fusion protein (Figures 3.8A, B). Cells expressing dCas9 fusions were infected again with constructs expressing gRNAs individually targeting promoters and enhancers of GATA6 and HAND1. The gRNA constructs were also tagged with either GFP or RFP frame for cell population enrichment. Finally, I detected spontaneous expression of GATA6 using dCas9-NANOG with gRNA targeting its promoter and enhancer, under self-renewal culture condition (Figure 3.8C). Similarly, but to a much less extend, HAND1was significantly induced by dCas9-OCT4 setup (Figure 3.8C). The difference in gene activation could be due to absence of extra BMP ligands in culture medium.

KNOCKOUT OF PLURIPOTENCY FACTOR IMPAIRS SMAD BINDING TO UPOPENED CHROMATIN

We next investigated whether the loss of NANOG or OCT4 leads to defects during DE or Meso differentiation, respectively. I generated inducible knockout cell lines for NANOG and OCT4 by knocking in an Auxin-inducible degron (AID) frame before stop codon (Natsume et al., 2016)

(Supplemental Figure 4.9A). In the presence of 500 μM indole-3acetic acid (IAA), a natural auxin, endogenous NANOG or OCT4 proteins rapidly degraded in hours (Supplemental Figures 4.9 B, C). Using the conditional knockout cell line, I re-visited the NANOG-SMAD2/3 binding events at two enhancers of GATA6 (Figure 3.9A) by day 1 of DE differentiation. ChIP-qPCR results demonstrated that loss of NANOG binding after IAA treatment, accompanied by decreased indirect binding of SMAD2/3 (Figures 3.9B, C). In addition, significant loss of H3K27ac deposition at these two GATA6 enhancers were observed (Supplemental Figure 4.10), suggesting a failed initiation of enhancer activation. Similarly, upon IAA induced OCT4 knockout, indirect SMAD1 binding to HAND1 enhancers are significantly decreased by day 1 of Meso (Figures 3,9D-F). Together, these results suggest a critical role of pluripotency factors regulating SMAD binding to cis-elements and hereafter the enhancer activation.

To more globally test whether loss of NANOG or OCT4 can disrupt SMAD binding profile, H3K27ac deposition to enhancers, and gene expression profile, we prepared a cohort of ChIP-seq and RNA-seq samples across DE and Meso cell types using the two degron cell lines. The samples are pending sequencing. These data will provide more mechanistic insights for future works.

CONCLUSION AND DISCUSSION

In this work, we mapped the chromatin accessibility dynamics across hESC, DE, Meso, and Ecto, and identified putative active enhancers most closely related to differentiation based on the binary open-closed chromatin and H3K4me1/H3K27ac deposition. Remarkably, these putative active enhancers are occupied by lineage specific OSN-SAMD complex. We simplified the tier of lineage specific OSN-SMAD as NANOG-SMAD2/3 for DE, OCT4-SMAD1 for Meso, and

SOX2 alone for Ecto, given the most pronounced protein-protein interactions among hESC and three germ layers. Our work extends the understanding of core pluripotency factors as cell fate specifiers, in collaboration with SMAD. In short, OSN bind to their canonical motifs at distal elements for pluripotency related gene during self-renewal. Upon BMP4/Activin-A signaling cues, OCT4 and NANOG are recruited by SAMD1 and SMAD2/3 to enhancers enriched for canonical SBE, respectively. OCT4 and NANOG are capable of directly binding to unopened chromatin, allowing the interacted SMAD1 and SMAD2/3 to bind to enhancers in Meso and DE, respectively. In addition, OCT4 and NANOG mediate dynamic enhancer-promoter interactions in Meso and DE, respectively.

An attractive model for how OSN act to trigger differentiation is the pioneer factor model, the mechanism of which is widely accepted in iPSC reprogramming. Indeed, OSN, individually or in combination, may act as pioneer factors binding to unopened chromatin which later becomes an active enhancer (Figure 3.3B, and Figures 3.5E, F). Intriguingly, OCT4-SOX2 transiently bind to fibroblast cell specific distal elements at the beginning of reprogramming, reflecting a progress of OCT4-SOX2 redistribution from differentiation loci to pluripotency loci (Chronis et al., 2017; Li et al., 2017). It appears that the choice between self-renewal and differentiation is decided by a reversible OSN binding shift. A key issue to the understanding of such choice is how OSN decide when to bind their canonical motif and when to bind to other sequences. Here, I will address three hypotheses.

In the first hypothesis, since ectopically forced expression of OSN is commonly used for human iPSC reprogramming (Yu et al., 2007), it supports the model of passive TF competition in chromatin remodeling (Figure 1.3A). In context of OSN-SMAD complex, a higher protein concentration of OSN favors their binding to canonical OSN motif in hESC. Conversely, higher

concentration of phosphorylated R-SMAD induced by BMP/Activin favors the SMAD binding to SMAD motif in Meso and DE, respectively (Figure 3.2D, Figure 3.3B, Supplemental Figure 4.6B). The balance shift seems to be determined by ratio of protein levels between each factor of OSN-SMAD. Given the mutual interactions between OSN and phosphorylated R-SMAD, the lineage specific OSN-SMAD complex is redistributed depending on the factor with dominating concentration ratio. As for Ecto, dual inhibition of SMAD leads to blocked expression of O, N, and de-phosphorylation of R-SMADs, S then is the sole factor out of OSN-SMAD complex, relocating to the pan-motifs for SOX family (Supplemental Figure 4.3, Supplemental 4.5). In traditional model of pluripotency, the protein levels of OSN needs to be kept within a subtle range to maintain self-renewal while preventing unwanted differentiation (see Chapter 1). These observations may be a result of intra complex interaction dynamics, evident by the fact that OSN each has an interaction preference with either SMAD1 or SMAD2/3 dependent on SMAD phosphorylation state (Supplemental Figure 4.6B and Supplemental Figure 4.7). To test this hypothesis, we can synthesize DNA oligos containing either one of the motifs for O, S, N, SMAD and titrate protein concentration for each factor during incubation *in vitro*. The difference in a banding shift assay will provide a mechanistic insight.

A second hypothesis is based on the interplay of trans-regulation mechanism (Figure 1.3C). We noted that during mitosis, interactions between OSN-SMAD are mostly observed outside of nucleus (Figure 3.4C), in concurrent with loss of chromosome compartments and TADs in metaphase (Naumova et al., 2013). One important question is how do OSN-SMAD re-entry the interphase precisely in two daughter cells and respond to differentiation cues quickly in late G1(Singh et al., 2015)? We show that OSN-SMAD are promiscuously enriched at promoter-proximal elements for both pluripotency and differentiation genes in self-renewal state

(Supplemental Figure 4.8). The promoter-proximal elements are known to be occupied by Mediator, cohesion, RNA PolII, of which Mediator is reported to form phase-separated condensates with OSN in ESCs (Boija et al., 2018). A recent study proposed a new enhancer-promoter regulation model that Mediator functions as a TF reservoir which diffuses active TFs to proximal regulatory loci within cohesion mediated loops (El Khattabi et al., 2019). Consistently, we show that SMC1, core component of cohesion, is overlapping with O, N in respective cell type at active enhancers where O, N mediated enhancer-promoter loops are formed (Figures 3.7A, C). Together, the results suggest a portion of OSN proteins are kept in a Mediator reservoir along the genome, possibly relying on the phase condensation property, which provides alternative explanation for canonical motif independent binding profile of OSN. We can test this hypothesis by knocking down Mediator and see whether OSN pre-marking at promoter-proximal regions are affected.

The third hypothesis is inspired by a previous work in our lab. Cliff et al. demonstrated a reducing-oxidative metabolic shift between hESC and DE/Meso cells (Cliff et al., 2017). Moreover, it is reported that key cysteine residues of OCT4 are subject to oxidation and disulfide bond formation (Marsboom et al., 2016). After evaluating the crystal structure of OCT4-DNA complex (Michael et al., 2020), I identified two candidate cysteine residues inside DNA binding domain. If a disulfide bond is formed under oxidation, the distance between the two cysteine residues is predicted to decrease from 18.9 Å to 3 Å, resulting in conformation changes at DNA binding surface. Therefore, the switch of cellular metabolic environment could possibly affect OCT4 binding to its canonical motif. This hypothesis can be validated by performing OCT4 ChIP experiments under reducing vs oxidating cellular environment. The underlying mechanism

regulating OSN redistribution may consist of either one, or mixed or additional biological process discussed above.

Our current work presents a simplified model to show OSN function as cell fate specifiers in response to SMAD regulated differentiation. However, there are several critical questions remain puzzled. First, Wnt is a shared signaling between DE and Meso differentiation, we omit Wnt signaling from the current model. However, we noticed that Wnt signaling can augment the effect of NANOG redistribution under minimal SMAD2/3 activation, whereas Wnt alone failed to induce DE differentiation (data not shown). One of major Wnt effectors, TCF4, also overlaps with OCT4-SMAD1 binding sites in Meso (Figure 3.3A). What are other Wnt effectors involved in OSN redistribution and SMAD binding to active enhancers? Do those Wnt effectors exhibit cell type specific binding property? What are other co-factors involved in diverging OSN-SMAD complex in respective cell fate? Such questions need further investigation using proteomics approaches.

Next, stepwise enhancer activation is a critical issue to understand stem cell differentiation. We showed that SMAD2/3 activation induced NANOG binding to one of the putative enhancers of GATA6 within one day (Figure 3.4D). However, there are two candidate loci devoid of increased NANOG binding (data not shown). These results indicate that enhancer may undergo transient activation or inactivation, or time dependent activation mechanism. Similar phenomenon is observed during human iPSC reprogramming and endoderm differentiation (Chronis et al., 2017; Li et al., 2017; Wang et al., 2015). The temporal difference in enhancer activation prompt us to speculate alternative parameters besides chromatin states.

In addition, CTCF motif is uniquely enriched in hESC-Poised loci (Supplemental Figure 4.3). Global mapping of CTCF and SMC1 ChIP-seq data revealed intersection of the two

architectural proteins at those loci (data not shown). The presence of both CTCF and SMC1 indicate the barrier of TAD domain shaping the neighbor gene regulation, rather than loop anchors for enhancers (Dowen et al., 2014). The results suggest a unique chromatin organization signature for self-renewal. Further investigation is needed to better understand the unknown characteristic of pluripotency.

Last but not least, the proposed model of pluripotency factors as cell fate specifiers could be a general principle by which pre-existing pioneer factors act in conjunction with cell type specific signaling effectors to initiate chromatin re-organization. DNA binding profiles of TEAD1 and OTX2 were investigated superficially in this study (data not shown). These TFs can also bind to DNA in both motif dependent and independent manners, in Meso and DE, respectively. Further analysis of their binding in later differentiation stages are needed to support our guess.

One caveat in this study is that proximity ligation assay only reflects averaged counts among OSN-SMAD interactions. To provide more convincing evidence, we need results showing direct OSN-SMAD interactions at validated locus, for example GATA6 in DE. A hybrid technique of proximity ligation assay and *in situ* hybridization is required to test the context of OSN-SMAD complex at defined anchors of promoter-enhancer loops. Additionally, we want to confirm our model is true at a physiological level. Due to the limitation of use of human embryo, we analyzed the chromatin accessibility data during mouse embryo gastrulation (Xiang et al., 2020) and identified candidate loci to test our model in mouse counterpart (data not shown). Further validation in mouse embryo sections would be valuable.

Figure 3.1

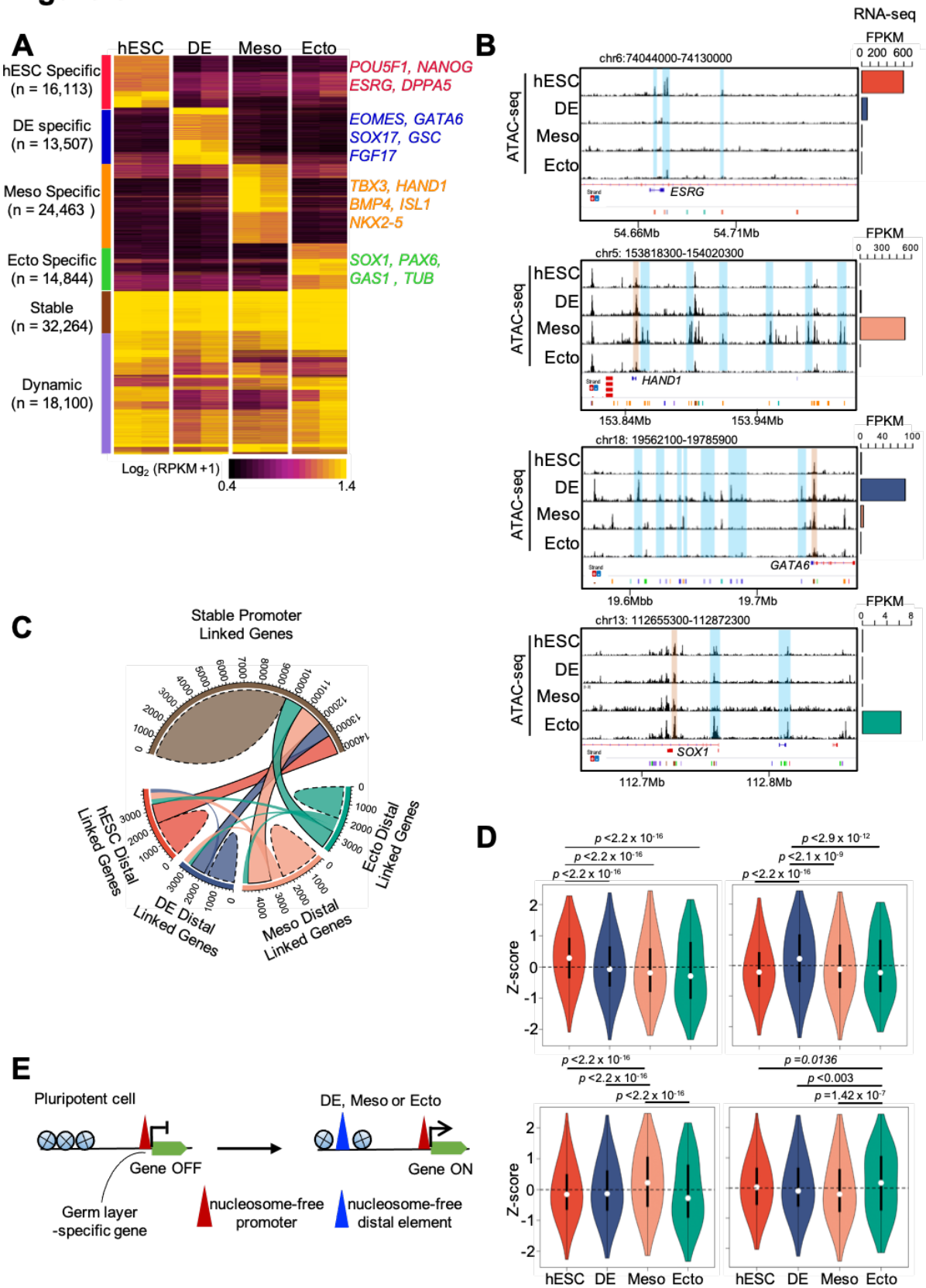

A
hESC
DE
Meso
Ecto
hESC Specific (n = 16,113)
DE specific (n = 13,507)
Meso Specific (n = 24,463)
Ecto Specific (n = 14,844)
Stable (n = 32,264)
Dynamic (n = 18,100)
POU5F1, NANOG
ESRG, DPPA5
EOMES, GATA6
SOX17, GSC
FGF17
TBX3, HAND1
BMP4, ISL1
NKX2-5
SOX1, PAX6,
GAS1 , TUB
Log2 (RPKM +1)
0.4
1.4
B
RNA-seq
FPKM
chr6:74044000-74130000
ATAC-seq
hESC
DE
Meso
Ecto
ESRG
54.66Mb
54.71Mb
chr5: 153818300-154020300
HAND1
153.84Mb
153.94Mb
chr18: 19562100-19785900
GATA6
19.6Mbb
19.7Mb
chr13: 112655300-112872300
SOX1
112.7Mb
112.8Mb
C
Stable Promoter Linked Genes
hESC Distal Linked Genes
DE Distal Linked Genes
Meso Distal Linked Genes
Ecto Distal Linked Genes
D
Z-score
p <2.2 x 10-16
p <2.9 x 10-12
p <2.1 x 10-9
p =0.0136
p <0.003
p =1.42 x 10-7
hESC DE Meso Ecto
E
Pluripotent cell
DE, Meso or Ecto
Gene OFF
Gene ON
Germ layer -specific gene
nucleosome-free promoter
nucleosome-free distal element

Figure 3.1 - Distal elements shape transcription profiles.

(A) Loci of open chromatin among different cell types were categorized into groups based on normalized ATAC-Seq signal intensity. Units were CPM values within 250bp flanking peak summits in log2 transformation.

(B) Representative examples of ATAC-Seq data in defined category from C1 to C4: ESGR (chr3: 54630000-54777000), GATA6 (chr18: 19562100-19785900, HAND1 (chr5: 153818300-154020300), SOX1 (chr13: 112655300-112872300). All genomic views are to the same height scale (0 - 25) in IGV browser. CTS loci are highlighted in blue boxes while promoters with Stable ATAC signals are highlighted in brown boxes. The RNA-Seq expression levels of representative genes were attached to the right. RNA-Seq expression units are in FPKM. Previously defined CTS ATAC peaks are color coded by a RGB track in the bottom.

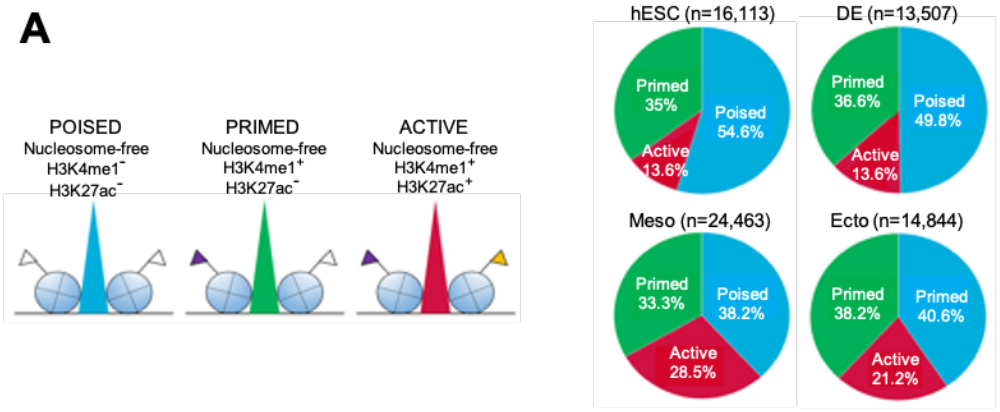
(C) Chord diagram of the intersection of gene numbers among genes linked by each ATAC-seq cluster. Promoter peaks are defined as occurring between -1kb and +1kb flanking a transcription start site (TSS). Distal peaks in this analysis are those within 10kb of a TSS and non-promoter peaks. The self-linked area in each sector represents the total number of genes linked by respective cluster of ATAC peaks.

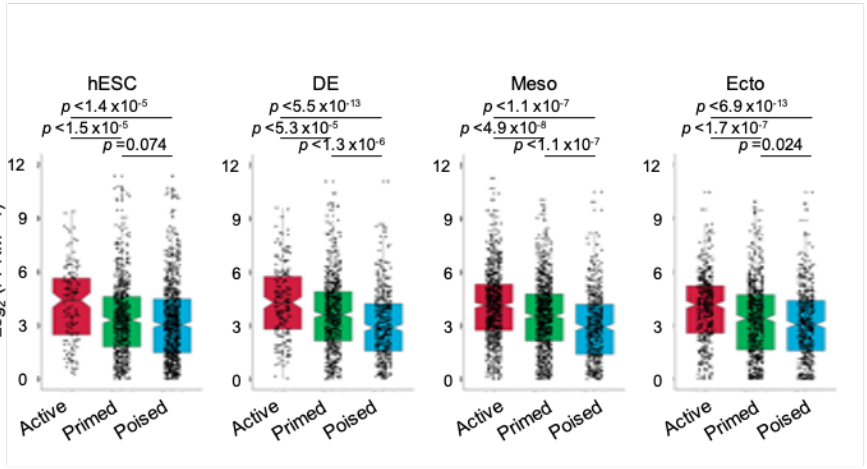
(D) Violin plots of the expression level of genes linked to distal ATAC peaks within 10kb of a nearest TSS. RNA-Seq data were converted to a row-wise Z score based on SD. *** means p value < 0.001, ** means p value < 0.01 using Mann-Whitney U test versus other cell types in each category.

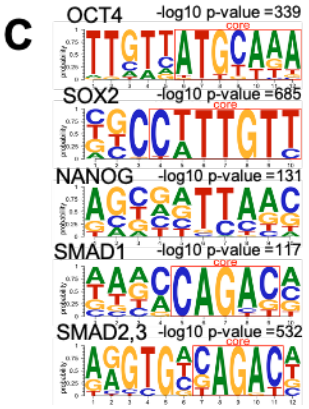
(E) Schematic of the accessibility dynamics during human ES cell differentiation. The developmental genes were depleted of nucleosomes at promoters and primed for expression.

Upon signaling cues and specific TF(s) binding, the distal *cis*-elements were “opened” and activate differentiation gene transcription

Figure 3.2


A
POISED
Nucleosome-free
H3K4me1-
H3K27ac-
PRIMED
Nucleosome-free
H3K4me1+
H3K27ac-
ACTIVE
Nucleosome-free
H3K4me1+
H3K27ac+
hESC (n=16,113)
Primed 35%
Poised 54.6%
Active 13.6%
DE (n=13,507)
Primed 36.6%
Poised 49.8%
Active 13.6%
Meso (n=24,463)
Primed 33.3%
Poised 38.2%
Active 28.5%
Ecto (n=14,844)
Primed 38.2%
Primed 40.6%
Active 21.2%


B
hESC
p <1.4 x10-5
p <1.5 x10-5
p =0.074
DE
p <5.5 x10-13
p <5.3 x10-5
p <1.3 x10-6
Meso
p <1.1 x10-7
p <4.9 x10-8
p <1.1 x10-7
Ecto
p <6.9 x10-13
p <1.7 x10-7
p =0.024
Log2 (FPKM+1)
0
3
6
9
12
Active
Primed
Poised


C
OCT4 -log10 p-value =339
SOX2 -log10 p-value =685
NANOG -log10 p-value =131
SMAD1 -log10 p-value =117
SMAD2,3 -log10 p-value =532

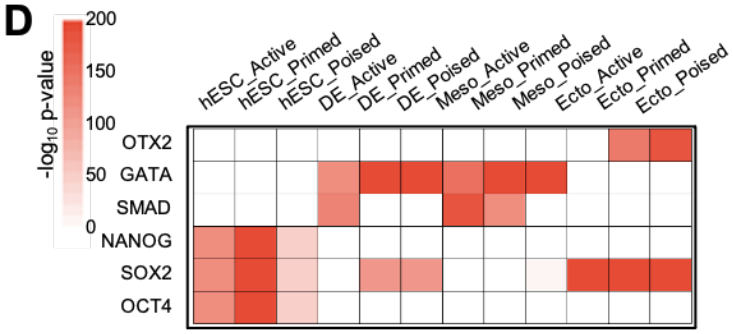

D
-log10 p-value
200
150
100
50
0
hESC_Active
hESC_Primed
hESC_Poised
DE_Active
DE_Primed
DE_Poised
Meso_Active
Meso_Primed
Meso_Poised
Ecto_Active
Ecto_Primed
Ecto_Poised
OTX2
GATA
SMAD
NANOG
SOX2
OCT4

Figure 3.2 – Motifs for important TFs mirror enhancer activity

(A) Classification of the Cell Type Specific (CTS) open chromatin loci based on binary presence of histone modification markers. See Supplemental Figure 4.2 for details.

(B) Box plot of the expression level of genes linked to sub-categorized ATAC peaks within 10kb of a nearest TSS. RNA-Seq readouts were FPKM values. P values were calculated using Wilcoxon test.

(C) *de novo* motifs generated from in-house ChIP-seq data.

(D) Representative motif enrichment in the CTS open chromatin loci, ranked by Active, Primed, and Poised chromatin epigenomic state. See Supplemental Figure 4.3 for a complete motif discovery list.

Figure 3.3

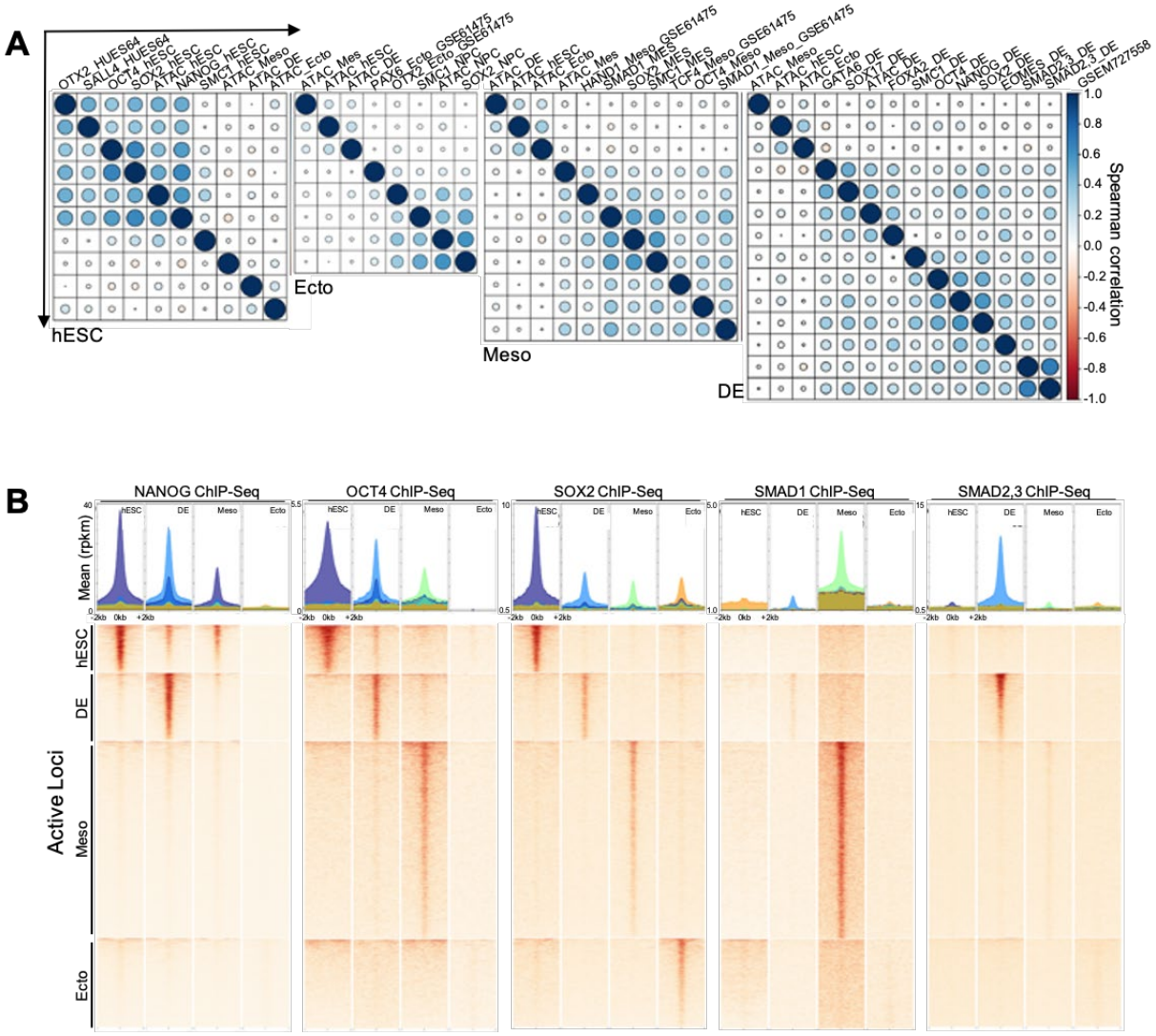

A
hESC
Ecto
Meso
DE
Spearman correlation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
B
NANOG ChIP-Seq
OCT4 ChIP-Seq
SOX2 ChIP-Seq
SMAD1 ChIP-Seq
SMAD2,3 ChIP-Seq
Mean (rpkm)
hESC
DE
Meso
Ecto
Active Loci
-2kb
0kb
+2kb

Figure 3.3 – Core pluripotency factors predominantly bind to SMAD targeted active enhancers for differentiated cells.

(A) Hierarchical clustering of selected TF ChIP-seq data at Active enhancer loci in respective cell type. The selection is based on significant clustering of ChIP-seq data with corresponding ATAC-seq data in the same cell type. Data were clustered based on a Euclidean distance matrix and complete linkage using Spearman correlation.

(B) Heatmap of NANOG, OCT4, SOX2, SMAD1, and SMAD2/3 over the Active enhancers across hESC, DE, Meso, and Ecto. For complete clustering of ChIP-seq and ATAC-seq signals, see Supplemental Figure 4.5

Figure 3.4

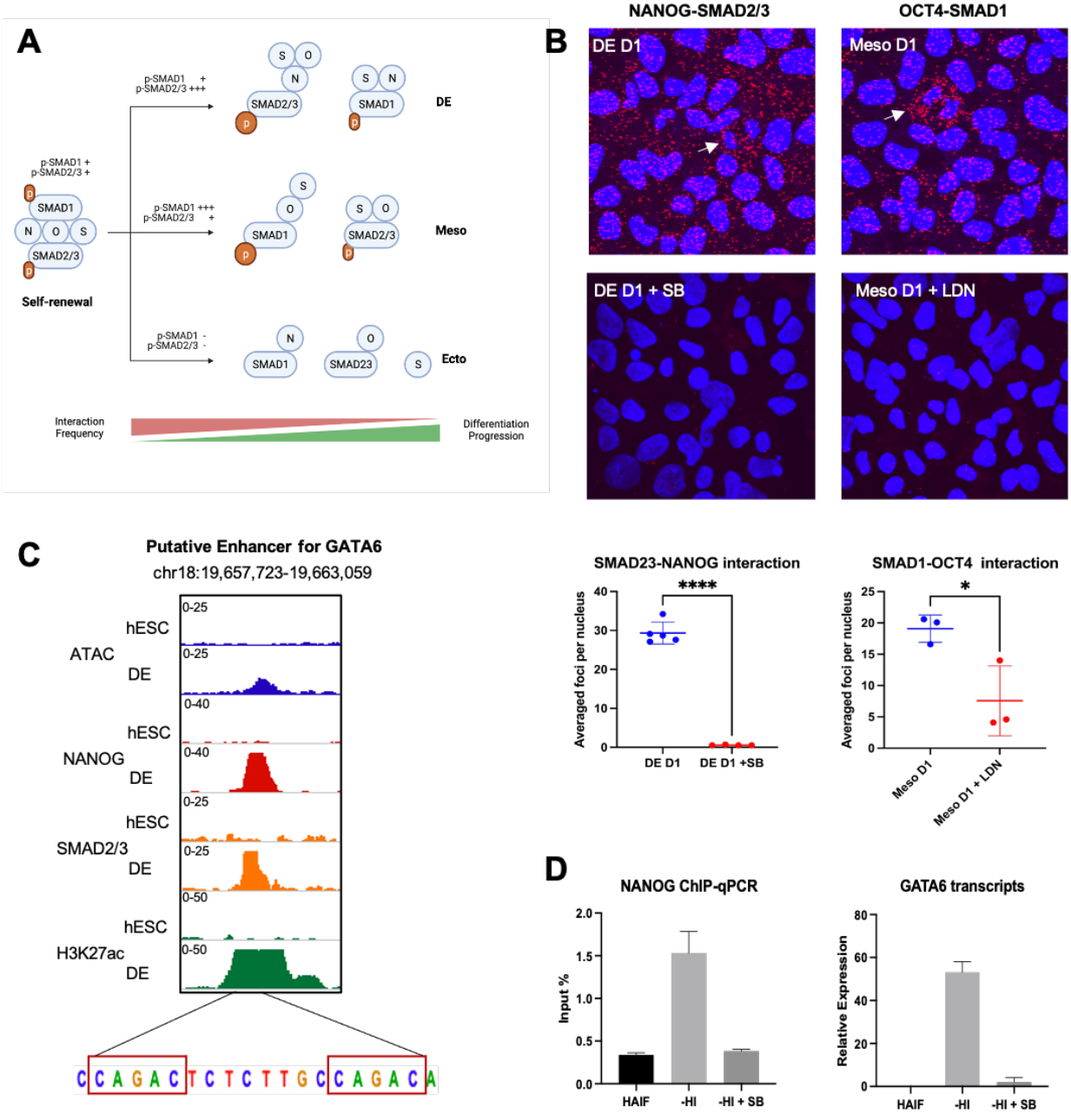

A
p-SMAD1 +
p-SMAD2/3 +++
DE
p-SMAD1 +
p-SMAD2/3 +
SMAD1
N O S
SMAD2/3
Self-renewal
p-SMAD1 +++
p-SMAD2/3 +
Meso
p-SMAD1 -
p-SMAD2/3 -
SMAD23
Ecto
Interaction Frequency
Differentiation Progression
B
NANOG-SMAD2/3
OCT4-SMAD1
DE D1
Meso D1
DE D1 + SB
Meso D1 + LDN
C
Putative Enhancer for GATA6
chr18:19,657,723-19,663,059
ATAC
NANOG
SMAD2/3
H3K27ac
hESC
DE
0-25
0-40
0-50
C C A G A C T C T C T T G C C A G A C A
SMAD23-NANOG interaction
Averaged foci per nucleus
****
DE D1
DE D1 +SB
SMAD1-OCT4 interaction
*
Meso D1
Meso D1 + LDN
D
NANOG ChIP-qPCR
Input %
HAIF
-HI
-HI + SB
GATA6 transcripts
Relative Expression

Figure 3.4 – Coordinated OSN-SMAD interaction and redistribution are dependent of SMAD phosphorylation.

(A) Cartoon summary of pairwise proximity ligation assay (PLA) from hESC differentiation into DE, Meso, and Ecto. In the cartoon, the intersection between two factors means that they have significant contact in respective cell type than the others. For details, see Supplemental Figures 4.6 and 4.7.

(B) PLA of NANOG-SMAD2/3 and OCT4-SMAD1 under SB and LDN treatment, respectively. Results were taken with a Zeiss710 confocal microscope. Each data point represents an imaging field containing around 20 – 50 cells. The PLA foci were counted if inside nucleus. Automatic foci counting was conducted using customized ImageJ script. The number of foci was quantified. The white arrow headed cells are during mitosis.

(C) An example of putative active enhancer for GATA6. This locus contains two canonical SBE sites.

(D) ChIP-qPCR of NANOG at putative GATA6 enhancer locus. NANOG binding event was detected after 24 hours of minimal DE induction with or without SB inhibitor. Relative GATA6 expression was measured accordingly.

Figure 3.5

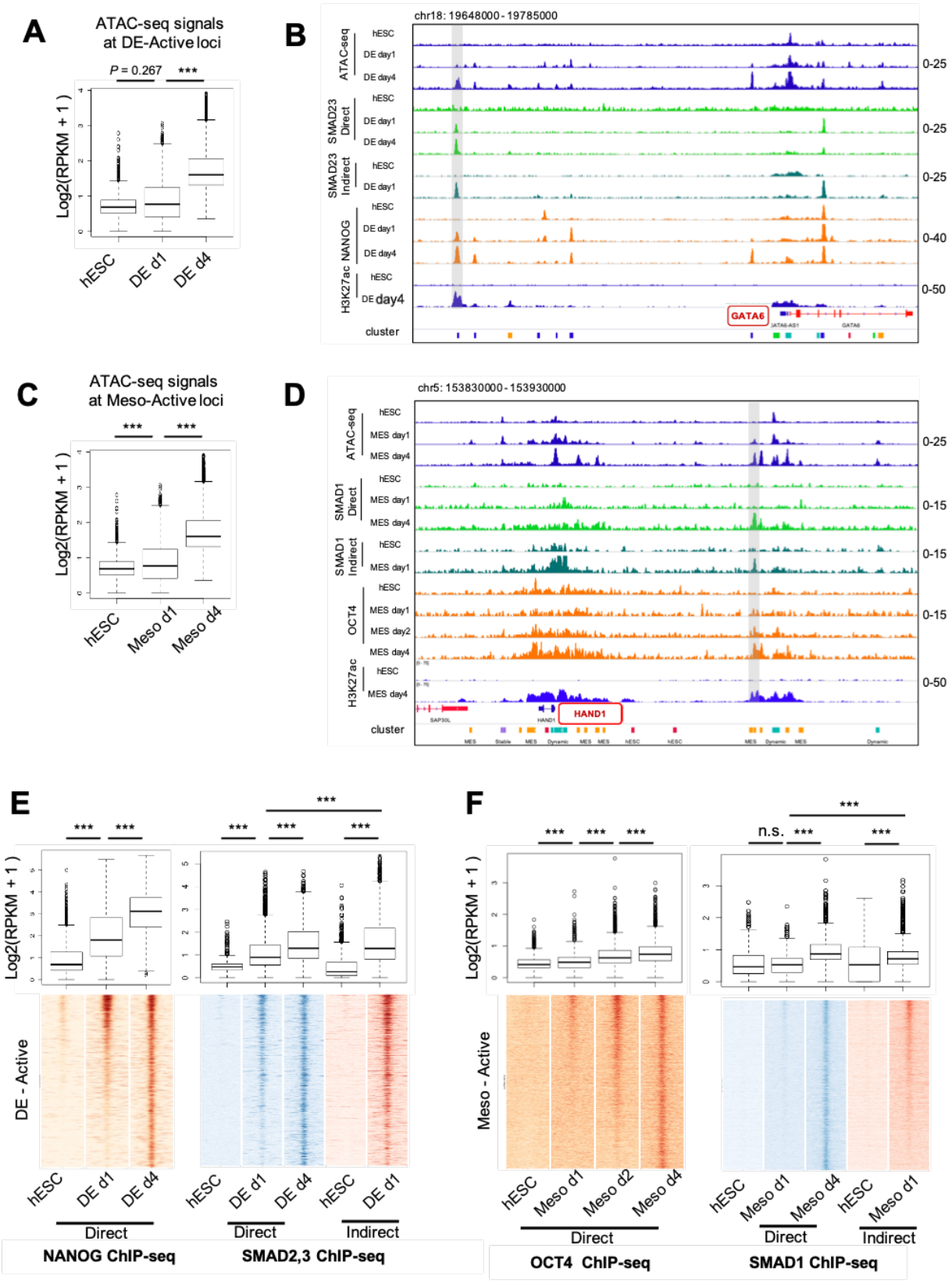

A
ATAC-seq signals at DE-Active loci
P = 0.267
***
Log2(RPKM + 1)
hESC
DE d1
DE d4
B
chr18: 19648000 - 19785000
ATAC-seq
SMAD23 Direct
SMAD23 Indirect
NANOG
H3K27ac
hESC
DE day1
DE day4
0-25
0-40
0-50
GATA6
cluster
C
ATAC-seq signals at Meso-Active loci
Log2(RPKM + 1)
hESC
Meso d1
Meso d4
D
chr5: 153830000 - 153930000
ATAC-seq
SMAD1 Direct
SMAD1 Indirect
OCT4
H3K27ac
MES day1
MES day2
MES day4
0-15
HAND1
cluster
E
Log2(RPKM + 1)
DE - Active
hESC
DE d1
DE d4
Direct
Indirect
NANOG ChIP-seq
SMAD2,3 ChIP-seq
F
n.s.
Meso - Active
Meso d1
Meso d2
Meso d4
OCT4 ChIP-seq
SMAD1 ChIP-seq

Figure 3.5 (continued)

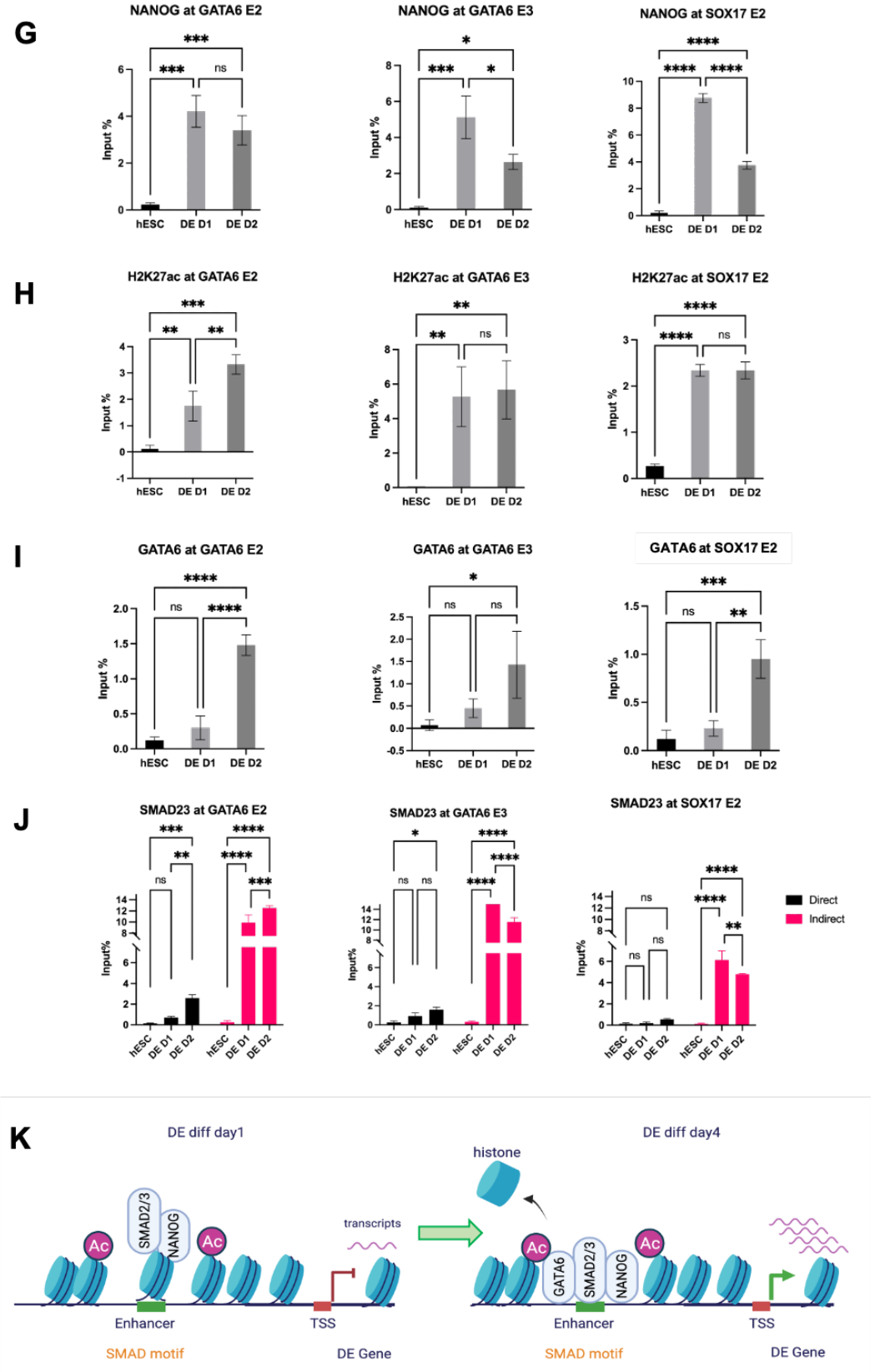


Figure 3.5 – OCT4 and NANOG bind to chromatinized DNA with much higher affinity than SMAD.

(A) boxplot of normalized ATAC-seq signals in DE differentiation at D0, D1, and D4.

(B) Genome view of the ATAC-seq and ChIP-seq data for GATA6. The enhancer of GATA6 is highlighted in shaded gray area.

(C) boxplot of normalized ATAC-seq signals in Meso differentiation at D0, D1, and D4.

(D) Genome view of the ATAC-seq and ChIP-seq data for HAND1. The enhancer of HAND1 is highlighted in shaded gray area.

(E) Heatmap of NANOG, direct SMAD2/3, and indirect SMAD2/3 ChIP-seq data in time course. P values were calculated using Wilcoxon test. *** means

(F) Heatmap of OCT4, direct SMAD1, and indirect SMAD1 ChIP-seq data in time course. P values were calculated using Wilcoxon test. *** means p value < 0.001.

(G-J) ChIP-qPCR for NANOG, H3K27ac, GATA6, and SMAD2/3 at three previously investigated enhancers which are chromatin open in DE cells. For SMAD2/3 ChIP, both direct ChIP and indirect ChIP using EGS dual crosslinking were conducted. In (G-I), p values were calculated using ordinary one-way ANOVA, **** means p value < 0.0001. In (J), p values were calculated using two-way ANOVA, **** means p value < 0.0001.

(K) Model scheme showing NANOG directly binds to nucleosome at enhancers for DE fate, whereas SMAD2/3 is incompetent for direct DNA contact due to nucleosome protection by D1 of DE.

Figure 3.6

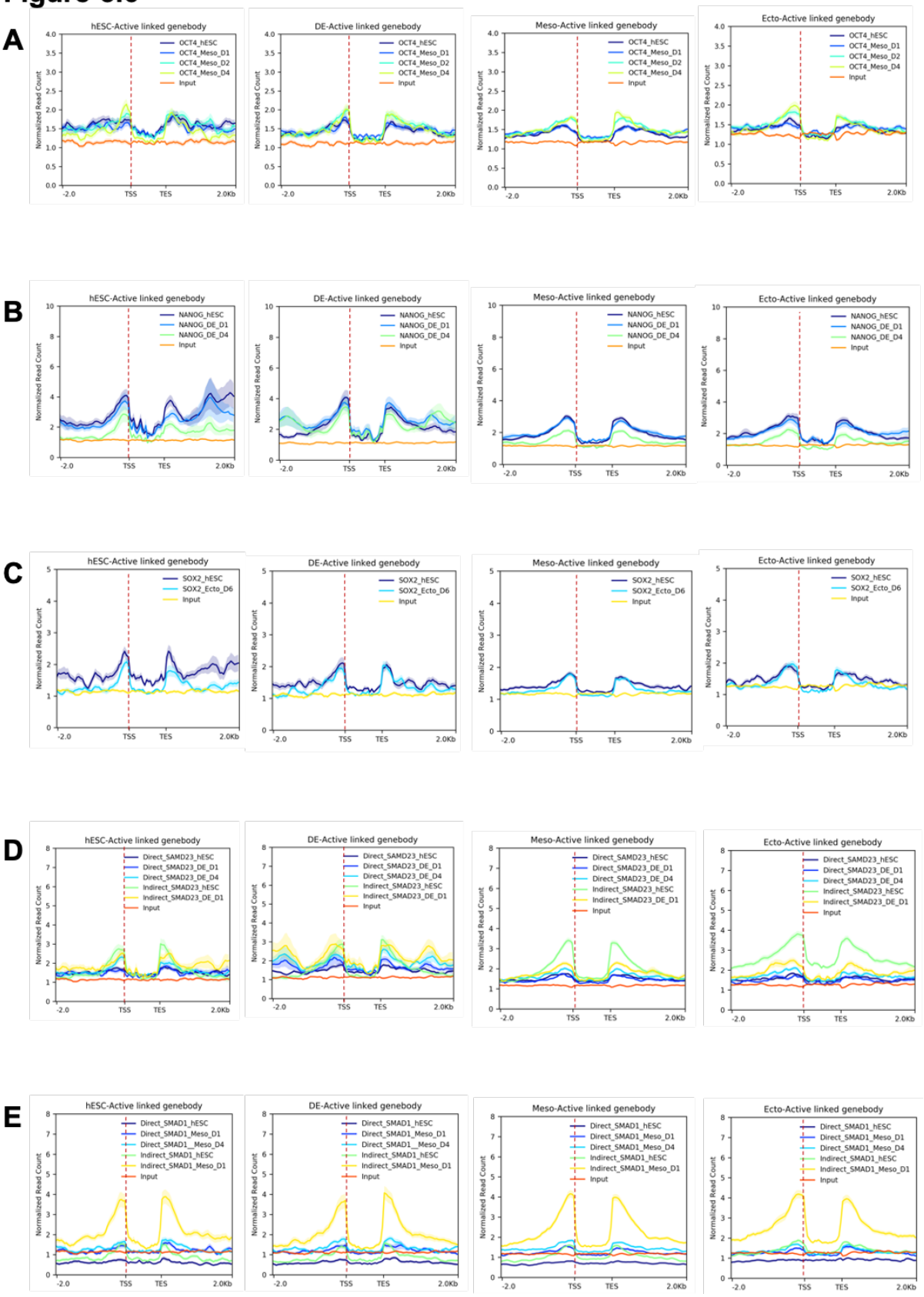

A
hESC-Active linked genebody
DE-Active linked genebody
Meso-Active linked genebody
Ecto-Active linked genebody
OCT4_hESC
OCT4_Meso_D1
OCT4_Meso_D2
OCT4_Meso_D4
Input
B
NANOG_hESC
NANOG_DE_D1
NANOG_DE_D4
Input
C
SOX2_hESC
SOX2_Ecto_D6
Input
D
Direct_SAMD23_hESC
Direct_SMAD23_DE_D1
Direct_SMAD23_DE_D4
Indirect_SMAD23_hESC
Indirect_SMAD23_DE_D1
Input
E
Direct_SMAD1_hESC
Direct_SMAD1_Meso_D1
Direct_SMAD1__Meso_D4
Indirect_SMAD1_hESC
Indirect_SMAD1_Meso_D1
Input
Normalized Read Count
-2.0
TSS
TES
2.0Kb

Figure 3.6 – OSN-SMAD2/3 pre-mark promoter proximal elements of differentiation genes.

(A) OCT4 ChIP-seq data from hESC, Meso D1, D2, and D4 were normalized and mapped along gene bodies, which were linked by Active enhancers.

(B) NANOG ChIP-seq data from hESC, DE D1, and D4 were normalized and mapped along gene bodies, which were linked by Active enhancers.

(C) SOX2 ChIP-seq data from hESC, Ecto D1, and D6 were normalized and mapped along gene bodies, which were linked by Active enhancers.

(D) SMAD2/3 ChIP-seq data from hESC, DE D1, and D4 were normalized and mapped along gene bodies, which were linked by Active enhancers.

(E) SMAD1 ChIP-seq data from hESC, Meso D1, and D4 were normalized and mapped along gene bodies, which were linked by Active enhancers.

Figure 3.7

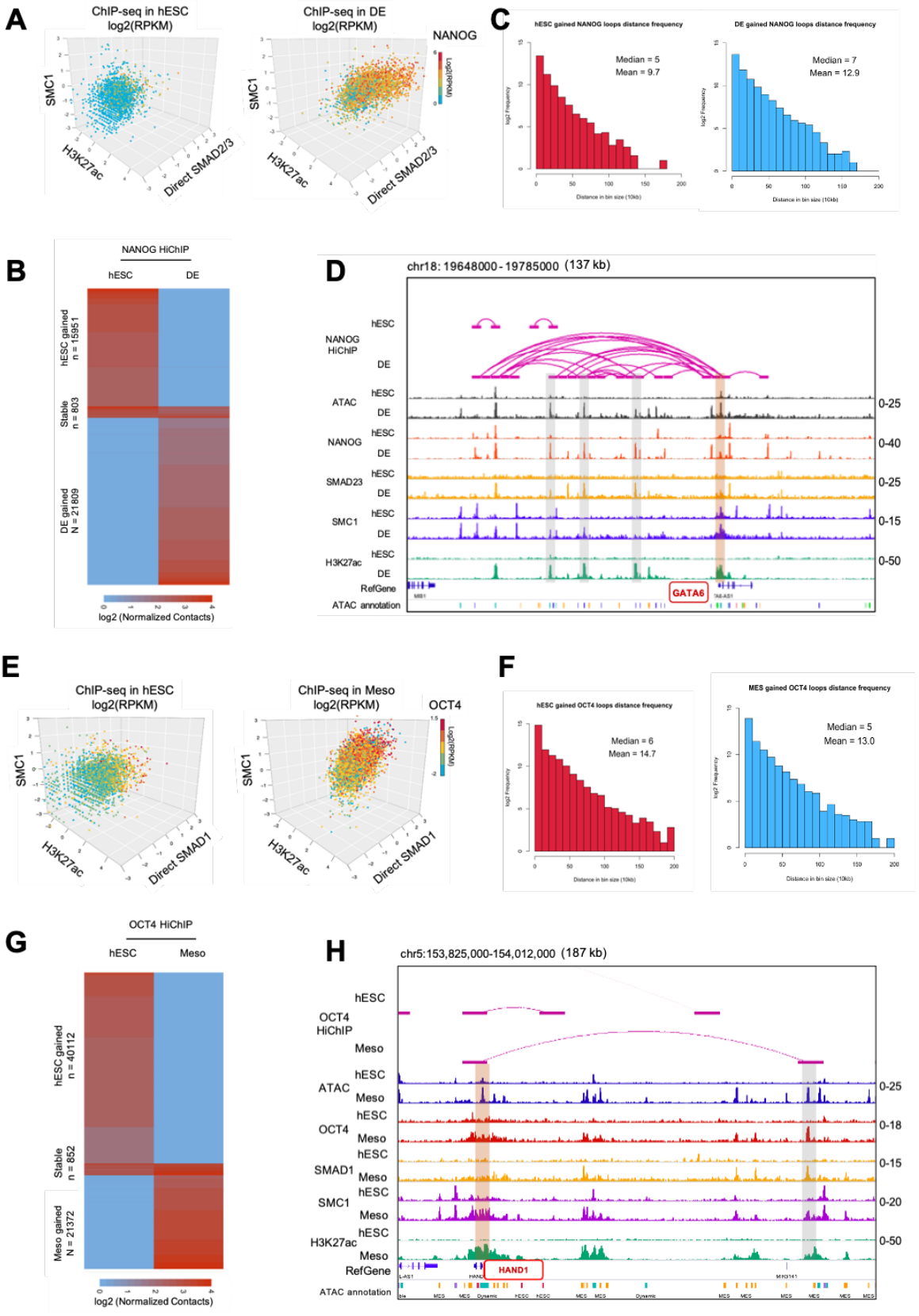

A
ChIP-seq in hESC
log2(RPKM)
ChIP-seq in DE
log2(RPKM)
NANOG
SMC1
H3K27ac
Direct SMAD2/3
B
NANOG HiChIP
hESC
DE
hESC gained
n = 15951
Stable
n = 803
DE gained
N = 21809
log2 (Normalized Contacts)
C
hESC gained NANOG loops distance frequency
Median = 5
Mean = 9.7
DE gained NANOG loops distance frequency
Median = 7
Mean = 12.9
D
chr18: 19648000 - 19785000 (137 kb)
NANOG HiChIP
ATAC
NANOG
SMAD23
SMC1
H3K27ac
RefGene
ATAC annotation
GATA6
0-25
0-40
0-25
0-15
0-50
E
ChIP-seq in hESC
log2(RPKM)
ChIP-seq in Meso
log2(RPKM)
OCT4
Direct SMAD1
F
hESC gained OCT4 loops distance frequency
Median = 6
Mean = 14.7
MES gained OCT4 loops distance frequency
Median = 5
Mean = 13.0
G
OCT4 HiChIP
hESC
Meso
hESC gained
n = 40112
Stable
n = 852
Meso gained
N = 21372
log2 (Normalized Contacts)
H
chr5:153,825,000-154,012,000 (187 kb)
OCT4 HiChIP
ATAC
OCT4
SMAD1
SMC1
H3K27ac
RefGene
ATAC annotation
HAND1
0-25
0-18
0-15
0-20
0-50

Figure 3.7 – Pluripotency factors mediate unique enhancer-promoter interactions in differentiated cells.

(A) Concordant mapping of NANOG, SMC1, SMAD2/3, and H3K27ac ChIP-seq signals over DE-Active open chromatin loci.

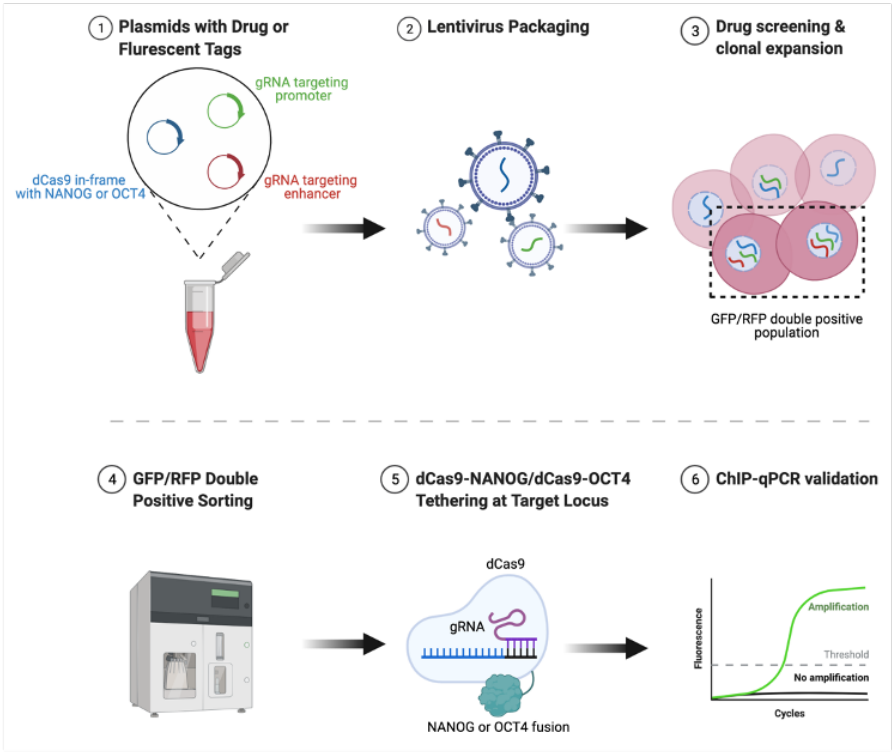
(B) Heat map of the differential NANOG HiChIP contacts during hESC differentiation to DE.

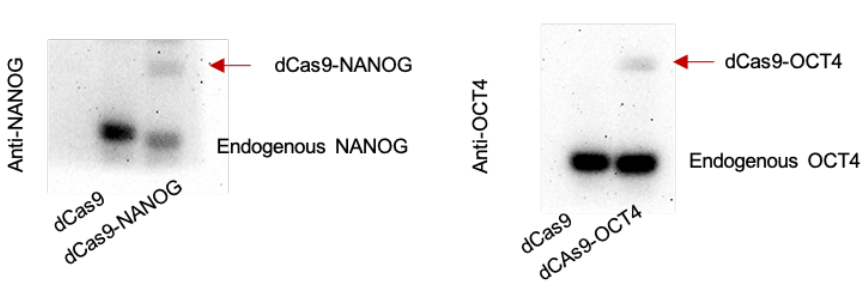
(C) Frequency of distance between NANOG loop anchors.

(D) Genome view of NANOG loops mediating enhancer-promoter interactions at GATA6 locus. Loop resolution is 10 kb. The Active enhancers overlapped with NANOG loop anchors are highlighted in the gray areas. The promoter of GATA6 is highlight in the brown area.

(E-H) OCT4 counterparts in hESC and Meso.

Figure 3.8


A
1 Plasmids with Drug or Flurescent Tags
2 Lentivirus Packaging
3 Drug screening & clonal expansion
gRNA targeting promoter
dCas9 in-frame with NANOG or OCT4
gRNA targeting enhancer
GFP/RFP double positive population
4 GFP/RFP Double Positive Sorting
5 dCas9-NANOG/dCas9-OCT4 Tethering at Target Locus
6 ChIP-qPCR validation
dCas9
gRNA
NANOG or OCT4 fusion
Amplification
Threshold
No amplification
Fluorescence
Cycles


B
Anti-NANOG
dCas9-NANOG
Endogenous NANOG
dCas9
dCas9-NANOG
Anti-OCT4
dCas9-OCT4
Endogenous OCT4
dCas9
dCAs9-OCT4

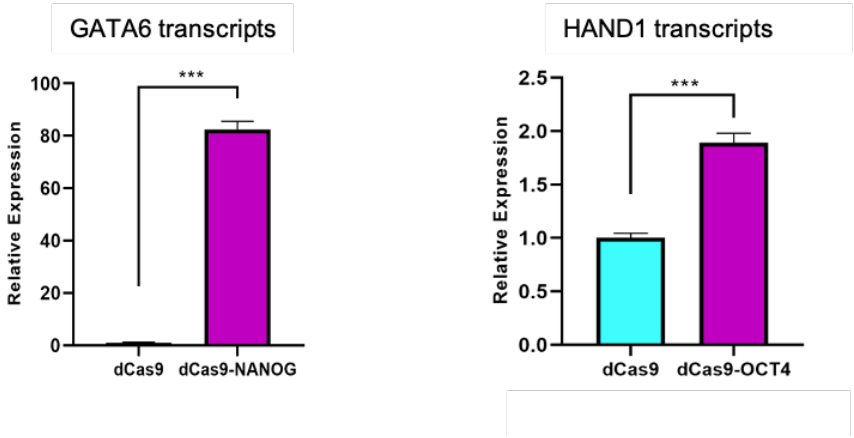
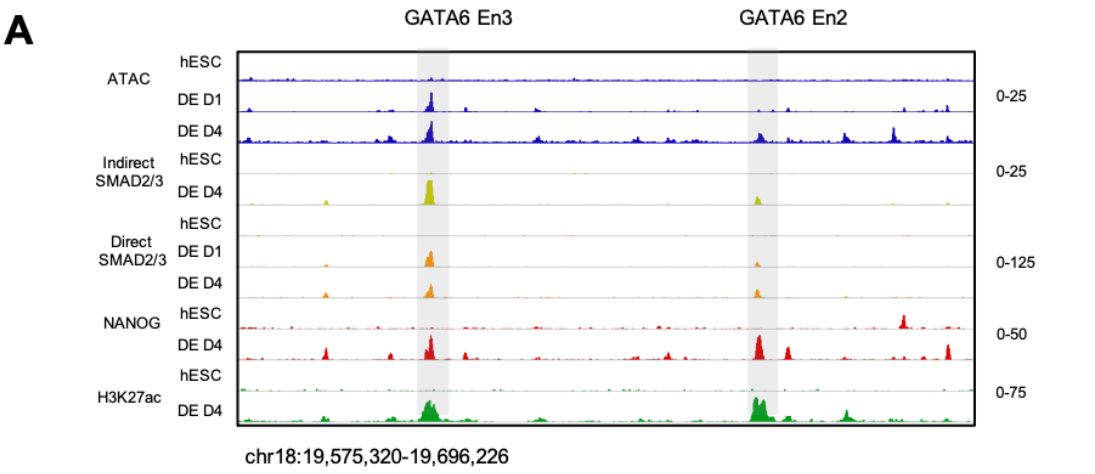

C
GATA6 transcripts
Relative Expression
100
80
60
40
20
0
***
dCas9
dCas9-NANOG
HAND1 transcripts
Relative Expression
2.5
2.0
1.5
1.0
0.5
0.0
***
dCas9
dCas9-OCT4

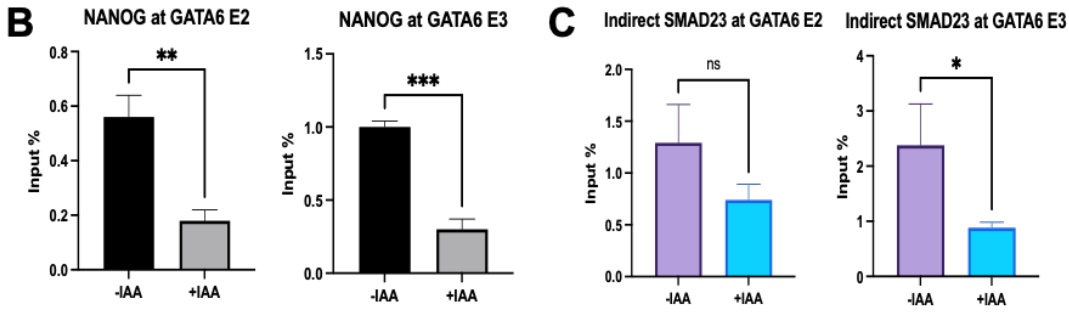
Figure 3.8 – Tethering of OCT4 or NANOG alone at closed enhancer is sufficient to activate SMAD target gene expression.

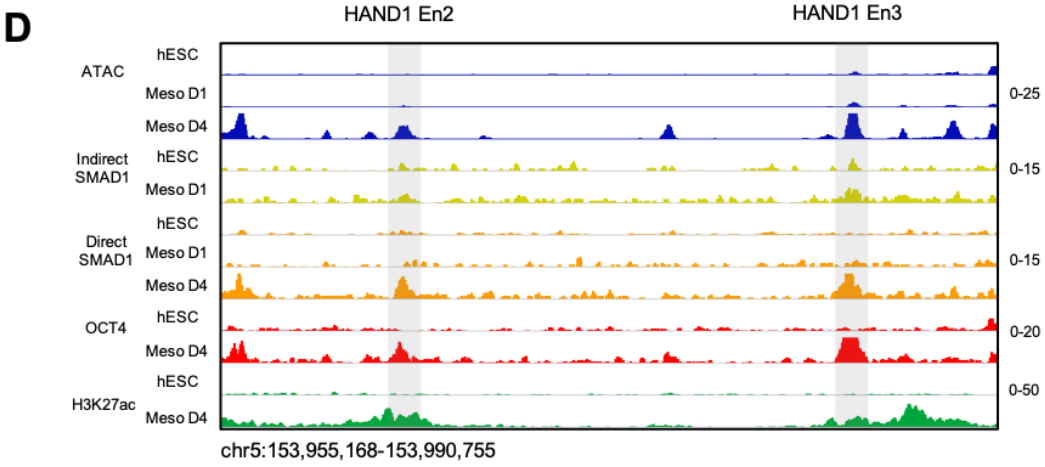
(A) Experimental design of dCas9 fusion protein tethering.

(B) Western blot results showing that H9-hESC transduced with lentivirus to stably express dCas9 and dCas9 fusion with either NANOG or OCT4.

(C) Relative expression of target genes in dCas9 fusion cell line expressing gRNAs targeting promoter and enhancer. Control cell lines express dCas9 alone.


A
GATA6 En3
GATA6 En2
ATAC
hESC
DE D1
DE D4
Indirect SMAD2/3
hESC
DE D4
Direct SMAD2/3
hESC
DE D1
DE D4
NANOG
hESC
DE D4
H3K27ac
hESC
DE D4
0-25
0-25
0-125
0-50
0-75
chr18:19,575,320-19,696,226


B
NANOG at GATA6 E2
NANOG at GATA6 E3
C
Indirect SMAD23 at GATA6 E2
Indirect SMAD23 at GATA6 E3
Input %
-IAA
+IAA
**
***
ns
*


D
HAND1 En2
HAND1 En3
ATAC
hESC
Meso D1
Meso D4
Indirect SMAD1
hESC
Meso D1
Direct SMAD1
hESC
Meso D1
Meso D4
OCT4
hESC
Meso D4
H3K27ac
hESC
Meso D4
0-25
0-15
0-15
0-20
0-50
chr5:153,955,168-153,990,755

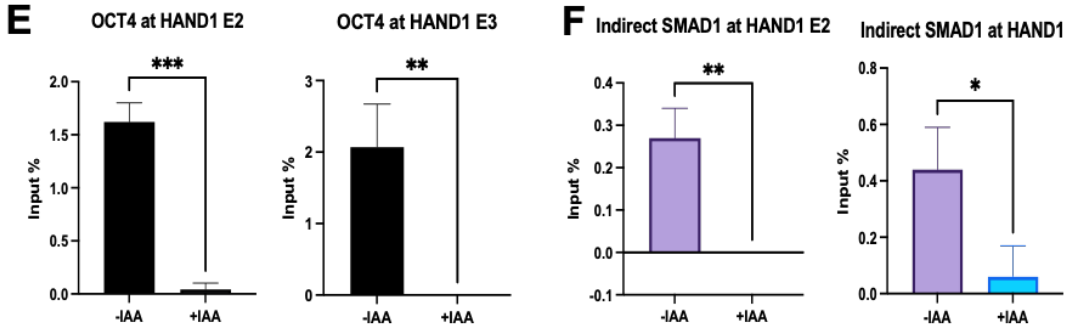

E
OCT4 at HAND1 E2
OCT4 at HAND1 E3
F
Indirect SMAD1 at HAND1 E2
Indirect SMAD1 at HAND1 E3
Input %
-IAA
+IAA
***
**
**
*

Figure 3.9: Pluripotency factor is required for SMAD binding to target loci.

(A) Genome view of the ATAC-seq and ChIP-seq data for GATA6. The putative enhancers of GATA6 are highlighted in shaded gray area.

(B) ChIP-qPCR detection of NANOG at the two highlighted enhancers for GATA6 by day 1 of DE differentiation.

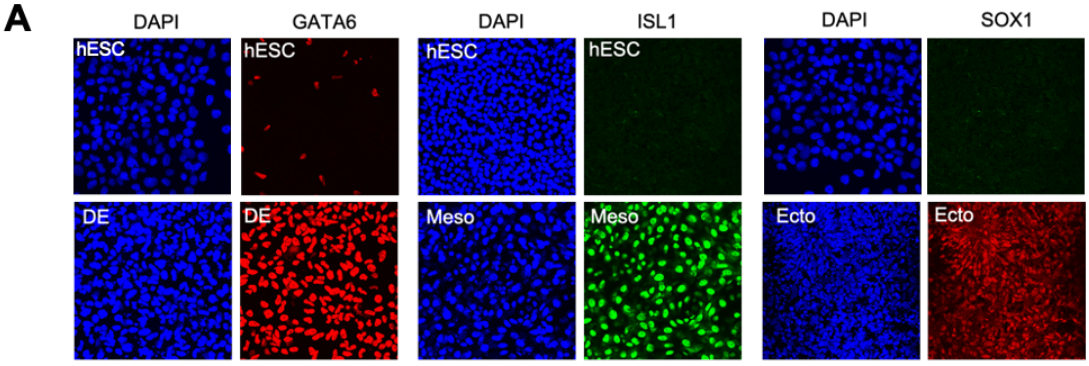
(C) Indirect ChIP-qPCR detection of SMAD2/3 at the two highlighted enhancers for GATA6 by day 1 of DE differentiation.

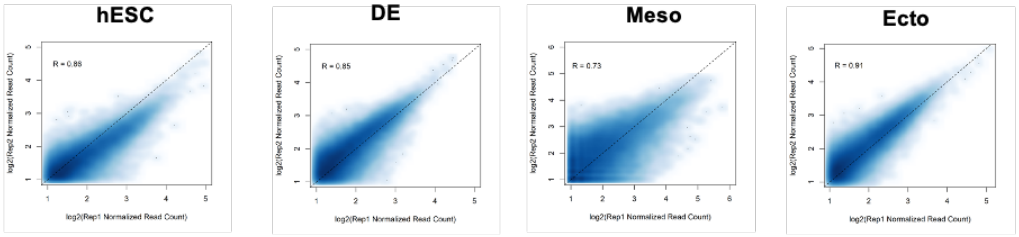
(D) Genome view of the ATAC-seq and ChIP-seq data for HAND1. The putative enhancers of HAND1 are highlighted in shaded gray area.

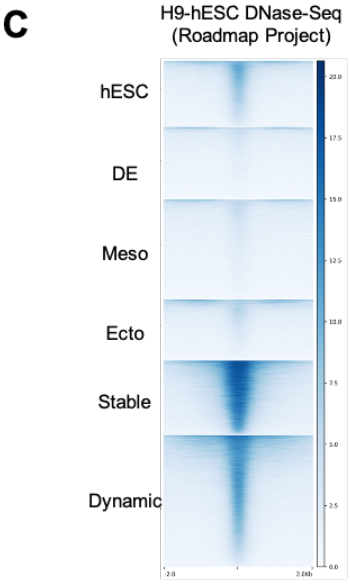
(E) ChIP-qPCR detection of OCT4 at the two highlighted enhancers for HAND1 by day 1 of Meso differentiation

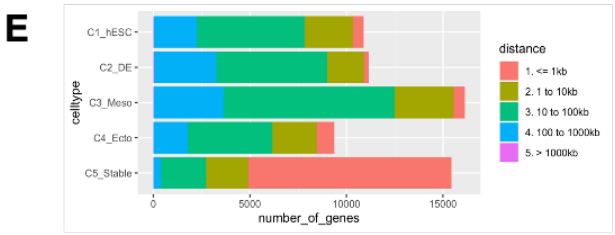
(F) Indirect ChIP-qPCR detection of SMAD1 at the two highlighted enhancers for HAND1 by day 1 of Meso differentiation

Supplemental Figure 4.1



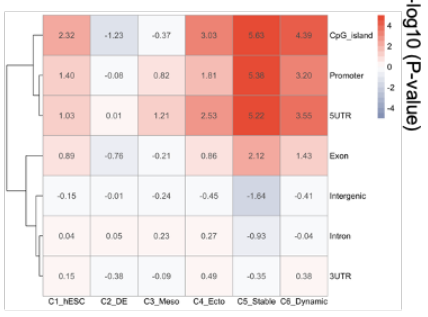






F

# Peaks per gene	Distal				Promoter
	C1 - hESC	C2 - DE	C3 - MES	C4 - NPC	C5-Stable
Mean	1.7	2.1	2.3	1.7	1.1
95% quantile	4	6	6	4	2



Supplemental Figure 4.1– related to Figure 3.1

(A) Immunostaining QC of human ES cell differentiation towards DE (GATA6), Meso (ISL1) and Ecto (SOX1).

(B) Scatter plot showing reproducibility of ATAC-Seq data. Each dot is log2 RPKM value within 250bp flanking an ATAC-Seq peak summit. R values are Pearson's correlation ecoefficiency.

(C) Heatmap of Roadmap H1-hESC DNase-seq data on H9-hESC ATAC-seq peaks.

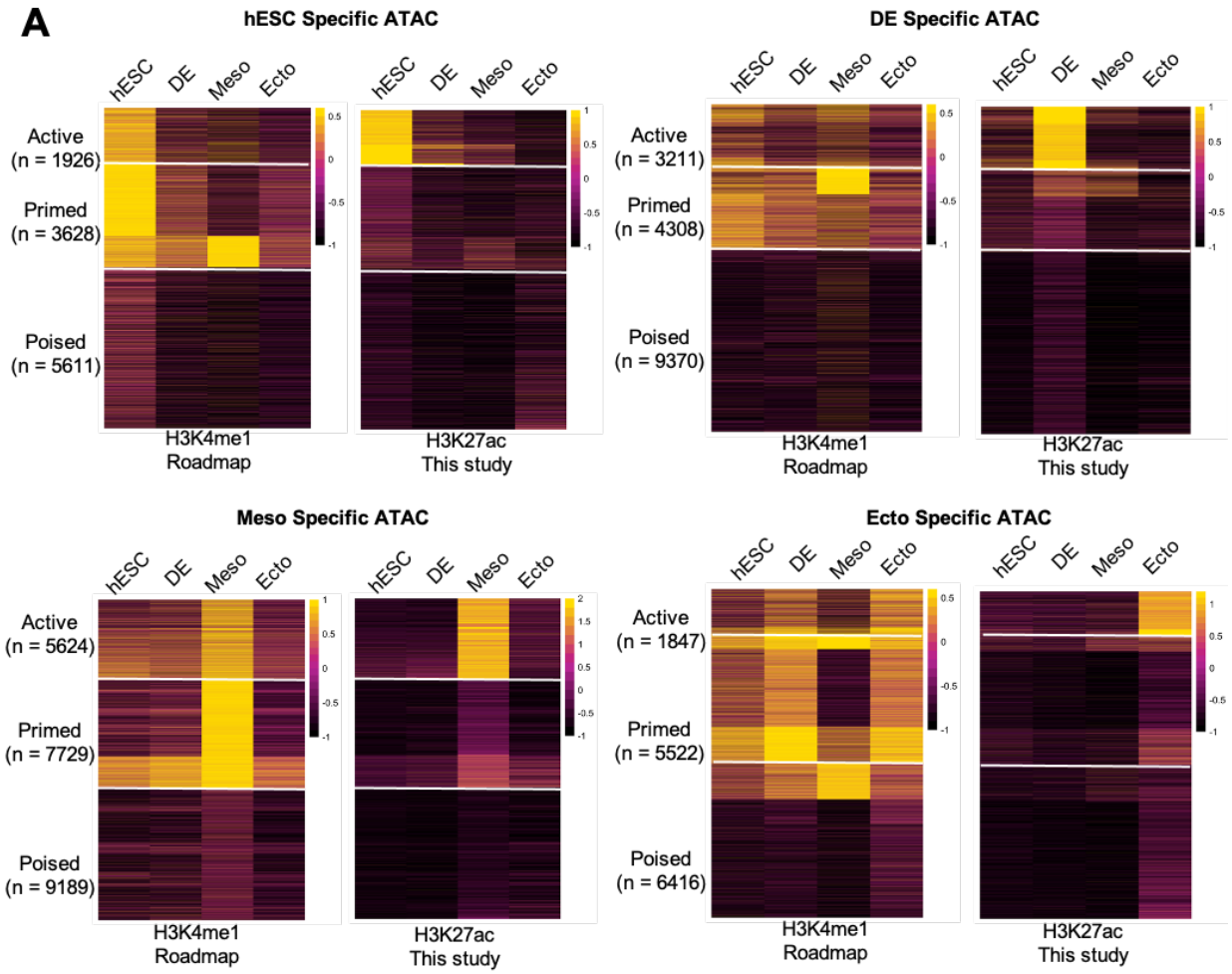
(D) Genomic feature enrichment analysis for all ATAC-Seq peaks in each category. Scores were calculated using HOMER function annotatePeaks with default parameters.

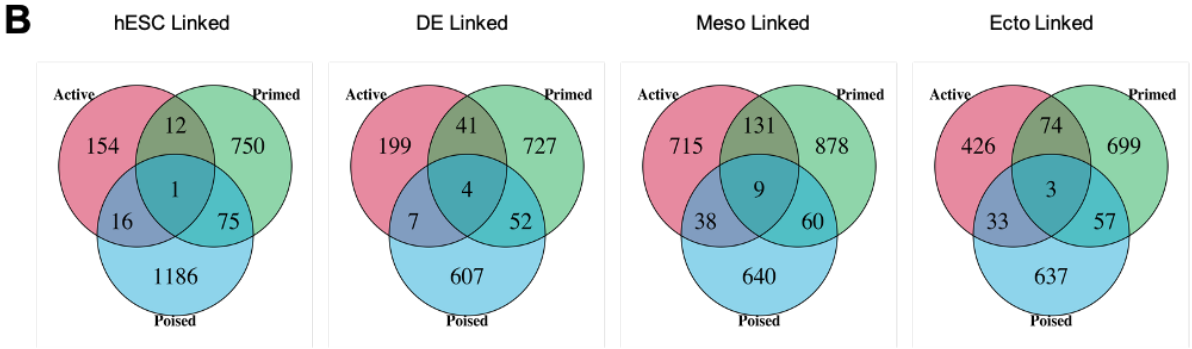
(E) Distance distribution of ATAC-seq peaks to nearest TSS.

(F) Number of ATAC peaks linked to a nearest TSS of gene. Promoter peaks are defined as occurring between -1kb and +1kb flanking a TSS. Distal peaks are all non-promoter peaks.

(G) Gene Ontology (GO) analysis (Enrichr) testing all genes within 100kb of an ATAC-Seq peak. Log2 (combined scores) were clustered based on Euclidean distances using complete linkage. Significant enrichment was marked with *.

Supplemental Figure 4.2



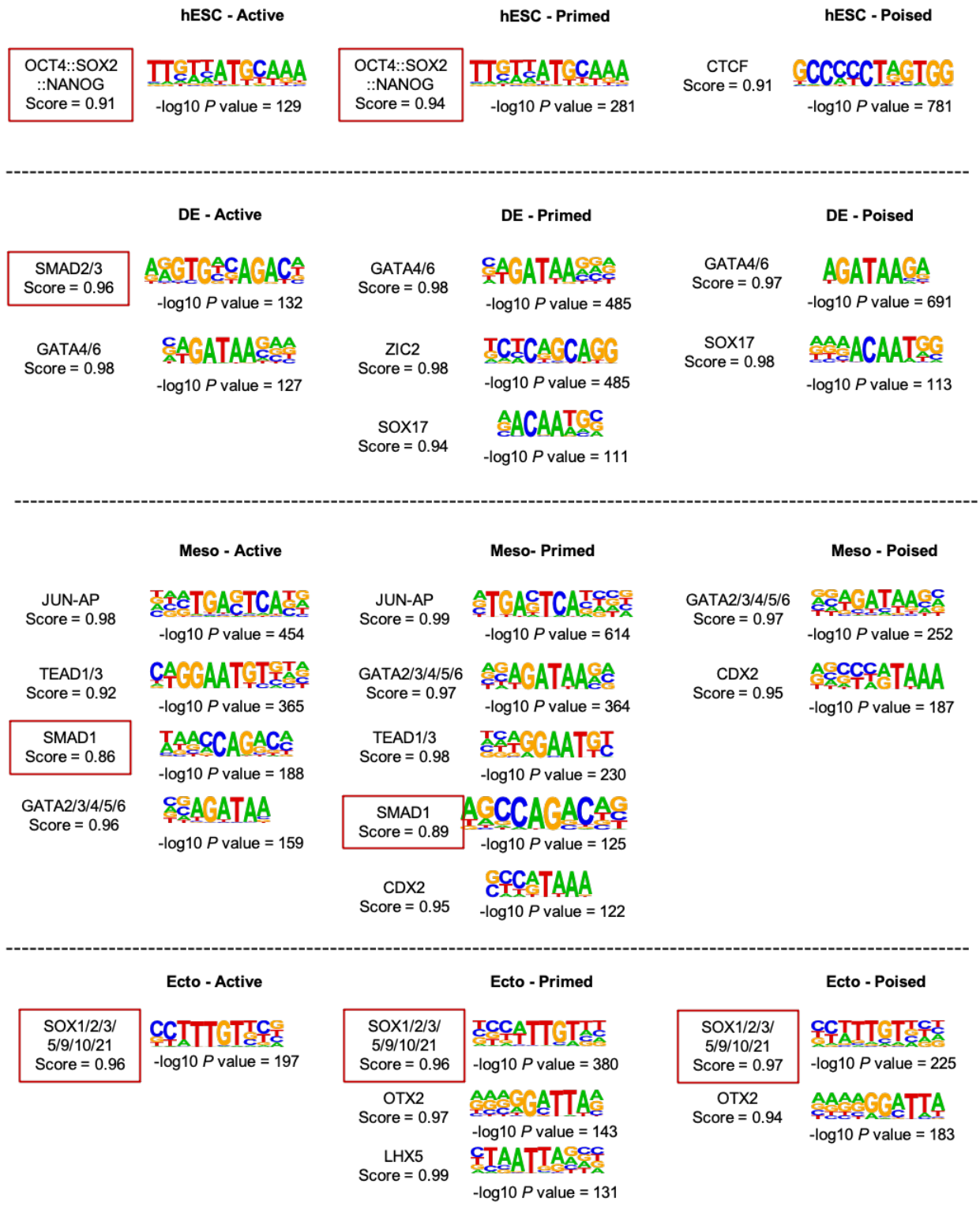


Supplemental Figure 4.2 – related to Figure 3.2

(A) Enhancer sub-categorization of the CTS open chromatin based on binary threshold of normalized histone mark, H3K4me1 and H3K27ac, ChIP-seq data. Units were CPM values within 500bp flanking peak summits in log2 transformation.

(B) Number of genes linked by Active, Primed, or Poised enhancers. The peak-to-gene linkage is assigned for peaks within 10 kb of nearest TSS.

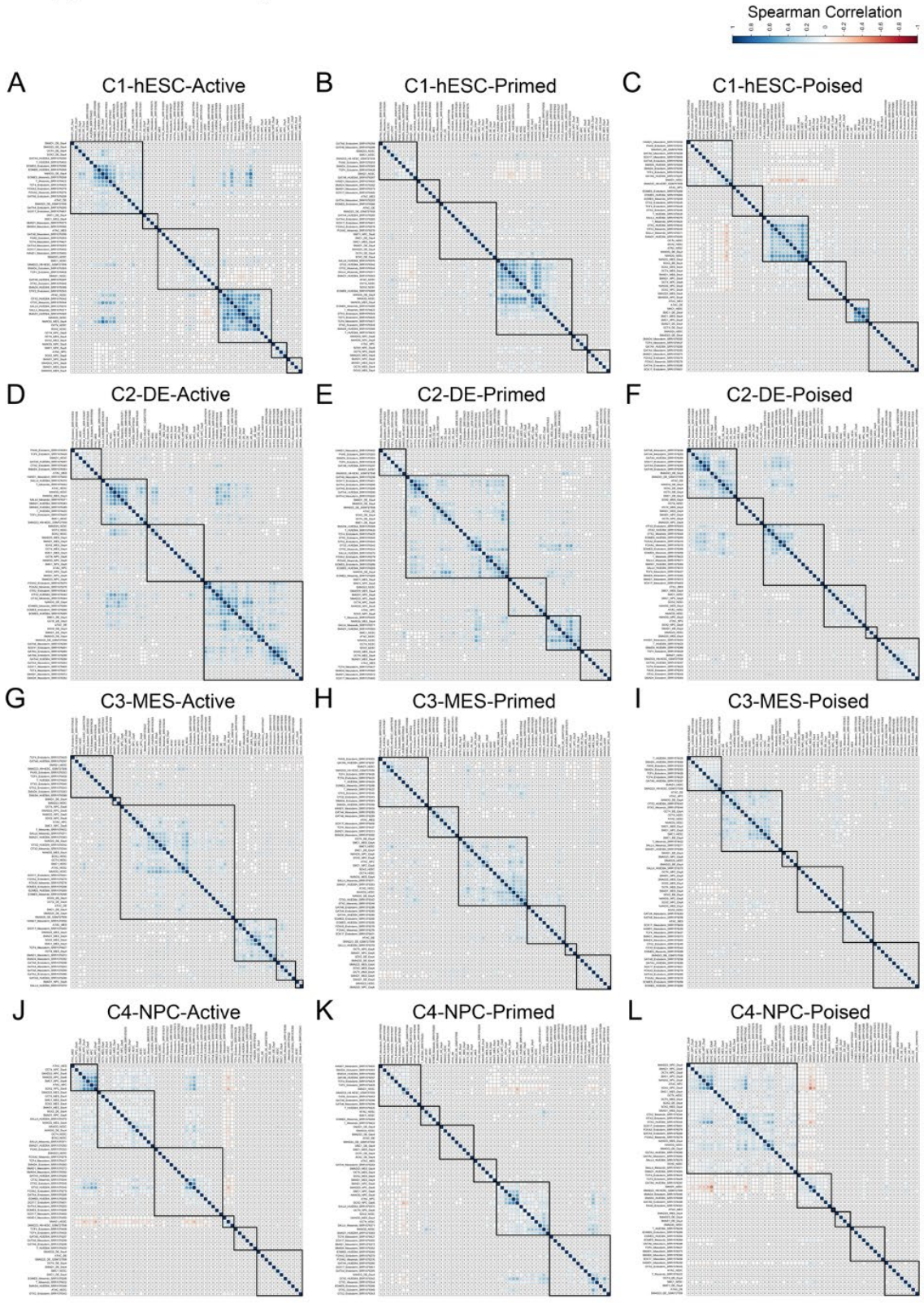
Supplemental Figure 3.3



Supplemental Figure 4.3 – related to Figure 3.2

Complete list of *de novo* motif discovery in Active, Primed, and Poised enhancers from respective cell type. HOMER was used with default settings. Filters for calling motif were set -log10(P-value) > 100 and match score > 0.9.

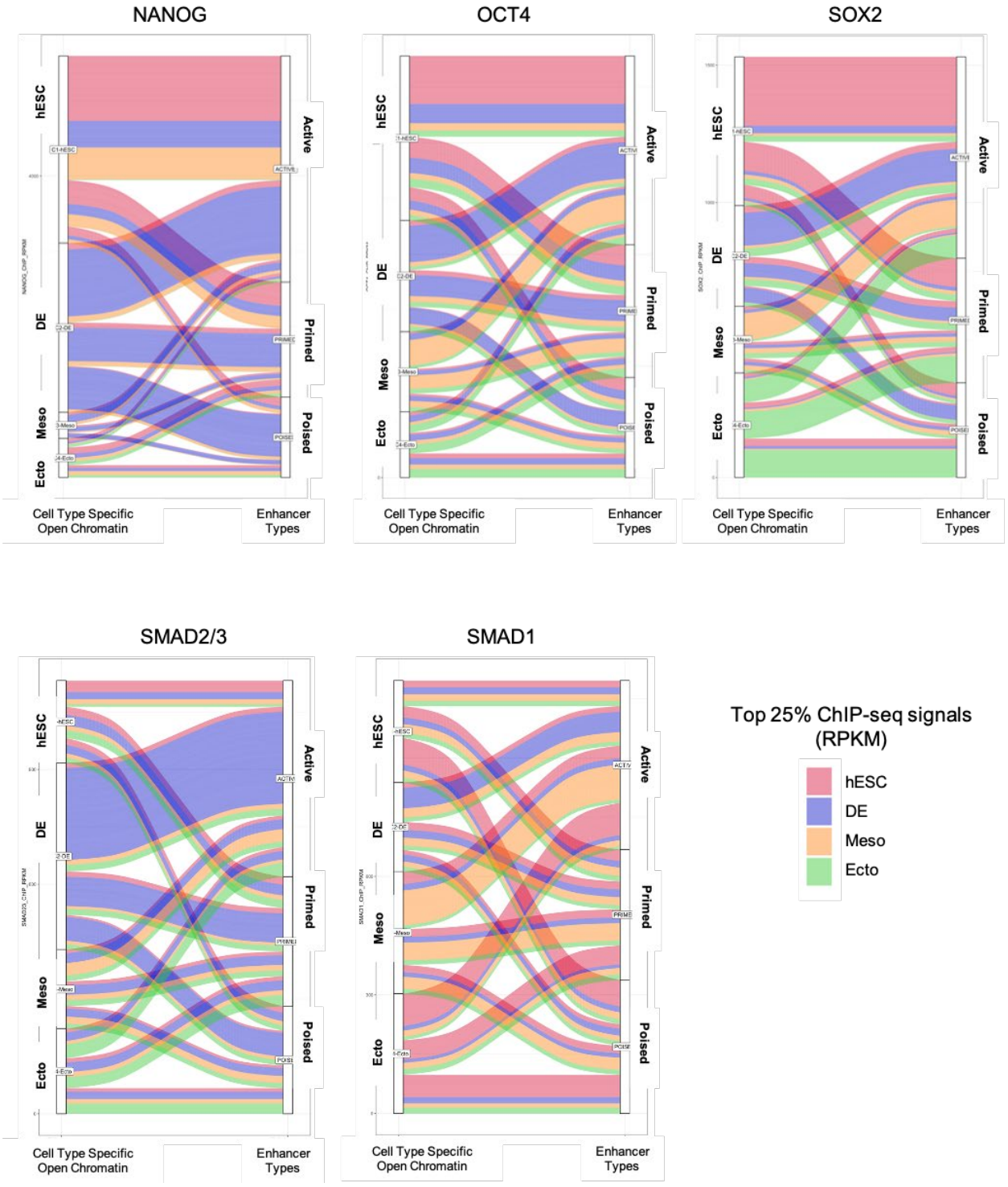
Supplemental Figure 4.4



Supplemental Figure 4.4 – related to Figure 3.3

(A-L) Hierarchical clustering of the pairwise correlation between ATAC-seq peaks and a selection of in-house and published ChIP-seq data (See Table 1 for details). Data were clustered based on a Euclidean distance matrix and complete linkage using Spearman correlation. Grey arrow star means $p > 0.05$.

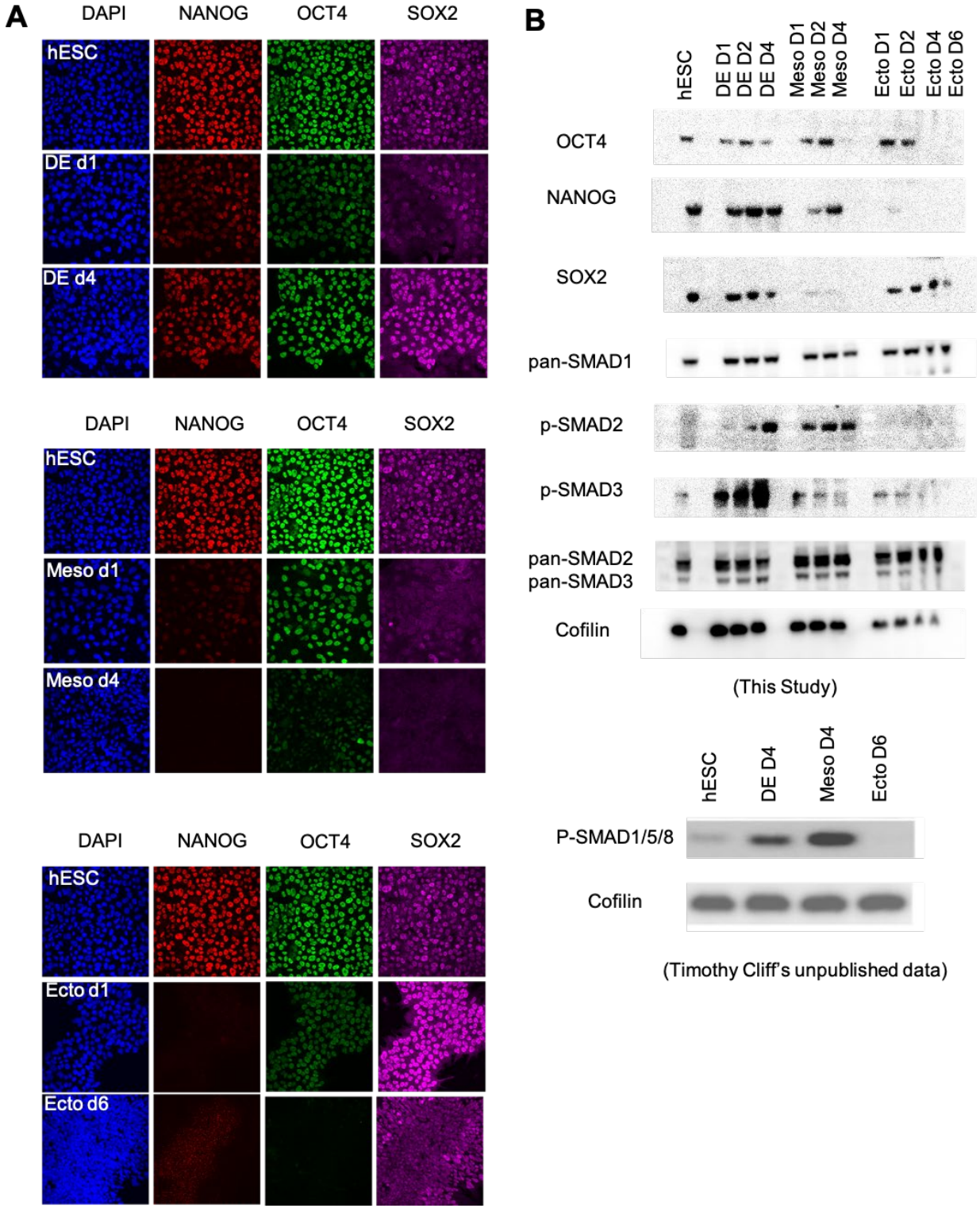
Supplemental Figure 4.5



Supplemental Figure 4.5 – related to Figure 3.3

Alluvial plot of top 25% ChIP-seq signals of NANOG, OCT4, SOX2, SMAD2/3, and SMAD1 from hESC and the three germ layers. ChIP peaks were intersected with the CTS open chromatin loci. ChIP signals were normalized 250 bp flanking the ATAC peak center. Next, each ATAC peak was annotated with either one of the Active, Primed, or Poised enhancer type. The height of each figure is the cumulated ChIP signals over intersected open chromatins. Band width is proportional to the weight of one ChIP signals from one sample out of the four cell types.

Supplemental Figure 4.6

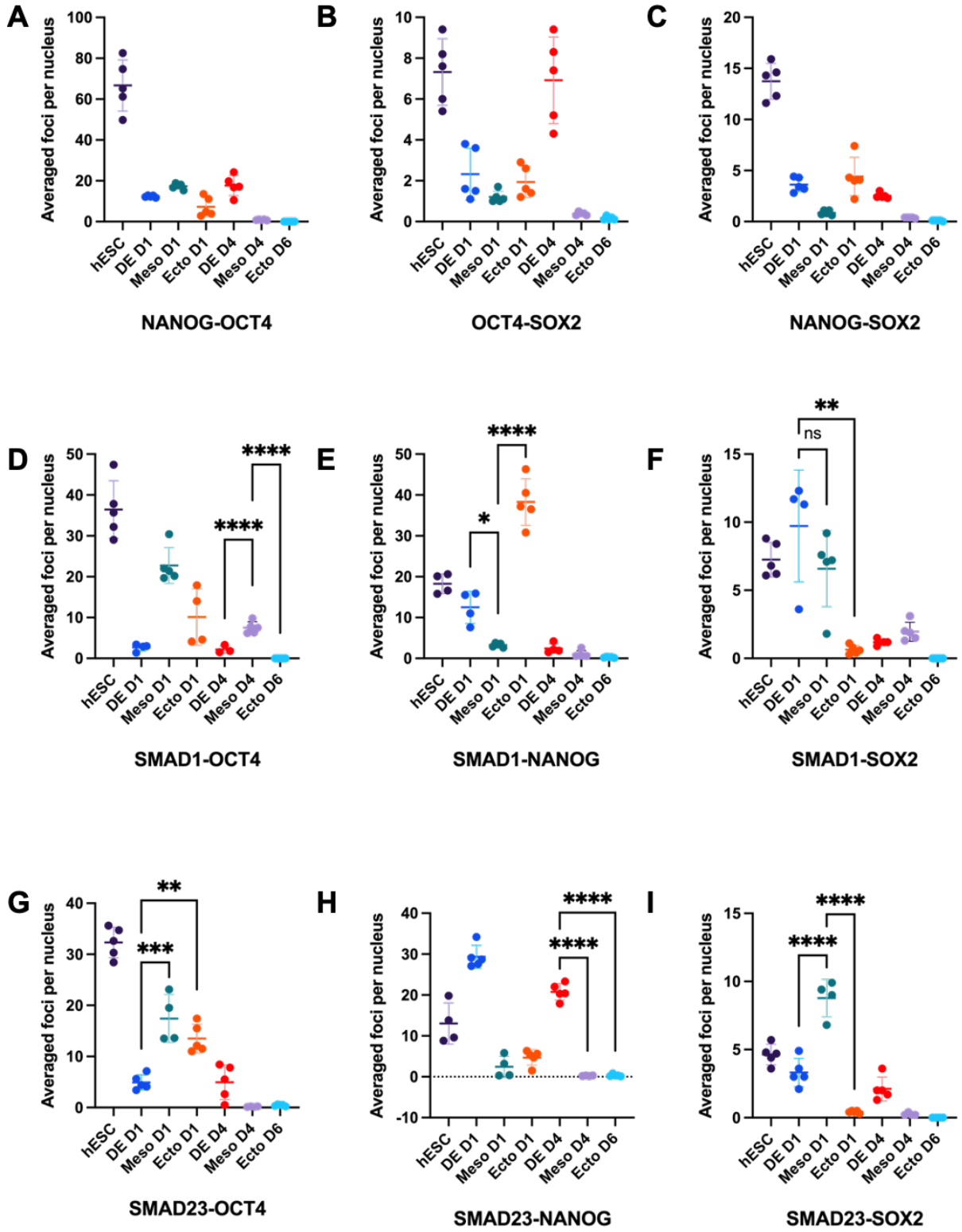


Supplemental Figure 4.6 – related to Figure 3.4

(A) Immunofluorescence of the core pluripotency factor OCT4, SOX2, and NANOG in hESC and derived DE, Meso, Ecto at D0, D1, and D4/6.

(B) Western blot for OCT4, SOX2, NANOG, p-SMAD2 (S465/467), p-SMAD3(S423/S425), p-SMAD1/5(S463/465)/8(S426/S428), pan-SMAD2/3, pan-SMAD1, and Cofilin (loading control).

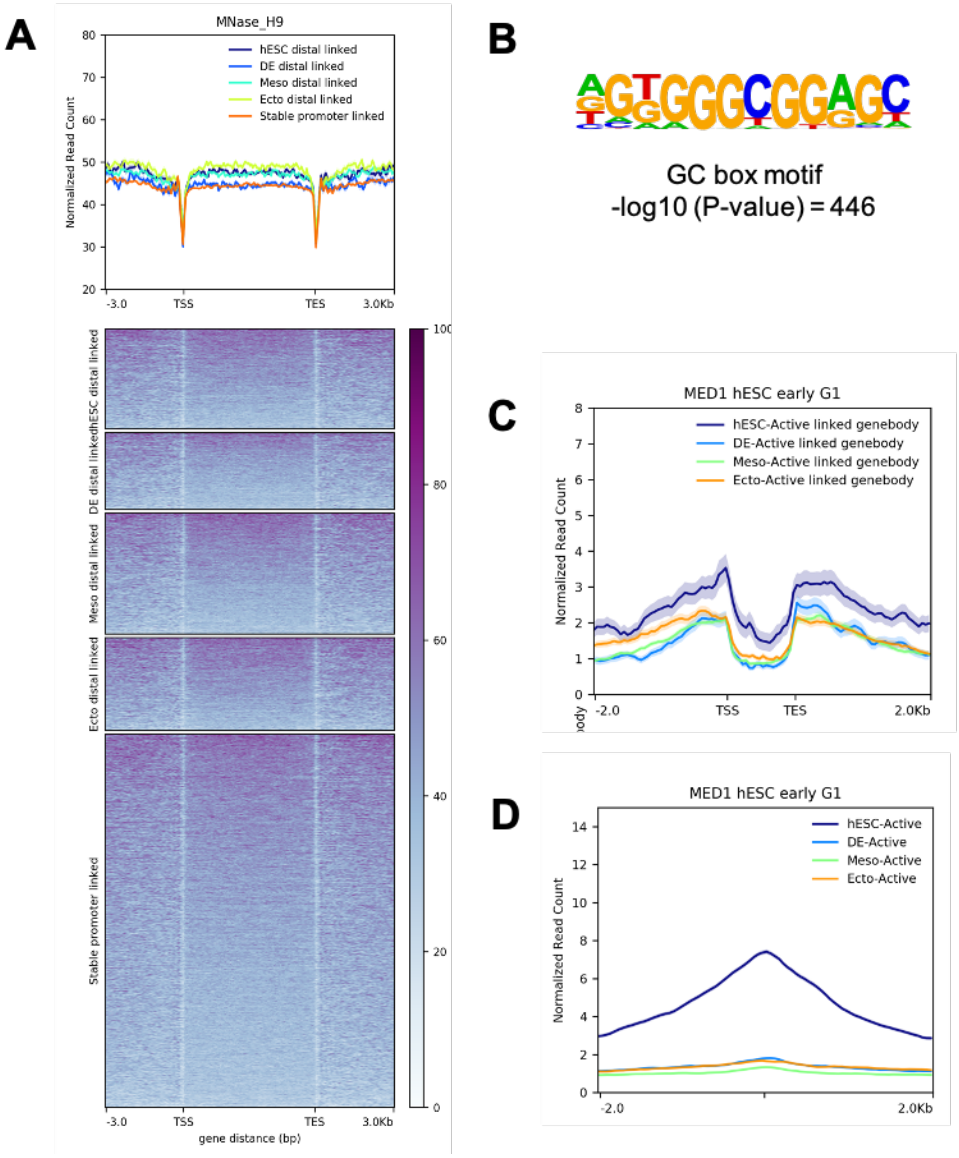
Supplemental Figure 4.7



Supplemental Figure 4.7 – related to Figure 3.4

Proximity ligation assay (PLA) was conducted for OCT4, SOX2, NANOG, SMAD1, and SMAD2/3 in a pairwise manner. Each pair of transcription factors was assessed the three germ layer differentiation at D0, D1, and D4 (DE/Meso)/6 (Ecto). Results were taken with a Zeiss confocal microscope. Each data point represents an imaging field containing around 20 – 50 cells. The PLA foci were counted if inside nucleus. Automatic data processing was conducted using customized ImageJ script.

Supplemental Figure 4.8



Supplemental Figure 4.8 – related to Figure 3.6

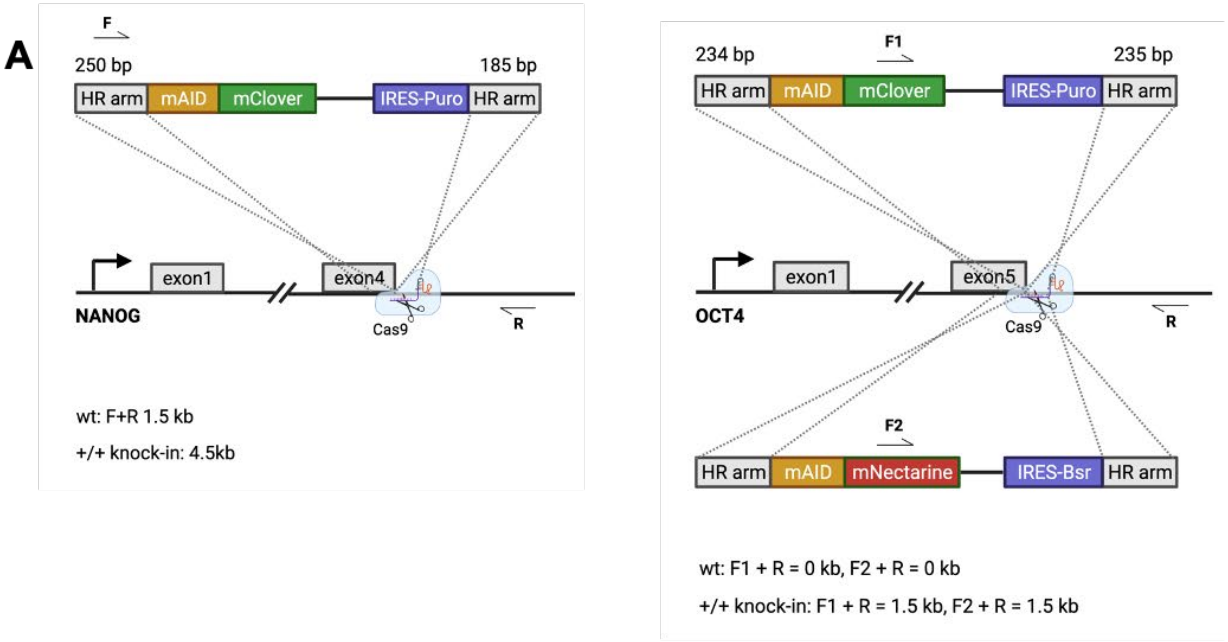
(A) Mapping of Mnase-seq data (GSM1194221) from H9-hESC over categorized open chromatin loci. Mnase signals represents nucleosome occupancy.

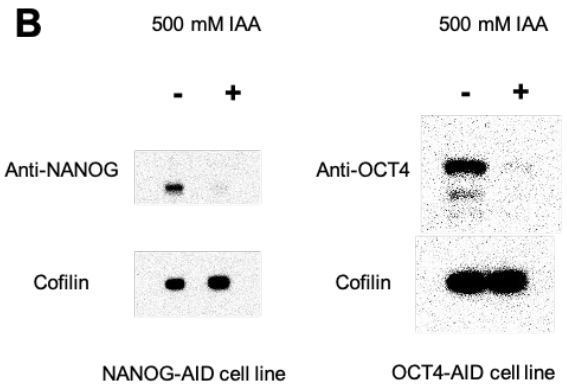
(B) GC box motif calculated from pooled promoter proximal regions.

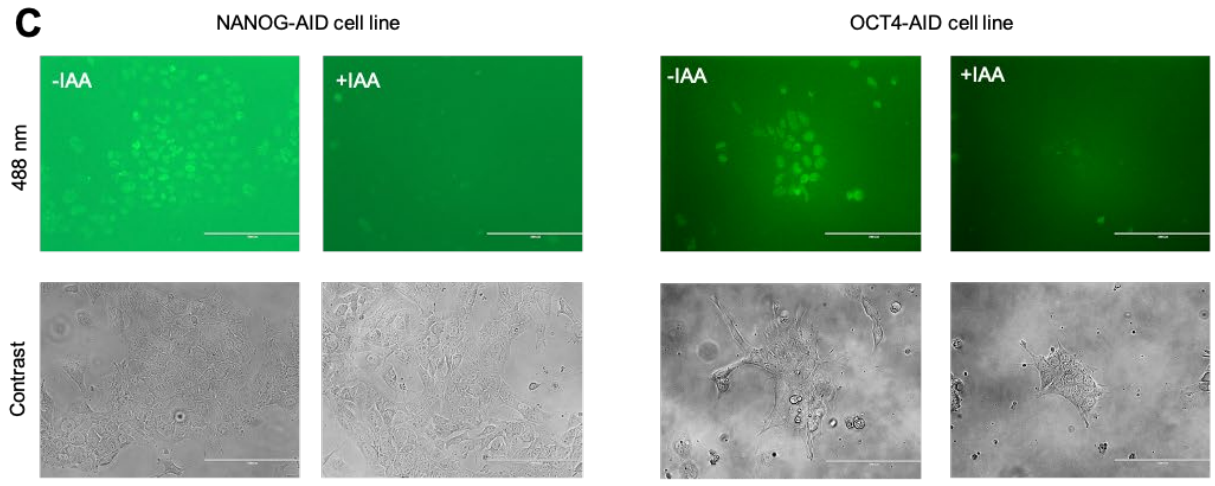
(C) MED1 ChIP-seq signals along open chromatin linked gene body.

(D) MED1 ChIP-seq signals centered at the Active open chromatins.

Supplemental Figure 4.9





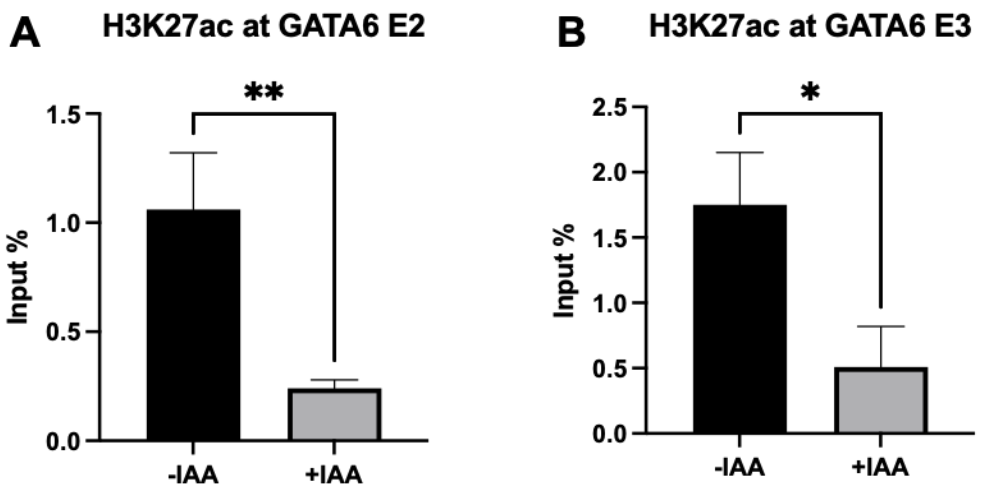


Supplemental Figure 4.9 – related to Figure 3.9

(A) Experimental scheme used to construct template for knocking in AID to the C-terminal of NANOG or OCT4. Both alleles of NANOG were knocked-in to create homozygous NANOG-AID-mClover. Genomic modified OCT4-AID fusion is heterozygous, one allele is OCT4-AID-mClover, the other is OCT4-AID-mNacterine.

(B) Western blot validation of IAA induced conditional knock-out of NANOG and OCT4 proteins. IAA treatment was 24 hours.

(C) Microscopic observation of positive clones of NANOG-AID and OCT4-AID. AID fusion proteins are localized in nucleus. Conditional knockout can be maintained for 24 hours in the presence of IAA in culture medium.



Supplemental Figure 4.10 – related to Figure 3.9

ChIP-qPCR experiments of H3K27ac at GATA6 enhancer 2 (A) and enhancer 3 (B).

Table 1: Published data analyzed

Data Type	Accession	Target	Cell Type	Species
Dnase-seq	SRR412248		hESC	Hs
Dnase-seq	SRR412247		hESC	Hs
Dnase-seq	GSM493384		CD34_primary_cell	Hs
Dnase-seq	GSM595924		Fetal_Lung	Hs
Dnase-seq	GSM701509		Fetal_Spleen	Hs
Dnase-seq	GSM817158		Fetal_Skin	Hs
Dnase-seq	GSM817220		Fetal_Hear	Hs
Dnase-seq	GSM878663		Fetal_Spinal_Cord	Hs
Dnase-seq	GSM1027310		Fetal_Adrenal_Gland	Hs
Dnase-seq	GSM1027326		Pancrease	Hs
Dnase-seq	GSM1027335		Pnacrease	Hs
Mnase-seq	GSM1194220		H1-HESC	Hs
Mnase-seq	GSM1194221		H9-hESC	Hs
ChIP-seq	SRR1576266	EOMES	Endoderm	Hs
ChIP-seq	SRR1576265	EOMES	HUES64	Hs
ChIP-seq	SRR1576268	EOMES	Mesendoderm	Hs
ChIP-seq	SRR1576279	FOXA2	Endoderm	Hs
ChIP-seq	SRR1576276	FOXA2	Mesendoderm	Hs
ChIP-seq	SRR1576285	GATA4	Endoderm	Hs
ChIP-seq	SRR1576280	GATA4	HUES64	Hs
ChIP-seq	SRR1576283	GATA4	Mesoderm	Hs
ChIP-seq	SRR1576298	GATA6	Endoderm	Hs
ChIP-seq	SRR1576297	GATA6	HUES64	Hs
ChIP-seq	SRR1576299	GATA6	Mesoderm	Hs
ChIP-seq	SRR1576450	HAND1	Mesoderm	Hs
ChIP-seq	SRR1576345	OTX2	Ectoderm	Hs
ChIP-seq	SRR1576343	OTX2	Endoderm	Hs
ChIP-seq	SRR1576342	OTX2	HUES64	Hs
ChIP-seq	SRR1576344	OTX2	Mesendoderm	Hs
ChIP-seq	SRR1576353	PAX6	Ectoderm	Hs
ChIP-seq	SRR1576370	SALL4	HUES64	Hs
ChIP-seq	SRR1576371	SALL4	Mesendoderm	Hs
ChIP-seq	SRR1576383	SMAD1	HUES64	Hs

ChIP-seq	SRR1576373	SMAD1	Mesoderm	Hs
ChIP-seq	SRR1576393	SMAD4	Ectoderm	Hs
ChIP-seq	SRR1576389	SMAD4	HUES64	Hs
ChIP-seq	SRR1576392	SMAD4	Mesoderm	Hs
ChIP-seq	GSM727558	SMAD23	Endoderm	Hs
ChIP-seq	GSM727558	SMAD23	hESC	Hs
ChIP-seq	SRR1576401	SOX17	Endoderm	Hs
ChIP-seq	SRR1576400	SOX17	Mesoderm	Hs
ChIP-seq	SRR1576420	T	HUES64	Hs
ChIP-seq	SRR1576421	T	Mesendoderm	Hs
ChIP-seq	SRR1576428	TCF4	Ectoderm	Hs
ChIP-seq	SRR1576425	TCF4	Endoderm	Hs
ChIP-seq	SRR1576427	TCF4	Mesoderm	Hs
ChIP-seq	SRR1036375	H3K27me3	Endoderm	Hs
ChIP-seq	SRR1067036	H3K27ac	Endoderm	Hs
ChIP-seq	SRR1067048	H3K4me1	Endoderm	Hs
ChIP-seq	SRR1067447	H3K9me3	Endoderm	Hs
ChIP-seq	SRR1067484	H3K9me3	Endoderm	Hs
ChIP-seq	SRR1067482	H3K4me3	Endoderm	Hs
ChIP-seq	SRR1067034	H3K27ac	Mesoderm	Hs
ChIP-seq	SRR1067462	H3K4me3	Mesoderm	Hs
ChIP-seq	SRR1067471	H3K4me1	Mesoderm	Hs
ChIP-seq	SRR1067475	H3K27me3	Mesoderm	Hs
ChIP-seq	SRR1067477	H3K9me3	Mesoderm	Hs
ChIP-seq	SRR1067483	H3K36me3	Mesoderm	Hs
ChIP-seq	SRR316161	H3K27ac	Neural Progenitor Cell	Hs
ChIP-seq	SRR316169	H3K4me3	Neural Progenitor Cell	Hs
ChIP-seq	SRR353319	H3K27me3	Neural Progenitor Cell	Hs
ChIP-seq	SRR353320	H3K36me3	Neural Progenitor Cell	Hs
ChIP-seq	SRR353325	H3K4me1	Neural Progenitor Cell	Hs
ChIP-seq	SRR353342	H3K9me3	Neural Progenitor Cell	Hs
ChIP-seq	GSM605310	H3K36me3	H9-hESC	Hs
ChIP-seq	GSM667626	H3K4me1	H9-hESC	Hs
ChIP-seq	SRR097966	H3K27me3	H9-hESC	Hs
ChIP-seq	SRR097971	H3K9me3	H9-hESC	Hs
ChIP-seq	SRR179684	H3K27me3	H9-hESC	Hs
ChIP-seq	SRR179689	H3K4m1	H9-hESC	Hs
ChIP-seq	SRR179703	H3K4me3	H9-hESC	Hs
ChIP-seq	SRR179706	H3K27ac	H9-hESC	Hs

ChIP-seq	SRR179734	H3K36me3	H9-hESC	Hs
ChIP-seq	SRR1576259	CTCF	HUES64	Hs
ChIP-seq	SRR1576260	CTCF	HUES64	Hs
ChIP-seq	ENCODE	CTCF	Endoderm	Hs
ChIP-seq	ENCODE	CTCF	Mesoderm	Hs
ChIP-seq	ENCODE	CTCF	Neural Progenitor Cell	Hs
ATAC-seq	SRR8454427		E65Epi	Mouse
ATAC-seq	SRR8454430		E65VE	Mouse
ATAC-seq	SRR8454432		Ect	Mouse
ATAC-seq	SRR8454435		PS	Mouse
ATAC-seq	SRR8454438		Mes	Mouse
ATAC-seq	SRR8454440		End	Mouse
ChIP-seq	SRR8454496	H3K27ac	E65Epi	Mouse
ChIP-seq	SRR8454499	H3K27ac	E65VE	Mouse
ChIP-seq	SRR8454503	H3K27ac	Ect	Mouse
ChIP-seq	SRR8454507	H3K27ac	PS	Mouse
ChIP-seq	SRR8454513	H3K27ac	Mes	Mouse
ChIP-seq	SRR8454519	H3K27ac	End	Mouse

Table 2: Primary antibody list

Antibody	Application	Dilution	Vendor	Catalog
GATA6	IF, ChIP	1:200, 5ug/5M cells	cst	5851
H3K27ac	ChIP	2.5 ug/5 M cells	active motif	39133
ISL1	IF	1:500	R&D	AF1837
NANOG	IF, PLA, ChIP	1:200, 3 ug/5M cells	R&D	AF1997
OCT4	WB, ChIP	1:1000, 5 ug/5M cells	abcam	19857
OCT4A	PLA	1:150	santa cruz	8628 (discountinued)
p-SMAD1/5/8	WB	1:1000	cst	9511s (discontinued)
p-Smad2 (Ser245/250/255)	WB	1:1000	cst	3104
p-SMAD3 (Ser423/425)	WB	1:1000	cst	9520s
SMAD1	IF, PLA	1:200	cst	6944s
SMAD1	ChIP	20 ul/5M cells	cst	9743
SMAD23	ChIP	5 ug/5M cells	R&D	AF3797
SMAD23	IF, PLA	1:100	cst	5678
SOX2	IF, WB, ChIP	1:100, 1:1000, 5 ug/5M cells	R&D	MAB2018
p-SMAD1(S206)	WB	1:1000	diagenode	C15410274
OCT4	WB	1:1000	cst	2840
SMC1	ChIP	5 ug/5M cells	Bethyl	A300-055A
MED1	ChIP	5 ug/5M cells	Bethyl	A300-793A
CTCF	ChIP	5 ug/5M cells	Milipore	07-729

Table 3: Synthesized DNA oligos

ChIP-qPCR primers	Strand	Sequencec	Locus	Note
GSC E2	Forward	TCTGAGCCTCCATCTTCCAT	chr14:95231168-95231255	ATAC cluster match, NANOG bound in hESC
	Reverse	TTTTCACTGCAAAGCCCTTC		
GSC E3	Forward	AAGCCAAGGAATGTGGATTG	chr14:95241903-95241977	ATAC cluster match
	Reverse	ACTTGGGAGGGCAACCTTAT		
EOMES_P	Forward	TGGAAACTTATGGGCTGTCA	chr3:27764198-27764292	
	Reverse	ACCAGCCAATAGGAGGGTCT		
EOMES E2	Forward	TCCCCAATTAGATTGCTGCT	chr3:27931005-27931114	dCas9 tethering, ATAC cluster match
	Reverse	GAGCTGAAGTGCCAGACAGA		
EOMES E3	Forward	GAGACAGAGGGGAAGCATGA	chr3:27812092-27812231	
	Reverse	GTACCCAGGACCCCAGAGAT		
GATA6 P	Forward	TTTTCTCTCCTCCCCTCGAT	chr18:19749331-19749415	
	Reverse	AGGCTGTGGGTCGGAACT		
GATA6 E1	Forward	TCCAACAGTCCCCTGATTTC	chr18:19739153-19739253	
	Reverse	CAAGCTGCTCCCAGATAAGC		
GATA6 E2	Forward	GCCAGACATCTCCTGTGGAT	chr18:19659837-19659911	nucleosome depletion assay, ATAC cluster match, dCas9 tethering
	Reverse	GCCAGAAAACCCTTTGATGA		

GATA6 E3	Forward	CAAATCCCCCAGCTCCTAGT	chr18:19606222-19606366	ATAC cluster match, dCas9 tethering
	Reverse	GGCACTCCAGGAGGTATTGA		
GATA6 E4	Forward	CTTGGCCTGGGACACTTTAC	chr18:19571198-19571302	Dynamic
	Reverse	CAGAGAAGAGCAGCCCTGAG		
SOX17 E1	Forward	TGCTGTTTGCTTCTCCACTG	chr8:55172564-55172676	ATAC cluster match
	Reverse	CCAAGTGTTGCTCAATGCTG		
SOX17 E2	Forward	GGACACTGGGTTTTCCTTGA	chr8:55181139-55181232	dynamic
	Reverse	GGGAACCAGGTTAGGGACAT		
SOX17 distal	Forward	TTGAGGTTGCATCAGTCTCG	chr8:55137459-55137550	near DE cluster
	Reverse	GCCACCTAATCAATGCCTGT		
TBX3 E1	Forward	GGAATGCACAGGCTGATTTA	chr12:115141617-115141694	ATAC cluster match, dCas9 tethering
	Reverse	ACCAAGGGTCATTTCTGGAG		
TBX3 E2	Forward	GGAGGTGGAGACGGTTCTTA	chr12:115596816-115596926	nucleosome depletion assay, ATAC cluster match, dCas9 tethering
	Reverse	AGAGGCAAACCCTTCATCAC		
HAND1 E1_3	Forward	CTCAGCCTGGCTCTCCACTA	chr5:153896855-153896938	ATAC cluster match, dCas9 tethering
	Reverse	TTTGTCCACAGATGGGTGTG		
HAND1 E2	Forward	CTGGCACACAGTAGGCACTC	chr5:153955950-153956049	ATAC cluster match, no SMAD motif
	Reverse	AAAGTGCCTGGAGCTGAAAT		
HAND1 E3	Forward	CTCTCTGTGCCCTCCTTTTC	chr5:153983265-153983377	not in cluster annotation track, dCas9 tethering
	Reverse	CCTCCAGGACACTTTGTGTG		
ISL1 E1	Forward	GGTATTTCTGGGCCCTTCTC	chr5:50694915-50695013	Near Meso specific ATAC
	Reverse	TCTAGACTCGCGACCTCACA		

ISL1 distal En	Forward	GGGCACTCTCAGGACTCTGT	chr5:50917970-50918099	HiChIP loop anchor
	Reverse	GCAAGGTGGCATTCACAAAT		
TEAD1 intron	Forward	GGCCTGTAATGGTTGCTGAC	chr11:12,836,539-12,836,616	ATAC cluster match
	Reverse	TGCGTCTTTCCCTTTTCAAC		
PITX1 E1	Forward	GCAAGCAGTCAATGTGGCTA	chr5:134,544,799-134,544,873	ATAC cluster match
	Reverse	TCTCCTCGCATTCATGTCCT		
PITX1 E2	Forward	TCCCAGGAGTACCAAGCATC	chr5:134,607,504-134,607,601	ATAC cluster match
	Reverse	GCCCCACTTTTGACATCACT		

REFERENCES

Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., and Hannett, N.M. (2018). Transcription factors activate genes through the phase-separation capacity of their activation domains. Cell *175*, 1842-1855. e1816.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., and Jenner, R.G. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. cell *122*, 947-956.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods *10*, 1213.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J.*, et al.* (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell *133*, 1106-1117.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative binding of transcription factors orchestrates reprogramming. Cell *168*, 442-459. e420.

Cliff, T.S., Wu, T., Boward, B.R., Yin, A., Yin, H., Glushka, J.N., Prestegaard, J.H., and Dalton, S. (2017). MYC Controls Human Pluripotent Stem Cell Fate Decisions through Regulation of Metabolic Flux. Cell Stem Cell *21*, 502-516.e509.

Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., and Cho, S.W. (2018a). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898.

Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., and Cho, S.W. (2018b). The chromatin accessibility landscape of primary human cancers. Science *362*.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K.*, et al.* (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374-387.

El Khattabi, L., Zhao, H., Kalchschmidt, J., Young, N., Jung, S., Van Blerkom, P., Kieffer-Kwon, P., Kieffer-Kwon, K.-R., Park, S., and Wang, X. (2019). A pliable mediator acts as a functional rather than an architectural bridge between promoters and enhancers. Cell *178*, 1145-1158. e1120.

Faial, T., Bernardo, A.S., Mendjan, S., Diamanti, E., Ortmann, D., Gentsch, G.E., Mascetti, V.L., Trotter, M.W., Smith, J.C., and Pedersen, R.A. (2015). Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. Development *142*, 2121-2135.

Festuccia, N., Osorno, R., Wilson, V., and Chambers, I. (2013). The role of pluripotency gene regulatory network components in mediating transitions between pluripotent cell states. Current Opinion in Genetics & Development *23*, 504-511.

Gaarenstroom, T., and Hill, C.S. (2014). TGF-β signaling to chromatin: how Smads regulate transcription during self-renewal and differentiation. Paper presented at: Seminars in cell & developmental biology (Elsevier).

Gorkin, David U., Leung, D., and Ren, B. (2014). The 3D Genome in Transcriptional Regulation and Pluripotency. Cell Stem Cell *14*, 762-775.

Griffiths, A.J., Miller, J.H., Suzuki, D.T., Lewontin, R.C., and Gelbart, W.M. (2000). Transcription: an overview of gene regulation in eukaryotes. an introduction to genetic analysis 7th edition New York: WH Freeman.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Hill, C.S. (2016). Transcriptional control by the SMADs. Cold Spring Harbor perspectives in biology *8*, a022079.

Jaenisch, R., and Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell *132*, 567-582.

Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nature Reviews Genetics *20*, 207-220.

Li, D., Liu, J., Yang, X., Zhou, C., Guo, J., Wu, C., Qin, Y., Guo, L., He, J., and Yu, S. (2017). Chromatin accessibility dynamics during iPSC reprogramming. Cell stem cell *21*, 819-833. e816.

Marsboom, G., Zhang, G.-F., Pohl-Avila, N., Zhang, Y., Yuan, Y., Kang, H., Hao, B., Brunengraber, H., Malik, A.B., and Rehman, J. (2016). Glutamine metabolism regulates the pluripotency transcription factor OCT4. Cell reports *16*, 323-332.

Michael, A.K., Grand, R.S., Isbel, L., Cavadini, S., Kozicka, Z., Kempf, G., Bunker, R.D., Schenk, A.D., Graff-Meyer, A., and Pathare, G.R. (2020). Mechanisms of OCT4-SOX2 motif readout on nucleosomes. Science.

Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nature methods *13*, 919.

Natsume, T., Kiyomitsu, T., Saga, Y., and Kanemaki, M.T. (2016). Rapid protein depletion in human cells by auxin-inducible degron tagging with short homology donors. Cell reports *15*, 210-218.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the mitotic chromosome. Science *342*, 948-953.

Ong, C.-T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Reviews Genetics *12*, 283-293.

Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.-T., Hookway, T.A., Guo, C., and Sun, Y. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281-1295.

Singh, A.M., Reynolds, D., Cliff, T., Ohtsuka, S., Mattheyses, A.L., Sun, Y., Menendez, L., Kulik, M., and Dalton, S. (2012). Signaling Network Crosstalk in Human Pluripotent Cells: A Smad2/3-Regulated Switch that Controls the Balance between Self-Renewal and Differentiation. Cell Stem Cell *10*, 312-326.

Singh, A.M., Sun, Y., Li, L., Zhang, W., Wu, T., Zhao, S., Qin, Z., and Dalton, S. (2015). Cell-Cycle Control of Bivalent Epigenetic Domains Regulates the Exit from Pluripotency. Stem cell reports *5*, 323-336.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555-568.

Teo, A.K.K., Arnold, S.J., Trotter, M.W., Brown, S., Ang, L.T., Chng, Z., Robertson, E.J., Dunn, N.R., and Vallier, L. (2011). Pluripotency factors regulate definitive endoderm specification through eomesodermin. Genes & development *25*, 238-250.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., and Vernot, B. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Tian, B., Yang, J., and Brasier, A.R. (2012). Two-step cross-linking for analysis of protein–chromatin interactions. In Transcriptional Regulation (Springer), pp. 105-120.

Tsankov, A.M., Gu, H., Akopian, V., Ziller, M.J., Donaghey, J., Amit, I., Gnirke, A., and Meissner, A. (2015). Transcription factor binding dynamics during human ES cell differentiation. Nature *518*, 344-349.

Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N.A., Muth, K., Palmer, J., Qiu, Y., and Wang, J. (2015). Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. Cell stem cell *16*, 386-399.

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

Xiang, Y., Zhang, Y., Xu, Q., Zhou, C., Liu, B., Du, Z., Zhang, K., Zhang, B., Wang, X., and Gayen, S. (2020). Epigenomic analysis of gastrulation identifies a unique chromatin state for primed pluripotency. Nature genetics *52*, 95-105.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., and Stewart, R. (2007). Induced pluripotent stem cell lines derived from human somatic cells. science *318*, 1917-1920.

Zaret, K.S., and Yamamoto, K.R. (1984). Reversible and persistent changes in chromatin structure accompany activation of a glucocorticoid-dependent enhancer element. Cell *38*, 29-38.

CHAPTER 4

DETAILED EXPERIMENTAL PROCEDURES

MATERIALS AND METHODS

Cell Culture and Differentiations

Human embryonic stem cell (hESC) line, H9 (WA09, WiCell), was passaged and differentiated as previously described (Cliff et al., 2017). Briefly, hESC were passaged routinely at 50,000 cells/cm^2 on Geltrex coated plastic dishes. Self-renewal was maintained using a chemically defined base medium (CDM) supplemented with 10 ng/ml human Heregulinβ -1 (Peprotech), 10 ng/ml Activin A (R&D), 200 ng/ml LONG^R R3 human IGF-I (Sigma), and 8 ng/ml human basic-FGF (R&D), termed HAIF media.

For DE differentiation, hESCs were passaged at 50,000 cells/cm^2. CDM was supplemented with 100 ng/ml Activin, 8 ng/ml bFGF for 4 days. During the first 24 hours, 25 ng/ml human Wnt-3a (R&D) was added. Media was changed every day.

Meso was generated by seeding 50,000 cells/cm^2 in HAIF media plus 100 ng/ml human BMP-4 (R&D) and 25 ng/ml Wnt-3a for 4 days. Media was changed daily.

Neural Ecto cells were generated by culturing hESC at a density of 90,000 cells/cm^2 for 6 days in CDM supplemented with 10 ng/ml human Heregulinβ -1, 200 ng/ml LONG^R R3 human IGF-I, 20 μM SB431542 (R&D), and 500 nM LDN193189 (Sigma). Media was changed every 24 hours.

qRT-PCR

Cells were harvested using Accutase and lysed with E.Z.N.Z RNA isolation kit (Omega) following the manufacturer's protocols. Isolated total RNA was quantitated with a Biotek Synergy 2 plate reader.cDNA was synthized using 1 µg of RNA via the Iscript cDNA synthesis kit (Bio-Rad) following the manufacturer's protocols. The cDNA was then diluted to a final volume of 500 µL with molecular grade water. ΔΔCt qRT-PCR analysis was performed on a ViiA7 Real-Time PCR System (Life Technologies) in a 384 well plate with a reaction of 5 µL TaqMan Universal PCR Master Mix No AmpErase UNG (appliedbiosystems), 0.5 µL TaqMan primer (Life Technologies), 0.5 µL molecular grade water and 4 µL cDNA. Expression of each transcript was normalized to 18s ribosome, performed in triplicate, and plotted as the mean ± standard deviation.

Western Blotting

Cells were washed with ice cold DPBS, collected with a cell scrapper, pelleted via centrifugation at 1000 rpm for 4 min, flash-frozen in liquid nitrogen and stored at -80 °C as cell pellets. Cell pellets were resuspended and lysed on ice for 30 min in 1x RIPA lysis buffer (Sigma), 1x protease inhibitor (Roche), 1x phosphatase inhibitor (Calbiochem) and 1 mM DTT. Lysates were centrifuged at 20,000 x g for 10 min at 4 °C with the supernatant collected. Protein concentrations within supernatants were determined via Bradford assay at 595 nm on a Biotek Synergy 2 and boiled with 1:1 with Laemmli buffer (Bio-Rad) at 95 °C for 5 min. For running bis-Tris precast gels (Life Technologies), ~30 µg of total protein were loaded into each lane. Blotted membrane was blocked with 2% non-fat blocking reagent (Bio-Rad) in 0.05% TBST (Tween 20) for 1 hour, followed by 1:1000 diluted primary antibody incubation overnight at 4

°C. Protein levels on membranes were detected by SuperSignal ECL reagent (Thermo) as manufacture described.

Immunostaining

Cells were washed with DPBS and fixed with a 4% paraformaldehyde (Electron Microscopy Sciences) in DPBS solution for 10 min. After 3 washes with ice-cold DPBS, cells were permeabilized with 0.01% PBST (Triton X100). After 3 more washes, cells can be store at 4 °C for 2 weeks.

For staining, slides were brought to RT and then blocked with a PBST containing 10% donkey serum (Equitech-Bio) for 1 hr at room temperature. Primary antibodies (see Table 4.2 for antibodies used and dilutions/concentrations) were prepared in blocking buffer and incubated overnight at 4 °C. After 3 washes with DPBS, secondary antibodies were incubated for 1 hr at room temperature in the dark. Following removal of secondary, 1 µg/mL DAPI 4 (Sigma) in DPBS was added to cells for 5 mins. After 3 washes in DPBS, coverslips were mounted to slides with ProLong Diamond Antifade (Invitrogen). A Leica DM6000B microscope and Zeiss LSM 710 confocal microscope were used to obtain images.

ATAC-seq

ATAC-seq was performed as previously described (Buenrostro et al., 2013; Buenrostro et al., 2015). Briefly, a total of 50,000 cells were washed once with 50 µl of cold PBS and resuspended in 50 µl of lysis buffer (10mMTris-HCl pH 7.4, 10mMNaCl, 3 mMMgCl2, 0.2% (v/v) IGEPAL CA-630, 0.1% Digitonin). The suspension of nuclei was then centrifuged for 10 min at 500 g at 4 C, followed by the addition of 50 µl of transposition reaction mix (25 µl TD buffer, 2.5 µl Tn5

transposase and 22.5 µl nuclease-free H2O) of Nextera DNA library Preparation Kit (96 samples) (FC-121-1031, Illumina). Samples were then PCR amplified and incubated at 37 C for 30min. DNA was isolated using a DNA concentrator column Kit (Zymo). ATAC-seq libraries were first subjected to 5 cycles of pre-amplification. To determine the suitable number of cycles required for the second round of PCR the library was assessed by quantitative PCR as described (Buenrostro et al., 2015) and the library was then PCR amplified for the determined number of cycles. Libraries were further purified with by Mag-bind beads (Omega) to remove excessive primers. Finally, the ATAC library was sequenced on a NextSeq500 using a NextSeq500 High Output SE75 flow cell according to manufacturer's instructions.

ChIP-qPCR and ChIP-Seq

For direct DNA binding detection, cells were fixed with 1% formaldehyde for 10 min at RT. For indirect DNA binding detection, cells were fixed with 2 mM EGS for 20 min, followed by 1% formaldehyde for 10 min. Fixation was quenched with 200 mM Glycine. Fixed cells were centrifuged and stored at -80 °C for long term.

At day 1 of ChIP, 50 µl of Dynabeads Protein G (Invitrogen) was washed for 3 times with 0.5% PBST (Tween-20) and blocked with PBST containing 0.5% (w/v) BSA (Sigma) for 30 min. For primary antibody (see Table xx) conjugation, 5 µg of antibody was incubated with 500 µl of PBST resuspended beads for over 2 hours at 4 °C. Crosslinked cell pellets were resuspended in 1 mL of lysis buffer, 10 mM Tris-HCl ph 8.0 (Invitrogen), 100 mM NaCl (Ambion), 1 mM EDTA (Ambion), 0.5 mM EGTA, 0.1% Na-Deoxycholate, 0.5% N-lauroylsarcosine and 1x protease inhibitors (Sigma), for 30 min on ice. Lysates were sonicated in a 1 mL milliTUBE (Covaris) on a Covaris S220 for 20 min at 200 cycles/burst, a peak power of

140 and a duty factor of 5%. Sonicated lysates were supplemented with 1% Triton X-100 and centrifuged at max speed for 5 min at 4 °C. For ChIP-qPCR, 1% of the supernatant was saved as input. Antibody conjugated beads were washed 3 times with washing buffer and then added to a sonicated lysate and incubated overnight at 4 °C while rotating.

At day 2 of ChIP, beads were then washed twice with 1 ml of 20 mM Tris-HCl ph 7.4 (Invitrogen), 150 mM NaCl, 0.1% SDS (KD Medical), 1 % Triton X-100 and 2 mM EDTA. Then washed twice with 1 mL of 10 mM Tris-HCl ph 7.4, 250 mM LiCl, 1% Triton X-100, 0.7% DOC (Sigma) and 1 mM EDTA. For ChIP-seq, beads were transferred to a clean tube and tagmentated with 1 µl of Tn5 (illumina) for 2 min at 37 °C. Finally, the beads were washed twice with 1 mL of 10 mM Tris-HCl ph 8.0, 0.2 % Tween-20 (Bio-Rad) and 1 mM EDTA. Crosslinking was reversed by incubating beads in 100 µl of 10 mM Tris-HCl pH 8.0, 0.5% SDS, 300 mM NaCl, 5 mM EDTA and 3.2 units of Proteinase K (New England Biolabs) for 1 hr at 55 °C followed by 8 hr at 65 °C.

Eluted DNA was recovered from collected with a DNA concentrator column (Zymo) following the manufacturer's protocol. For ChIP-qPCR, resultant DNA was diluted to a final volume of 150 µl with molecular grade water. ChIP-DNA was quantified via qPCR on a ViiA7 real time PCR system in a reaction consisting of 4 µL DNA, 5 µl SYBR Green (Bio-Rad), 0.5 µl of 10 mM right primer, 0.5 µl of 10 mM left primer and 0.2 µl of Rox Low (Bio-Rad). See Table 2 for ChIP-qPCR primers.

For ChIP-seq, the recovered DNA was eluted with 16.5 µl of H_2O. 1 µl was amplified by adapter primers to determine number of cycles for library amplification. The ready-to-run libraries were submitted to Georgia Genome Facility for downstream sequencing.

Proximity Ligation Assay

Experiments were conducted according to the commercially available Duolink PLA kit (Sigma). Antibody used for this experiment was listed in Table 2. The resultant PLA foci were analyzed using a customized Image J script (Abràmoff et al., 2004).

dCas9 Tethering Assay

The dCas9 fusion constructs were generated by modifying the pAW90.dCas9-YY1 plasmid (Addgene #104373). Briefly, total RNAs were collected from H9 human embryonic stem cells and reverse transcribed using oligo dT primers. The cDNA of NANOG(NM_024865) and OCT4A (NM_002701) were amplified using custom primers:

BamHI-NANOG Forward:

5'-TATTGGATCCGGACGGGCTAGTGTGGACCCAGCTTGTCCCCAAAGCT-3'

BsrGI-NANOG Reverse:

5'-CGCTTGTACAGTTAATCACGTCTTCAGGTTGCATG-3'

BamHi-OCT4 Forward:

5'-TATTGGATCCGGACGGGCTGCGGGACACCTGGCTTCGGATTTC-3'

BsrGI-OCT4 Reverse:

5'-CGCTTGTACAGTTAATGTTTGAATGCATGGGAGAG-3'

Next, the YY1 ORF was removed and replaced by either NANOG or OCT4A cDNA to generate Lenti-dCas9-NANOG and Lenti-dCas9-OCT4 plasmids. The pAW91-dCas9(Addgene #104372) alone was used as an empty control.

For Virus production, HEK293 cells were grown on a 15 cm dish to reach 70% confluency and then transfected with lipofectamine 2000 (ThermoFisher 11668019) and a mixture of 10 μg of pAW90 or pAW91, psPAX2 (Addgene #12260) and pMD2.G (Addgene

#12259) (ratio of 3:2:1). After 12 hours, media was replaced. The supernatant containing lentivirus was collected 36 hours post transfection. The lentivirus was concentrated by Lenti-X concentrator (Clontech 631231) according to manufacture instructions. Concentrated virus was added to H9-hESCs in the presence of polybrene at 12 μg/ml. 48 hours post viral transduction, cells were treated with Blasticidin-HCl (Sigma SBR00022) at 1.5 μg/ml until all non-transduced cells died.

To tether the dCas9 fusions to desired genomic loci, different gRNAs targeting promoters and putative enhancers (Table Sx) were cloned into pAW12.lentiguide-GFP (Addgene #104374) and pAW13.lengtiguid-mCherry (Addgene #104375). Lentivirus were prepared as described above. Transduced hESCs were expanded and the GFP/mCherry double positive population was sorted by Beckman Coulter MoFlo platform at Cytometry Core Facility at University of Georgia. The resultant cell lines were analyzed by Western Blot, ChIP-qPCR and RT-qPCR as described previously in the methods.

Auxin-AID Inducible Knock-Out Cell Lines

We used px330 plasmid to construct CRISPR/Cas9 vectors as described (Natsume et al., 2016). The gRNA oligos for NANOG or OCT4 are as following:

NANOG gRNA F1: CAC CGC TCA ATT TCA GTC TGG ACA C

NANOG gRNA R1AAA CGT GTC CAG ACT GAA ATT GAG C

NANOG gRNA F2: CAC CGG TCT TCA GGT TGC ATG TTC A

NANOG gRNA R2: CAC CGG TCT TCA GGT TGC ATG TTC A

OCT4 gRNA F1: CAC CGC CCT TCT AGG AAT GGG GGA C

OCT4 gRNA R1: AAA CGT CCC CCA TTC CTA GAA GGG C

OCT4 gRNA F2: CAC CGC CCT GTC TCC GTC ACC ACT C

OCT4 gRNA R2: AAA CGA GTG GTG ACG GAG ACA GGG C

An experimental overview of the construction of donor vectors is presented in Supplemental Figure 3.9A. To construct the donor vectors, short homology arms were synthesized and cloned into pUC19 backbone by BamHI digestive site. To deliver the gRNA template bearing pX330 and donor vectors into hESC, a mixture of 3 μg of vectors (gRNA1: gRNA2: donor = 1:1:1) were transfected using Neon electroporation system (Thermo). 48 hours post transfection, cells were screened using 0.5 ng/ml of puromycin or 100 μg/ml of G418 according to donor design. Once confluence, cells were sorted using GFP channel or in combination of mCherry channel as single cells into a 96-well plate for clonal expansion. Each clone was tested by genomic PCR for correct editing.

Cartoon figures

Cartoon figures were created with BioRender.com.

SEQUENCING DATA PROCESSING

ATAC-seq & ChIP-seq Data Processing and Peak Calling

ATAC-seq and ChIP-seq fastq data were mapped onto hg19 genome build using bowtie2 (Langmead and Salzberg, 2012) with '--local -N 1 --phred33-quals' options. Reads mapped to unwanted regions, such as ChrM were removed using 'sed '/chrM/d;/random/d;/chrUn/d;/chrY/d' '. After removal of these regions, samtools (Li et al., 2009) was used to sort and isolate uniquely mapped reads using '-bS -F 4 -h' options. Resultant BAM files were converted to bigwig files for visualization using deeptools or called for peaks using MACS2 (Feng et al., 2012).

For peak calling, each sample was processed by MACS2 callpeak command with parameters '-keep-dup 1 --nomodel --shift 75 --extsize 150 –nolambda'. The peak summits were extended by 250 bp on either side, filtered to remove ENCODE hg19 blacklist (https://sites.google.com/site/anshulkundaje/projects/blacklists). To obtain the reproducible ATAC peaks, individual replicates were merged using bedtools. Next we measured ATAC signals in CPM (counts per million reads) using EAseq (Lerdrup et al., 2016). Read counts within 500 bp window centered on merged peaks. To remove background noise, a Tn5-tagmented whole genome sample from hESC was introduced (analogs to input sample in ChIP-seq experiments). A signal-to-background ratio 2 was set as the cutoff to exclude false positive ATAC peaks. Reproducible peaks were those with non-zero CPM values in both replicates and were used for downstream analysis.

Motif discovery analysis

Motif analysis was performed using HOMER (Heinz et al., 2010)using default settings. Motifs were only kept if the -log10(P-value) > 100 and match score > 0.9.

HiChIP Data Processing

HiChIP fastq data were first mapped to hg19 using HiC-Pro with default settings (Servant et al., 2015). HiChIP loops were called from merged allvalidpairs generated by HiC-Pro using FitHiChIP (Bhattacharyya et al., 2019), the loops anchors must overlap with respective ChIP-seq peak by at least 1 bp. Loops were called using '-Peak2ALL -L20000 –U200000 -P2PBckgr_0 -Coverage_Bias_1' options. Bedtools (Quinlan and Hall, 2010) was used to intersect loops with open chromatin peaks using '-type either' option. To visualize HiChIP loops, the bedpe files

were loaded to the IGV browser (Thorvaldsdóttir et al., 2013). For APA analysis (Durand et al., 2016), juicer tools were used as '-r 5000 -k VC_SQRT -u '.

REFERENCES

Abràmoff, M.D., Magalhães, P.J., and Ram, S.J. (2004). Image processing with ImageJ. Biophotonics international *11*, 36-42.

Bhattacharyya, S., Chandra, V., Vijayanand, P., and Ay, F. (2019). Identification of significant chromatin contacts from HiChIP data by FitHiChIP. Nature communications *10*, 1-14.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods *10*, 1213.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC‐seq: A Method for Assaying Chromatin Accessibility Genome‐Wide. Current Protocols in Molecular Biology, 21.29. 21-21.29. 29.

Cliff, T.S., Wu, T., Boward, B.R., Yin, A., Yin, H., Glushka, J.N., Prestegaard, J.H., and Dalton, S. (2017). MYC Controls Human Pluripotent Stem Cell Fate Decisions through Regulation of Metabolic Flux. Cell Stem Cell *21*, 502-516.e509.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell systems *3*, 95-98.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nature protocols *7*, 1728-1740.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Lerdrup, M., Johansen, J.V., Agrawal-Singh, S., and Hansen, K. (2016). An interactive environment for agile analysis and visualization of ChIP-sequencing data. Nature structural & molecular biology *23*, 349-357.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078-2079.

Natsume, T., Kiyomitsu, T., Saga, Y., and Kanemaki, M.T. (2016). Rapid protein depletion in human cells by auxin-inducible degron tagging with short homology donors. Cell reports *15*, 210-218.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome biology *16*, 1-11.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics *14*, 178-192.