# DISSECTING TRANSCRIPTIONAL NETWORK IN MOUSE EMBRYONIC STEM CELLS

**FANG FANG**

**(*M.Sci., Wuhan University*)**

**(*B.Sci., Wuhan University*)**

**A THESIS SUBMITTED**

**FOR THE DEGREE OF DOCTOR OF PHILOSOPHY**

**IN COMPUTATION AND SYSTEMS BIOLOGY (CSB)**

**SINGAPORE-MIT ALLIANCE**

**NATIONAL UNIVERSITY OF SINGAPORE**

**2011**

# ACKNOWLEDGEMENTS

I am indebted to many people for their constant support during my PhD training. I can’t even imagine myself write this part of thesis without their helps along this exciting but strenuous journey.

Special thanks to my advisor, Paul Matsudaira, who has been a tremendous supportive, encouraging and inspiring advisor. I would have quitted my PhD career if he was not so helpful and supportive when I was in my career crash. I am also grateful for his guidance and patience for ensuring the continuous progress of my research progress.

I would like to thank my co-advisor, Harvey Lodish, who is, although, for most of time, thousands of miles away from Singapore, for his encouragement and instruction for my research and future career planning.

I have also received a lot of critical and helpful advice from my thesis committee members, consisting of Gong Zhiyuan, Yu Hao, Chan Woon Khiong and Neil Digby Clarke.

Thanks also go to my labmates, in particular Chen Xi and Xu Yifeng, who have been inspiring mentors and good friends and contributed a lot to the work described here.

My PhD training is supported by funding and a graduate fellowship from the Singapore-MIT Alliance and I am grateful for all the administrative support during last four and half years. Especially the independent research funding for PhD students provide us a lot of freedom to carry out our own ideas.

Last but not the least; I am deeply grateful to my family, who has unconditionally given me invaluable love and support for me to overcome all the difficulties during PhD trainings. I can only say no success in life would have been possible without them.

# TABLE OF CONTENTS

# SUMMARY

Embryonic stem (ES) cells are featured by their ability of self-renewal and pluripotency. Although external signalling pathways as well as epigenetic signatures have been shown necessary for ES cells maintenance, considerable evidence indicates that naïve pluripotency of ES cells is dependent on their specific transcription network that regulate the gene expression programs in a spatially and temporally orchestrated and precise pattern. Delineating the transcription network within ES cell system should be a fascinating science challenge that would provide new insights into the fundamental nature of pluripotency as well as advance its application in regenerative medicine. My thesis project has applied computational and systems biology tools to dissect transcriptional network of mouse ES cells, and has extensively expanded our knowledge of the network by introducing novel self-renewal and pluripotency associated transcription factors into the known core regulatory circuit. Furthermore, I looked into coactivators that facilitate the functions of transcription factors and further linked coactivator regulation to higher-order chromatin structure. This is the first study of in vivo higher-order chromatin organization that is unique to pluripotent cells based on the binding sites of transcription factors and coactivators, adding a new content to the list of unusual findings regarding the chromatin structure in ES cells as well as a new layer to the ES cell specific transcriptional network.

## LIST OF TABLES

# LIST OF FIGURES

# CHAPTER I: Literature review

## 1.1 Derivation and culture of pluripotent stem cells

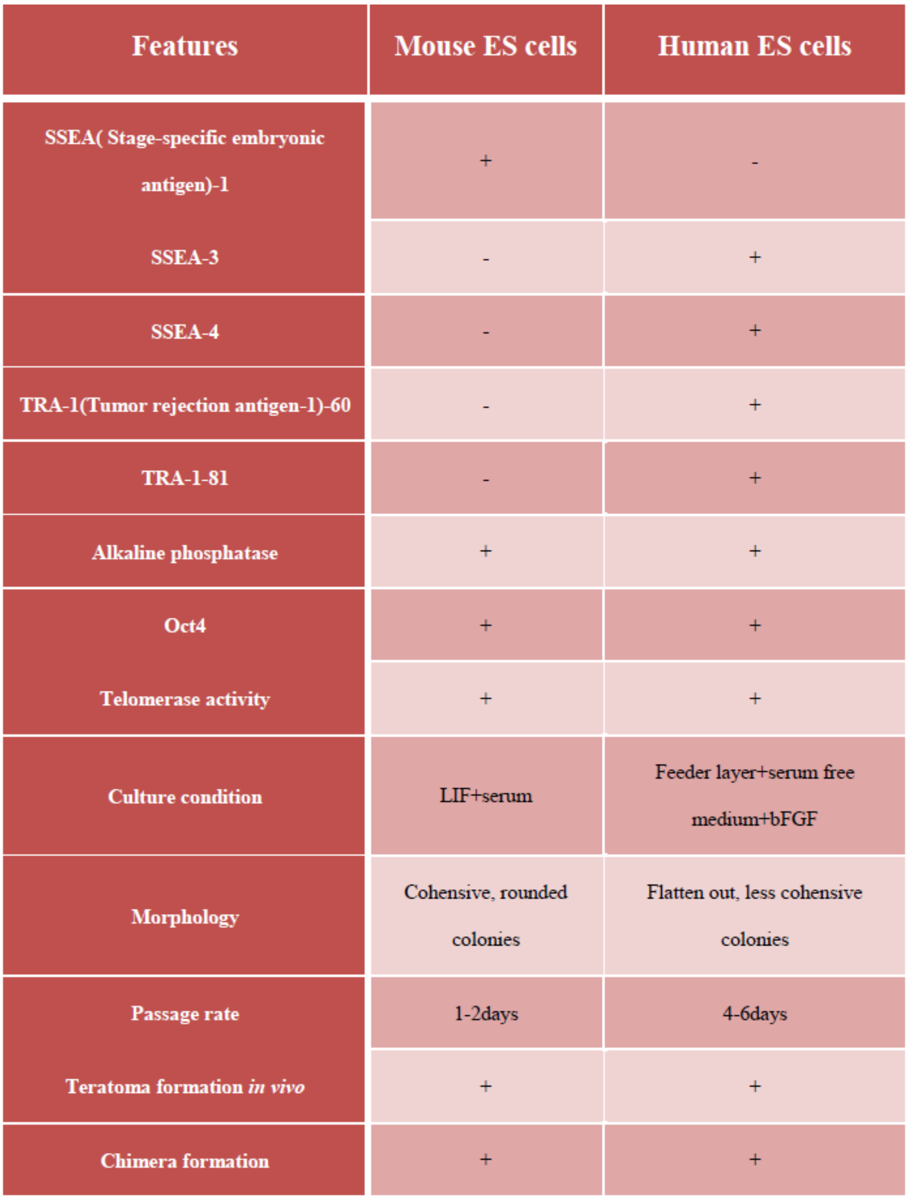
Although the era of embryonic stem (ES) cells is considered to begin officially in 1981, when mouse ES cells were first isolated and successfully cultured *in vitro* as self-renewal and pluripotent cell lines by two groups (Evans and Kaufman, 1981; Martin, 1981), the adventure to search for exogenous cells that are capable of recapitulating early embryogenesis had started earlier. At the beginning, researchers had tried to manipulate early mouse embryogenesis by embryonal carcinoma cells (EC) cells. EC cells are the pluripotent stem cells from teratocarcinomas, which are highly malignant tumors that occasionally occur in a gonad of a fetus and are comprised of a mixture of a large population of undifferentiated cells and differentiated cells of multiple lineages. EC cells could be maintained indefinitely with mitotically inactivated embryonic fibroblast *in vitro*, and is able to give rise to cells of multiple lineages (Finch and Ephrussi, 1967). However, further studies have found out that EC cells were karyotypically abnormal or unable to differentiate normally (Berstine et al., 1973; Papaioannou et al., 1975), which led to the efforts to isolate a new type of stem cells, embryonic stem cells, from the mouse embryo.

Embryonic stem cell lines are derived from the inner cell mass (ICM) of the mouse blastocyst at embryonic day 3.5 (E3.5). These cells were initially maintained in culture as self-renewal and pluripotent cell lines in either EC cell-conditioned medium (Martin, 1981), or in a co-culture system in which cells were grown on a layer of mitotically inactivated mouse embryonic fibroblast (MEF) feeder population in the presence of blood serum (Evans and Kaufman, 1981). Since medium conditioned by feeder cells is sufficient to sustain the self-renewal and pluripotent state of mouse ES cells, the presence of a diffusible factor has been postulated. Further research has found out that under serum-free culture conditions, specific cytokines promoted the maintenance of ES cells. Leukaemia inhibitory factor (LIF) and bone morphogenetic protein 4 (BMP4), a member of the transforming growth factor (TGF) β family, were required to sustain ES cells indefinitely in culture (Chambers and Smith, 2004; Ying et al., 2003), as in the absence of them, ES cells identity cannot be preserved, which led to profound differentiation.

Similar to mice, ES cells have also been established from other primates (Thomson and Marshall, 1998), and the extensive studies and characterization of these ES cells finally led to the derivation of human ES cells, which hold tremendous potential for the development of cell transplantation therapies for regenerative medicine and the treatment of various human diseases. The first successful human ES cell line was derived by Thomson group (Thomson et al., 1998). They isolated human ES cells from the blastocyst derived from day 5 to day 8 blastocysts after *in vitro* fertilization (IVF) and plated them onto mitotically inactivated MEF cells. Under *in vitro* conditions, they exhibit the prolonged undifferentiated proliferation and differentiation potential, which are the two basic characteristics of ES cells. Two years later, Reubinoff et al. (Reubinoff et al., 2000) confirmed that human ES cells could be efficiently derived from surplus embryos. Since then, rapid progress has been achieved and numerous studies have described the derivation of new human ES cell lines and optimized the methods of growing undifferentiated human ES cells.

Similar to mouse ES cells, human ES cells can also be cultured under feeder free conditions, however, instead the requirement of LIF and BMP4, human ES cells rely on Activin and FGF2 for the maintenance, suggesting that mouse ES cells may not be equivalent to human ES cells in the developmental stage. In fact, besides culture conditions, human and mouse ES cells differ in a few other aspects, such as morphology, gene expression profile and epigenetic landscapes, as shown in Table 1.

**Table 1.1: Comparison of mouse and human ES cells.**



| Features | Mouse ES cells | Human ES cells |
|---|---|---|
| SSEA( Stage-specific embryonic antigen)-1 | + | - |
| SSEA-3 | - | + |
| SSEA-4 | - | + |
| TRA-1(Tumor rejection antigen-1)-60 | - | + |
| TRA-1-81 | - | + |
| Alkaline phosphatase | + | + |
| Oct4 | + | + |
| Telomerase activity | + | + |
| Culture condition | LIF+serum | Feeder layer+serum free medium+bFGF |
| Morphology | Cohensive, rounded colonies | Flatten out, less cohensive colonies |
| Passage rate | 1-2days | 4-6days |
| Teratoma formation *in vivo* | + | + |
| Chimera formation | + | + |

Recently, Brons et al., (2007) demonstrated that pluripotent stem cells may be derived from the late epiblast layer (embryonic day 5.5–7.5) of post implantation mouse embryos, and these cells are called EpiSCs (post-implantation epiblast derived stem cells) (Brons et al., 2007). These cells display profound differences from mouse ES cells in the combination of growth factors that maintain their pluripotent states. They can only be cultured using chemically defined media supplemented with FGF2 and Activin, and they display flat colony morphology, which resemble the culture conditions and morphology of human ES cells. More importantly, upon stimulation by Activin and Fgf2, mouse ES cells can develop to EpiSCs, indicating that EpiSCs are in a more advanced developmental stage than are ES cells and it may be at the same developmental stage as human ES cells. Although EpiSCs are able to form teratomas, they contribute very little to the germline in chimeric mice.

FAB-SCs, another form of pluripotent stem cells, can be derived from mouse blastocysts in the presence of bFGF, activin, BIO (which is a GSK3 kinase inhibitor) and an anti-LIF antibody (Chou et al., 2008). These cells cannot differentiate as mouse ES cells unless stimulated by LIF and BMP4 or force expression of E-cadherin, suggesting that these cells are in a latent state of pluripotency.

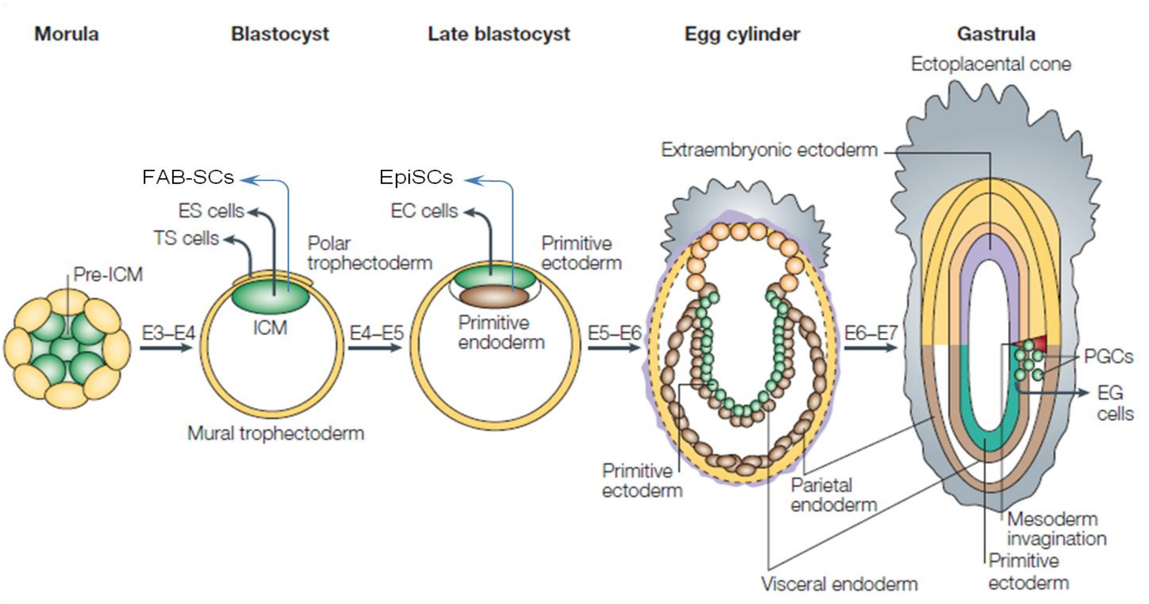


**Figure 1.1. Origin of stem cells during mammalian embryogenesis.** In this figure, the pluripotent cells of the embryo are tracked in green. From left to right, the morula-stage mouse embryo (embryonic day 2.5; E2.5) holds a core of pre-ICM cells that turn into ICM cells at cavitation/blastulation (E3–E4). At this stage, ES cell and Trophoblast stem (TS) cell lines can be derived *in vitro*, and implantation occurs *in vivo*. FAB-SCs can be derived from mouse blastocysts in combination of bFGF, activin, BIO and an anti-LIF antibody. As the blastocyst fully expands (and undergoes implantation *in vivo*), the ICM delaminates giving rise to a primitive ectoderm and a primitive endoderm layer. At this stage, pluripotent cell lines that are known as embryonal carcinoma (EC) cells can be derived from the primitive ectoderm. EpiSCs are derived from E5.5–E5.75 post-implantation epiblasts in the presence of activin and bFGF. At E6 and subsequent stages, the experimental ability to derive ES Cells, TS cells and EC cells from the mouse embryo is progressively lost, and the *in vivo* embryo will start gastrulating. (Adapted from Bioani and Sholer et al. 2005)

Another source of pluripotent stem cells is provided by induced pluripotent stem cells (iPSCs) from somatic cells by enforced expression of a few pluripotency-associated transcription factors. The discovery of induced pluripotency can be traced back to the work of somatic cell nuclear transfer (SCNT) that first established by Briggs and King (Briggs and King, 1952; King and Briggs, 1955). The cloning of Dolly sheep further showed that the genome of even terminally differentiated cells preserve the potential to develop into an entire organism (McLaren, 2000; Wilmut et al., 1997). However, SCNT is technically challenging and the cloned animals always exhibit abnormalities in gene expression and phenotype. An alternative approach is developed by *in vitro* hybridization between somatic and pluripotent cells. The hybrid cells by fusion of EC cells with somatic cells, such as thymocytes, resemble EC cells in terms of biochemical properties and differentiation potential, while lose the features of somatic cells (Miller and Ruddle, 1976, 1977), indicating that some soluble regulatory factors in EC cells confer a pluripotent state to somatic cells. However, hybrid cells lack therapeutic potential because of their abnormal ploidy and the presence of nonautologous genes from the pluripotent parent. A great breakthrough was achieved by Yamanaka and Takahashi in 2006 (Takahashi and Yamanaka, 2006). The original idea was to induce pluripotency from somatic cells by enforced expression of specific transcription factors, which was based on the observation that lineage-associated transcription factors were able to change the cell fate when ectopically expressed in certain heterologous cells (Davis et al., 1987; Laiosa et al., 2006; Xie et al., 2004; Zhou et al., 2008). To induce pluripotency, they performed an elegant screen for factors within a pool of 24 pluripotency-associated candidate genes and came out a core set of four genes, Oct4, Sox2, Klf4 and c-Myc, called “Yamanaka genes”, which are minimally required to be enforced expressed for reprogram mouse fibroblasts to iPSCs. The resultant mouse iPSCs have passed the most stringent test of pluripotency, tetraploid complementation, a technique in which iPSCs are injected into a tetraploid blastocyst and are shown to contribute to the generation of an entire living mouse (Kang et al., 2009; Zhao et al., 2009). The iPSCs field has progressed at a breathtaking pace in the last 5 years, including derivation of iPSCs from other species, such as human; optimization of the efficiency of iPSCs generation; development of virus-free factors delivery system and establishment of disease-specific iPSCs. In addition to being an exciting academic research model to study cellular development, iPSCs hold significant therapeutic potential for regenerative medicine, disease modeling and drug development. Notwithstanding these achievements, iPSCs technology remains in its infancy and a better understanding of the reprogramming process is required in order to develop more efficient strategies for pluripotency induction and a careful analysis of the genomic and epigenomic characteristics of iPSCs, as well as the development of a robust protocol for directed differentiation are required for future utilities of iPSCs in clinic medicine.

Although different types of pluripotent stem cells have been generated and broadly expand our knowledge for pluripotency, the biggest challenge remains to produce mature, functional and pure derivatives of cell types that can be utilized for transplantation purposes. To facilitate these developments, a large amount of efforts is put to get a comprehensive understanding of the biology of ES cells including genes that are important for the maintenance of ES cells, especially human ES cells. However, due to the ethical challenge of the source of human ES cells and the inability to test pluripotency of human ES cells by chimera formation, extensive work has been carried out initially on mouse ES cells. Mouse ES cells are easier to manipulate and have been extensively characterized for 20 more years than human ES cells; therefore the discovery on mouse ES cells will eventually shed light on the understanding of human

ES cells. In my thesis work, I focus all my studies on mouse ES cells, and particularly on the transcriptional regulation of these cells, to understand the molecular mechanisms underlying pluripotency.

### 1.2 Characteristics of mouse embryonic stem (ES) cells

Mouse ES cells are well known for two distinguished properties: self-renewal and pluripotency. Self-renewal is the ability of ES cells to proliferate continuously in culture in undifferentiated state (Smith and Benchimol, 1988). More importantly, unlike EC cells and other primary cell lines that can only be passaged for several times before senesce, these cells can be passaged for years while maintaining normal karyotypes (Keller, 2005).

The second property of ES cells is that they recapitulate full developmental potential when injected into mouse blastocysts, contributing cells to all three germ layers and to the germline of chimeric animals. It is known as pluripotency, which has attracted huge interest of numerous researchers because of its promising applications in regenerative medicine. The golden rule to judge pluripotency of ES cells is by their ability to integrate into the ICM of the blastocysts and contribution to germline formation. So far, pluripotency has only be proven conclusively in mouse ES cells, as they can completely integrate into the blastocyst, after transplantation, and exhibit high efficiency of chimera formation and germline transmission. ES cells can also be induced to differentiate *in vitro* by a number of strategies. By cultivation *in vitro* as 3D aggregates called embryoid bodies (EBs), ES cells can differentiate into derivatives of endoderm, mesoderm, and ectoderm. Removal from the self-renewing environment by taking out cytokines, such as LIF or BMP4, from culture medium triggers intrinsic differentiation programs that resembles a developmental course that was interrupted when the ICM was extracted from the blastocyst. Moreover, adding in soluble molecules, such as retinoic acid, will stimulate ES cell differentiation as well.

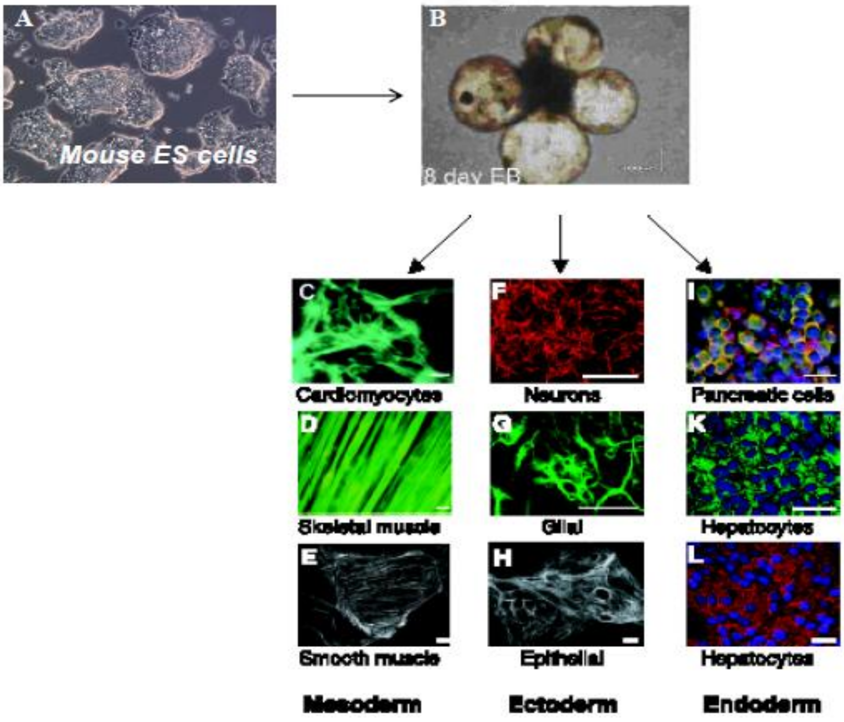


**Figure 1.2. Differentiation of mouse ES cells by EB formation.** (A) Mouse ES cells cultured under feeder free condition; (B) Embryoid bodies (EB) are formed 8 days after suspension culture. (C-E) Examples of mesoderm lineage: Cardiomyocytes (C), Skeletal muscles (D) and Smooth muscles (E); (F-H) Examples of ectoderm lineage: Neurons (F), Glial (G) and Epithelial (H); (I-L) Examples of endoderm lineage: Pancreatic cells (I), Hepatocytes (K-L).

However, the therapeutic use of ES cells will require more precise control over this process in order to make these cells differentiate efficiently and strictly to a specific lineage. Intensive work has been conducted to the field of directed differentiation to influence the lineage commitment of ES cells *in vitro*. Various strategies involving supplementation of growth factor cocktail, cell co-cultures, conditioned medium and specific gene transfection are used to drive lineage specific emergence (Fair et al., 2003; Ogawa et al., 2005; Wells and Melton, 1999; Zhou et al., 2007c). Nevertheless, the improved knowledge of the molecular mechanisms governing ES cell maintenance and differentiation towards specific lineage are desired to better facilitate direct differentiation of ES cells for therapeutic applications.

### 1.3 Application of ES cells

As we have discussed above, the most extraordinary property of ES cells is their ability to re-enter embryogenesis. Indeed, a major interest of ES cells to the scientific community is their utility as cellular vehicles for engineering of the mouse genome. Mouse ES cells can be injected into the blastocysts and integrate into the ICM cells to produce viable chimeras. The derivation of transgenic mice from genetically modified ES cells was first reported in 1984 (Bradley et al., 1984). Afterwards, ES cell technology has been most often used to produce null mutants (gene knockouts) through homologous recombination (Thomas and Capecchi, 1986) for the *in vivo* study of gene function during development and this can even be achieved in a conditional knockout manner. Moreover, they can also be used to introduce subtle genetic modifications down to the level of single nucleotide mutation in endogenous mouse genes. Transgenic mice derived from ES cells has not only revolutionized basic biological research through the creation of genetically altered animals, but also permits the evaluation of therapeutic strategies in models of human disease, as well as the investigation of disease progression in a manner not possible in human subjects.

The discovery of human ES cells has been considered as the key tool for understanding most of the fundamental questions in both basic and clinical human biology. Human ES cells may allow scientists to investigate how early human cells become committed to specific lineages and differentiated into the myriad functional cell types that build up tissues and organs of the entire body. The knowledge gained will greatly accelerate our understanding of the causes of birth defects and thus lead directly to their possible prevention. Human ES cells can also be applied as a valuable *in vitro* model system to study diseases that only occur in human or have significant difference between human and other species, such as HIV, HCV. In the clinic trail, they could be used to create an unlimited supply of cells, tissues, or even organs that could be used to restore function. Human ES cell-derived progeny have been successfully exploited in animal models of spinal cord injury (Keirstead et al., 2005; Sharp et al., 2010), retinopathies (Lamba et al., 2009), and Parkinson’s disease (Yang et al., 2008). And this idea is greatly promoted by the generation of patient-specific iPSCs. Disease-specific iPSCs have already been created from patients suffering from amyotrophic lateral sclerosis (Dimos et al., 2008), juvenile onset type 1 diabetes mellitus (Park et al., 2008a), Parkinson’s disease and spinal muscular atrophy (SMA) (Ebert et al., 2009). Critically, the pathophysiology of SMA could be recapitulated in motor neurones derived from patient-specific iPSCs. In the long run, these patient-specific iPSCs may be ideally suited for cellular therapy, given that they are derived from the patient to be treated, thus minimizing the risk of immune rejection. However, it is noteworthy that these iPSCs, however, are only the starting point for the preparation of cells for clinic trials, as therapeutic cells should be differentiated cell lines with the characteristics proper of the various tissues (muscle, neural, epithelial, haematic, germinal, etc.). Methods for obtaining therapeutic cells from human ES cells or iPSCs are still being studied and even if successful for some specific cell types, a testing assay to certain that the inoculation or therapeutic implant was free of stem cells is also crucial, as the remnant stem cells may result in tumors.

### 1.4. Molecular characteristics of ES cells

The maintenance of ES cells engages complex and precisely controlled molecular and cellular regulatory machinery. While self-renewal and pluripotency associated genes are up-regulated to maintain the undifferentiated state of ES cells, genes that induce differentiation are suppressed but poised for subsequent expression during cellular differentiation. Tremendous effort has been applied to uncover the molecular mechanisms governing self-renewal and pluripotency in ES cells, and based on our current knowledge, the balanced state of ES cells is achieved through the complex interplay of cell cycle regulation, signaling pathways, epigenetic modification, small regulatory RNAs as well as ES-specific transcriptional network.

#### 1.4.1. *Cell cycle regulation*

Cell cycle program of mouse ES cells is characterized by extraordinarily rapid proliferation rate and a pluripotent cell specific cell cycle structure, which is controlled by an unusual mode of cell cycle regulation. The work from the last few years has revealed the importance of cell cycle regulation to the maintenance of ES cells, as the process of self-renewal requires the coordination of cell cycle progression and cell-fate determination (self-renewal versus commitment). A few transcription factors as well as cell cycle regulators appear to be critical to this regulation.

Mouse ES cells have relatively short cell cycle period compared with differentiated cells, with ~8 to 10 hours total generation time, and an unusual cell cycle structure, with a reduction in the duration of G1 phase. Although human ES cells share a similar cell cycle structure, their generation time is significantly slower (~32-38 hours; (Dalton, 2009; Ohtsuka and Dalton, 2008) indicating that a short division may not be a pre-requisite for pluripotency. This is supported by the study showing that slowing cell cycle of mouse ES cells with chemical inhibitors has no measurable impact on the maintenance of ES cells (Stead et al., 2002). Instead, other observations suggest that mechanisms making up the specific cell cycle structure are more crucial to the ES cell maintenance. The short G1 phase allows ES cells to be less responsive to the differentiation signals sent by certain mitogenic signaling pathways that are active and act as potent differentiation inducer during G1 phase in somatic cells. It has been shown that mitogenic signaling pathways inhibit mouse ES cells self-renewal and promote their differentiation, while self-renewal of mouse ES cells is enhanced by the addition of inhibitors of mitogenic signaling pathways to the culture medium (Burdon et al., 2002; Burdon et al., 1999). Furthermore, the extended S phase may also shield cells from extrinsic differentiation signals by maintain chromatin in an “open” euchromatic state to facilitate rapid activation or repression of genes (Filipczyk et al., 2007;

Herrera et al., 1996).

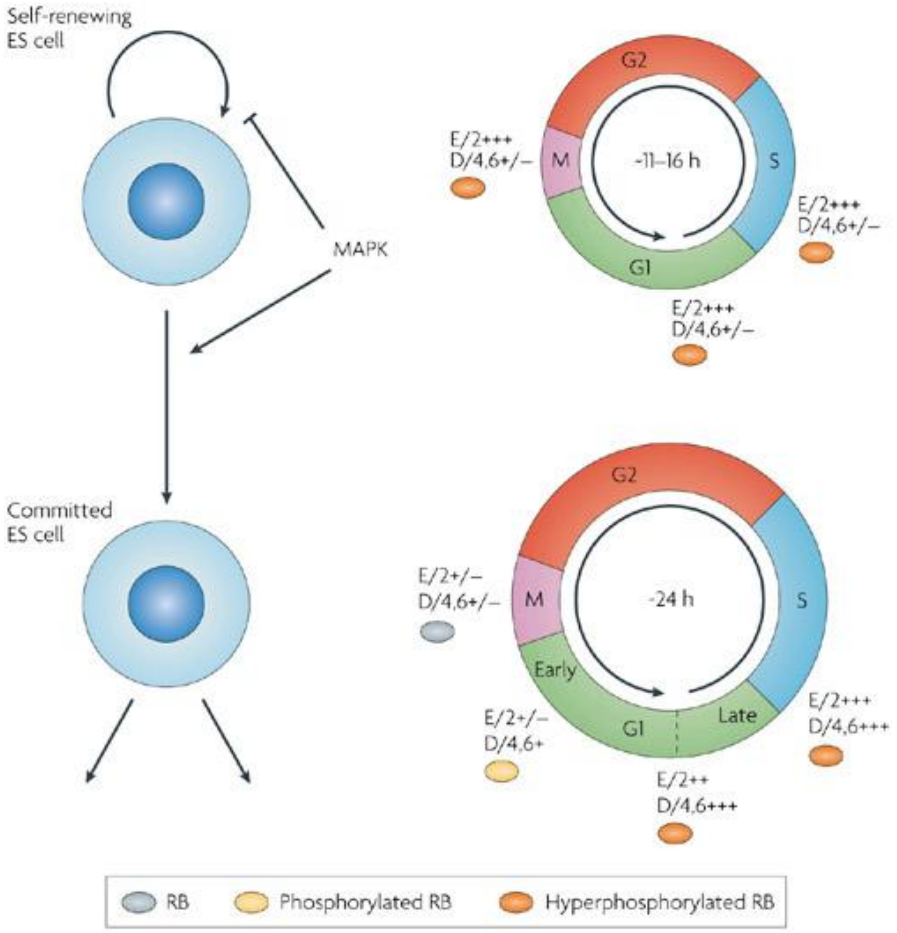


**Figure 1.3. The cell cycle of ES cells.** The cell cycle of ES cells is shortened relative to that of most other cells, which is due to an abbreviated G1 phase. For most cells, the transition through early G1 phase requires the accumulation of cyclin D, resulting in the hyperphosphorylation of the retinoblastoma tumour suppressor protein (RB) by cyclin D–CDK4 or cyclin D–CDK6 complexes (D/4,6). Inactivation of RB by hyperphosphorylation results in the mitogen-independent activity of cyclin E–CDK2 complexes, the defining characteristic of late G1 phase. In ES cells, cyclin E–CDK2 (E/2) is constitutively active throughout the cell cycle, which allows the transition of ES cells from M phase directly to late G1. The resulting absence of the cyclin D-dependent early G1 phase shortens the G1 phase and the entire cell cycle. + refers to cyclin–CDK activity: +/-, negligible; +, low; ++, intermediate; +++, high (Adapted from Orford and Scadden et al., (Orford and Scadden, 2008)).

A direct relationship between cell cycle regulation and master regulators of ES cells has recently been described. Oct4 and Sox2 are shown to regulate miR-

302, which targets cyclin D1, Rb, E2F1 and p130 (Card et al., 2008) and Nanog is suggested to be a regulator of G1 to S transition in ES cells through regulation of CDK6 and CDC25A, which are key players in the G1 cell cycle (Zhang et al., 2009).

The role of cell cycle regulation in maintaining ES cell identity is further emphasized by the study of reprogramming and iPSCs derivation. Myc is one of the four "Yamanaka factors" for iPSCs generation. Although subsequent studies have demonstrated that Myc is dispensable for the iPSCs recipe, it is shown to be critical for the early stages and high efficiency of reprogramming as it maintains the cells in a proliferative state in which they respond better to the other exogenous factors (Knoepfler, 2008; Zhao and Daley, 2008). Unlike other transcription factors in the reprogramming recipe, Oct4, Sox2 and Klf4, which have significant functions for maintaining self-renewal and pluripotency in ES cells, there is no much evidence indicating the direct relationship between the expression level of Myc to the state of ES cells, as no developmental defects have been observed in c/N-Myc knockout mice. However, there is considerable evidence linking Myc to the cell cycle regulation in ES cells. Elevated c-Myc expression accelerates progression through G1 by positively regulating cyclin-Cdk activity, whereas ES cells lost its specific cell cycle structure during differentiation while the expression of Myc is downregulated (Cartwright et al., 2005; White and Dalton, 2005). All these data place Myc at the center of a regulatory network linking fundamental self-renewal and pluripotency mechanisms to the cell cycle machinery in ES cells.

### 1.4.2. *Small regulatory RNAs*

Recent research have discovered a large populations of non-coding RNAs (ncRNAs), which comprise a large fraction of transcriptome in the cell, and many of them have been shown to have important biological functions in a wide range of cellular processes. Small ncRNAs are not functioning as translated proteins, but able to influence gene expression at post-transcriptional level. They are mainly in charge of gene suppression or silence through partial complementary to one or more messenger RNA (mRNA) molecules, generally in 3' UTRs.

There are three types of ncRNAs have been identified so far, including microRNAs, piRNAs and siRNAs, and among them, microRNA is most extensively studied in ES cells. MicroRNAs are ~22nt small RNAs found in all eukaryotic cells. They suppress gene expression by degradation of target mRNA transcripts or inhibition of mRNA translation (Kloosterman and Plasterk, 2006). Profiling microRNAs expression in ES cells have identified a unique repertoire of microRNAs (Houbaviy et al., 2003; Suh et al., 2004), which are not present or exist at very low levels in somatic cells. These ES cell specific microRNAs are down-regulated as ES cells differentiate (Viswanathan et al., 2008), suggesting their role in the maintenance of ES cells. The function of microRNAs in ES cells can be first learned in the knockout studies. Dicer (an RNase III-family nuclease critical for microRNA generation) knockout mice die at early stages of embryogenesis and Dicer-deficient ES cells are defective in differentiation (Kanellopoulou et al., 2005; Murchison et al., 2005). ES cells deficient for DGCR8, which results in a complete absence of mature microRNAs, fail to differentiate properly in response to differentiation signals. All these reinforce the important roles of microRNAs pathways in maintaining pluripotency of ES cells. Interestingly, members of this set of ES cell specific microRNAs possess the same or similar seed motif, indicating common target mRNAs (Gangaraju and Lin, 2009). Significantly, recent studies have shed light on the molecular and functional interaction between microRNA and core transcriptional circuity in maintaining the ‘stemness’ of ES cells. Many of the microRNAs are shown to be directly regulated by important transcription factors, Oct4, Sox2, Nanog, c-Myc and Tcf3 in ES cells (Marson et al., 2008). In turn, these microRNAs were shown to inhibit the epigenetic silencing of pluripotency factors (Gangaraju and Lin, 2009).

Despite the importance of microRNAs in pluripotency and self-renewal, detailed mechanisms and the crosstalk with transcription network remain elusive. During differentiation, different set of microRNAs might be induced to facilitate the process by down-regulation of pluripotency associated gene. Further mechanism studies are required to elucidate the functions of microRNAs in both the maintenance and inhibition of pluripotency.

### 1.4.3. *Epigenetic regulations*

Epigenetic regulation is specifically defined as heritable changes in the chromatin structure by mechanisms independent of changes in the primary DNA sequence (Surani et al., 2007). As substrate of transcription, chromatin is subjected to various forms of epigenetic regulation, including histone modification, histone variants, chromatin remodeling, and DNA methylation. The crucial role of epigenetics in modulating the transcriptional outcome and thereby regulating cell fate decisions has emerged over the last decade.

Examination of the epigenetic status of ES cells has identified a number of pluripotent cell specific properties that maintain undifferentiated state while preserving the ability to respond rapidly to differentiated signals.

A distinct feature of ES cell chromatin is called ‘poised’ state. It is featured at specific regulatory sites, particularly those at lineage specific transcription factor loci, which appear to be in a silent status but poised for activation in response of subsequent signal for differentiation. These loci are characteristically associated with both repression marker, histone H3 lysine 27 tri-methlation (H3K27me) and activation marker, histone H3 lysine 4 tri-methylation (H3K4me), consisting a ‘bivalent domains’ (Bernstein et al., 2006). Upon differentiation, the repression marker H3K27me was lost at lineage specific transcription factors loci and the expression of those genes was activated; whereas the activation marker H3K4me was erased from loci that remain silent to eventually repress the expression from those genes. Thus, ‘bivalent domains’ provide a hyperdynamic and plastic chromatin structure to ES cells. It is believed that polycomb group (PcG) proteins are responsible for maintaining the repressive state in the ‘bivalent domains’. In general, PRC2, which is composed of EZH2, SUZ12 and EED, is the complex that initiates transcription repression. Loss of Ezh2 or Suz12 causes deficiency in cell proliferation in the inner cell mass and early embryonic lethality (Lee et al., 2006). Genome wide mapping studies of PcG proteins in both human and mouse ES cells has demonstrated that the genes regulated by the PcG proteins are co-occupied by H3K27me3 markers. These genes are transcriptionally repressed in ES cells and are preferentially activated when differentiation is induced. Interestingly, the pluripotency factors Oct4, Sox2 and Nanog co-occupy a significant fraction of the PcG protein regulated genes (Boyer et al., 2006a; Lee et al., 2006). The histone variant, H2AZ, is also required for gene repression in ES cells. In ES cells, H2AZ is enriched at silenced promoters targeted by PcG proteins and H3K27me3 and plays an important role in silencing lineage promoting genes (Creyghton et al., 2008).

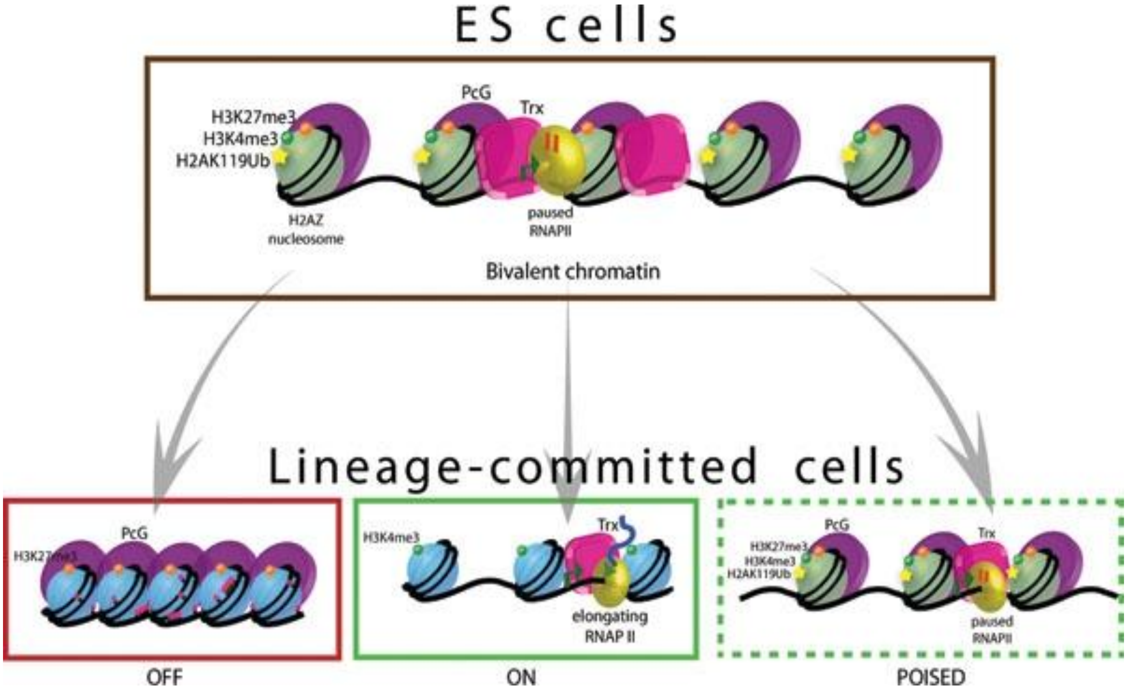


**Figure 1.4. Bivalent chromatin domains in ES cells.** Bivalent domains mark the promoters of developmentally important genes in pluripotent ES cells. PcG proteins proteins catalyze the tri-methylation of histone H3 on lysine 27. As such, bivalent genes are said to be silent, yet poised for activation. H2AZ is highly enriched in a manner that is remarkably similar to PRC2 and may also be an important regulatory component at bivalent genes. Upon differentiation, the bivalent histone marks can be resolved to monovalent modifications in which the gene is “ON” or “OFF”. Bivalent domains can also be maintained or newly established in lineage-committed cells (Adapted from Sha, K. and Boyer, L. A. StemBook, 2009).

Besides histone covalent modifiers and histone variant, ATP dependent chromatin remodeling enzyme also regulate ES cell chromatin structure in a self-renewal and pluripotent state. On the basis of domain structure, the ATP-dependent remodeling factors can be grouped into four families (SWI/SNF, ISWI, Mi-2/CHD, and INO80), with each family having broad functions in diverse biological processes and cell types (Boyer et al., 2000; Wang et al., 2007). Recent studies implicate SWI/SNF components as important regulators of ES cells. Knockout Brg1, which is the ATPase for SWI/SNF complex, results in lethality at the blastocyst stage, thus no ES cells can be derived from Brg1 deficient embryo (Bultman et al., 2000). In addition, knocking down Brg1 in ES cells led to ES cell differentiation (Ho et al., 2009). Genome wide mapping studies has shown that Brg1 interacts with master regulators Oct4, Sox2 and Nanog to control the expression of pluripotency associated genes (Liang et al., 2008). Other studies have revealed that the composition of the BAF complex varies during development (Lessard et al., 2007; Yan et al., 2008) and that an ES cell specific BAF (esBAF) complex is required for pluripotency and self-renewal (Ho et al., 2009). Downregulation of CHD1, which is one of the ATPase subunits of the chromodomain helicase DNA-binding (CHD) family, compromises ES cell self-renewal (Gaspar-Maia et al., 2009). The NuRD component Mbd3 is required for maintainance of ES cell pluripotency, but not self renewal (Kaji et al., 2006). However, interestingly, a unique NuRD complex called NODE that lacks Mbd3 but contains Mta1/2 and Hdac1/2 has been shown to interact with Nanog and Oct4 in ES cells and is recruited to Nanog/Oct4 target genes, independently of Mbd3 (Liang et al., 2008). In addition, ES cells depleted of Tip60-p400 subunits, which contains a bipartite SWI/SNF like ATPase as well as intrinsic acetyltransferase activities, exhibit altered morphology and are impaired in their ability to self renew (Fazzio et al., 2008).

### 1.4.4. *Signaling pathways*

ES cells can be maintained in an undifferentiated state in culture, but are poised for rapid differentiation. Extracellular signals provided by several soluble factors have been identified that exert either positive or negative effects on ES cell maintenance.

One approach to elucidate the requirement of ES cells for extrinsic stimulation has been to refine the culture medium conditions. When ES cells were first isolated from the blastocyst, they were cultured on a feeder layer of mitotically inactivated fibroblasts together with fetal bovine serum (FBS) (Smith and Benchimol, 1988; Williams et al., 1988). These feeder cells and FBS, create the very first extrinsic environment for ES cells. However, it is too complex to dissect the critical signaling pathways in feeder cultured ES cells as the complex communication between feeder cells and ES cells as well as undefined multifactorial components in serum. A key advance was the discovery of LIF (leukaemia inhibitory factor), which is able to sustain ES cells maintenance in the absence of feeder cells. LIF is known to function through binding to its receptor, LIFR (leukemia inhibitory factor receptor), to dimerize with gp130 on the cell membrane, resulting in the phophorylation of STAT3 (Signal transducer and activator of transcription 3) via JAK (Janus kinase) activation (Burdon et al., 2002; Niwa et al., 1998). Phosphorylated STAT3 dimerizes and translocates to the nucleus to activate a variety of downstream genes to maintain ES cell specific gene expression profiles.

However, LIF is only able to sustain the undifferentiated state of mouse ES cells in the presence of medium, suggesting that additional factors in the medium are required for ES cell maintenance. BMP4 (bone morphogenetic protein 4) is considered to be a key factor derived from serum in culture to influence the undifferentiated status of ES cells. In combination of LIF signal, BMP4 can support ES cell culture in the absence of serum by activating the expression of SMAD1 (MAD homolog 1), which, in turn, upregulates the expression of Id (inhibitor of differentiation) to suppress differentiation (Ying et al., 2003). By contrast, in the absence of LIF, BMP4 induce non-neural differentiation by interacting with different SMAD (SMAD1, 5, 8), which, in the contrary, repress the expression of Id (Rajan et al., 2003; Ying et al., 2003).

Although LIF is required for preserve the pluripotent state of ES cells for *in vitro* feeder free culture, *in vivo* ICM cells are able to develop into ES cells in the absence of LIF signaling, indicating that alternative pathways might be involved. Recent studies have challenged our knowledge of regulation by signaling pathways in ES cells that based on empirical configurations of the culture environment. They proposed that ES cells are intrinsically self-maintaining if shielded effectively from inductive differentiation stimuli including FGF4 (fibroblast growth factor-4) and GSK (glycogen synthase kinase-3) signaling pathways. In the mice embryo, FGF4 is produced in the ICM cells and are firstly postulated to promote proliferation of the ICM. In ES cells, FGF4 are secreted in an autocrine manner, which stimulate a RAS-ERK signaling cascade, results in a massive accumulation of phosphorylated ERK1/2. FGF4 as well as ERK2 deficient ES cells are resistant to differentiation along the neural and mesodermal lineage (Kunath et al., 2007; Stavridis et al., 2007), indicating that FGF4/ERK pathway is responsible for the exit of undifferentiated state and differentiation into neural or mesodermal lineage. However, neither LIF nor BMP4 has been shown to block the activation of FGF4/ERK signaling (Ying et al., 2003). In combination with LIF, inhibitors of either FGF receptor tyrosine kinase or ERK cascade can replace the requirement for serum/BMP4 and supports robust long-term ES-cell propagation (Ying et al., 2008). Though inhibiting FGF/ERK signaling reduces differentiation, two inhibitors compromise the viability and proliferation of ES cells. ES-cell propagation has been reported to be enhanced by an inhibitor of glycogen synthase kinase-3 (GSK3) (Sato et al., 2004). Importantly, combination of these three inhibitors is able to support derivation and proliferation of ES cells bypassing both LIF and BMP pathways, suggesting that LIF and BMP pathways act downstream of FGF/ERK pathway to block cell commitment (Ying et al., 2008).

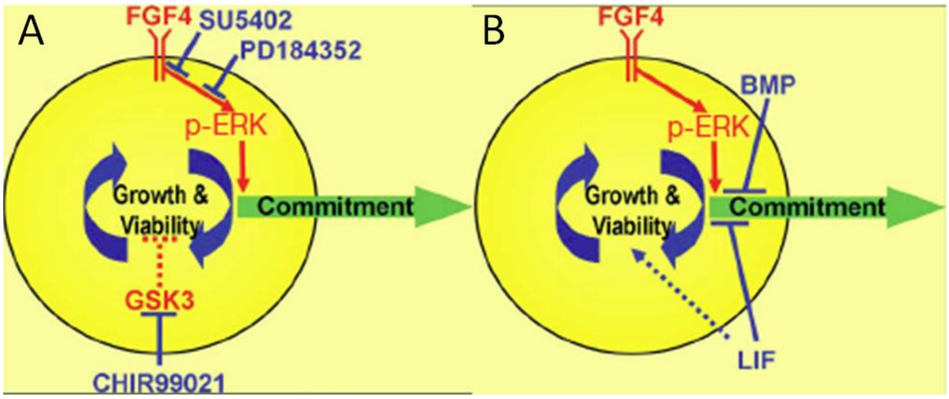


**Figure 1.5. Blocking FGF4/ERK and GSK3 signaling pathways are able to maintain ES cell.** phospho-ERK signaling is either inhibited upstream by chemical antagonists (A) or counteracted downstream by LIF and BMP (B). (Adapted from Ying et al., (Ying et al., 2008))

GSK3 was initially identified as the kinase responsible for phosphorylation and inhibition of glycogen synthase. It acts as a downstream regulatory switch for numerous signaling pathways and involved in the regulation of a variety of biological processes. GSK3 is negatively regulated by PI3K (Phosphatidylinositol 3-Kinase)-mediated activation of Akt/PKB (Protein Kinase-B) and it has a further role in the canonical WNT signaling pathway (Clodfelder-Miller et al., 2005). Inhibition of GSK3 using small molecules stimulates the activation of canonical WNT signaling (Doble and Woodgett, 2003).

In the canonical WNT pathway, Wnt proteins bind to cell-surface receptors of the Frizzled family, which inhibit a 'β-catenin destruction protein complex', composed of axin/GSK3/APC (adenomatosis polyposis coli). This stabilizes the pool of β-catenin and enables it to translocate into the nucleus and interact with TCF (transcription factor 3)/LEF (Lymphoid enhancer-binding factor) family transcription factors to promote specific gene expression (Wu and Pan, 2010). It seems that WNT pathways have dual functions in ES cells. Numerous studies have reported that WNT signaling contribute to the maintenance of pluripotency (Wang and Wynshaw-Boris, 2004). For example, Wnt signalling has been found to specifically inhibit neural differentiation (Aubert et al., 2002; Haegele et al., 2003). However, several other studies have implicated a role of WNT signaling in differentiation process. Repression of Apc in ES cells casues differentiation defects both *in vitro* and in teratomas (Kielman et al., 2002) and similar phenotype was observed when a dominant negative β-catenin without phosphorylation sites was stablized (Kielman et al., 2002). The contradictory conclusion may due to the interplay of WNT signaling with other signaling pathways or the function of its downstream transcription factors.

### 1.4.5. *Transcription network*

Extrinsic signaling pathways eventually lead to the nucleus and result in the transcriptional responses to sustain the ‘stemness’ of ES cells by either activation or repression of specific sets of genes. A major advance in understanding the gene expression profiling in ES cells has come with the identification of a transcription network that centered by three master transcription factors, Oct4, Sox2 and Nanog (Boyer et al., 2005; Chen et al., 2008; Loh et al., 2006).

Oct4 is encoded by *Pou5f1* gene and belongs to the POU family transcription factor. During embryogenesis, it is expressed in the pluripotent cells of the ICM and epiblast, but repressed in trophectodermal cells (Nichols et al., 1998; Palmieri et al., 1994; Rosner et al., 1990; Scholer et al., 1990). Oct4-deficient mouse embryo die following implantation due to a lack of ICM cells (Nichols et al., 1998). In ES cells, Oct4 acts as a gatekeeper to prevent ES cell from differentiation. However, the dosage of Oct4 is critical for pluripotency, as loss of Oct4 lead to differentiation into trophectoderm by interaction with Cdx2, which is a trophectodermal marker; while a twofold increase of Oct4 cause cell differentiated into a mixed population of mesodermal and endodermal cells (Niwa et al., 2005).

Oct4 has been reported to regulate diverse downstream targets by forming heterodimers with Sox2 (SRY-related HMG box 2). Sox2 is an HMG domain-containing transcription factor that has a similar expression pattern to that of Oct4 during mouse preimplantation development (Chew et al., 2005; Kuroda et al., 2005; Rodda et al., 2005). Sox2-null mice embryo fails to develop beyond implantation and have primary defects in the pluripotent epiblast. Similar to Oct4-null blastocysts, Sox2-null blastocysts are incapable of giving rise to pluripotent ES cells (Avilion et al., 2003; Nichols et al., 1998). In ES cells, Sox2 difficient ES cells differentiated mainly to trophectodermal lineage (Maruyama et al., 2005), whereas a two-fold overexpression of Sox2 resulted in the differentiation of ES cells into a mixture of lineages except endoderm (Kopp et al., 2008). Interestingly, forced expression of Oct4 is able to rescue the pluripotency of Sox2-null ES cells (Masui et al., 2007).

Another master regulator residing in the same complex with Oct4 (Wang et al., 2006) is Nanog, an NK-2 class homeobox transcription factor, whose expression is activated at 8-cell stage and later highly restricted to ICM and epiblast (Chambers et al., 2003; Mitsui et al., 2003). Nanog knockout embryos fail to form epiblasts, and are mostly composed of disorganized extraembryonic tissue (Chambers et al., 2003; Mitsui et al., 2003). More recently, it has been shown that although downregulation of Nanog predisposes ES cells towards differentiation, ES cells can however self-renew in the complete absence of Nanog (Chambers et al., 2007). This finding suggests that Nanog plays a major role in stabilizing the “stemness” state of ES cells.

Oct4, Sox2 and Nanog are not working alone, and instead, they are found to form an interconnected autoregulatory network. They bind to their own *cis*-regulatory elements (eg., promoter, enhancer) and the *cis*-regulatory elements of the other two genes to collaboratively regulate their own expressions. Furthermore, genome wide mapping studies have found out that Oct4, Sox2 and Nanog share a substantial fraction of target genes across the mouse and human genome (Boyer et al., 2005; Chen et al., 2008; Loh et al., 2006), including both transcriptionally active genes and repressed genes. In addition, their binding sites are in close proximity, which indicates that these proteins work in concert.

Recent studies have begun to provide new insights to add in more components into the current regulatory map, expanding our knowledge to the understanding of ES cells. Noval transcriptional regulators have been uncovered, such as Esrrb (Ivanova et al., 2006; Loh et al., 2006), Tbx3 (Ivanova et al., 2006), Sall4 (Elling et al., 2006; Sakaki-Yumoto et al., 2006; Wu et al., 2006; Zhang et al., 2006), Zfx (Galan-Caridad et al., 2007), Zic3 (Lim et al., 2007), Klf family (Jiang et al., 2008), and Ronin (Dejosez et al., 2008). These transcription factors are preferentially up-regulated in the undifferentiated ES cells. Depletion of these factors impairs the ability of ES cells to proliferate or maintain pluripotency.

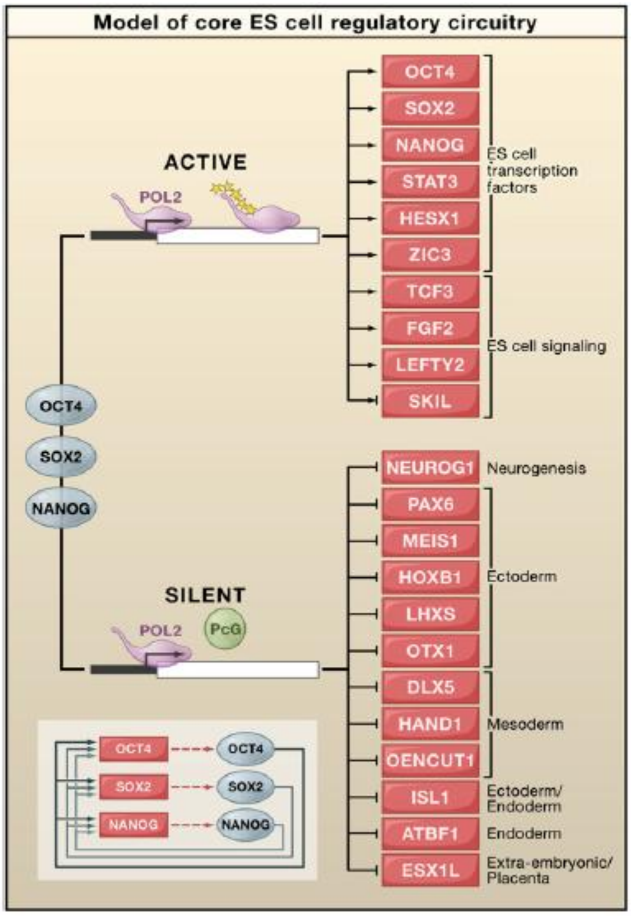


**Figure 1.6. Model of core ES cell regulatory circuit.** Oct4, Sox2, and Nanog occupy actively transcribed genes, including transcription factors and signaling components necessary to maintain the ES cell state. The three regulators also occupy silent genes encoding transcription factors that, if expressed, would promote other more differentiated cell states. PcG proteins co-occupy at this latter set of genes to inhibit RNA polymerase II (POL2) to produce complete transcripts. The interconnected autoregulatory loop, where Oct4, Nanog, and Sox2 bind together at each of their own promoters, is shown (bottom left). (Jaenisch and Young, 2008)

Another critical finding to appreciate the importance of transcription factors in ES cell regulation is provided by the generation of iPSCs. Introducing specific transcription factors into somatic cells, initially as Oct4, Sox2, Klf4 and Myc, is able to reprogram the differentiated state to pluripotent state, completely converting the cell cycle and epigenetic landscape to pluripotent cell specific manner. Subsequent studies have shown that the combination of transcription factors for reprogramming can be varied; for example, Oct4, Nanog, Sox2 together with Lin 28 is also able to generate successful iPSCs (Park et al., 2008b), which is emphasizing the potential significance of novel transcription factors and encouraging the study of identification of novel key transcription factors in ES cells.

# CHAPTER II: Zfp143 regulates *Nanog* through modulation of Oct4 binding

Part of this chapter is published as: Chen, X. [*], Fang, F. [*], Liou, Y.C., and Ng, H.H. (2008). Zfp143 regulates Nanog through modulation of Oct4 binding. Stem Cells 26, 2759-2767.

[*] These authors contribute equally to this work.

**My contribution to this project:**

Molecular study of Zfp143 as an ES cell regulator was initiated by Chen Xi. I worked closely with him when I first started my Phd work. I was responsible for RNAi rescue experiments, Electrophoretic Mobility Shift Assays to confirm the binding motif of Zfp143, luciferase reporter assays to demonstrate that Zfp143 regulate Nanog promoter activity, and microarray data analysis. I also worked with Chen Xi to construct all the manuscript figures as well as the writing of manuscript text. In addition, I took the main responsibility to answer reviewer's questions and did the supplementary data, including knockdown on D3 ES cells and secondary plating experiments.

## 2.1 SUMMARY OF CHAPTER II

Identification of transcriptional regulators governing the transcriptional network to maintain the identity of embryonic stem (ES) cells is crucial to the understanding of ES cell biology. In this work, we identified a zinc finger protein, Zfp143 as a novel regulator for self-renewal of ES cells. Depletion of *Zfp143* by RNAi causes loss of self-renewal of ES cells. We characterized *Nanog* as one of the downstream targets of Zfp143, as Zfp143 directly binds to *Nanog* proximal promoter and regulate its expression. Chromatin immunoprecipitation and EMSA show the direct binding of Zfp143 to *Nanog* proximal promoter. Knockdown of *Zfp143* or mutation of Zfp143 binding motif significantly down-regulates *Nanog* proximal promoter activity. Importantly, enforced expression of *Nanog* is able to rescue the *Zfp143* knockdown phenotype, indicating that *Nanog* is one of the key downstream effectors of Zfp143. More interestingly, we further show that Zfp143 regulates *Nanog* expression through modulation of Oct4 binding. Co-immunoprecipitation experiments revealed that Zfp143 and Oct4 physically interact with each other. This interaction is important because Oct4 binding to *Nanog* promoter is promoted by Zfp143. Furthermore, besides *Nanog*, Zfp143 co-occupy other targets with Oct4 as well, including genes that are known to be essential for ES cells, such as *Trp53* and *Jarid2*, indicating that Zfp143 may act as an activator to recruit and modulate Oct4 binding at specific loci in the ES cell genome, thus promote the expression of ES-specific gene expression and control ES cell self renewal. Our study reveals a novel regulator functionally important for the self-renewal of ES cells and provides

new insights into the expanded regulatory circuitry that maintains ES cell pluripotency.

## 2.2 INTRODUCTION

Embryonic stem (ES) cells are isolated from the inner cell mass (ICM) of blastocysts at day 3.5 of mouse development(Loebel et al., 2003). These cells are considered self renewal, as they can regenerate themselves as stem cells continuously and pluripotent as they exhibit the ability to differentiate into most specialized cell types found in the adult mouse(Geijsen et al., 2004; Schmitt et al., 2004). The derivation and manipulation of these cells, particularly human ES cells, hold great promise for both basic biological research and regenerative therapeutic medicine. The maintenance of ES cells is cooperatively controlled by external signaling pathways such as LIF/STAT3 pathway (Matsuda et al., 1999; Niwa et al., 1998; Raz et al., 1999) and BMP pathway (Ying et al., 2003), the intrinsic transcriptional network, centered around the core transcription factors, Oct4, Sox2 and Nanog(Loh et al., 2006), epigenetic modifications by chromatin-modifying enzymes as well as small regulatory RNA molecules.

Genetic studies and genomic mapping studies have shown that Oct4, Sox2 and Nanog are essential regulators of embryogenesis and ES cell identity (Chambers et al., 2003; Chambers and Smith, 2004; Mitsui et al., 2003; Nichols et al., 1998) and considered as the base to establish the transcriptional network. Oct4 is encoded by *Pou5f1* gene and belongs to the POU family transcription factor. It is preferably expressed in the pluripotent cells of the

ICM and epiblast, acts as a gatekeeper to prevent ES cell differentiation. It has been reported to regulate diverse downstream targets by forming heterodimers with Sox2 (SRY-related HMG box 2). Sox2 is an HMG domain-containing transcription factor that has a similar expression pattern to that of Oct4 during mouse preimplantation development (Chew et al., 2005; Kuroda et al., 2005; Rodda et al., 2005). Both *Oct4*- and *Sox2*-null mice have primary defects in the pluripotent epiblast and both *Oct4*- and *Sox2*-null blastocysts are incapable of giving rise to pluripotent ES cells (Avilion et al., 2003; Nichols et al., 1998). Interestingly, the forced expression of Oct4 is able to rescue the pluripotency of Sox2-null ES cells (Masui et al., 2007). Another key regulator residing in the same complex with Oct4 (Wang et al., 2006) is Nanog, an NK-2 class homeobox transcription factor, whose expression is also highly restricted to ICM and epiblast (Chambers et al., 2003; Mitsui et al., 2003). *Nanog* knockout embryos fail to form epiblasts, and are mostly composed of disorganized extraembryonic tissue (Chambers et al., 2003; Mitsui et al., 2003). More recently, it has been shown that although downregulation of Nanog predisposes ES cells towards differentiation, ES cells can however self-renew in the complete absence of Nanog (Chambers et al., 2007). This finding suggests that Nanog plays a major role in stabilizing the “stemness” state of ES cells. Strikingly, recent work has found that introducing specific transcription factors to somatic cells, initially as Oct4, Sox2, Klf4 and Myc, is able to reprogram the differentiated state to pluripotent state, which has draw a significant attention to the importance of transcription network in ES cells. Subsequent studies have shown that the combination of transcription factors for reprogramming can be varied, emphasizing the potential significance of novel transcription factors and encouraging the study of identification of novel key transcription factors in ES cells.

In this study, we report Zfp143 (a selenocysteine tRNA gene transcription-activating factor) as a novel regulator that maintains the undifferentiated state of ES cells by regulating the transcription of *Nanog*. Depletion of *Zfp143* by RNA interference (RNAi) resulted in cellular differentiation and a significant reduction in *Nanog* expression. ChIP and luciferase assays revealed Zfp143 binding at the proximal promoter region of *Nanog*. Furthermore, we found that it interacts with Oct4 on this *cis*-regulatory element of *Nanog*. Our data extend knowledge of the transcription network in ES cells by integrating Zfp143 as an upstream activator of *Nanog*, through modulation of Oct4 binding.

## 2.3 MATERIALS AND METHODS

**Cell culture** – E14 or D3 mouse ES cells, cultured under feeder-free conditions were maintained in Dulbecco’s Modified Eagle-Medium (DMEM, GIBCO), with 15 % heat-inactivated ES qualified fetal bovine serum (FBS, GIBCO), 0.055 mM β-mercaptoethanol (GIBCO), 2mM L-glutamine, 0.1 mM MEM non-essential amino acid, 5,000 units/ml penicillin/streptomycin and 1,000 units/ml of LIF (Chemicon). 293T cells were cultured in DMEM with 10 % FBS and maintained at 37 $^{o}$C with 5 % $CO_2$. The human ES cell-line (H1, WiCell) was cultured as described previously (Brandenberger et al., 2004). HEK293T (293) cells were cultured in DMEM supplemented with 10% FBS. In differentiation experiments, cells were treated with 1 μM of retinoic acid (RA) for mouse ES cells and 10 μM RA for human ES cells for 5 days.

**Transfection and short hairpin RNA mediated knockdown** – The 19 nucleotides targeted by the siRNAs are GGCAGATGGTGACAATTTA (for *Zfp143-1*) and GCAGTACGCAGCAAAGGTA (for *Zfp143-2*); we obtained similar results for the 2 RNAi constructs for *Zfp143* knockdown. Transfection of shRNA and overexpression plasmids was performed using Lipofectamine 2000 (Invitrogen) according to manufacturer's instructions. Briefly, 1.5 μg plasmid DNA was transfected into ES cells on 60 mm plates for RNA and protein extraction. Detection of alkaline phosphatase, which is indicative of the non-differentiated state of ES cells, was carried out using a commercial ES cell characterization kit (Chemicon).

**Secondary replating assay** - After 3 days of puromycin selection, shRNA-transfected cells were trypsinized and resuspended in medium. Ten-thousand cells were plated onto newly gelatin-coated 60-mm plates to form secondary ES cell colonies. After 4 days, emerging colonies were stained for alkaline phosphatase activity. For all the data shown (unless indicated otherwise), the cells were harvested and analyzed after 4 days of puromycin selection. Detection of alkaline phosphatase, which is indicative of the undifferentiated state of ES cells, was carried out using a commercial ES Cell Characterization Kit from Chemicon (catalog no. SCR001).

**ChIP and RNA expression analysis** – ChIP was performed as described previously(Loh et al., 2006) with Zfp143 antibody (H00007702-M01, Abnova); Oct4 antibody (sc-8628, Santa Cruz); HA antibody (sc-7392, Santa Cruz); Sox2 antibody (sc-17320, Santa Cruz) or RNA polymerase II antibody (05-623, Upstate). RNA extraction, reverse transcription and quantitative realtime PCR were carried out as described previously (Loh et al., 2006).

**Co-immunoprecipitation** – Transfected cells were lysed in cell lysis buffer (50 mM Tris HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 1% NP40, 10% glycerol with protease inhibitor cocktail) for 1 h. Whole cell extracts were collected and precleared. Beads coated with Oct4 **(**sc-8628, Santa Cruz) or HA (sc-7392, Santa Cruz) antibody were incubated with the precleared whole cell extracts at 4 $^{o}$C for overnight. The beads were washed with cell lysis buffer 4 times. Finally, the beads were boiled in 2x sample buffer for 10 min. The eluents were analyzed by either protein staining or Western blot.

**Sequential ChIP**- Oct4 antibody was crosslinked to protein G sepharose beads using DMP to prevent the leaching of antibody during SDS elution. The beads were then incubated with chromatin extracts overnight. Subsequently, the beads were washed and eluted with 1% SDS elution buffer at 37 ℃ for 45 minutes. The eluate was diluted to a final SDS concentration of 0.1% and incubated with fresh antibody-bound beads for the second IP. For the final round of IP, washed beads were eluted with 1% SDS elution buffer at 68 ℃ for 30 minutes. Eluate was decrosslinked in the presence of pronase and heated at 68 ℃ for 6 hours and DNA was purified by phenol:chloroform extraction.

**GST pulldown assay** – Full-length *Zfp143* and various deletion fragments were cloned into pET42b (Novagen). The plasmids were transformed into BL21 *E coli*. The Zfp143 proteins were expressed and purified with GSH-sepharose beads (Amersham) followed by Ni-NTA beads (Qiagen). The purified proteins were bound to GSH beads and incubated with Oct4 overexpressed cell lysates for 2 h in 4 $^{o}$C. The beads were washed 6 times with cell lysis buffer. The eluents were analyzed by Western blot.

**Electrophoretic mobility shift assays (EMSAs)** – DNA binding domain of Zfp143 was cloned into pET42b (Novagen). The plasmid was transformed into BL21 *E coli*. The DNA binding domain of Zfp143 protein was expressed and purified with GSH-sepharose beads (Amersham) followed by Ni-NTA beads (Qiagen). The purified protein was dialyzed against dialysis buffer (20mM HEPES, pH 7.9, 20% glycerol, 100mMKCl, 0.83mM EDTA, 1.66mM DTT, Protease Inhibitor Cocktail (Roche)) at $4^{o}$C for 4h. The concentration of the protein was measured with a Bradford assay kit (Bio-Rad). Double-stranded DNA oligonucleotides (Proligo) labeled with biotin at the 5' termini of the sense strands were annealed with reverse strands in annealing buffer (10 mM Tris-HCl, pH 8.0, 50 mM NaCl, 1 mM EDTA) and purified with an agarose gel DNA extraction kit (Qiagen). The sense strand sequence is shown in Figure 2.3A. EMSA was performed in a 10-µl reaction mixture containing 10mM HEPES, pH 7.5, 10mM KCl, 10mM $MgCl_2$, 1mM DTT, 1 mM EDTA, 10% glycerol, 3 ng of biotin-labeled oligonucleotide, 1 µg of poly(dI-dC) (Amersham) and 100ng recombinant Zfp143 DNA binding domain protein. Binding reaction mixtures were incubated for 10 min at room temperature and then subjected to electrophoresis on pre-run 5% native PAGE gels in 0.5x TBE buffer. Gels were transferred to Biodyne B nylon membranes (Pierce Biotechnologies) and detected with LightShift Chemiluminescent EMSA kit (Pierce Biotechnologies).

**Luciferase assay** – E14 embryonic stem cells were transfected with reporter constructs by Lipofectamine 2000 (Invitrogen) following the manufacture's protocol. A *Renilla* luciferase plasmid (pRL-SV40 from Promega) was co-transfected as an internal control. Cells were harvested after 36 h and the luciferase activity of the cell lysates was measured with the Dual-luciferase Reporter Assay System (Promega).

**Plasmids** – The promoter region of murine *Nanog* was amplified from genomic DNA. Primers for amplification, with restriction sites for cloning purposes indicated in lowercase, were:

GTCTGTagatctAATGGAAGAGGAAACTCAGATCC (*Nanog* promoter forward);

CCACACacatgtCAGTGTGATGGCGAGGGAAGGG (*Nanog* promoter reverse);

Products were cloned into pGL3 vector and sequence-verified. For the *Nanog* proximal promoter luciferase construct containing deletion of Zfp143 binding site, the sequence CCTCTTTTTGGG was deleted.

**DNA microarray** – Illumina kits was applied for cDNA expression profiling. mRNAs derived from *Zfp143* knockdown shRNA and *Luc* shRNA treated ES cells were reverse transcribed, labeled and analyzed using Illumina microarray platform (Sentrix Mouse-6 Expresion BeadChip v1.1). Arrays were processed as per manufacturer's instructions. Three biological repeats of profiles (each for control and knockdown) were used to generate statistically significant gene lists. The microarray data were analyzed by SAM. The thresholds for the differentially expressed genes were (I) more than 2 fold change and (II) q-value of less than 0.05.

**Primer sequence**

For quantitative PCR

| Gene | Primer | Sequence |
|---|---|---|
| ***Zfp143*** **(Mouse)** | F | CAGGTCAAGGTGATGATGTTCTTAAAGGGT |
| | R | GGCCTGCATGTCAGCTTGAGATATG |
| ***ZFP143*** **(Human)** | F | CAGGTCAAGGTGAAGATGTTCTTAAAGGGT |
| | R | GGCCTGCATGTCAGCTTGAGATATGTTGAC |
| ***Pou5f1*** | F | TTGGGCTAGAGAAGGATGTGGTT |
| | R | GGAAAAGGGACTGAGTAGAGTGTGG |
| ***Sox2*** | F | GCACATGAACGGCTGGAGCAACG |
| | R | TGCTGCGAGTAGGACATGCTGTAGG |
| ***Nanog*** | F | GGCTATCTGGTGAACGCATCTGGAAG |
| | R | AACTGTACGTAAGGCTGCAGAAAGTCCTC |
| ***Esrrb*** | F | GCCTCAAAGTGGGGATGCTGAAGGAAGGTG |
| | R | GCCAATTCACAGAGAGTGGTCAGGGCCTTG |
| ***Bmp2*** | F | CCAAGATGAACACAGCTGGTCACAGATAAGGC |
| | R | AGGTGGTCAGCAAGGGGAAAAGGACACTCC |
| ***Gata4*** | F | AAGCTCCATGGGGTTCCCAGGCCTCTTGCAAT |
| | R | TGAATGTCTGGGACATGGAGCTGCTGTGCC |
| ***Fgf5*** | F | GAGAGTGGTACGTGGCCCTGAACAAGAGAG |
| | R | CTTCAGTCTGTACTTCACTGGGCTGGGACT |
| ***Brachyury*** | F | CCAACCTATGCGGACAATTCATCTGC |
| | R | GTGTAATGTGCAGGGGAGCCTCGAA |
| ***Cdx2*** | F | CGCAGAACTTTGTCAGTCCTCCGCAGTACC |
| | R | GTATTCGGCGGGGCTGCTGTAGCCCATAGC |
| ***Hand1*** | F | CCTGCCCAAACGAAAAGGCTCAGGACCCAA |
| | R | CGACCGCCATCCGTCTTTTTGAGTTCAGCC |

| | | |
|---|---|---|
| ***Cdh3*** | F | CTCCGAAACGATGTAGTGCCAACCTTC |
| | R | CTCGTAGTCAAAAACCAGCAGGGAGTCGTA |
| ***Esx1*** | F | GAGGCCTTTTTCCAGCGCGTCCAGTACCC |
| | R | ATGTTTCTGAATGCCTGTGCCCGCCGAAGT |

For ChIP assays on *Nanog* promoter

| **Symbol** | **Primer** | **Sequence** |
|---|---|---|
| **1** | F | ATTTCTTCTTCCATTGCTTAGACGGCTGAG |
| | R | CTACCACCATGCCCAATTTAAGGAGTGTTT |
| **2** | F | CCAGGTTTCCCAATGTGAAGAGCAAGCAA |
| | R | TGGCGATCTCTAGTGGGAAGTTTCAGGTCA |
| **3** | F | GGGTCACCTTACAGCTTCTTTTGCATTA |
| | R | GGCTCAAGGCGATAGATTTAAAGGGTAG |
| **4** | F | CTCTTTCTGTGGGAAGGCTGCGGCTCACTT |
| | R | CATGTCAGTGTGATGGCGAGGGAAGGGA |
| **5** | F | GCGGGTGTCCTTATCACTCTTCTGGAAA |
| | R | TCCAAGCTAGGATGTTAGGTCTCCCTGCTA |
| **6** | F | AGCTCAGTGCTCCTTCCAAACCCCAAACAA |
| | R | ACACCCGAGCATCACAACACGCACCT |

F, forward; R, reverse

## 2.4 RESULTS

### Zfp143 maintains the undifferentiated state of ES cells

As Oct4, Nanog and several other key regulators of ES cells are predominantly expressed in the ICM, the identification of genes that preferentially expressed in ICM would be promising to provide a potential list of regulators for the maintenance of ES cells. A list of 48 genes predominantly expressed in ICM has been uncovered by Yoshikawa et al. using whole mount *in situ* hybridization (Yoshikawa et al., 2006). Their specific expression patterns suggest their potential roles in regulating early embryogenesis and the identity of ES cells as well. *Zfp143* is one of the genes in the list. To check its expression pattern in both mouse and human ES cells and during differentiation, mouse E14 cells and human H1 cells were treated with retinoic acid (RA) to induce differentiation. The expression of *Zfp143* was downregulated, resembling the expression pattern of *Pou5f1* and *Nanog*, indicating the positive correlation between lost of its expression and lost of ES cell identity (Figure 2.1).

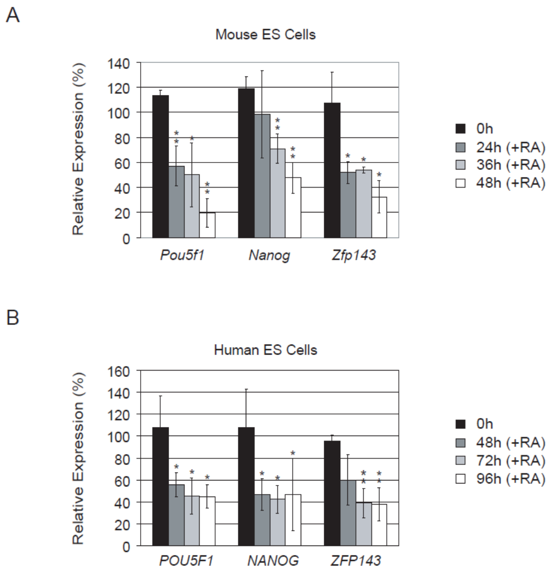


**Figure 2.1. *Zfp143* expression is downregulated in both human and mouse ES cells upon RA induced differentiation.** Real-time PCR analysis of *Zfp143* expression in human and mouse ES cells upon RA induced differentiation. 10uM Retinoid Acid (RA) was used to induce mouse (A) or human (B) ES cells differentiation. *Zfp143* expression was downregulated as both mouse and human ES cells differentiate. Data are presented as the mean ± SEM. *, $P$<0.05; **, $P$<0.005.

To assess the functional role of *Zfp143* in ES cells, we depleted endogenous *Zfp143* by RNAi. Two short-hairpin RNAi constructs targeting different regions of *Zfp143* coding sequence were used to ensure that the effects were specific. Both RNAi constructs were effective in reducing the transcript level of *Zfp143* compared with empty vector and control *luciferase* RNAi (Figure. 2.2B). Strikingly, *Zfp143* knockdown cells lost the typical mouse ES colony morphology. Alkaline phosphatase (AP) staining of pluripotent ES cells (red color) was reduced dramatically in the *Zfp143* knockdown cells, indicative of differentiation (Figure. 2.2A). RNAi depletion of three other ICM-specific transcripts (*Etv5*, *Mll3*, *4930548G07Rik*) (Yoshikawa et al., 2006) did not result in a differentiation phenotype (data not shown). This indicates that not all genes which are preferentially expressed in ICM will be important in the maintenance of ES cells.

To gain insights into the molecular alteration induced by *Zfp143* knockdown, the expression of pluripotency and lineage marker genes was analyzed. The expression of *Nanog* and *Esrrb* was reduced to 50% and 65 % respectively relative to the control, while the expression of *Pou5f1*and *Sox2* did not show appreciable changes (Figure. 2.2C). *Fgf5* and *Cdx2*, which are markers for primitive ectoderm and trophectoderm lineage respectively, were up-regulated (Figure. 2.2D).

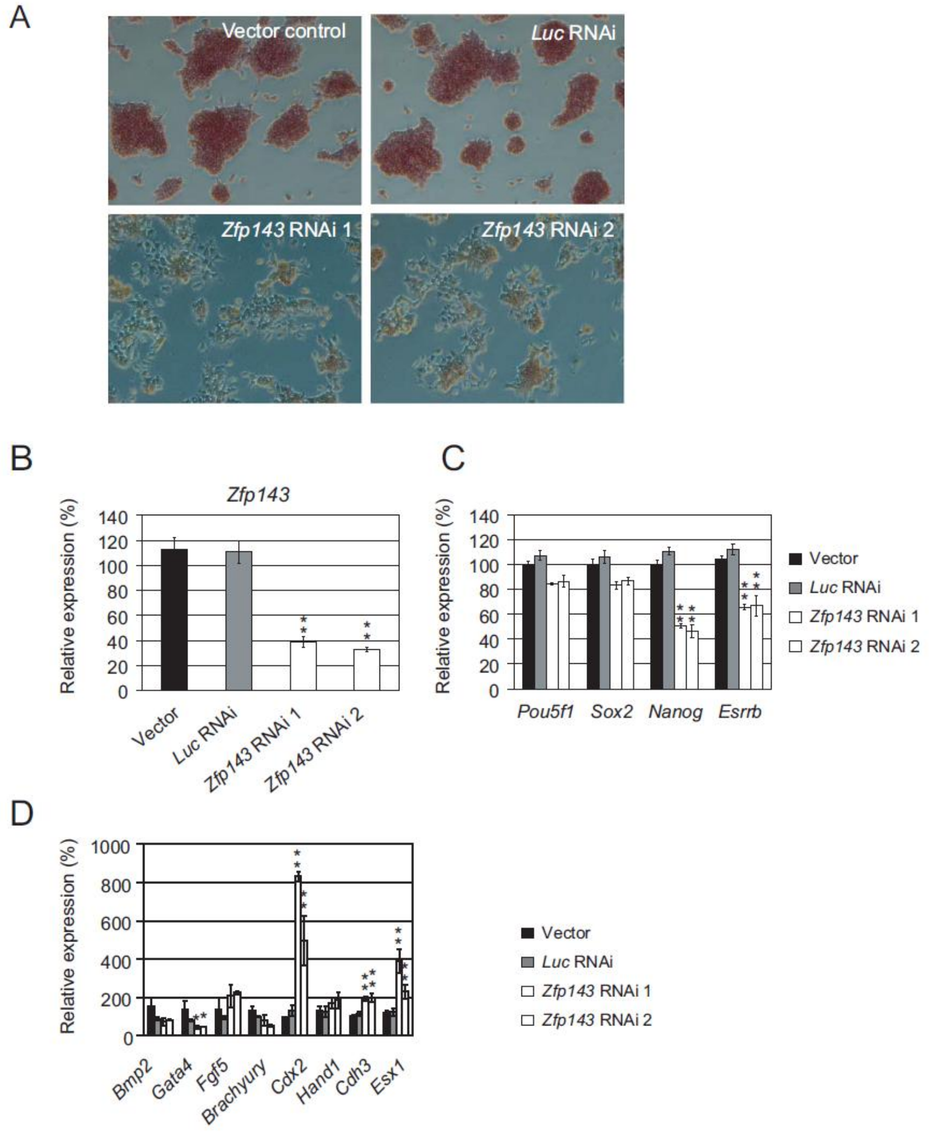


**Figure 2.2. Zfp143 is required for the maintenance of undifferentiated state of ES cells.** (A) *Zfp143* knockdown induced ES cells differentiation. Flattened fibroblast-like cells lacking alkaline phosphatase activity were formed when *Zfp143* was depeleted by RNAi. In empty vector and *luciferase* shRNA-transfected cells, normal undifferentiated ES colonies with positive alkaline phosphatase staining (red color staining) were maintained. (B) Quantitative real-time PCR analysis of *Zfp143* expression after knockdown using two shRNA constructs targeting different regions of the *Zfp143* coding sequence. The levels of the transcripts were normalized against control empty vector transfection. (C) Realtime PCR analysis of ES cell-associated gene expression in *Zfp143* knockdown ES cells. The levels of the transcripts were normalized against control empty vector transfection. (D) Real-time PCR analysis of lineage-specific marker gene expression in *Zfp143* knockdown cells. The levels of the transcripts were normalized against control empty vector transfection.

Knockdown experiments were repeated in D3 ES cells using the same conditions as described for E14 ES cells and similar results were gotten (Figure 2.3), suggesting the general roles of Zfp143 in mouse ES cells.

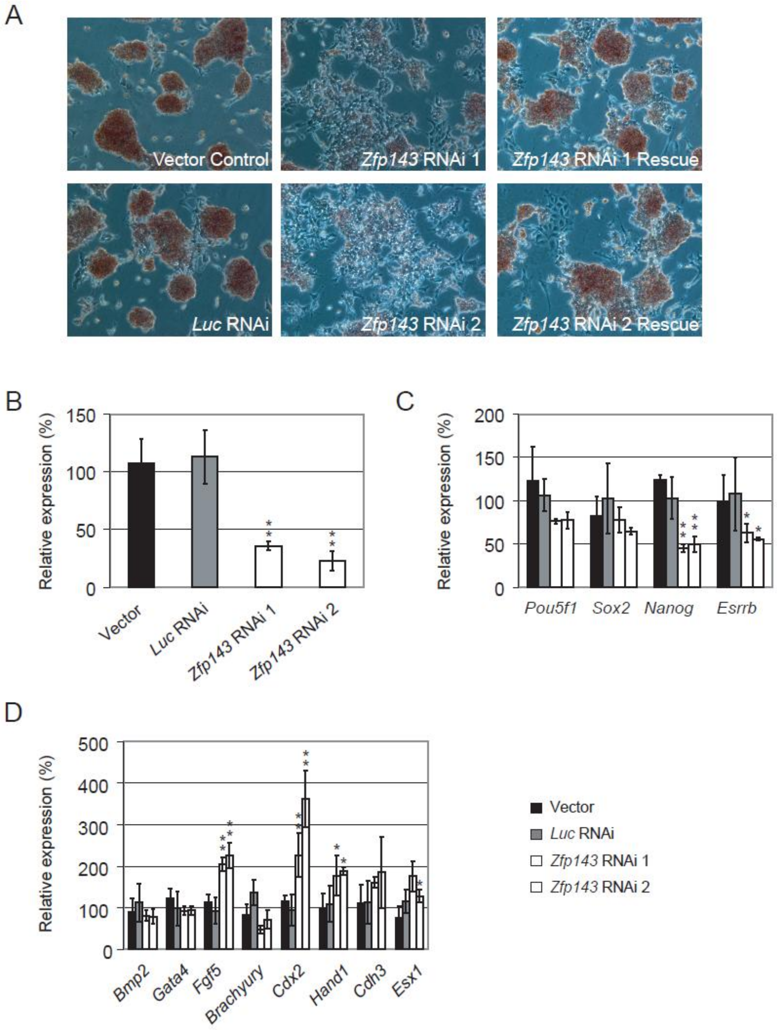


**Figure 2.3. Zfp143 is required for the maintenance of undifferentiated state of D3 ES cells.** (A) *Zfp143* knockdown induced D3 ES cells differentiation. Flattened fibroblast-like cells lacking alkaline phosphatase activity were formed when *Zfp143* was depeleted by RNAi. In empty vector and *luciferase* shRNA-transfected cells, normal undifferentiated ES colonies with positive alkaline phosphatase staining (red color staining) were maintained. Co-expression of RNAi-resistant *Zfp143* could rescue the differentiation phenotype. Typical colony morphology of ES cells with positive alkaline phosphatase staining (red) was restored. (B) Quantitative real-time PCR analysis of *Zfp143* expression after knockdown using two shRNA constructs targeting different regions of the respective transcripts. The levels of the transcripts were normalized against control empty vector transfection. (C) Real-time PCR analysis of ES cell-associated gene expression in *Zfp143* knockdown D3 ES cells. The levels of the transcripts

were normalized against control empty vector transfection. (D) Real-time PCR analysis of lineage-specific marker gene expression in *Zfp143* knockdown D3 ES cells. The levels of the transcripts were normalized against control empty vector transfection. Data are presented as the mean ± SEM and derived from three independent experiments. *, *P*<0.05; **, *P<0.005*.

To further characterize the *Zfp143*-depleted ES cells, we analyzed their ability to form colonies in a replating assay. Transfected cells were dissociated with trypsin and replated to allow the ES cells to expand into colonies. *Zfp143* knockdown reduced the number of ES cell colony-forming units (CFUs) by fourfold to 19-fold compared with control knockdown, rescue experiments could significantly restored the colony forming ability (Figure 2.4). These results suggest that Zfp143 plays a role in maintaining the self-renewal of ES cells.

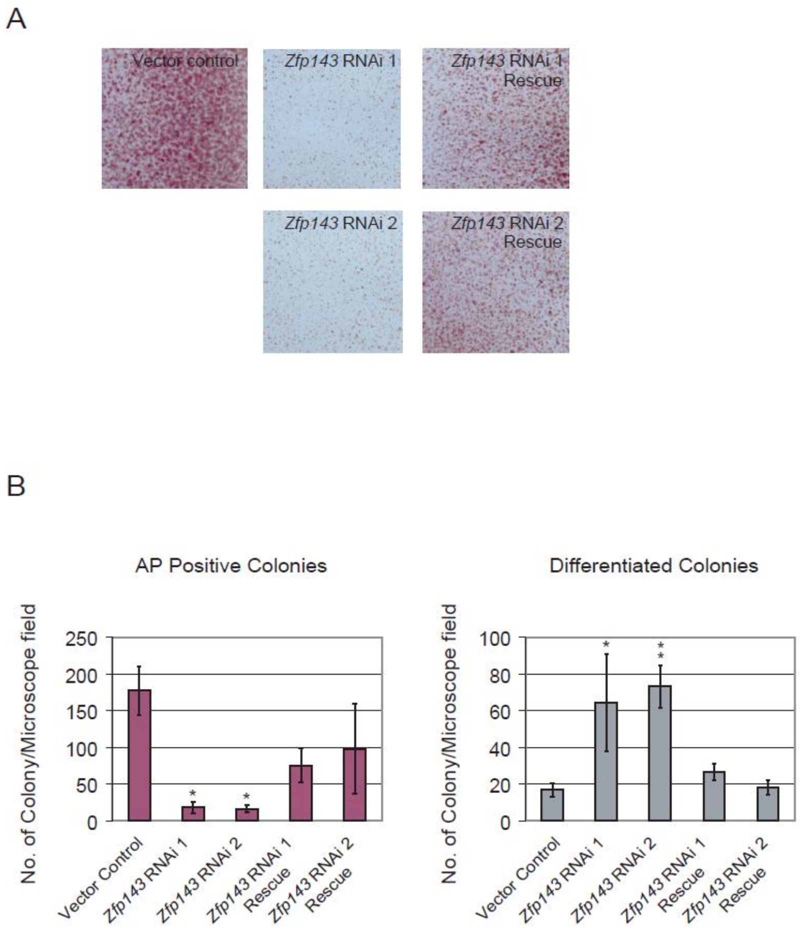


**Figure 2.4. *Zfp143* knockdown reduced ES cell capacity to form colonies in replating assay.** (A) Replating showed that *Zfp143* knockdown ES cells had significantly reduced capacity to form alkaline phosphatase-positive colonies whereas co-expression of RNAi-resistant *Zfp143* could rescue the capacity. (B) Counting of alkaline phosphatase-positive and differentiated colonies (per microscopy field) of the replated cells. Numbers of colonies counted are average of 5 different fields from three independent experiments. *, *P*<0.05; **, *P*<0.005.

To exclude off-target effects of shRNA, we performed rescue experiments. *Zfp143* RNAi-immune construct was made by introducing four silent mutations in the shRNA targeted region of *Zfp143* open reading frame. The rescue expression construct was co-transfected with *Zfp143* shRNA.

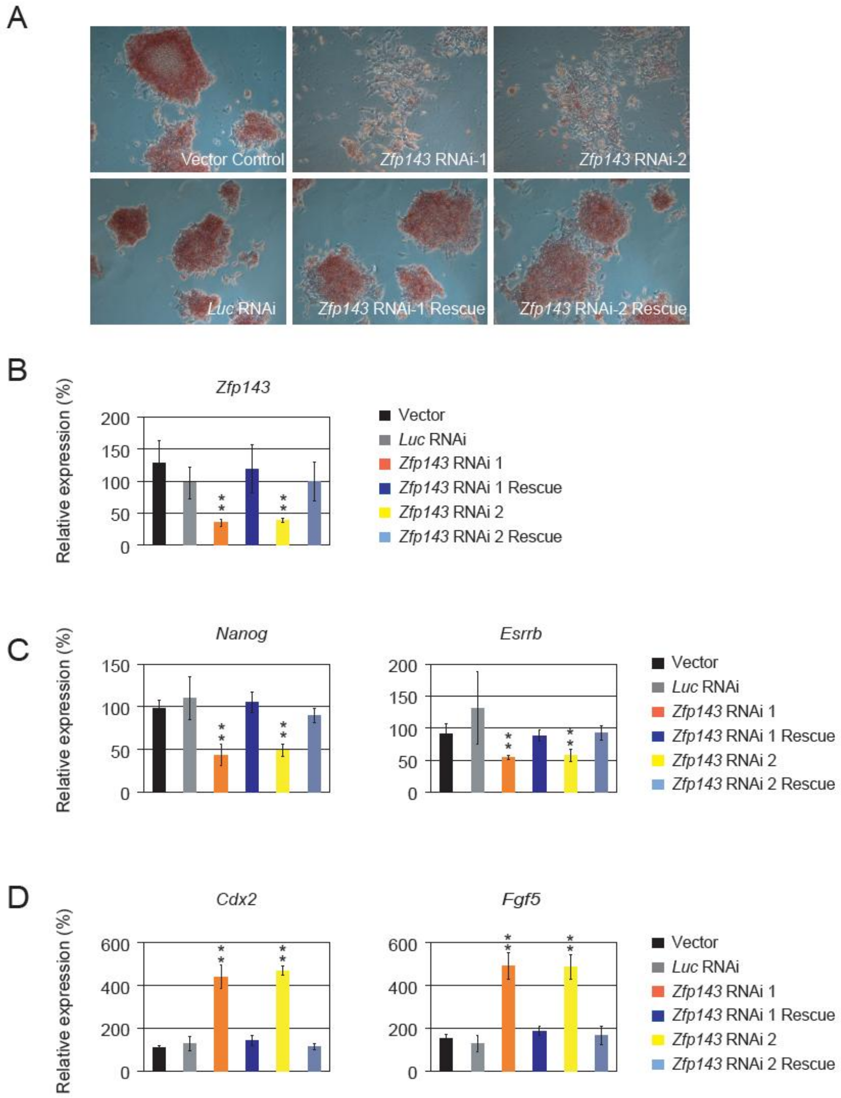


**Figure 2.5. Rescue of differentiation phenotype induced by *Zfp143* RNAi.** (A) Co-expression of RNAi-resistant *Zfp143* could rescue the differentiation phenotype induced by *Zfp143* knockdown. RNAi-resistant *Zfp143*expression constructs were co-transfected with corresponding *Zfp143* RNAi construct into mouse ES cells. Typical colony morphology of ES cells with positive alkaline phosphatase staining (red) was restored. (B) Co-expression of RNAi-resistant *Zfp143* rescued the down-regulation of *Zfp143* upon knockdown by two RNAi constructs. (C) RNAi-resistant *Zfp143* rescued the down-regulation of *Nanog* and *Esrrb* upon *Zfp143* knockdown by two RNAi constructs. (D) RNAi-resistant *Zfp143* restored the differentiation markers *Fgf5* and *Cdx2* to normal ES cell level. Data are presented as the mean ± SEM and derived from three independent experiments. *, *p*<0.05; **, *p<0.005*.

The result showed that RNAi-resistant Zfp143 was able to rescue the differentiation phenotype induced by *Zfp143* depletion (Figure 2.5A). The expression of pluripotency and lineage marker genes was comparable to normal ES cell levels (Figure 2.5B, C and D). The second RNAi-immnue *Zfp143* expression construct was also able to rescue the second shRNA targeting different region of *Zfp143* transcript. All these data demonstrate that the *Zfp143* knockdown phenotype is indeed caused by the *Zfp143* RNAi.

As depletion of Zfp143 by RNAi led to differentiation, DNA microarray experiments were performed to capture the transcriptome change in the whole genome level after transfection of shRNA expression constructs. 167 genes were upregulated and 259 genes were downregulated to more than two folds. The expression of a few essential self-renewal-related genes including *Nanog*, *Sox2*, *Tcfcp2l1* and *Jmjd1b* were reduced (Figure 2.6), consistent with real time PCR results, indicating that Zfp143 can positively regulate the expression of many self-renewal genes.

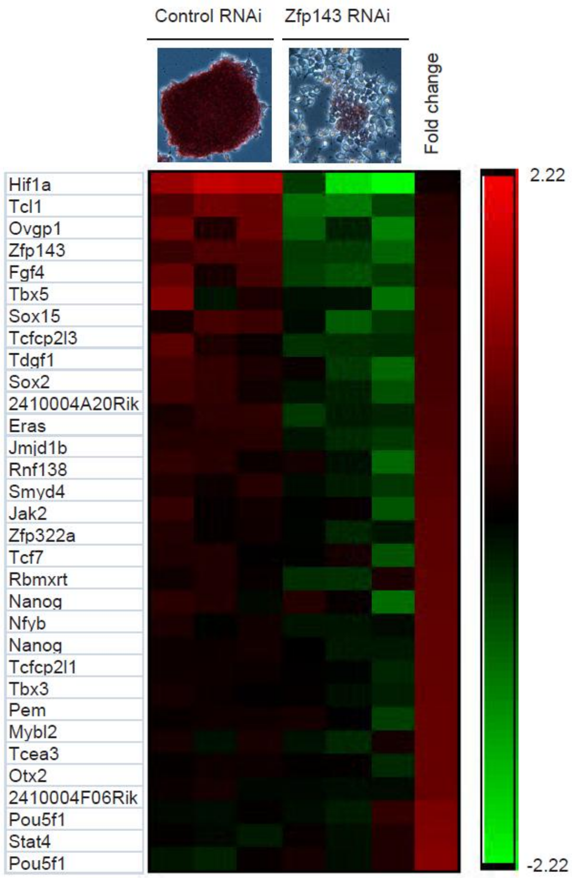


**Figure 2.6. Global gene expression changes after knockdown of Zfp143.** DNA microarray analyses were performed to measure gene expression changes after Zfp143 knockdown. The morphology and alkaline phosphatase staining are shown for both the control and Zfp143-depleted cells. Microarray heatmaps depicting expression changes of selected self renewal-associated marker genes are shown. Red indicates increased expression compared to control samples, whereas green means decreased expression. The genes expression levels were mean centred to show their relative change.

### Zfp143 binds to and regulates *Nanog*

Next, we investigated how Zfp143 maintains the undifferentiated state of ES cells. As *Nanog* was down-regulated when *Zfp143* was reduced by RNAi, we therefore asked if *Nanog* is a direct target of Zfp143. ChIP assay was performed using a monoclonal antibody raised against Zfp143.

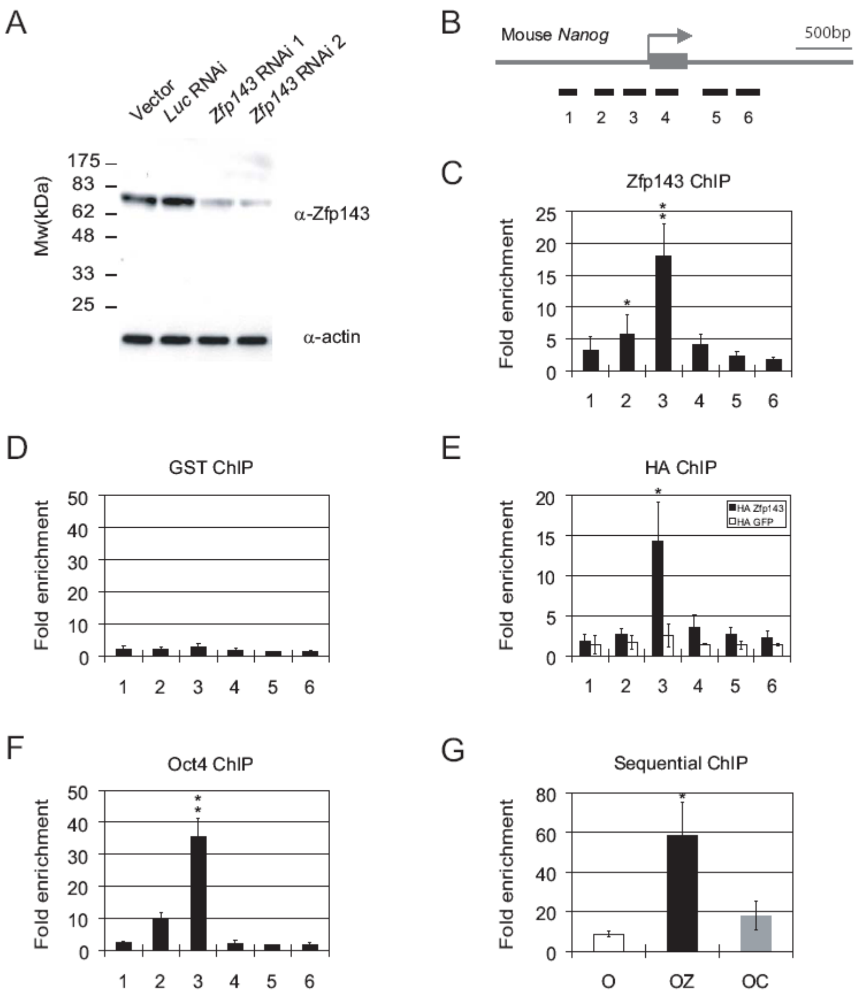


**Figure 2.7. Zfp143 and Oct4 co-occupy *Nanog* proximal promoter.** (A) Specificity of Zfp143 monoclonal antibody. Western blotting analysis of *Zfp143* knockdown and control ES lysates were carried out using anti-Zfp143 monoclonal antibody. β-actin served as loading control. (B) The locations of the amplified products (black boxes) along *Nanog* proximal promoter. (C) Zfp143 binds to *Nanog* proximal promoter. ChIP assay was performed using anti-Zfp143 monoclonal antibody to detect enriched fragments. Fold enrichment is the relative abundance of DNA fragments at the amplified region over a control amplified region. (D) GST antibody was used as mock ChIP control. (E) 3HA tagged Zfp143 construct was transiently transfected into ES cells, chromatin was extracted and subject to ChIP analysis using anti-HA antibody. 3HA tagged GFP served as mock control. (F) Oct4 binds to *Nanog* proximal promoter. ChIP assay was performed using an anti-Oct4 antibody to detect enriched fragments. (G) Zfp143 and Oct4 co-occupy *Nanog* proximal promoter. Sequential ChIP was performed using the anti-Oct4 antibody first (O). The eluants were then subjected to a second ChIP using anti-Zfp143 antibody (OZ) or a control antibody (OC). Data are presented as the mean ± SEM and derived from three independent experiments. *, *p*<0.05; **, *p<0.005*

The specificity of the antibody was characterized by western blotting using whole cell lysates transfected with control vector or *Zfp143* RNAi constructs (Figure 2.7A). Real-time PCR was used to quantify the ChIP-enriched DNA along *Nanog* proximal promoter. The result showed that Zfp143 occupied the *Nanog* proximal promoter (Figure 2.7B, C) while mock GST ChIP did not show any significant enrichment in this region (Figure 2.7D). In addition, ChIP assay using HA antibody against ectopically expressed HA-Zfp143 showed the same binding profile at the *Nanog* proximal promoter in ES cells (Figure 2.7E). These data independently confirm that Zfp143 binds to the *Nanog* proximal promoter *in vivo*. Interestingly, the profile of Zfp143 binding mirrored that of Oct4 binding at the *Nanog* proximal promoter (Figure 2.7F). Whether Zfp143 and Oct4 co-occupy this region of *Nanog* or in a mutually exclusive manner is of interest. To address this issue, we performed sequential ChIP assay. Chromatin extracts were first immunoprecipitated using the anti-Oct4 antibody. The eluents were then subjected to a second ChIP using the anti-Zfp143 antibody or a control antibody. Further enrichment after the second ChIP indicated that Zfp143 and Oct4 co-occupied the same molecule of DNA (Figure. 2.7G, OZ). An anti-GFP antibody used as a control in the second round of ChIP did not show any further enrichment of the *Nanog* sequence (Figure. 2.7G, OC). Thus, we conclude that Oct4 and Zfp143 co-occupy the *Nanog* proximal promoter.

A Zfp143 consensus binding site can be found at the peak region revealed by Zfp143 ChIP (Figure 2.7C, E). Using the electrophoretic mobility shift assay (EMSA), we further showed that the DNA binding domain of *Zfp143* could interact with this sequence (Figure 2.8A). Furthermore, mutagenesis of the

DNA probe revealed that the CCCA sequence which was reported to be critical for Zfp143 binding (Schaub et al., 1997) was required for this interaction (Figure 2.8A). These results showed that Zfp143 directly binds to *Nanog* proximal promoter through a conserved binding motif.

Having established the interaction between Zfp143 and the *Nanog* proximal promoter, we sought to understand the functional roles of Zfp143 on this promoter. A mutation in the Zfp143 binding motif was introduced into a luciferase reporter construct driven by *Nanog* promoter. Reporter assays showed that the mutation reduced *Nanog* promoter activity (Figure 2.8B). To further dissect the functional roles of Zfp143, *Zfp143* RNAi construct was co-transfected with the *Nanog* proximal promoter driving luciferase reporter into normal ES cells. *Nanog* depletion by RNAi was used as a positive control. The depletion of *Zfp143* reduced the *Nanog* proximal promoter activity to the same extent as mutating the Zfp143 motif (Figure 2.8C). Taken together, these data demonstrate that Zfp143 directly binds to *Nanog* proximal promoter and regulates *Nanog* expression.

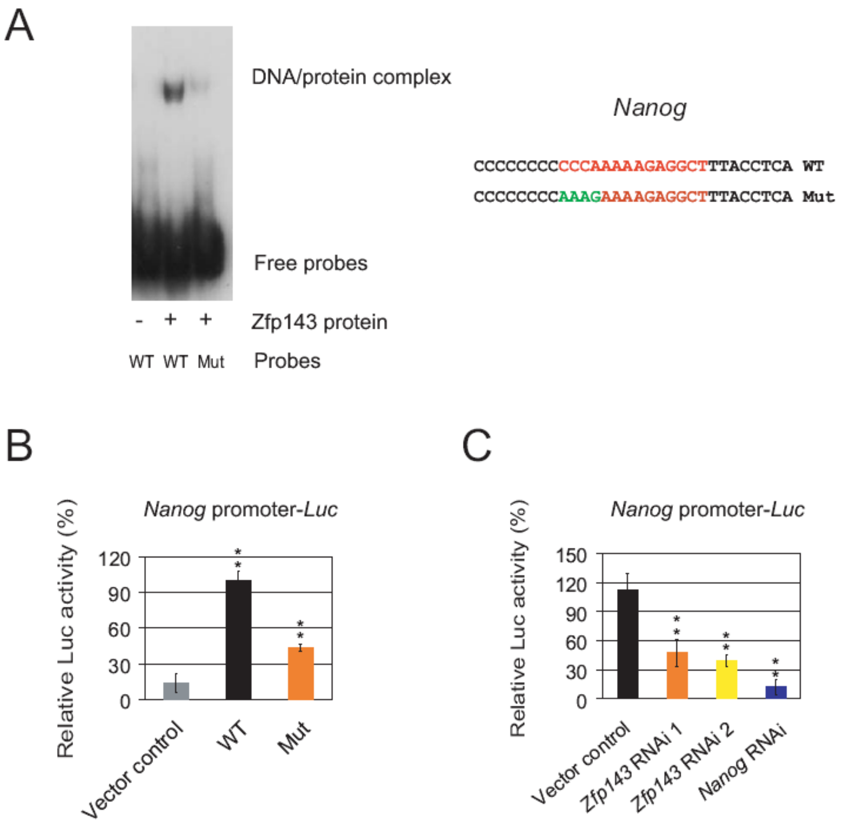


**Figure 2.8. Zfp143 regulates *Nanog* proximal promoter.** (A) Zfp143 directly binds to *Nanog* proximal promoter region. Electrophoretic mobility shift assays (EMSA) were used to analyze the binding of Zfp143 on *Nanog* proximal promoter. Purified recombinant DNA binding domain of Zfp143 was used for EMSAs. EMSA with the wild-type probe detected specific Zfp143/DNA complex. The effect of mutation on the Zfp143 DBD/DNA complex was also shown. The right panel shows the sequence of the *Zfp143* element (shown in red) and corresponding mutation (shown in green) used in this study. (B) Zfp143 binding site is crucial for *Nanog* promoter activity. Zfp143 binding site was mutated in the *Nanog* proximal promoter reporter (Mut) and tested for promoter activity in ES cells. (C) Depletion of Zfp143 attenuates *Nanog* promoter activity. *Nanog* promoter-*Luc* reporter or control vector was co-transfected with *Zfp143* RNAi construct into mouse ES cells and the luciferase activities were assayed. All luciferase activities were measured relative to the *Renilla* luciferase internal control. Data are presented as the mean ± SEM and derived from three independent experiments. **, *p<0.005*.

### *Nanog* is one of the key downstream effectors of Zfp143 for the maintenance of ES cells

As *Nanog* is a direct target regulated by Zfp143, we next investigate whether enforced expression of *Nanog* will rescue the effects induced by *Zfp143* knockdown. To test this hypothesis, *Nanog* was co-expressed in *Zfp143* knockdown ES cells.

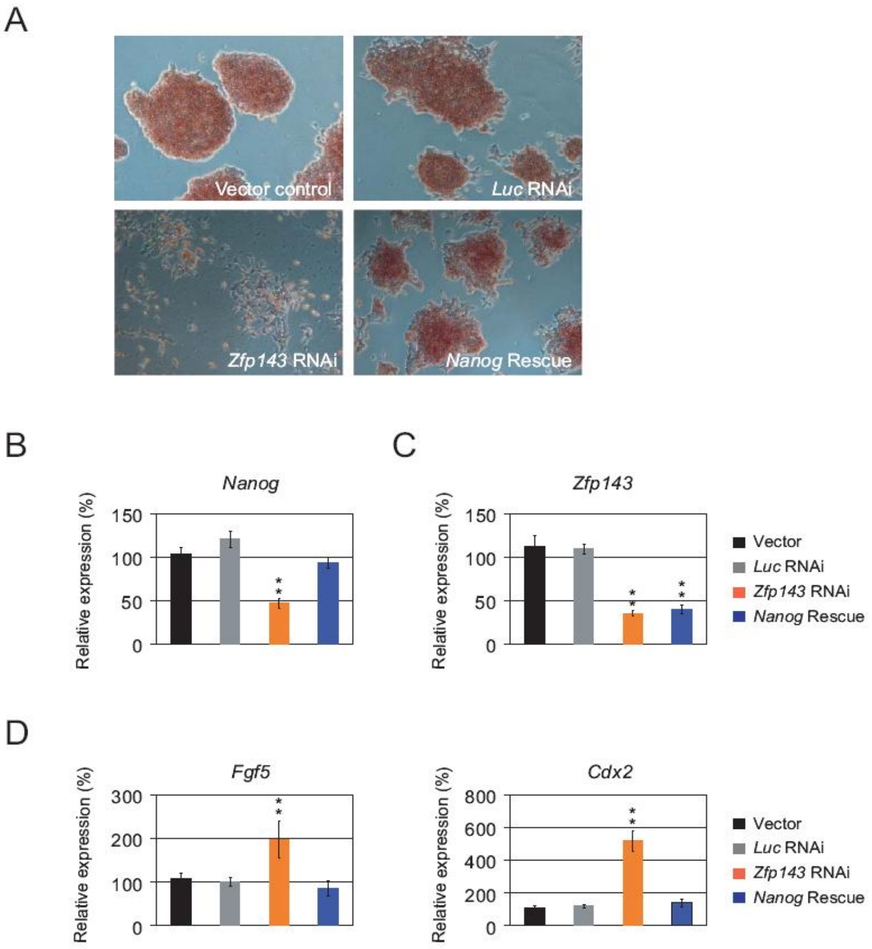


**Figure 2.9. *Nanog* is a key downstream effector of Zfp143 for maintaining ES cells.** (A) Enforced expression of *Nanog* could rescue the differentiation phenotype induced by *Zfp143* knockdown. Typical colony morphology of ES cells with positive alkaline phosphatase staining (red) was restored in *Zfp143* knockdown cells with enforced expression of *Nanog*. ES cells were co-transfected with a *Nanog*-expression vector and *Zfp143* RNAi construct. The cells were stained for alkaline phosphatase activity and the morphologies were examined by microscopy. (B) Enforced expression of *Nanog* rescued the down-regulation of *Nanog* induced by *Zfp143* knockdown. (C) Enforced expression of *Nanog* did not affect *Zfp143* knockdown efficiency. (D) Enforced expression of *Nanog* rescued the up-regulation of *Fgf5* and *Cdx2* induced by *Zfp143* knockdown to normal ES cell level. Quantitative real-time PCR was used to determine the expression. Data are presented as the mean ± SEM and derived from three independent experiments. **, *p<0.005*.

ES cells co-transfected with control vector and *Zfp143* RNAi construct differentiated, as observed by the loss of ES colony and alkaline phosphatase staining (Figure 2.9A). However, ES cells co-transfected with *Nanog* expression vector and *Zfp143* RNAi construct were able to retain the undifferentiated phenotype of ES cells evidenced by morphology and alkaline phosphatase staining (Figure 2.9A, B). The depletion of *Zfp143* was not affected by the enforced expression of *Nanog*, excluding the possibility that the rescued phenotype was due to inefficient depletion of *Zfp143* (Figure 2.9C). We further analyzed the transcripts of pluripotency and lineage markers affected by *Zfp143* depletion. With the enforced expression of *Nanog* in *Zfp143* knockdown cells, *Fgf5* and *Cdx2* were not induced (Figure 2.9D). These data suggest that *Nanog* is one of the key effectors of Zfp143 and can compensate for the depletion of *Zfp143* in ES cells.

### Zfp143 is a novel Oct4 interacting transcription factor

Since Oct4 and Zfp143 co-occupy *Nanog* proximal promoter and share the same binding pattern, we tested for potential interaction between the two proteins. Co- immunoprecipitation experiments were performed using ES cell nuclear extracts. Zfp143 was found to co-precipitate with Oct4 (Figure 2.10A). The reciprocal co-immunoprecipitation could not be performed as we found that our Zfp143 antibody could not efficiently immunoprecipitate Zfp143 from nuclear extract. Hence, we transfected a construct expressing HA-tagged Zfp143 into ES cells to do the reverse co-immunoprecipitation experiment. Using an anti-HA monoclonal antibody to immunoprecipitate HA-Zfp143, we showed that Oct4 co-immunoprecipitated with HA-Zfp143

(Figure 2.10B). Co-IP results obtained in a heterogeneous cell type 293T cells overexpressing Oct4 and HA-Zfp143 independently confirmed this interaction (Figure 2.10C, D).

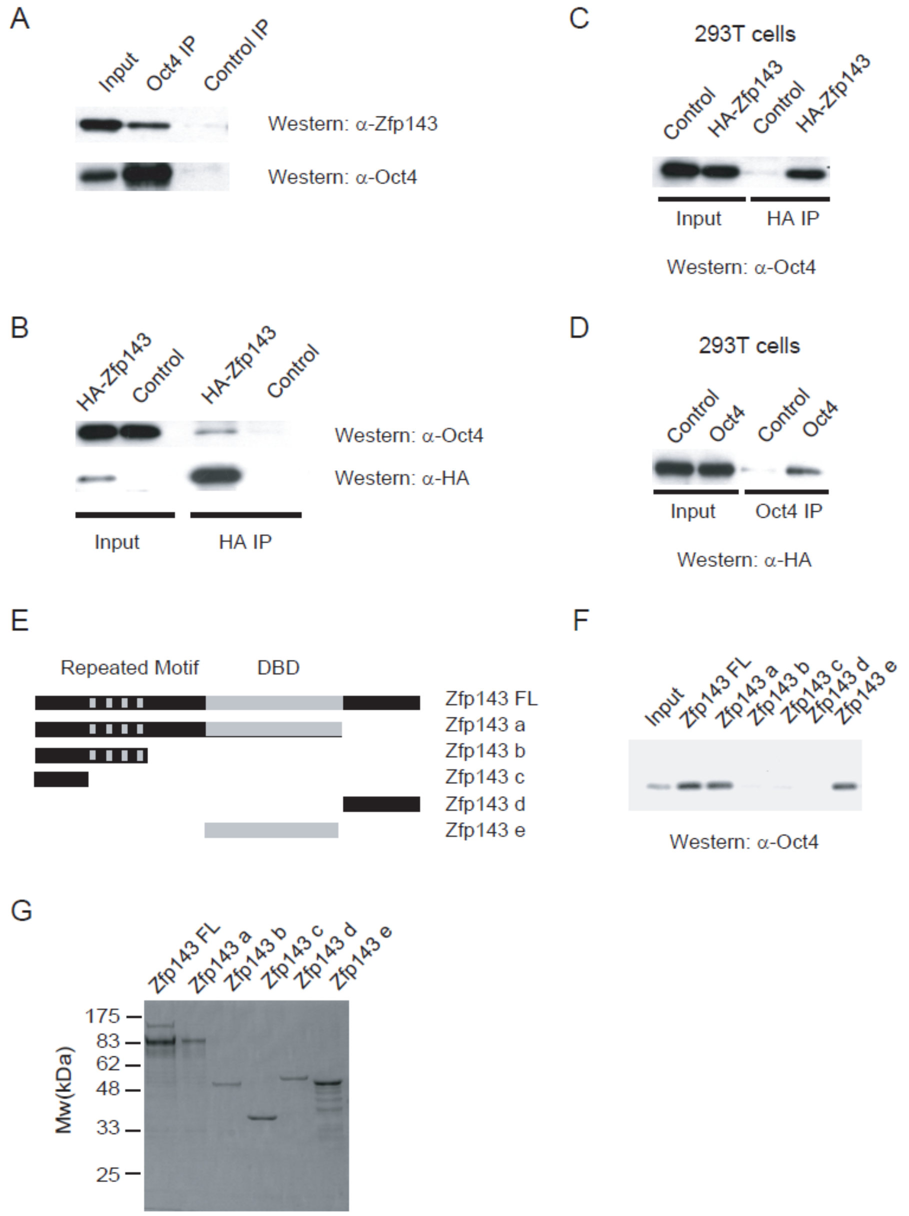


**Figure 2.10. Zfp143 is an Oct4 interacting protein.** (A) Co-IP using ES cell nuclear extracts was performed using anti-Oct4 antibody. Western blotting was carried out with anti-Zfp143 antibody. Control IP was performed using anti-GFP antibody. The affinity of anti-Oct4 antibody to pull down Oct4 was detected in lower panel. (B) Reverse co-IP using the ES cell lysates transiently expressing 3HA tagged Zfp143 was performed using anti-HA antibody.

Western blotting was carried out with anti-Oct4 antibody. Control HA IP was performed using ES cell lysates transiently expressing 3HA tagged GFP. (C) Co-IP using 293T cell lysates overexpressing Oct4 and 3HA tagged Zfp143 was performed using anti-HA antibody. Western blot was carried out with anti-Oct4 antibody. 293 cell lysates expressing 3HA tagged GFP and Oct4 served as control. (D) Reverse co-IP using 293T cell lysates overexpressing Oct4 and 3HA tagged Zfp143 was performed using anti-Oct4 antibody to confirm the interaction between Zfp143 and Oct4. Western blotting was carried out with anti-HA antibody. 293 cell lysates expressing 3HA tagged GFP and Oct4 served as control. (E) Schematic diagram of full length and truncated forms of Zfp143 protein. (F) GST pull down assay was carried out using GST-tagged Zfp143 proteins and 293T cell lysates overexpressing Oct4. Western blot was performed with anti-Oct4 antibody. (G) GST-tagged full length and different truncated forms of Zfp143 proteins.

To determine the region of Zfp143 that interacts with Oct4, full length and truncated fragments of Zfp143 were expressed and purified as recombinant GST-fusion proteins (Figure 2.10E, G). These proteins were immobilized onto GSH-sepharose beads and incubated with extracts harvested from 293T cells overexpressing Oct4. Zfp143 containing only the N-terminal repeats failed to pull down Oct4 (Figure 2.10F). However, the fragment containing only the DNA binding domain of Zfp143 was able to pull down Oct4. This demonstrates that the DNA binding domain of Zfp143 interacts with Oct4.

### Zfp143 modulates the binding of Oct4 at *Nanog* promoter

We have demonstrated that Zfp143 and Oct4 interact with each other and co-occupy *Nanog* proximal promoter to regulate *Nanog*. However, whether they work independently is not clear. To gain insights into the molecular mechanism of this regulation, we depleted *Zfp143* and examined the occupancy of Oct4 at different genomic sites. ES cells transfected with *Zfp143* shRNA constructs were crosslinked and the extracts were used for ChIP assay.

The protein level of Oct4 was not altered by *Zfp143* depletion (Figure 2.11A). As expected, the binding of Zfp143 on *Nanog* proximal promoter was reduced upon depletion of *Zfp143* (Figure 2.11B). Interestingly, we observed the reduction in Oct4 binding on the same region as well (Figure 2.11C) when Sox2 binding was not affected (Figure 2.11D). Oct4 occupancy at *Oct4* enhancer which is not occupied by Zfp143 (data not shown) was however not affected (Figure 2.11E). To further investigate the relationship between Zfp143 and the basal transcription machinery, we performed ChIP against RNA polymerase II. Consistent with the reduction in *Nanog* transcript after *Zfp143* depletion, the binding of RNA polymerase II to *Nanog* proximal promoter was also significantly reduced (Figure 2.11F). These data indicate that Zfp143 controls the binding of Oct4 at *Nanog* promoter.

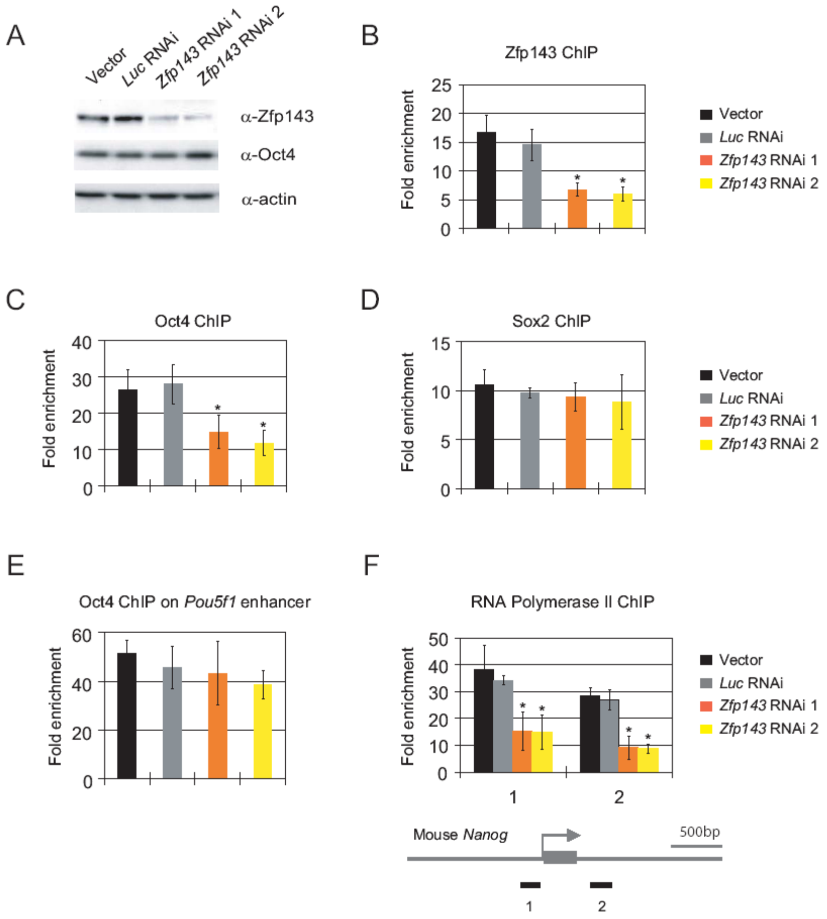


**Figure 2.11. The binding of Oct4 to chromatin is dependent on Zfp143.** (A) *Zfp143* knockdown did not affect Oct4 protein level. Western blotting analysis of control ES lysates and *Zfp143* knockdown lysates were carried out using anti-Zfp143 monoclonal antibody and anti-Oct4 antibody. β-actin served as loading control. (B) Zfp143 binding was reduced upon *Zfp143* knockdown. Chromatin extracts from control ES cells or *Zfp143* knockdown cells were subjected to ChIP using anti-Zfp143 antibody. (C) Oct4 binding was reduced upon *Zfp143* knockdown. Chromatin extracts from control ES cells or *Zfp143* knockdown cells were subjected to ChIP using anti-Oct4 antibody. (D) Sox2 binding was not affected upon *Zfp143* knockdown. Chromatin extracts from control ES cells or Zfp143 knockdown cells were subjected to ChIP using anti-Sox2 antibody. The primers used to detect ChIP-enriched DNA in (B-D) were the peak pair of primers numbered 3 in Figure 2B. (E) Oct4 binding at *Pou5f1* enhancer was not altered upon *Zfp143* knockdown. ChIP using chromatin from control ES cells or *Zfp143* knockdown cells was performed using anti-Oct4 antibody to detect Oct4 binding at enhancer region of mouse *Pou5f1*. (F) RNA polymerase II binding was reduced upon *Zfp143* knockdown. Chromatin extracts from control ES cells or *Zfp143* knockdown cells were subjected to ChIP using anti-RNA polymerase II antibody. The primers used were schematically shown in the lower panel. Data are presented as the mean ± SEM and derived from three independent experiments. *, $p<0.05$.

### Zfp143 co-occupy other targets with Oct4 in the genome of ES cell

Our data has revealed that Zfp143 co-occupy *Nanog* promoter with Oct4 and regulate *Nanog* expression. However, besides *Nanog*, it would be interesting to know whether Zfp143 targets other downstream genes that are essential for ES cells, and more importantly, whether Zfp143 and Oct4 interaction and modulation is also applied for other gene regulation is of our interest. A group has performed a large scale analysis to evaluate the binding of ZNF143 to mammalian promoters containing the consensus binding element and variants therein by bioinformatics in the human genome (Myslinski et al., 2006). Their data provided a list of putative genes whose promoters contain ZNF143 binding motif. Based on this list, we have identified two genes that have been shown critical for ES cell maintenance, *Jarid2* and *Trp53*. We validated the binding of Zfp143 on the promoter of *Jarid2* and *Trp53* by ChIP in ES cells

(Figure 2.12B, E). Furthermore, we found Oct4 bind to their promoter as well (Figure 2.12C, F), resembling the binding profile of Zfp143 and Oct4 on *Nanog* promoter. Therefore, we propose the protein partner Zfp143-Oct4 is not just a specific regulatory unit for *Nanog* promoter, but specifically regulate a few important genes in ES cells and thus maintain the entire transcriptional circuit controlling self renewal of ES cells.

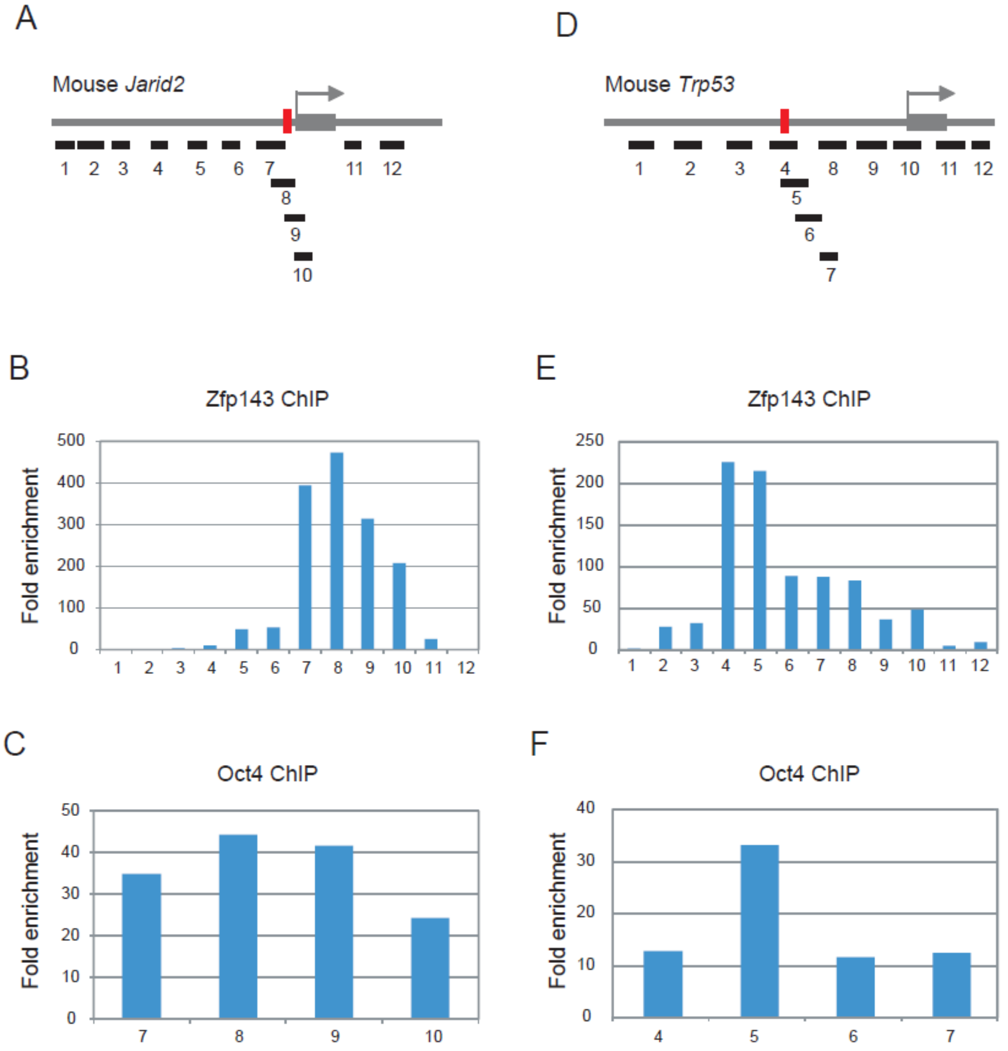


**Figure 2.12. Zfp143 and Oct4 co-occupy other targets that are important for ES cells.** (A, D)The locations of the amplified products (black boxes) along *Jarid2* (A) and *Trp53* (D) promoter. (B, E) Zfp143 ChIP on *Jarid2* (B) and *Trp53* (E) promoter. (C,F) Oct4 ChIP on *Jarid2* (C) and *Trp53* (F) promoter.

## 2.5 DISCUSSION

Zfp143, also known as STAF, is a zinc finger protein that was originally identified in *Xenopus laevis* as the transcriptional activator of the tRNA$^{Sec}$ gene and was then be found to be the transcription activator of snRNA and snRNA-type genes transcribed by RNA Pol II and III(Schaub et al., 1997; Schuster et al., 1998; Schuster et al., 1995). Zfp143 also plays a critical role in basal and tissue-specific expression of transaldolase and regulating the metabolic network controlling cell survival and differentiation(Grossman et al., 2004). In addition, Zfp143 is inducible by DNA damaging agents such as gamma-irradiation, etoposide and adriamycin and activates gene expression in response to DNA damage and binds to cisplatin-modified DNA(Ishiguchi et al., 2004). During mouse early embryogenesis, *Zfp143* is one of the 48 genes expressed predominantly in the inner cell mass (ICM)(Yoshikawa et al., 2006), which suggests it might be a good candidate for further analysis of its role in preimplantation development and cellular pluripotency.

Differentiation induced by *Zfp143* knockdown indicates that it plays a role in the maintenance of ES cells. The entire transcriptional program has been disturbed due to the knockdown, a number of self renewal and pluripotency associated genes were downregulated, while multiple lineage marker genes were significantly upregulated. More importantly, the depletion of *Zfp143* leads to a reduction of *Nanog* level, while the level of *Pou5f1* and *Sox2* is modestly affected. This suggests that Zfp143 is directly regulating *Nanog*. Using ChIP and EMSA assays, we showed that Zfp143 binds to *Nanog* proximal promoter and maintains its activity. Enforced expression of *Nanog* rescued the differentiation induced by *Zfp143* depletion, suggesting that *Nanog* is one of the key effectors of Zfp143. Besides *Nanog*, we also found two other self-renewal and pluripotency associated genes, *Jarid2* and *Trp53*, are bound and regulated by Zfp143. Although we cannot exclude the possibility that there could be other important self-renewal or pluripotency associated genes directly controlled by Zfp143, our findings that *Nanog* is able to compensate for the depletion of *Zfp143* in ES cells highlight the importance of *Nanog* in the whole transcription network. It is also conceivable that overexpression of Nanog renders the ES cells more resistant to differentiation (Chambers et al., 2007).

Zfp143 (Staf) has seven contiguous zinc-finger repeats for DNA binding located in the middle of the protein and two transactivation domains at the amino-terminal portion(Schuster et al., 1998). Binding site selection experiments identified the 18-bp DNA sequence TACCCATAATGCATYGCG as its consensus binding sites (Schaub et al., 1999a; Schaub et al., 1997). However, known Staf-binding sites revealed a high degree of divergence (Schaub et al., 1999b). At *Nanog* proximal promoter, we have identified a 15bp binding motif for Zfp143 at the peak region of its binding, CCCAAAAAGAGGCT, which is consistent with the previous consensus motif and the first CCCA is shown to be the key sequence for the binding affinity.

The co-occurrence of an octamer and Zfp143 motif is often found in the distal sequence element of a large number of RNA polymerase II and III transcribed snRNA-type genes (Schaub et al., 1997). It has been shown that the trans-activation function of the distal sequence element is mediated essentially by

Oct-1 and Zfp143 binding at the octamer and staf motif (Schaub et al., 1997). Our results here indicate that this octamer/Zfp143 motif regulation model is not restricted to sn-RNA type genes only, but also used to control the gene expression of an ES cell-specific gene. Disruption of either motif, Oct4 or Zfp143, as shown by EMSA and luciferase assays, downregulates the transcription of *Nanog*. It has been shown that addition of an octamer element in the vicinity of a Zfp143 binding site in the *Xenopus* Pol II UIb2 and Pol III U6 genes produced a synergistic effect on transcriptional activation, thus suggesting a functional cooperativity between the two DNA-bound factors. However, the molecular mechanism of how octamer binding protein and Zfp143 collaborate to control the transcription was not explained. Here, we demonstrate for the first time that Zfp143 interacts with Oct4 through its DNA binding domain.

Knocking down of *Zfp143* significantly reduced Oct4 binding on the *Nanog* proximal promoter while Oct4 transcription and protein level as well as its binding on other *cis*-regulatory elements remained unaltered. Our data show that Oct4 binding at the *Nanog* promoter is dependent on Zfp143. It should be noted that *Zfp143* depletion did not lead to a complete loss of Oct4 binding. It is possible that there exists other Zfp143 independent binding of Oct4 at other nearby sites. Although Oct4 and Sox2 heterodimer extensively co-occupies genome-wide targets (Boyer et al., 2005; Loh et al., 2006), this study shows that Oct4 can interact with factors other than Sox2 to assist in its binding to chromatin and regulate transcription.

*Nanog* expression is restricted to pluripotent ES cells and precisely controlled during mouse embryogenesis. To date, two *cis*-regulatory regions have been uncovered (Figure 2.13). The first region is an enhancer 5 kb upstream of the transcription start site. This site is reported to be bound and positively regulated by STAT3, T (brachyury), Nanog-Sall4 complex and Klf transcription factors (Klf2, Klf4 and Klf5) (Jiang et al., 2008; Loh et al., 2006; Suzuki et al., 2006). The second important regulatory region is the proximal promoter which is bound and regulated by the Oct4-Sox2 heterodimer and FoxD3 (Kuroda et al., 2005; Pan et al., 2006; Rodda et al., 2005). Oct4-Sox2 complex and FoxD3 directly bind to *Nanog* proximal promoter and promote *Nanog* expression. However, the precise balance between the maintenance of Nanog level and the ability of lineage differentiation requires negative regulators in the transcription network. Several repressors have also been characterized to down-regulate *Nanog* expression. Tcf3, a transcription factor downstream of Wnt signaling pathway, directly binds to *Nanog* promoter and represses its expression. The depletion of *Tcf3* delayed ES cells differentiation and up-regulated Nanog protein level (Pereira et al., 2006). p53, another negative regulator of *Nanog*, represses *Nanog* expression after phosphorylation of the Ser315 which is induced upon differentiation (Lin et al., 2005). GCNF (Gu et al., 2005), an orphan nuclear receptor, mediates *Nanog* repression upon RA-induced ES cells differentiation by binding to *Nanog* promoter and 3'UTR region. Most of the transcription activator and repressors binding sites at *Nanog* regulatory regions have been discovered and validated for direct binding (Table 2.1). Recent study mapping 13 transcription factor and 2 regulators binding sties in ES cells further reveals extensive co-localization of Nanog, Smad1, STAT3, Klf4, Esrrb and Tcfcp2l1 on *Nanog* enhancer and n-Myc, c-Myc Zfx and E2f1 on the proximal promoter (Chen et al., 2008). Other than sequence-specific DNA binding transcription factors, recent studies have begun to uncover novel pathways involved in the regulation of *Nanog*. Jmjd2c, a histone demethylase that converts H3K9 trimethylation to dimethylation, has recently been identified to regulate the H3K9Me3 status of *Nanog*(Loh et al., 2007). Specific demethylation of H3K9Me3 by Jmjd2c at the *Nanog* promoter could inhibit the binding of transcription co-repressors such as HP1 and KAP1 and sustain *Nanog* expression (Loh et al., 2007). At the post-transcriptional regulation level, microRNA has been reported to exert functional roles in regulating *Nanog* expression. For instance, *miR-134* is found to specifically attenuate the translation of *Nanog*(Tay et al., 2008). Altogether, these studies highlight the intricacy in modulating the expression of *Nanog* through positive and negative regulation at both the transcriptional and post-transcriptional levels.

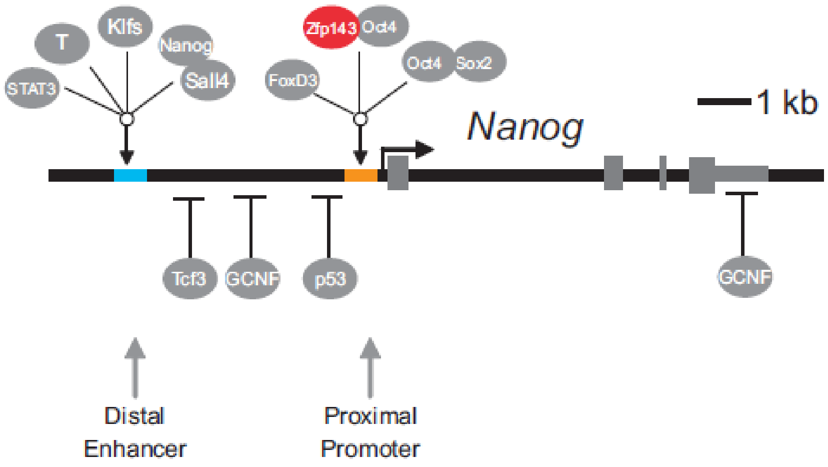
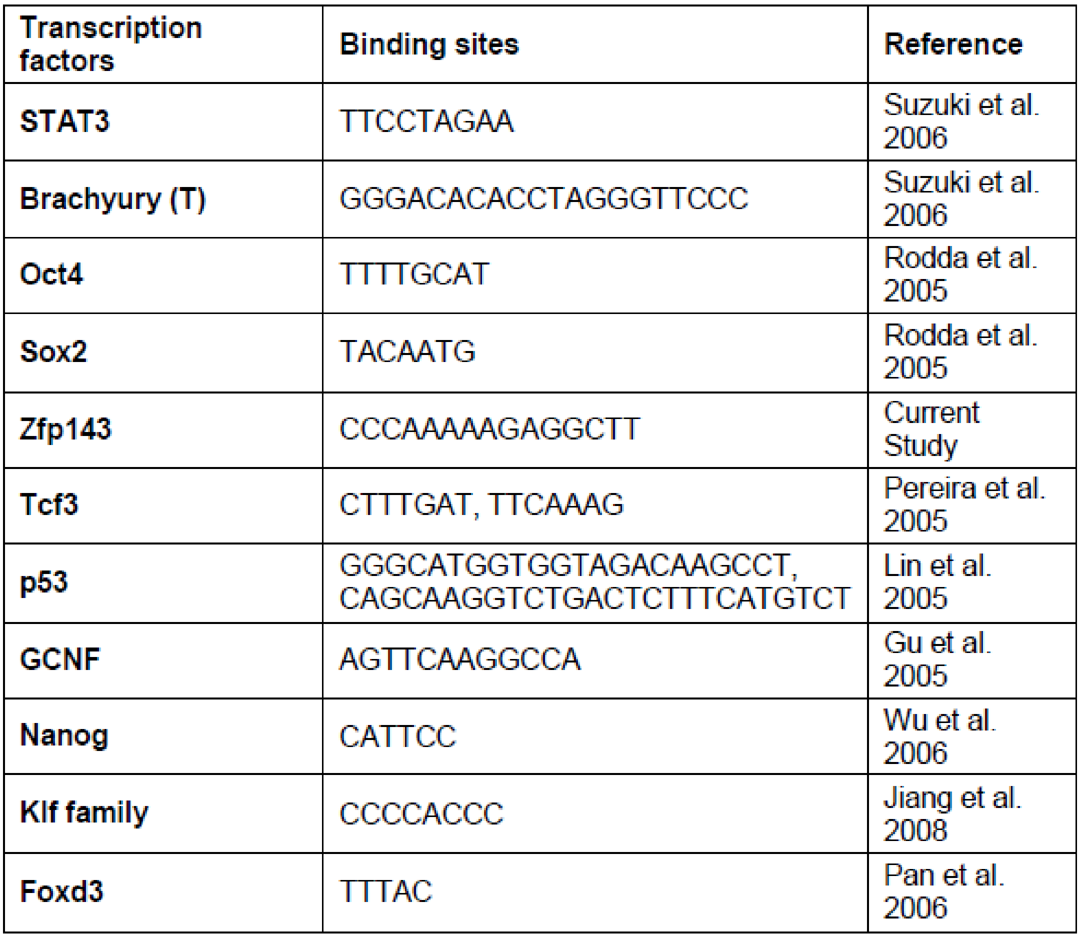


**Figure 2.13. A model depicting the different transcriptional regulators that interact with Nanog cis-regulatory regions.** Nanog is regulated by both activators and repressors. The enhancer and promoter regions of Nanog are shown as the blue bar and orange bar, respectively. STAT3, T, Klf transcription factors (Klf2, Klf4, and Klf5), and Nanog-Sall4 complex occupy the Nanog enhancer and activate Nanog transcription. The Oct4-Sox2 complex, FoxD3, and Zfp143-Oct4 complex positively regulate the proximal

promoter. Tcf3 and p53 occupy the promoter and exert repressive roles. GCNF binds to two regulatory elements located at 2.5 kb upstream of the transcription start site and 3'-untranslated region to repress Nanog expression upon retinoic acid-induced differentiation. Abbreviations: GCNF, germ cell nerve factor; kb, kilobase.

**Table 2.1. Known transcription activator and repressor binding sites at the *Nanog* regulatory regions.**



| Transcription factors | Binding sites | Reference |
|---|---|---|
| **STAT3** | TTCCTAGAA | Suzuki et al. 2006 |
| **Brachyury (T)** | GGGACACACCTAGGGTTCCC | Suzuki et al. 2006 |
| **Oct4** | TTTTGCAT | Rodda et al. 2005 |
| **Sox2** | TACAATG | Rodda et al. 2005 |
| **Zfp143** | CCCAAAAAGAGGCTT | Current Study |
| **Tcf3** | CTTTGAT, TTCAAAG | Pereira et al. 2005 |
| **p53** | GGGCATGGTGGTAGACAAGCCT,<br>CAGCAAGGTCTGACTCTTTCATGTCT | Lin et al. 2005 |
| **GCNF** | AGTTCAAGGCCA | Gu et al. 2005 |
| **Nanog** | CATTCC | Wu et al. 2006 |
| **Klf family** | CCCCACCC | Jiang et al. 2008 |
| **Foxd3** | TTTAC | Pan et al. 2006 |

# CHAPTER III: Dissecting early differentially expressed genes in a mixture of differentiating embryonic stem cells

Part of the chapter is published as: Hong, F., Fang, F., He, X., Cao, X., Chipperfield, H., Xie, D., Wong, W.H., Ng, H.H., and Zhong, S. (2009). Dissecting early differentially expressed genes in a mixture of differentiating embryonic stem cells. PLoS Comput Biol 5, e1000607.

**My contribution to this project:**

It is a collaboration project with a bioinformatic group at UIUC. I was responsible for leading the experimental validation for this project. I have validated by RNA interference for ten genes from the predication list and two of them were shown in the manuscript. I also worked with Sheng Zhong to construct most of the manuscript figures as well as the writing of manuscript text.

## 3.1 SUMMARY FOR CHAPTER III

Cellular differentiation is the process by which a less specified cell becomes a more specialized cell type, characterized by the constant change of gene expression pattern during the differentiation process. Identifying the subset of genes that initiate the differentiation process is critical to dissect the molecular mechanisms underlying differentiation and further control and manipulate the progress for application to clinical medicine. The current methods of identifying differentially expressed genes by comparing different cell types inevitably include a large portion of genes that respond to, rather than regulate, the differentiation process. We demonstrate through the use of biological replicates and a novel statistical approach that the gene expression data obtained without prior separation of cell types are informative for detecting differentially expressed genes at the early stages of differentiation. Applying the proposed method to analyze the differentiation of mouse embryonic stem (ES) cells to embryoid bodies (EB), we successfully generated a gene list of transcription regulators during differentiation, and further identified and experimentally verified *Smarcad1* as novel regulators of pluripotency and self-renewal. Furthermore, using the genes identified by our method, we constructed a gene regulatory network that strongly indicates the importance of Notch signaling pathway in triggering the early differentiation of ES cells. Our statistical approach can be formalized as a statistical test that can be generally applicable to analyze other differentiation processes.

## 3.2 INTRODUCTION

Differentiation process is initiated by gradual loss of differentiation inhibiting genes and induction of differentiation genes, thus lead to the change of highly-controlled modifications in gene expression. The search for marker genes is widely pursued in almost every differentiation process, although a principled approach is still missing. The current practice is to separate distinguishable cell types, measure gene expression from each cell type, and then identify differentially expressed genes. Such methods require the expression data for both cell types to be available. A limitation of these methods is that by the time the cell types are distinguishable, for example by morphology; many genes have already shown differential expression. This set of differentially expressed genes may include the class of “early marker genes” that are enriched for markers of early differentiating cell lineages as well as genes whose down-regulation triggers differentiation. However, the set of differentially expressed genes will also include a second, larger class of genes in which gene expression is not important to the regulation of the differentiation process but in which genes are simply characteristic of the fully differentiated cell types. Traditional sample comparison procedures are not designed to separate the two classes differentially expressed genes and as a result, the large lists of differentially expressed genes usually do not provide direct guidance for dissecting underlining mechanisms of differentiation. As a benchmark experiment, Zhou et al. used fluorescence activated cell sorting (FACS) to obtain the subset of differentiating mouse ES cells that express a GFP under the control of an Oct4 promoter (Oct4+) and the subset of cells that do not express Oct4-promoter controlled GFP (Oct4-) (Zhou et al., 2007a).

Oct4 is one of the master regulators of self-renewal of mouse ES cells, and its expression level is extensively used as the indicator of the differentiation state (Ivanova et al., 2006). Differentially expressed genes between Oct4+ and Oct4- cells reported by Zhou et al. were used as a benchmark gene list (Zhou et al., 2007a).

Recognizing early marker genes enables separation of cell types at an early stage of differentiation; in turn, separating cell types at an early stage of differentiation enables identification of early marker genes. However, neither piece of the puzzle is currently available to a study of a new differentiation process. We demonstrate that, contrary to common belief, early marker genes can be detected by measuring the average expression of a mixture of cell types, provided that enough biological replicates have been measured and statistical test based on variance ratio has been used. In this study, we provide a novel statistical method to identify early marker genes during differentiation based on the theoretical reasoning, and applying this method to analyze the process of mouse ES cell differentiation, we further performed two validation experiments.

## 3.3 MATERIALS AND METHODS

### The statistical model for the Differentiation-Test

**Model for cell-level transcript copy numbers.** Let $y_{gtrc}$ denote the gene expression level (copy number) of gene transcript $g$ in cell $c$ of biological replicate (sample) $r$ at time $t$. Without loss of generalizability, assume that during the first differentiation event, a parental cell population becomes a mixture of two cell types. For a cell, let $d = d(c) \in \{0,1\}$ denote its cell type: 0 for the parental and 1 for a descendent cell type. Suppose there are $n_r$ cells in biological replicate (sample) $r$. Let $X_{tr}$ denote the proportion of the cells that belong to a differentiated cell type ($d = 1$). The copy number of transcript g can be expressed as:

$$\begin{aligned} y_{gtr\bullet} &= \sum_{c=1}^{n_r} y_{gtrc} \\ &= n_r \mu_{gt} + n_{1r} \beta_{gt} \qquad (1) \\ &= n_r \left( \mu_{gt} + X_{tr} \beta_{gt} \right) \end{aligned}$$

where $n_{1tr}$ are the number of cells of type 1, $\mu_{gt}$ is the mean copy number of transcript $g$ in the parental cell type ($d = 1$), and $\beta_{gt}$ is the difference of the mean copy numbers between the descendent cell type ($d = 1$) and the parental cell type ( $d = 0$ ).The mean of the copy number of transcript $g$ is $n_r \mu_{gt} + n_{1r} \beta_{gt} = n_r \left( \mu_{gt} + X_{tr} \beta_{gt} \right)$.

**Model for raw microarray data.** The raw microarray readouts are the fluorescence intensities of fluorophores attached to the hybridized RNA molecules. These readouts are monotone transformations of the transcript copy numbers with measurement noise. A commonly accepted model between transcript copy number and fluorescence intensity is given by Rocke and Durbin (Rocke and Durbin, 2001):

$$\begin{aligned} w_{gtr} &= f\left( y_{gtr\bullet}, \varepsilon_{gtr} \right) = d + c_g y_{gtr\bullet} + \varepsilon_{gtr} \\ &= d + c_g n_r \left( \mu_{gt} + X_{tr} \beta_{gt} \right) \exp\left( \varepsilon_{gtr} \right) + \varepsilon_{gtr} \end{aligned} \qquad (2)$$

where $\varepsilon_{gtr}$ is a multiplicative error term with $\varepsilon_{gtr} \sim N\left(0,\sigma^2_{\varepsilon,g}\right)$ ; $\varepsilon_{gtr}$ is an additive background noise error term with $\varepsilon_{gtr} \sim N\left(0,\sigma^2_{\varepsilon,g}\right)$; and $c_g$ is a ‘‘unit-conversion’’ constant. Except for low-abundance transcripts, the multiplicative error dominates the additive error and thus the latter can be ignored (Rocke and Durbin, 2001). This practice is consistent with the observation that the microarray readouts are approximately linear to the targeted transcripts (Cope et al., 2004; Irizarry et al., 2003). After normalization and log transformation of the raw data, a normal error model can be derived from (2), which has general support from independent literature (Durbin et al., 2002; Huang et al., 2004):

$$\begin{aligned} Z_{gtr} &= \log\left(w_{gtr}\right) \\ &= \log\left(c_g n\left(\mu_{gt} + X_{tr}\beta_{gt}\right)\right) + \varepsilon_{gtr} \qquad (3) \\ &= \log n + \log c_g + \log\left(\mu_{gt} + X_{tr}\beta_{gt}\right) + \varepsilon_{gtr} \end{aligned}$$

where $Z_{gtr}$ is the normalized and log transformed microarray readout. The normalization removes the differences of cell numbers and overall fluorescence intensities across samples, and therefore the subscript $r$ in $n_r$ was dropped. The independence of mean $\log n + \log c_g + \log\left(\mu_{gt} + X_{tr}\beta_{gt}\right)$ and the technical noise $\varepsilon_{gtr}$ in model (3) was often assumed in published analyses, because the log transformation of the raw data usually removes the dependences between the mean and the variance of the raw array data (see (2)). Nevertheless, to ensure such an independence, the authors recommend first applying the variance stabilization normalization (VSN) (Huber et al., 2003) before performing the following tests.

**The test statistic.** Within the model for raw microarray data, the search for differentially expressed genes is turned into a gene-by-gene test of its differentiation effect:

$$H_0: \quad \beta_{gt} = 0 \quad vs \quad H_1: \quad \beta_{gt} \neq 0 \qquad (4)$$

at time $t$ for gene $g$. To identify an appropriate test statistic, we examine the behavior of the variance of measured data. Given transcript $g$ and time $t$, the variance of its microarray measurement (6) across the replicates is:

$$\begin{aligned} \mathrm{Var}\left(Z_{gtr}\right) &= \mathrm{Var}\left(\log n\right) + \mathrm{Var}\left(\log c_g\right) + \lambda \mathrm{Var}\left(X_{tr}\right)\beta_{gt}^2 + \mathrm{Var}\left(\varepsilon_{gtr}\right) \\ &= 0 + 0 + \lambda \mathrm{Var}\left(X_{tr}\right)\beta_{gt}^2 + \sigma_{\varepsilon,g}^2 \\ &= \lambda \mathrm{Var}\left(X_{tr}\right)\beta_{gt}^2 + \sigma_{\varepsilon,g}^2 \end{aligned} \qquad (5)$$

where $\lambda = 1 \Big/ \left(\mu_{gt} + E\left(X_{tr}\right)\beta_{gt}\right)^2$ is the factor derived by the Delta method of variance calculation (Casella G, 2002). $\log n$ represents the average intensity of the log transformed microarray readouts of the $r^{\text{th}}$ sample, which was adjusted to be the same by almost all normalization procedures, and therefore its variance is 0. Equation (5) shows that the variation of the log transformed microarray readout stems from at least two sources, one being the difference of the proportions of cell types across biological replicates ($\mathrm{Var}\left(X_{tr}\right)$), the other being the measurement error ( $\sigma_{\varepsilon,g}^2$ ). The differentiation effect $\beta_{gt}$ contributes to the first term $\mathrm{Var}\left(X_{tr}\right)\beta_{gt}^2$ in (9). Under the null hypothesis $\beta_{gt} = 0$, this term is 0. Under the alternative hypothesis, this term is positive and contributes to a larger variation of the measurements $Z_{gtr}$. However, a large variation of the measurements $Z_{gtr}$ does not necessarily favor the alternative hypothesis, because it might be confounded by a large measurement error $\sigma^2_{\varepsilon,g}$ . To adjust for the measurement error, the Differentiation-Test uses the ratio of measurement variances across time as the test statistic:

$$DT_{gt} = \frac{\mathrm{Var}\left(Z_{gt}\right)}{\mathrm{Var}\left(Z_{g0}\right)} = \frac{\frac{1}{R-1}\sum_{r=1}^{R}\left(Z_{gtr} - \bar{Z}_{gt\bullet}\right)^2}{\frac{1}{R-1}\sum_{r=1}^{R}\left(Z_{g0r} - \bar{Z}_{g0\bullet}\right)^2} \quad (6)$$

where $\mathrm{Var}\left(Z_{g0}\right)$ is the sample variance of the initial time point. If we assume the differentiation effect is the least manifested at the first time point, the test statistic $DT$ can be used to rank genes for their differentiation effect at time $t$. Under the null hypothesis, the test statistic follows an F-distribution: $DT_{gt} \sim F(R_t - 1, R_0 - 1)$ , where $R_t$ and $R_0$ are the number of biological replicates at time $t$ and time 0, respectively. With the null distribution, the Differentiation-Test reports both the p-value and the q-value (related to false discovery rate) (Storey and Tibshirani, 2003) for every gene. With a q-value cutoff of 0.1, Differentiation-Test reported 137 and 116 genes in 4-day and 8-day EBs, respectively. The overlap of the two gene lists contained 31 genes ( $\text{p-value} = 1.28 \times 10^{-30}$ ). The p-value was generated from the Fisher's Exact Test for enrichment analysis.

**Construction of the gene regulatory network.** The gene regulatory network in 4-day EBs is constructed as follows:

1) Node selection. The Differentiation-Test was applied to 4-day EB and 0-day ES data, and the genes with a q-value threshold of 0.1 were selected.

These genes should express different amounts of transcripts between the ES and the differentiated cells. Among these genes, the ones with Gene Ontology annotation of Transcriptional Regulation (GO: 0003700) and Signal Transduction (GO: 0007165) were selected as nodes of the gene regulatory network.

2) Regulatory relationship. From whole genome transcription factor (TF) or histone modification factor binding data (ChIPseq (Chen et al., 2008)and ChIP-chip (Boyer et al., 2006b)), if one node from step 1 binds to the genomic neighborhood region of another node, then a tentative regulatory relationship is drawn as an undirected edge between the two nodes (Figure 9). Furthermore, gene knockdown followed by microarray analysis data (Ivanova et al., 2006)were merged to the tentative regulatory relationships. When a tentative regulatory relationship is supported by the change of target gene expression after the knockdown of the putative regulatory node, the undirected edge is subsequently changed into a directed edge, with an activation or a repression sign to reflect the concordant or reverse directions of expression changes between the regulator and the target gene.

**Determining the binding site distribution of the transcription factor RBP-J.** 10kb upstream sequences (5k upstream and 5k downstream of the transcription start site) were collected for every gene in Figure 10. The position specific weight matrix (PSWM) of RBP-J was obtained from Transfac database. A sliding window with the same length of the PSWM was used to scan the upstream sequences, on both strands, and a likelihood ratio score was recorded for each sliding window(Jensen and Liu, 2004). Two sets of scores were computed for every upstream sequence.

1. Upstream binding affinity. An upstream binding affinity is a sum of the all the likelihood ratios that are larger than a threshold δ:

$$T_\delta = \sum_{i=1}^{N} LR_i \times I(LR_i \geq \delta) \ ,$$

where LRi is the likelihood ratio of the i$^{th}$ sliding window. N is the total number of sliding windows on a sequence. I(.) is a 0-1 indicator function. When δ is very small, the upstream binding affinity is the same as what's used in Conlon et al. (Conlon et al., 2003). Increasing δ will filter out false positive binding sites from the computation of upstream binding affinity.

2. Number of putative binding sites. The number of putative binding sites of RBP-J is computed by:

$$N_\delta = \sum_{i=1}^{N} I(LR_i \geq \delta) \ ,$$

It is basically the counts of sliding windows that reached the threshold δ.

The average $T_\delta$ and $N_\delta$ were computed from genes in the differentiation module and those in the pluripotency module (Figure 9).

**Transcription profiling.** Total RNA for transcriptional profiling was obtained from B6 mouse ES cells at 0 day (undifferentiated), 4 days and 8 days of spontaneous differentiation. B6 mouse ESC were cultured on mouse embryonic feeders (MEFs) using standard methods as previously described (Ramalho-Santos et al., 2002) in 15% FCS supplemented with LIF. Undifferentiated ES cell samples were obtained by trypsinising near confluent plates of ES cells and depleting the MEFs by plating the cells onto gelatin coated plates for 2 ×20 min. The ES on gelatin samples were MEF depleted ES cells seeded on gelatin coated dishes and cultured until they reached, 70% confluency. To ensure the undifferentiated ES cell samples were free from MEF contamination, MEF depleted ES cells that passaged once on gelatin were used as 0-day ES cell samples. To make EBs, the ES cells on gelatin were seeded into non-adherent petri dishes, and LIF was withdrawn to induce differentiation. Half of the EB media was changed every 3–4 days. The formation of EBs was consistent with previous studies (Doetschman et al., 1985; Robbins et al., 1990). After 8 days, numerous cystic structures were observed and became progressively larger over time. After about 10 days, beating foci of cardiac myocytes could be observed in some EBs, indicating the terminal differentiation of some cell types. Total RNA was extracted from the different samples using the RNeasy kit (Quiagen) and amplified using a two-round linear amplification strategy as previously described (Ramalho-Santos et al., 2002). The labeled RNA was then hybridized to Affymetrix MgU74A microarrays according to the manufacturer’s instructions. Normalization and probe-level modeling were done with dChip software (Li and Wong, 2003).

**Short hairpin RNA mediated knockdown.** Feeder-free E14 mouse ES cells were cultured at 37 ℃ with 5%CO2. All cells were maintained on gelatin-coated dishes in DMEM (Gibco), supplemented with 15% heat-inactivated FBS (Gibco), 0.055 mM b-mercaptoethanol (Gibco), 2 mM l-glutamine, 0.1 mM MEM nonessential amino acid, 5,000 units per ml penicillin–streptomycin, and 1,000 units per ml LIF (Chemicon), as described previously. Transfection of shRNA constructs was performed using Lipofectamine 2000

(Invitrogen) according to manufacturer's instructions. Briefly, 1.5 mg plasmid DNA was transfected into ES cells on 60 mm plates for RNA extraction. Puromycin (Sigma) selection was introduced 1 day after transfection at 1.0 mg/ml, and maintained for 2 and 4 days before harvesting. Detection of alkaline phosphatase, which is indicative of the nondifferentiated state of ES cells, was carried out using a commercial ES cell characterization kit (Chemicon). shRNA targeting specific genes was designed as previously described (Reynolds et al., 2004; Ui-Tei et al., 2004). The 19-nucleotide hairpin-type shRNAs with a 9-nucleotide loop were cloned into pSUPER.puro (Bgl II and Hind III sites, Oligoengine). Three shRNA, targeting different regions of respective transcripts, were designed for each gene to ensure specificity. pSuperpuro constructs expressing shRNA against *luciferase* (Firefly) were used as controls. The 19 nucleotide sequence for each gene is listed below:

***Smarcad1:***

GAAGCTCTGTTTACAAAGA

GAAGAGCGTAAGCAAATTA

GTATGAGGATTACAATGTA

***Pias2:***

GCCCTGCGGTTCAGATTAA

GCCTTCGACTTCAATTACA

GTTCAAGTGTCTTTAGTAA

**RNA extraction, reverse transcription, and quantitative real-time PCR.** Total RNA was extracted using TRIzol Reagent (Invitrogen) and purified with the RNAeasy Mini Kit (Qiagen). Reverse transcription was performed using SuperScript II Kit (Invitrogen). DNA contamination was removed by DNase (Ambion) treatment, and the RNA was further purified by an RNeasy column (Qiagen). Quantitative PCR analyses were performed in real time using an ABI PRISM 7900 sequence detection system and SYBR green master mix, as previously described (Ng et al., 2003). For all the primers used, each gave a single product of the correct size. In all controls lacking reverse transcriptase, no signal was detected. Each RNAi experiment was repeated at least three times with different batches of ES cells.

## 3.4 RESULTS

**The rationale behind the Differentiation-Test.** During the early stages of differentiation, a parental population of cells gives rise to at least one descendent cell type, generating a mixed population of both parent and descendent cells (Figure 3.1).

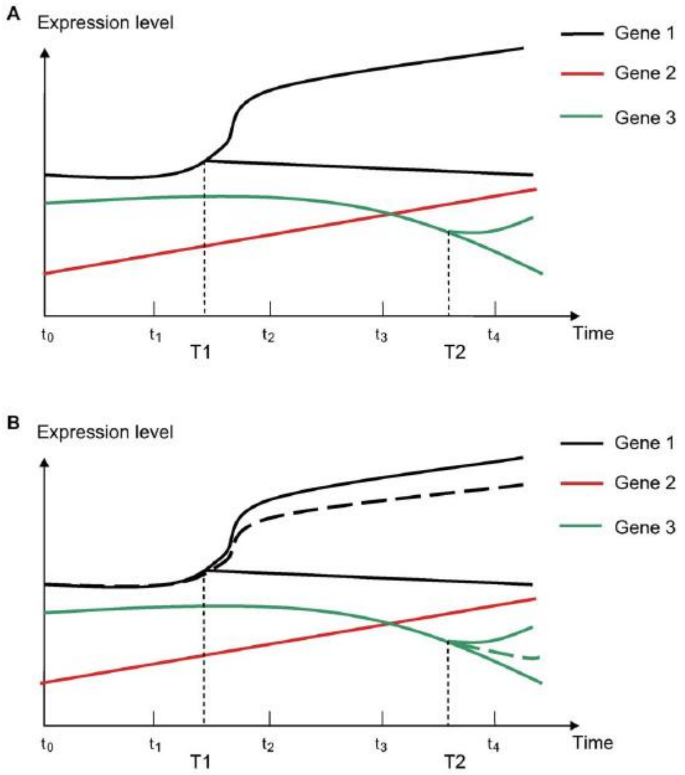


**Figure 3.1**. **A toy example of gene expression levels during a cellular differentiation process.** (A) Two differentiation events happened at T1 and T2, respectively. From T1, Gene 1 has two expression levels in two subsets of cells in the cell mixture. Gene expression data are available at t0 to t4. (B) The solid black and green lines are not observed after T1 and T2, respectively;

instead, the dotted lines are observed as mean expression levels of the cell mixture from microarray data.

In a general experimental design, the average expression of a gene in the cell mixture is measured, for example by microarrays, at a few time points (≥2) during the differentiation process. Biological replicates (≥3) are available for every time point. Our task is to identify the earliest group of genes that have differential expression patterns. For a toy example (Figure 3.1), this group of genes includes Gene 1 only, although all three genes have changed expression values over time. After time T1, the average expression level in a mixed cell population is measured for Gene 1 (dotted line, Figure 3.1B). After T1, the variance of measured expression of Gene 1 across biological replicates should inflate as compared to its variance before T1. The reason for this variance inflation is that the percentage of descendent cells is not identical across biological replicates (Figure 3.2).

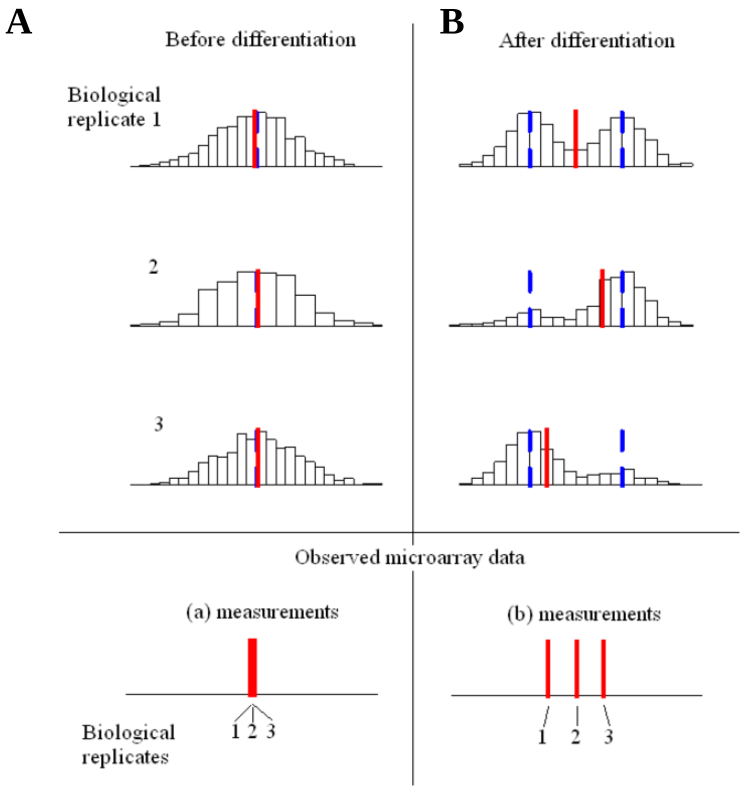


**Figure 3.2. An illustration of the inter-replicate variations of the average expressions of a gene.** An average expression of a gene is measured by microarray profiling in a parent population (A) and a mixture of parental and descendent populations (B). The histograms are for the (unobserved) cell level expressions of a gene. The three biological replicates after differentiation have different mixture proportions of cell types.

For example, at t2, biological replicate 1 may have 50% parental cells and 50% descendent cells, whereas biological replicate 2 may have an 80%–20% split of parental and descendent cells in the mixture (see Fig 5B of Dietrich and Hirragi et al. (Dietrich and Hiiragi, 2007)as an example). In contrast to a nearly 100% parental cell population at t0 for all biological replicates, the difference in percentage of sub-populations after differentiation is a signal that can be utilized in a statistical method, hereafter referred to as Differentiation-Test (Methods).

At the starting point of a differentiation process, nearly 100% of the cells in all biological replicates come from the parental cell population. In other words, almost every cell in any biological replicate at time 0 (Figure 3.1) takes a parental cell type. The gene expression level in each cell has the same (parental) mean, while the actual cell level expression values fluctuate around the mean due to the cell-to-cell variation. Thus a histogram of the cell level expression values of a certain gene in a parental population will show a uni-modal distribution (Figure 3.2A). Suppose at certain stage of the differentiation process, the cell population has been divided into multiple groups. For simplicity of illustration, we assume there are two cell groups after differentiation. A gene differentially expressed in the two groups will have two different mean expression levels (Figure 3.2B). A histogram of cell level expression values of this gene may take a bi-modal shape. These assumptions are supported with experimental data on biological replicate embryos (Dietrich and Hiiragi, 2007). We do not observe the histograms of cell level expressions with microarrays due to the difficulty of conducting the single cell experiments. With microarrays, we can only observe the average expression of all the cells in the population. With a single measurement of the average gene expression it is impossible to distinguish whether the underlying histogram of the cell level expressions is a mixture or uni-modal. Now, suppose we have replicates of cell populations with different mixture proportions (Figure 3.2B1~B3). Since the underlying distribution is a mixture and the mixture proportion is different across the replicates, the observed average expression value (denoted by the red bars in the histogram) varies across the replicates. For the biological replicates before differentiation, the variation due to different mixture proportion of cell types is much less, because all replicates have nearly 100% the parental cell population (Figure 3.2A1~A3).

Now consider two genes. Gene 1 is differentially expressed after differentiation and Gene 2 is not (Figure 3.1). Gene 1 would have an extra source of variation of its mean values across biological replicates compared to Gene 2. In real applications there can be more than two cell types in the cell mixture after differentiation, however the principle holds: a differentially expressed gene would have one more source of variation than a non-differentially expressed gene. Although the description of rationales above has various simplified assumptions, inflation of variance is intrinsic to unsynchronized differentiation events across biological replicates. Neither the model nor the applications assume the parental population is homogeneous.

### Analysis of differentiation of mouse embryonic stem cells.

We used this approach to study the differentiation of mouse ES cells into embryoid bodies (EB). Very early in this differentiation process, different subsets of mouse ES cells start to show different expression changes that then bias the development towards different lineages. These early marker genes are probably small in number, and the timing of their changes in early differentiating cells may be stochastic and exhibit large variation in replicate experiments. As differentiation continues, there will be further changes in the expression of these genes as well as in a larger number of other genes characteristic of the fully differentiated states of the various lineages (e.g., ectoderm, mesoderm, visceral and definitive endoderm). Strictly speaking, a time dependent mixture of two or more cell populations, as formulated in the Methods section and the above titration experiment, is too simplistic to model the setting of mouse ES to EB differentiation. However, the Differentiation-Test derived from such a model should still be applicable in this setting.

At an early time point, such as 4 days after differentiation, the stochastic timing of the changes in an early marker gene will lead to increased variability of its measured expression level in biological replicates. The Differentiation-Test was designed to detect exactly this increased variability. To test this idea, we differentiated mouse ES cells spontaneously into EBs (Figure 3.3). Gene expression of six biological replicates of undifferentiated mouse ES cells (0-day), as well as 4-day, 8-day and 14-day EBs was measured by Affymetrix microarrays (Methods).

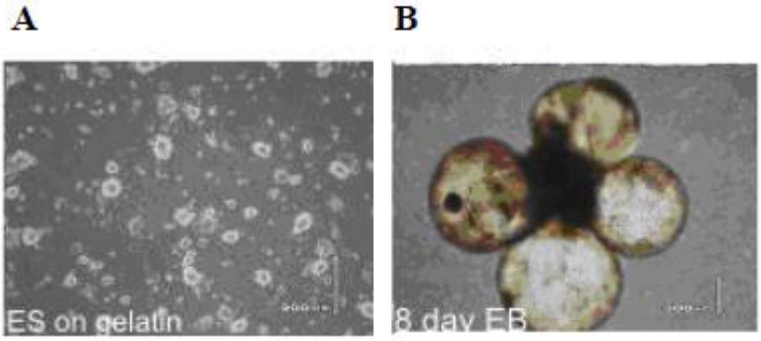


**Figure 3.3. Phase contrast micrographs of differentiating mouse ES cells on gelatin.** (A) 0-day ES cells. (B)8-day EB.

These time points represented early stages of mouse ES differentiation because after 8 days, numerous cystic structures were observed to become progressively larger over time. As an exploratory analysis of data quality, we plotted the scatter plot of standard deviation ( $s_g$ ) vs. mean for every gene at each time point and fitted LOWESS (Locally Weighted Scatterplot Smoothing) regression curves (Ivanova et al., 2006) (Figure 3.4). These plots show that $s_g$ is not influenced by the mean expression value. We therefore did not perform variance stabilization normalization to this dataset.

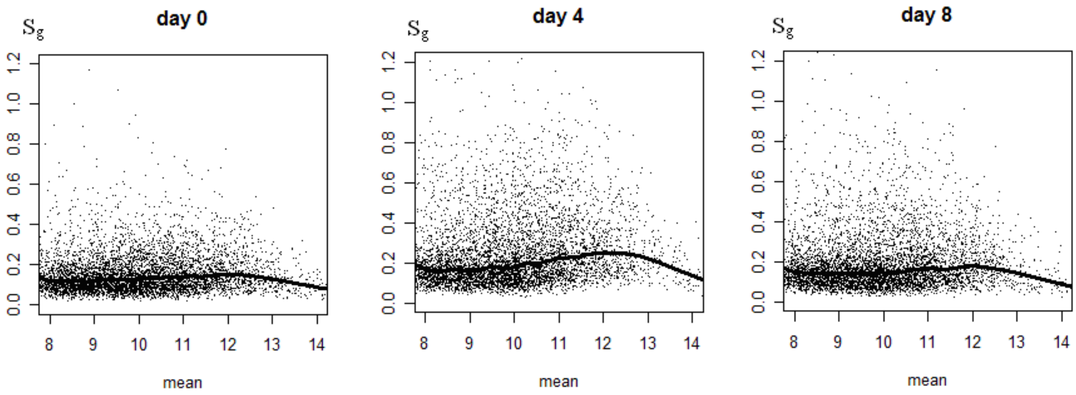


**Figure 3.4. Scatter plots of standard deviation vs. mean.** The mean expression value (x-axis) of a gene across replicate samples is plotted against its standard deviation (y-axis). LOWESS regression curves are shown in the scatter plots.

The variances of 4-day and 8-day EBs were respectively compared to the variance of 0-day ES cells (Figure 3.5). More genes with larger variances were found in 4-day and 8-day samples than in 0-day samples, indicating differential expression might be detectable at 4-day and 8-day stages. As a control, the variance of 0-day samples were compared to that of Oct4+ cells from data of Zhou et al (Zhou et al., 2007a) (Figure 3.5C). An increased number of genes with larger variances were not observed in either 0-day samples or Oct4+ cells.

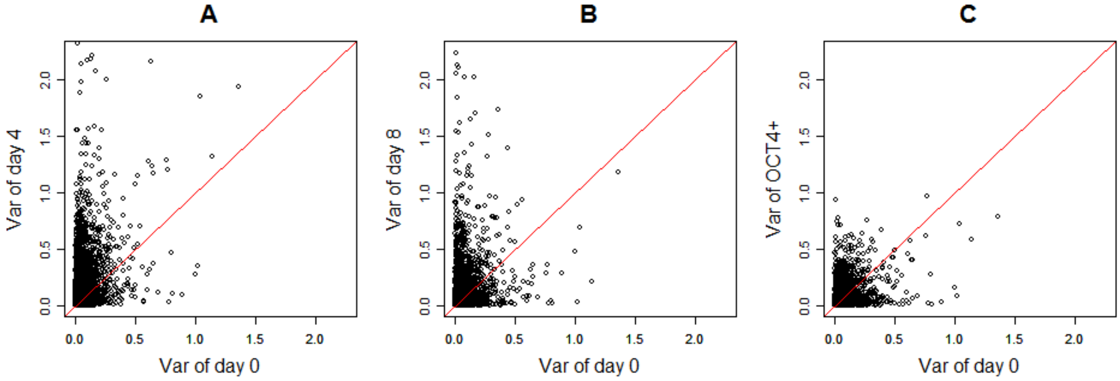


**Figure 3.5. Variance comparison.** Each dot represents a transcript, with its x and y axes representing the variances of the microarray measurements of this transcript at different time points. An increased number of genes with larger variances was observed in 4-day (A) and 8-day EBs (B) as compared to in 0-day ES cells. In contrast, a balanced distribution of variances was observed between 0-day ES cells and Oct4-GFP positive sorted cells (C).

Then, we applied the Differentiation-Test to this dataset and identified the top 200 differentially expressed genes of 4-day and 8-day EBs. The statistical significance of the overlap between the Differentiation-Test reported gene lists and the benchmark genes from Zhou et al was assessed by Fisher's Exact Test, generating p-values of $3.8 \times 10^{-8}$ and $1.7 \times 10^{-9}$ for 4-day and 8-day EBs, respectively. These small p-values were not due to a particular cutoff of the number of top-ranking genes reported (Table 3.1).

**Table 3.1 Fisher's Exact Tests between top-ranked genes of the Differentiation-Test and benchmark gene list.**

| | Day 4 | Day 4 | Day 8 | Day 8 |
|---|---|---|---|---|
| Reported Genes | Number of overlaps | p-value | Number of overlaps | p-value |
| top 100 | 23 | 1.4E-03 | 31 | 3.2E-07 |
| top 200 | 52 | 3.8E-08 | 55 | 1.7E-09 |
| top 300 | 70 | 2.5E-08 | 77 | 4.0E-11 |
| top 400 | 85 | 8.1E-08 | 99 | 6.4E-13 |
| top 500 | 102 | 3.6E-08 | 120 | 2.1E-14 |
| top 600 | 120 | 7.0E-09 | 140 | 1.3E-15 |
| top 700 | 133 | 2.8E-08 | 158 | 3.2E-16 |
| top 800 | 146 | 8.2E-08 | 179 | 4.0E-18 |
| top 900 | 158 | 3.1E-07 | 194 | 8.8E-18 |
| top 1000 | 174 | 1.3E-07 | 210 | 6.8E-18 |

In contrast, in testing 10,000 random lists of 200 genes each against the benchmark list, none (0%) of these reached p-values as significant as $3.8\times10^{-8}$ and $1.7\times10^{-9}$ (Figure 3.6).

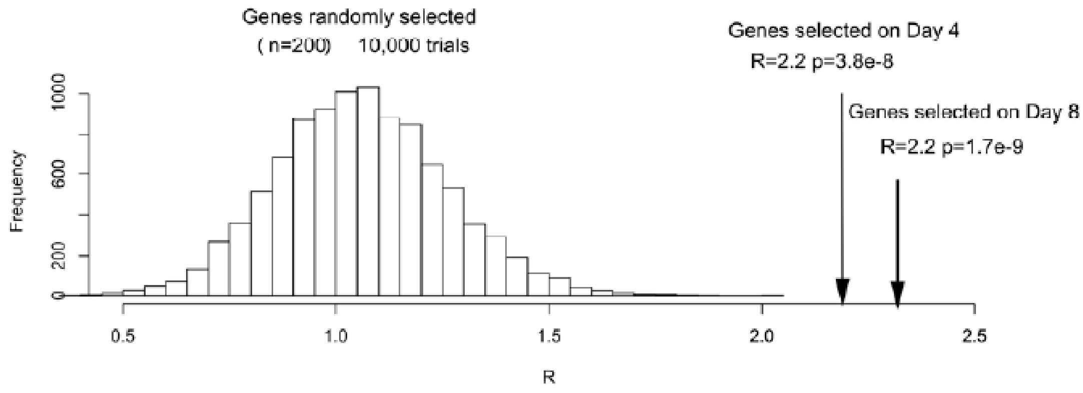
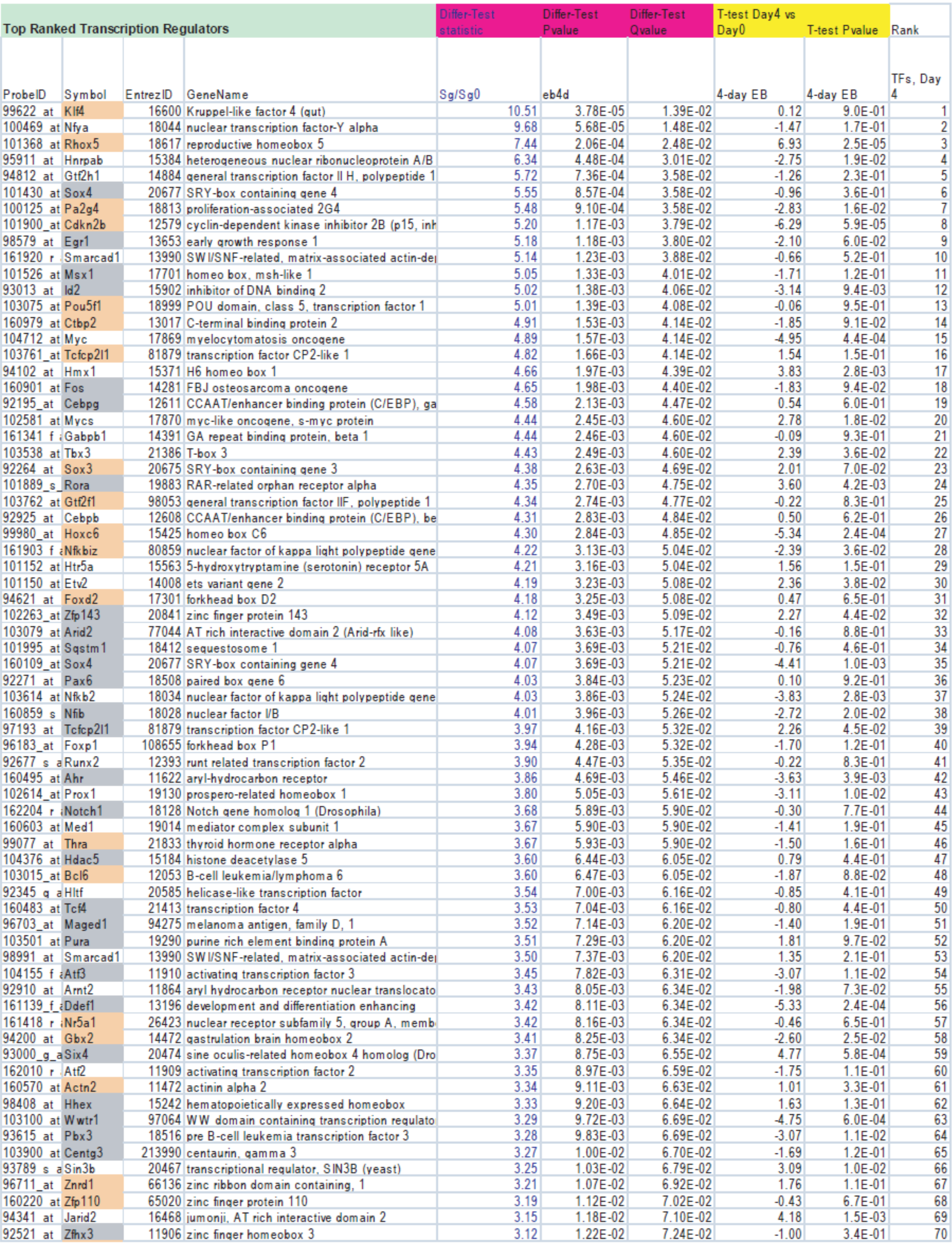


**Figure 3.6 Significance calibration from 10,000 random gene lists.** 10,000 randomly picked gene lists of 200 genes each were compared to the benchmark gene list. A histogram of calculated R values is shown. R = K/E (K), where K is the number of overlapped genes between a random list and the benchmark list, and E(K) is its expectation. Out of the 10,000 R values, only

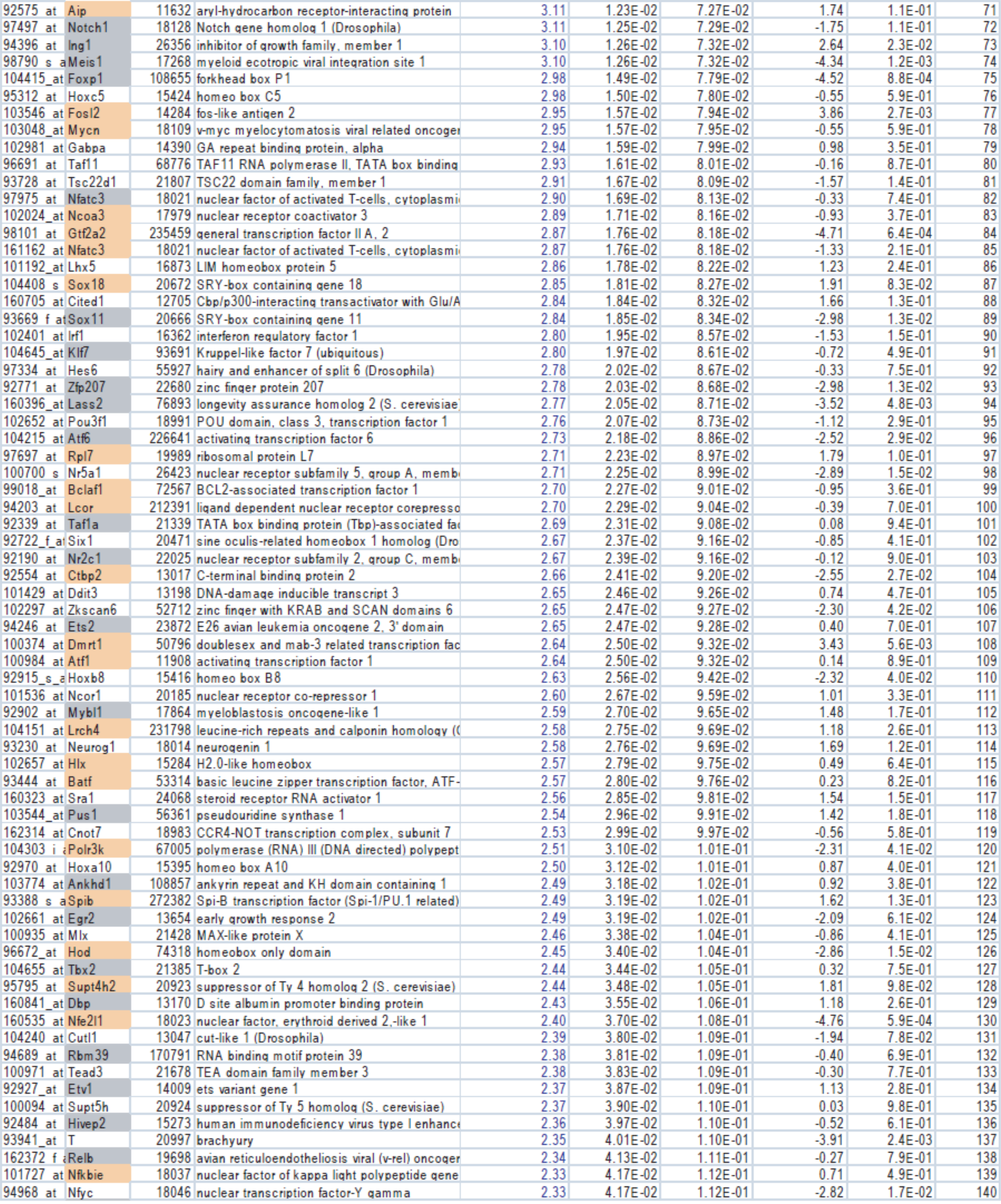
one was greater than the Differentiation-Test's 4-day R value (=2.2); none of them was greater than the Differentiation-Test's 8-day R value (=2.3).

In fact, the Differentiation-Test's top-ranked transcription regulators in 4-day EBs (Table 3.2) included a number of markers of early differentiation, including *Sox4*, *Egr1*, *Id2*, and *Pax6* (ranked as 6, 9, 12, and 36, respectively), as well as known self-renewal regulators of mouse ES cells, including *Klf4* (Jiang et al., 2008), and *Pou5f1* (Nichols et al., 1998; Niwa et al., 2000)(ranked 1 and 13, respectively). In contrast, a traditional T-test between 4-day EBs and undifferentiated mouse ES cells failed to reveal any of these differentially expressed genes because 4-day EBs still had a similar mean expression of the marker genes as 0-day mouse ES cells (Column H, Table 2). For example, T-test p-values for *Klf4* and *Pou5f1* are 0.90 and 0.95, respectively. These test results suggest that the Differentiation-Test detected differentially expressed genes in a very early stage of the differentiation process, generating consistent results to those obtained from a laborious experimental procedure of cell sorting. Cell sorting requires prior knowledge of a marker gene that is differentially expressed which may not be available for every differentiation process in future studies.

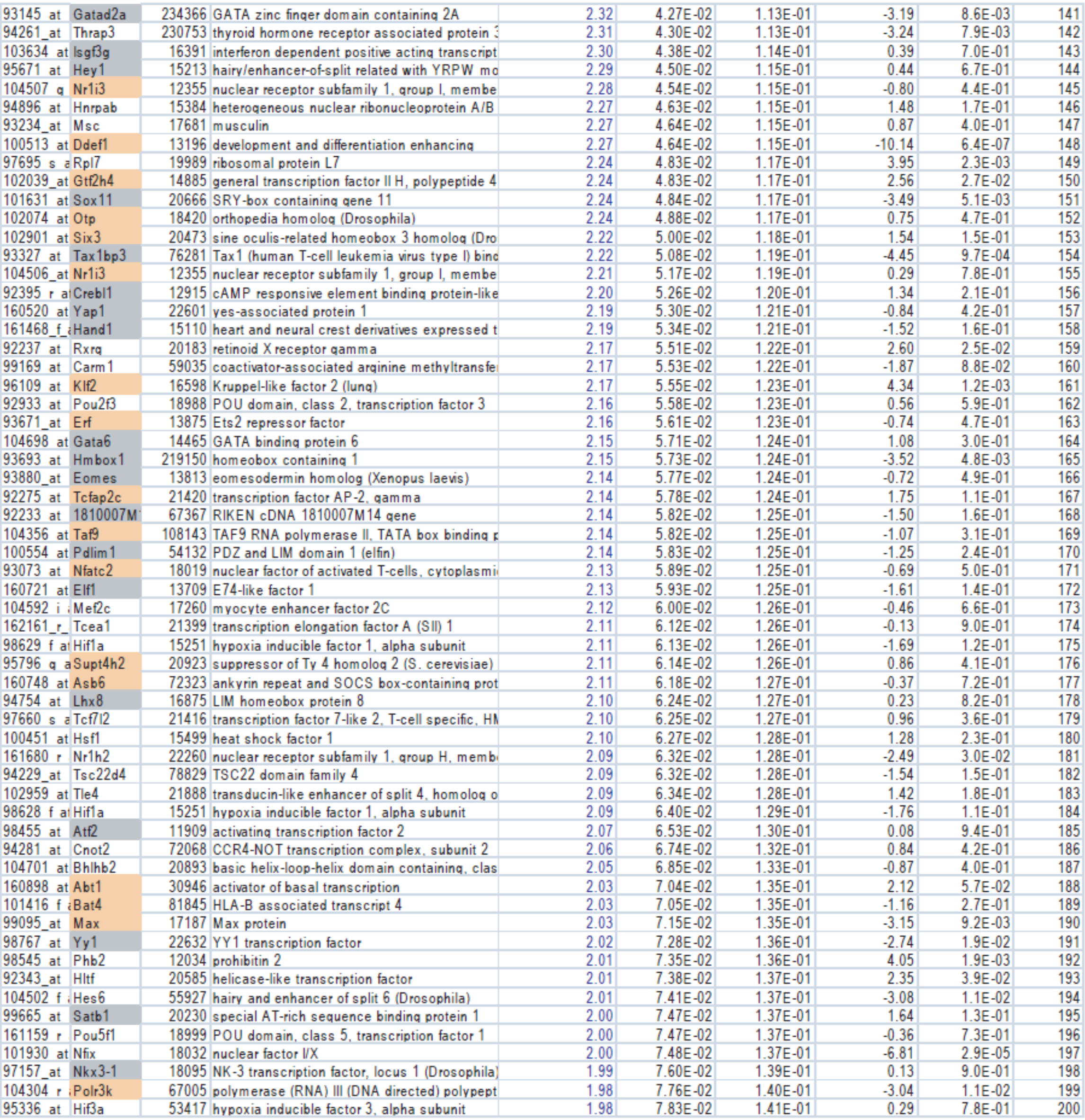
**Table 3.2. 200 top-ranked differentially expressed transcription regulators from the Differentiation-Test in 4-day EBs.**



| Top Ranked Transcription Regulators | | | | Differ-Test statistic | Differ-Test Pvalue | Differ-Test Qvalue | T-test Day4 vs Day0 | T-test Pvalue | Rank |
|---|---|---|---|---|---|---|---|---|---|
| ProbeID | Symbol | EntrezID | GeneName | Sg/Sg0 | eb4d | | 4-day EB | 4-day EB | TFs, Day 4 |
| 99622 at | Klf4 | 16600 | Kruppel-like factor 4 (gut) | 10.51 | 3.78E-05 | 1.39E-02 | 0.12 | 9.0E-01 | 1 |
| 100469 at | Nfya | 18044 | nuclear transcription factor-Y alpha | 9.68 | 5.68E-05 | 1.48E-02 | -1.47 | 1.7E-01 | 2 |
| 101368 at | Rhox5 | 18617 | reproductive homeobox 5 | 7.44 | 2.06E-04 | 2.48E-02 | 6.93 | 2.5E-05 | 3 |
| 95911 at | Hnrpab | 15384 | heterogeneous nuclear ribonucleoprotein A/B | 6.34 | 4.48E-04 | 3.01E-02 | -2.75 | 1.9E-02 | 4 |
| 94812 at | Gtf2h1 | 14884 | general transcription factor II H, polypeptide 1 | 5.72 | 7.36E-04 | 3.58E-02 | -1.26 | 2.3E-01 | 5 |
| 101430 at | Sox4 | 20677 | SRY-box containing gene 4 | 5.55 | 8.57E-04 | 3.58E-02 | -0.96 | 3.6E-01 | 6 |
| 100125 at | Pa2g4 | 18813 | proliferation-associated 2G4 | 5.48 | 9.10E-04 | 3.58E-02 | -2.83 | 1.6E-02 | 7 |
| 101900_at | Cdkn2b | 12579 | cyclin-dependent kinase inhibitor 2B (p15, inh | 5.20 | 1.17E-03 | 3.79E-02 | -6.29 | 5.9E-05 | 8 |
| 98579 at | Egr1 | 13653 | early growth response 1 | 5.18 | 1.18E-03 | 3.80E-02 | -2.10 | 6.0E-02 | 9 |
| 161920 r | Smarcad1 | 13990 | SWI/SNF-related, matrix-associated actin-de | 5.14 | 1.23E-03 | 3.88E-02 | -0.66 | 5.2E-01 | 10 |
| 101526 at | Msx1 | 17701 | homeo box, msh-like 1 | 5.05 | 1.33E-03 | 4.01E-02 | -1.71 | 1.2E-01 | 11 |
| 93013 at | Id2 | 15902 | inhibitor of DNA binding 2 | 5.02 | 1.38E-03 | 4.06E-02 | -3.14 | 9.4E-03 | 12 |
| 103075 at | Pou5f1 | 18999 | POU domain, class 5, transcription factor 1 | 5.01 | 1.39E-03 | 4.08E-02 | -0.06 | 9.5E-01 | 13 |
| 160979 at | Ctbp2 | 13017 | C-terminal binding protein 2 | 4.91 | 1.53E-03 | 4.14E-02 | -1.85 | 9.1E-02 | 14 |
| 104712 at | Myc | 17869 | myelocytomatosis oncogene | 4.89 | 1.57E-03 | 4.14E-02 | -4.95 | 4.4E-04 | 15 |
| 103761_at | Tcfcp2l1 | 81879 | transcription factor CP2-like 1 | 4.82 | 1.66E-03 | 4.14E-02 | 1.54 | 1.5E-01 | 16 |
| 94102 at | Hmx1 | 15371 | H6 homeo box 1 | 4.66 | 1.97E-03 | 4.39E-02 | 3.83 | 2.8E-03 | 17 |
| 160901 at | Fos | 14281 | FBJ osteosarcoma oncogene | 4.65 | 1.98E-03 | 4.40E-02 | -1.83 | 9.4E-02 | 18 |
| 92195_at | Cebpg | 12611 | CCAAT/enhancer binding protein (C/EBP), ga | 4.58 | 2.13E-03 | 4.47E-02 | 0.54 | 6.0E-01 | 19 |
| 102581 at | Mycs | 17870 | myc-like oncogene, s-myc protein | 4.44 | 2.45E-03 | 4.60E-02 | 2.78 | 1.8E-02 | 20 |
| 161341 f | Gabpb1 | 14391 | GA repeat binding protein, beta 1 | 4.44 | 2.46E-03 | 4.60E-02 | -0.09 | 9.3E-01 | 21 |
| 103538 at | Tbx3 | 21386 | T-box 3 | 4.43 | 2.49E-03 | 4.60E-02 | 2.39 | 3.6E-02 | 22 |
| 92264 at | Sox3 | 20675 | SRY-box containing gene 3 | 4.38 | 2.63E-03 | 4.69E-02 | 2.01 | 7.0E-02 | 23 |
| 101889_s_ | Rora | 19883 | RAR-related orphan receptor alpha | 4.35 | 2.70E-03 | 4.75E-02 | 3.60 | 4.2E-03 | 24 |
| 103762 at | Gtf2f1 | 98053 | general transcription factor IIF, polypeptide 1 | 4.34 | 2.74E-03 | 4.77E-02 | -0.22 | 8.3E-01 | 25 |
| 92925 at | Cebpb | 12608 | CCAAT/enhancer binding protein (C/EBP), be | 4.31 | 2.83E-03 | 4.84E-02 | 0.50 | 6.2E-01 | 26 |
| 99980_at | Hoxc6 | 15425 | homeo box C6 | 4.30 | 2.84E-03 | 4.85E-02 | -5.34 | 2.4E-04 | 27 |
| 161903 f | Nfkbiz | 80859 | nuclear factor of kappa light polypeptide gene | 4.22 | 3.13E-03 | 5.04E-02 | -2.39 | 3.6E-02 | 28 |
| 101152 at | Htr5a | 15563 | 5-hydroxytryptamine (serotonin) receptor 5A | 4.21 | 3.16E-03 | 5.04E-02 | 1.56 | 1.5E-01 | 29 |
| 101150 at | Etv2 | 14008 | ets variant gene 2 | 4.19 | 3.23E-03 | 5.08E-02 | 2.36 | 3.8E-02 | 30 |
| 94621 at | Foxd2 | 17301 | forkhead box D2 | 4.18 | 3.25E-03 | 5.08E-02 | 0.47 | 6.5E-01 | 31 |
| 102263_at | Zfp143 | 20841 | zinc finger protein 143 | 4.12 | 3.49E-03 | 5.09E-02 | 2.27 | 4.4E-02 | 32 |
| 103079 at | Arid2 | 77044 | AT rich interactive domain 2 (Arid-rfx like) | 4.08 | 3.63E-03 | 5.17E-02 | -0.16 | 8.8E-01 | 33 |
| 101995 at | Sqstm1 | 18412 | sequestosome 1 | 4.07 | 3.69E-03 | 5.21E-02 | -0.76 | 4.6E-01 | 34 |
| 160109_at | Sox4 | 20677 | SRY-box containing gene 4 | 4.07 | 3.69E-03 | 5.21E-02 | -4.41 | 1.0E-03 | 35 |
| 92271 at | Pax6 | 18508 | paired box gene 6 | 4.03 | 3.84E-03 | 5.23E-02 | 0.10 | 9.2E-01 | 36 |
| 103614 at | Nfkb2 | 18034 | nuclear factor of kappa light polypeptide gene | 4.03 | 3.86E-03 | 5.24E-02 | -3.83 | 2.8E-03 | 37 |
| 160859 s | Nfib | 18028 | nuclear factor I/B | 4.01 | 3.96E-03 | 5.26E-02 | -2.72 | 2.0E-02 | 38 |
| 97193 at | Tcfcp2l1 | 81879 | transcription factor CP2-like 1 | 3.97 | 4.16E-03 | 5.32E-02 | 2.26 | 4.5E-02 | 39 |
| 96183_at | Foxp1 | 108655 | forkhead box P1 | 3.94 | 4.28E-03 | 5.32E-02 | -1.70 | 1.2E-01 | 40 |
| 92677 s a | Runx2 | 12393 | runt related transcription factor 2 | 3.90 | 4.47E-03 | 5.35E-02 | -0.22 | 8.3E-01 | 41 |
| 160495 at | Ahr | 11622 | aryl-hydrocarbon receptor | 3.86 | 4.69E-03 | 5.46E-02 | -3.63 | 3.9E-03 | 42 |
| 102614_at | Prox1 | 19130 | prospero-related homeobox 1 | 3.80 | 5.05E-03 | 5.61E-02 | -3.11 | 1.0E-02 | 43 |
| 162204 r | Notch1 | 18128 | Notch gene homolog 1 (Drosophila) | 3.68 | 5.89E-03 | 5.90E-02 | -0.30 | 7.7E-01 | 44 |
| 160603 at | Med1 | 19014 | mediator complex subunit 1 | 3.67 | 5.90E-03 | 5.90E-02 | -1.41 | 1.9E-01 | 45 |
| 99077 at | Thra | 21833 | thyroid hormone receptor alpha | 3.67 | 5.93E-03 | 5.90E-02 | -1.50 | 1.6E-01 | 46 |
| 104376 at | Hdac5 | 15184 | histone deacetylase 5 | 3.60 | 6.44E-03 | 6.05E-02 | 0.79 | 4.4E-01 | 47 |
| 103015_at | Bcl6 | 12053 | B-cell leukemia/lymphoma 6 | 3.60 | 6.47E-03 | 6.05E-02 | -1.87 | 8.8E-02 | 48 |
| 92345 g a | Hltf | 20585 | helicase-like transcription factor | 3.54 | 7.00E-03 | 6.16E-02 | -0.85 | 4.1E-01 | 49 |
| 160483 at | Tcf4 | 21413 | transcription factor 4 | 3.53 | 7.04E-03 | 6.16E-02 | -0.80 | 4.4E-01 | 50 |
| 96703_at | Maged1 | 94275 | melanoma antigen, family D, 1 | 3.52 | 7.14E-03 | 6.20E-02 | -1.40 | 1.9E-01 | 51 |
| 103501 at | Pura | 19290 | purine rich element binding protein A | 3.51 | 7.29E-03 | 6.20E-02 | 1.81 | 9.7E-02 | 52 |
| 98991 at | Smarcad1 | 13990 | SWI/SNF-related, matrix-associated actin-de | 3.50 | 7.37E-03 | 6.20E-02 | 1.35 | 2.1E-01 | 53 |
| 104155 f | Atf3 | 11910 | activating transcription factor 3 | 3.45 | 7.82E-03 | 6.31E-02 | -3.07 | 1.1E-02 | 54 |
| 92910 at | Arnt2 | 11864 | aryl hydrocarbon receptor nuclear translocato | 3.43 | 8.05E-03 | 6.34E-02 | -1.98 | 7.3E-02 | 55 |
| 161139_f_ | Ddef1 | 13196 | development and differentiation enhancing | 3.42 | 8.11E-03 | 6.34E-02 | -5.33 | 2.4E-04 | 56 |
| 161418 r | Nr5a1 | 26423 | nuclear receptor subfamily 5, group A, memb | 3.42 | 8.16E-03 | 6.34E-02 | -0.46 | 6.5E-01 | 57 |
| 94200 at | Gbx2 | 14472 | gastrulation brain homeobox 2 | 3.41 | 8.25E-03 | 6.34E-02 | -2.60 | 2.5E-02 | 58 |
| 93000_g_a | Six4 | 20474 | sine oculis-related homeobox 4 homolog (Dro | 3.37 | 8.75E-03 | 6.55E-02 | 4.77 | 5.8E-04 | 59 |
| 162010 r | Atf2 | 11909 | activating transcription factor 2 | 3.35 | 8.97E-03 | 6.59E-02 | -1.75 | 1.1E-01 | 60 |
| 160570 at | Actn2 | 11472 | actinin alpha 2 | 3.34 | 9.11E-03 | 6.63E-02 | 1.01 | 3.3E-01 | 61 |
| 98408 at | Hhex | 15242 | hematopoietically expressed homeobox | 3.33 | 9.20E-03 | 6.64E-02 | 1.63 | 1.3E-01 | 62 |
| 103100 at | Wwtr1 | 97064 | WW domain containing transcription regulato | 3.29 | 9.72E-03 | 6.69E-02 | -4.75 | 6.0E-04 | 63 |
| 93615 at | Pbx3 | 18516 | pre B-cell leukemia transcription factor 3 | 3.28 | 9.83E-03 | 6.69E-02 | -3.07 | 1.1E-02 | 64 |
| 103900 at | Centg3 | 213990 | centaurin, gamma 3 | 3.27 | 1.00E-02 | 6.70E-02 | -1.69 | 1.2E-01 | 65 |
| 93789 s a | Sin3b | 20467 | transcriptional regulator, SIN3B (yeast) | 3.25 | 1.03E-02 | 6.79E-02 | 3.09 | 1.0E-02 | 66 |
| 96711_at | Znrd1 | 66136 | zinc ribbon domain containing, 1 | 3.21 | 1.07E-02 | 6.92E-02 | 1.76 | 1.1E-01 | 67 |
| 160220 at | Zfp110 | 65020 | zinc finger protein 110 | 3.19 | 1.12E-02 | 7.02E-02 | -0.43 | 6.7E-01 | 68 |
| 94341 at | Jarid2 | 16468 | jumonji, AT rich interactive domain 2 | 3.15 | 1.18E-02 | 7.10E-02 | 4.18 | 1.5E-03 | 69 |
| 92521 at | Zfhx3 | 11906 | zinc finger homeobox 3 | 3.12 | 1.22E-02 | 7.24E-02 | -1.00 | 3.4E-01 | 70 |



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 92575 at | Aip | 11632 | aryl-hydrocarbon receptor-interacting protein | 3.11 | 1.23E-02 | 7.27E-02 | 1.74 | 1.1E-01 | 71 |
| 97497 at | Notch1 | 18128 | Notch gene homolog 1 (Drosophila) | 3.11 | 1.25E-02 | 7.29E-02 | -1.75 | 1.1E-01 | 72 |
| 94396 at | Ing1 | 26356 | inhibitor of growth family, member 1 | 3.10 | 1.26E-02 | 7.32E-02 | 2.64 | 2.3E-02 | 73 |
| 98790 s a | Meis1 | 17268 | myeloid ecotropic viral integration site 1 | 3.10 | 1.26E-02 | 7.32E-02 | -4.34 | 1.2E-03 | 74 |
| 104415_at | Foxp1 | 108655 | forkhead box P1 | 2.98 | 1.49E-02 | 7.79E-02 | -4.52 | 8.8E-04 | 75 |
| 95312 at | Hoxc5 | 15424 | homeo box C5 | 2.98 | 1.50E-02 | 7.80E-02 | -0.55 | 5.9E-01 | 76 |
| 103546 at | Fosl2 | 14284 | fos-like antigen 2 | 2.95 | 1.57E-02 | 7.94E-02 | 3.86 | 2.7E-03 | 77 |
| 103048_at | Mycn | 18109 | v-myc myelocytomatosis viral related oncoger | 2.95 | 1.57E-02 | 7.95E-02 | -0.55 | 5.9E-01 | 78 |
| 102981 at | Gabpa | 14390 | GA repeat binding protein, alpha | 2.94 | 1.59E-02 | 7.99E-02 | 0.98 | 3.5E-01 | 79 |
| 96691 at | Taf11 | 68776 | TAF11 RNA polymerase II, TATA box binding | 2.93 | 1.61E-02 | 8.01E-02 | -0.16 | 8.7E-01 | 80 |
| 93728 at | Tsc22d1 | 21807 | TSC22 domain family, member 1 | 2.91 | 1.67E-02 | 8.09E-02 | -1.57 | 1.4E-01 | 81 |
| 97975 at | Nfatc3 | 18021 | nuclear factor of activated T-cells, cytoplasmi | 2.90 | 1.69E-02 | 8.13E-02 | -0.33 | 7.4E-01 | 82 |
| 102024_at | Ncoa3 | 17979 | nuclear receptor coactivator 3 | 2.89 | 1.71E-02 | 8.16E-02 | -0.93 | 3.7E-01 | 83 |
| 98101 at | Gtf2a2 | 235459 | general transcription factor II A, 2 | 2.87 | 1.76E-02 | 8.18E-02 | -4.71 | 6.4E-04 | 84 |
| 161162 at | Nfatc3 | 18021 | nuclear factor of activated T-cells, cytoplasmi | 2.87 | 1.76E-02 | 8.18E-02 | -1.33 | 2.1E-01 | 85 |
| 101192_at | Lhx5 | 16873 | LIM homeobox protein 5 | 2.86 | 1.78E-02 | 8.22E-02 | 1.23 | 2.4E-01 | 86 |
| 104408 s | Sox18 | 20672 | SRY-box containing gene 18 | 2.85 | 1.81E-02 | 8.27E-02 | 1.91 | 8.3E-02 | 87 |
| 160705 at | Cited1 | 12705 | Cbp/p300-interacting transactivator with Glu/A | 2.84 | 1.84E-02 | 8.32E-02 | 1.66 | 1.3E-01 | 88 |
| 93669 f at | Sox11 | 20666 | SRY-box containing gene 11 | 2.84 | 1.85E-02 | 8.34E-02 | -2.98 | 1.3E-02 | 89 |
| 102401 at | Irf1 | 16362 | interferon regulatory factor 1 | 2.80 | 1.95E-02 | 8.57E-02 | -1.53 | 1.5E-01 | 90 |
| 104645_at | Klf7 | 93691 | Kruppel-like factor 7 (ubiquitous) | 2.80 | 1.97E-02 | 8.61E-02 | -0.72 | 4.9E-01 | 91 |
| 97334 at | Hes6 | 55927 | hairy and enhancer of split 6 (Drosophila) | 2.78 | 2.02E-02 | 8.67E-02 | -0.33 | 7.5E-01 | 92 |
| 92771 at | Zfp207 | 22680 | zinc finger protein 207 | 2.78 | 2.03E-02 | 8.68E-02 | -2.98 | 1.3E-02 | 93 |
| 160396_at | Lass2 | 76893 | longevity assurance homolog 2 (S. cerevisiae) | 2.77 | 2.05E-02 | 8.71E-02 | -3.52 | 4.8E-03 | 94 |
| 102652 at | Pou3f1 | 18991 | POU domain, class 3, transcription factor 1 | 2.76 | 2.07E-02 | 8.73E-02 | -1.12 | 2.9E-01 | 95 |
| 104215 at | Atf6 | 226641 | activating transcription factor 6 | 2.73 | 2.18E-02 | 8.86E-02 | -2.52 | 2.9E-02 | 96 |
| 97697 at | Rpl7 | 19989 | ribosomal protein L7 | 2.71 | 2.23E-02 | 8.97E-02 | 1.79 | 1.0E-01 | 97 |
| 100700 s | Nr5a1 | 26423 | nuclear receptor subfamily 5, group A, memb | 2.71 | 2.25E-02 | 8.99E-02 | -2.89 | 1.5E-02 | 98 |
| 99018_at | Bclaf1 | 72567 | BCL2-associated transcription factor 1 | 2.70 | 2.27E-02 | 9.01E-02 | -0.95 | 3.6E-01 | 99 |
| 94203 at | Lcor | 212391 | ligand dependent nuclear receptor corepresso | 2.70 | 2.29E-02 | 9.04E-02 | -0.39 | 7.0E-01 | 100 |
| 92339 at | Taf1a | 21339 | TATA box binding protein (Tbp)-associated fa | 2.69 | 2.31E-02 | 9.08E-02 | 0.08 | 9.4E-01 | 101 |
| 92722_f_a | Six1 | 20471 | sine oculis-related homeobox 1 homolog (Dro | 2.67 | 2.37E-02 | 9.16E-02 | -0.85 | 4.1E-01 | 102 |
| 92190 at | Nr2c1 | 22025 | nuclear receptor subfamily 2, group C, memb | 2.67 | 2.39E-02 | 9.16E-02 | -0.12 | 9.0E-01 | 103 |
| 92554 at | Ctbp2 | 13017 | C-terminal binding protein 2 | 2.66 | 2.41E-02 | 9.20E-02 | -2.55 | 2.7E-02 | 104 |
| 101429 at | Ddit3 | 13198 | DNA-damage inducible transcript 3 | 2.65 | 2.46E-02 | 9.26E-02 | 0.74 | 4.7E-01 | 105 |
| 102297 at | Zkscan6 | 52712 | zinc finger with KRAB and SCAN domains 6 | 2.65 | 2.47E-02 | 9.27E-02 | -2.30 | 4.2E-02 | 106 |
| 94246 at | Ets2 | 23872 | E26 avian leukemia oncogene 2, 3' domain | 2.65 | 2.47E-02 | 9.28E-02 | 0.40 | 7.0E-01 | 107 |
| 100374 at | Dmrt1 | 50796 | doublesex and mab-3 related transcription fac | 2.64 | 2.50E-02 | 9.32E-02 | 3.43 | 5.6E-03 | 108 |
| 100984 at | Atf1 | 11908 | activating transcription factor 1 | 2.64 | 2.50E-02 | 9.32E-02 | 0.14 | 8.9E-01 | 109 |
| 92915_s_a | Hoxb8 | 15416 | homeo box B8 | 2.63 | 2.56E-02 | 9.42E-02 | -2.32 | 4.0E-02 | 110 |
| 101536 at | Ncor1 | 20185 | nuclear receptor co-repressor 1 | 2.60 | 2.67E-02 | 9.59E-02 | 1.01 | 3.3E-01 | 111 |
| 92902 at | Mybl1 | 17864 | myeloblastosis oncogene-like 1 | 2.59 | 2.70E-02 | 9.65E-02 | 1.48 | 1.7E-01 | 112 |
| 104151 at | Lrch4 | 231798 | leucine-rich repeats and calponin homology (( | 2.58 | 2.75E-02 | 9.69E-02 | 1.18 | 2.6E-01 | 113 |
| 93230 at | Neurog1 | 18014 | neurogenin 1 | 2.58 | 2.76E-02 | 9.69E-02 | 1.69 | 1.2E-01 | 114 |
| 102657 at | Hlx | 15284 | H2.0-like homeobox | 2.57 | 2.79E-02 | 9.75E-02 | 0.49 | 6.4E-01 | 115 |
| 93444 at | Batf | 53314 | basic leucine zipper transcription factor, ATF- | 2.57 | 2.80E-02 | 9.76E-02 | 0.23 | 8.2E-01 | 116 |
| 160323 at | Sra1 | 24068 | steroid receptor RNA activator 1 | 2.56 | 2.85E-02 | 9.81E-02 | 1.54 | 1.5E-01 | 117 |
| 103544_at | Pus1 | 56361 | pseudouridine synthase 1 | 2.54 | 2.96E-02 | 9.91E-02 | 1.42 | 1.8E-01 | 118 |
| 162314 at | Cnot7 | 18983 | CCR4-NOT transcription complex, subunit 7 | 2.53 | 2.99E-02 | 9.97E-02 | -0.56 | 5.8E-01 | 119 |
| 104303 i a | Polr3k | 67005 | polymerase (RNA) III (DNA directed) polypept | 2.51 | 3.10E-02 | 1.01E-01 | -2.31 | 4.1E-02 | 120 |
| 92970 at | Hoxa10 | 15395 | homeo box A10 | 2.50 | 3.12E-02 | 1.01E-01 | 0.87 | 4.0E-01 | 121 |
| 103774 at | Ankhd1 | 108857 | ankyrin repeat and KH domain containing 1 | 2.49 | 3.18E-02 | 1.02E-01 | 0.92 | 3.8E-01 | 122 |
| 93388 s a | Spib | 272382 | Spi-B transcription factor (Spi-1/PU.1 related) | 2.49 | 3.19E-02 | 1.02E-01 | 1.62 | 1.3E-01 | 123 |
| 102661 at | Egr2 | 13654 | early growth response 2 | 2.49 | 3.19E-02 | 1.02E-01 | -2.09 | 6.1E-02 | 124 |
| 100935 at | Mlx | 21428 | MAX-like protein X | 2.46 | 3.38E-02 | 1.04E-01 | -0.86 | 4.1E-01 | 125 |
| 96672_at | Hod | 74318 | homeobox only domain | 2.45 | 3.40E-02 | 1.04E-01 | -2.86 | 1.5E-02 | 126 |
| 104655 at | Tbx2 | 21385 | T-box 2 | 2.44 | 3.44E-02 | 1.05E-01 | 0.32 | 7.5E-01 | 127 |
| 95795 at | Supt4h2 | 20923 | suppressor of Ty 4 homolog 2 (S. cerevisiae) | 2.44 | 3.48E-02 | 1.05E-01 | 1.81 | 9.8E-02 | 128 |
| 160841_at | Dbp | 13170 | D site albumin promoter binding protein | 2.43 | 3.55E-02 | 1.06E-01 | 1.18 | 2.6E-01 | 129 |
| 160535 at | Nfe2l1 | 18023 | nuclear factor, erythroid derived 2,-like 1 | 2.40 | 3.70E-02 | 1.08E-01 | -4.76 | 5.9E-04 | 130 |
| 104240 at | Cutl1 | 13047 | cut-like 1 (Drosophila) | 2.39 | 3.80E-02 | 1.09E-01 | -1.94 | 7.8E-02 | 131 |
| 94689 at | Rbm39 | 170791 | RNA binding motif protein 39 | 2.38 | 3.81E-02 | 1.09E-01 | -0.40 | 6.9E-01 | 132 |
| 100971 at | Tead3 | 21678 | TEA domain family member 3 | 2.38 | 3.83E-02 | 1.09E-01 | -0.30 | 7.7E-01 | 133 |
| 92927_at | Etv1 | 14009 | ets variant gene 1 | 2.37 | 3.87E-02 | 1.09E-01 | 1.13 | 2.8E-01 | 134 |
| 100094 at | Supt5h | 20924 | suppressor of Ty 5 homolog (S. cerevisiae) | 2.37 | 3.90E-02 | 1.10E-01 | 0.03 | 9.8E-01 | 135 |
| 92484 at | Hivep2 | 15273 | human immunodeficiency virus type I enhance | 2.36 | 3.97E-02 | 1.10E-01 | -0.52 | 6.1E-01 | 136 |
| 93941_at | T | 20997 | brachyury | 2.35 | 4.01E-02 | 1.10E-01 | -3.91 | 2.4E-03 | 137 |
| 162372 f a | Relb | 19698 | avian reticuloendotheliosis viral (v-rel) oncoger | 2.34 | 4.13E-02 | 1.11E-01 | -0.27 | 7.9E-01 | 138 |
| 101727 at | Nfkbie | 18037 | nuclear factor of kappa light polypeptide gene | 2.33 | 4.17E-02 | 1.12E-01 | 0.71 | 4.9E-01 | 139 |
| 94968 at | Nfyc | 18046 | nuclear transcription factor-Y gamma | 2.33 | 4.17E-02 | 1.12E-01 | -2.82 | 1.7E-02 | 140 |



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 93145 at | Gatad2a | 234366 | GATA zinc finger domain containing 2A | 2.32 | 4.27E-02 | 1.13E-01 | -3.19 | 8.6E-03 | 141 |
| 94261_at | Thrap3 | 230753 | thyroid hormone receptor associated protein 3 | 2.31 | 4.30E-02 | 1.13E-01 | -3.24 | 7.9E-03 | 142 |
| 103634 at | Isgf3g | 16391 | interferon dependent positive acting transcript | 2.30 | 4.38E-02 | 1.14E-01 | 0.39 | 7.0E-01 | 143 |
| 95671 at | Hey1 | 15213 | hairy/enhancer-of-split related with YRPW mo | 2.29 | 4.50E-02 | 1.15E-01 | 0.44 | 6.7E-01 | 144 |
| 104507 g | Nr1i3 | 12355 | nuclear receptor subfamily 1, group I, membe | 2.28 | 4.54E-02 | 1.15E-01 | -0.80 | 4.4E-01 | 145 |
| 94896 at | Hnrpab | 15384 | heterogeneous nuclear ribonucleoprotein A/B | 2.27 | 4.63E-02 | 1.15E-01 | 1.48 | 1.7E-01 | 146 |
| 93234_at | Msc | 17681 | musculin | 2.27 | 4.64E-02 | 1.15E-01 | 0.87 | 4.0E-01 | 147 |
| 100513 at | Ddef1 | 13196 | development and differentiation enhancing | 2.27 | 4.64E-02 | 1.15E-01 | -10.14 | 6.4E-07 | 148 |
| 97695 s a | Rpl7 | 19989 | ribosomal protein L7 | 2.24 | 4.83E-02 | 1.17E-01 | 3.95 | 2.3E-03 | 149 |
| 102039_at | Gtf2h4 | 14885 | general transcription factor II H, polypeptide 4 | 2.24 | 4.83E-02 | 1.17E-01 | 2.56 | 2.7E-02 | 150 |
| 101631 at | Sox11 | 20666 | SRY-box containing gene 11 | 2.24 | 4.84E-02 | 1.17E-01 | -3.49 | 5.1E-03 | 151 |
| 102074 at | Otp | 18420 | orthopedia homolog (Drosophila) | 2.24 | 4.88E-02 | 1.17E-01 | 0.75 | 4.7E-01 | 152 |
| 102901 at | Six3 | 20473 | sine oculis-related homeobox 3 homolog (Dro | 2.22 | 5.00E-02 | 1.18E-01 | 1.54 | 1.5E-01 | 153 |
| 93327 at | Tax1bp3 | 76281 | Tax1 (human T-cell leukemia virus type I) binc | 2.22 | 5.08E-02 | 1.19E-01 | -4.45 | 9.7E-04 | 154 |
| 104506_at | Nr1i3 | 12355 | nuclear receptor subfamily 1, group I, membe | 2.21 | 5.17E-02 | 1.19E-01 | 0.29 | 7.8E-01 | 155 |
| 92395 r at | Crebl1 | 12915 | cAMP responsive element binding protein-like | 2.20 | 5.26E-02 | 1.20E-01 | 1.34 | 2.1E-01 | 156 |
| 160520 at | Yap1 | 22601 | yes-associated protein 1 | 2.19 | 5.30E-02 | 1.21E-01 | -0.84 | 4.2E-01 | 157 |
| 161468_f_a | Hand1 | 15110 | heart and neural crest derivatives expressed t | 2.19 | 5.34E-02 | 1.21E-01 | -1.52 | 1.6E-01 | 158 |
| 92237 at | Rxrg | 20183 | retinoid X receptor gamma | 2.17 | 5.51E-02 | 1.22E-01 | 2.60 | 2.5E-02 | 159 |
| 99169 at | Carm1 | 59035 | coactivator-associated arginine methyltransfe | 2.17 | 5.53E-02 | 1.22E-01 | -1.87 | 8.8E-02 | 160 |
| 96109 at | Klf2 | 16598 | Kruppel-like factor 2 (lung) | 2.17 | 5.55E-02 | 1.23E-01 | 4.34 | 1.2E-03 | 161 |
| 92933 at | Pou2f3 | 18988 | POU domain, class 2, transcription factor 3 | 2.16 | 5.58E-02 | 1.23E-01 | 0.56 | 5.9E-01 | 162 |
| 93671_at | Erf | 13875 | Ets2 repressor factor | 2.16 | 5.61E-02 | 1.23E-01 | -0.74 | 4.7E-01 | 163 |
| 104698 at | Gata6 | 14465 | GATA binding protein 6 | 2.15 | 5.71E-02 | 1.24E-01 | 1.08 | 3.0E-01 | 164 |
| 93693 at | Hmbox1 | 219150 | homeobox containing 1 | 2.15 | 5.73E-02 | 1.24E-01 | -3.52 | 4.8E-03 | 165 |
| 93880_at | Eomes | 13813 | eomesodermin homolog (Xenopus laevis) | 2.14 | 5.77E-02 | 1.24E-01 | -0.72 | 4.9E-01 | 166 |
| 92275 at | Tcfap2c | 21420 | transcription factor AP-2, gamma | 2.14 | 5.78E-02 | 1.24E-01 | 1.75 | 1.1E-01 | 167 |
| 92233 at | 1810007M1 | 67367 | RIKEN cDNA 1810007M14 gene | 2.14 | 5.82E-02 | 1.25E-01 | -1.50 | 1.6E-01 | 168 |
| 104356 at | Taf9 | 108143 | TAF9 RNA polymerase II, TATA box binding p | 2.14 | 5.82E-02 | 1.25E-01 | -1.07 | 3.1E-01 | 169 |
| 100554 at | Pdlim1 | 54132 | PDZ and LIM domain 1 (elfin) | 2.14 | 5.83E-02 | 1.25E-01 | -1.25 | 2.4E-01 | 170 |
| 93073 at | Nfatc2 | 18019 | nuclear factor of activated T-cells, cytoplasmi | 2.13 | 5.89E-02 | 1.25E-01 | -0.69 | 5.0E-01 | 171 |
| 160721 at | Elf1 | 13709 | E74-like factor 1 | 2.13 | 5.93E-02 | 1.25E-01 | -1.61 | 1.4E-01 | 172 |
| 104592 i a | Mef2c | 17260 | myocyte enhancer factor 2C | 2.12 | 6.00E-02 | 1.26E-01 | -0.46 | 6.6E-01 | 173 |
| 162161_r_ | Tcea1 | 21399 | transcription elongation factor A (SII) 1 | 2.11 | 6.12E-02 | 1.26E-01 | -0.13 | 9.0E-01 | 174 |
| 98629 f at | Hif1a | 15251 | hypoxia inducible factor 1, alpha subunit | 2.11 | 6.13E-02 | 1.26E-01 | -1.69 | 1.2E-01 | 175 |
| 95796 g a | Supt4h2 | 20923 | suppressor of Ty 4 homolog 2 (S. cerevisiae) | 2.11 | 6.14E-02 | 1.26E-01 | 0.86 | 4.1E-01 | 176 |
| 160748 at | Asb6 | 72323 | ankyrin repeat and SOCS box-containing prot | 2.11 | 6.18E-02 | 1.27E-01 | -0.37 | 7.2E-01 | 177 |
| 94754 at | Lhx8 | 16875 | LIM homeobox protein 8 | 2.10 | 6.24E-02 | 1.27E-01 | 0.23 | 8.2E-01 | 178 |
| 97660 s a | Tcf7l2 | 21416 | transcription factor 7-like 2, T-cell specific, HM | 2.10 | 6.25E-02 | 1.27E-01 | 0.96 | 3.6E-01 | 179 |
| 100451 at | Hsf1 | 15499 | heat shock factor 1 | 2.10 | 6.27E-02 | 1.28E-01 | 1.28 | 2.3E-01 | 180 |
| 161680 r | Nr1h2 | 22260 | nuclear receptor subfamily 1, group H, memb | 2.09 | 6.32E-02 | 1.28E-01 | -2.49 | 3.0E-02 | 181 |
| 94229_at | Tsc22d4 | 78829 | TSC22 domain family 4 | 2.09 | 6.32E-02 | 1.28E-01 | -1.54 | 1.5E-01 | 182 |
| 102959 at | Tle4 | 21888 | transducin-like enhancer of split 4, homolog o | 2.09 | 6.34E-02 | 1.28E-01 | 1.42 | 1.8E-01 | 183 |
| 98628 f at | Hif1a | 15251 | hypoxia inducible factor 1, alpha subunit | 2.09 | 6.40E-02 | 1.29E-01 | -1.76 | 1.1E-01 | 184 |
| 98455 at | Atf2 | 11909 | activating transcription factor 2 | 2.07 | 6.53E-02 | 1.30E-01 | 0.08 | 9.4E-01 | 185 |
| 94281 at | Cnot2 | 72068 | CCR4-NOT transcription complex, subunit 2 | 2.06 | 6.74E-02 | 1.32E-01 | 0.84 | 4.2E-01 | 186 |
| 104701 at | Bhlhb2 | 20893 | basic helix-loop-helix domain containing, clas | 2.05 | 6.85E-02 | 1.33E-01 | -0.87 | 4.0E-01 | 187 |
| 160898 at | Abt1 | 30946 | activator of basal transcription | 2.03 | 7.04E-02 | 1.35E-01 | 2.12 | 5.7E-02 | 188 |
| 101416 f a | Bat4 | 81845 | HLA-B associated transcript 4 | 2.03 | 7.05E-02 | 1.35E-01 | -1.16 | 2.7E-01 | 189 |
| 99095_at | Max | 17187 | Max protein | 2.03 | 7.15E-02 | 1.35E-01 | -3.15 | 9.2E-03 | 190 |
| 98767 at | Yy1 | 22632 | YY1 transcription factor | 2.02 | 7.28E-02 | 1.36E-01 | -2.74 | 1.9E-02 | 191 |
| 98545 at | Phb2 | 12034 | prohibitin 2 | 2.01 | 7.35E-02 | 1.36E-01 | 4.05 | 1.9E-03 | 192 |
| 92343_at | Hltf | 20585 | helicase-like transcription factor | 2.01 | 7.38E-02 | 1.37E-01 | 2.35 | 3.9E-02 | 193 |
| 104502 f a | Hes6 | 55927 | hairy and enhancer of split 6 (Drosophila) | 2.01 | 7.41E-02 | 1.37E-01 | -3.08 | 1.1E-02 | 194 |
| 99665 at | Satb1 | 20230 | special AT-rich sequence binding protein 1 | 2.00 | 7.47E-02 | 1.37E-01 | 1.64 | 1.3E-01 | 195 |
| 161159 r | Pou5f1 | 18999 | POU domain, class 5, transcription factor 1 | 2.00 | 7.47E-02 | 1.37E-01 | -0.36 | 7.3E-01 | 196 |
| 101930 at | Nfix | 18032 | nuclear factor I/X | 2.00 | 7.48E-02 | 1.37E-01 | -6.81 | 2.9E-05 | 197 |
| 97157_at | Nkx3-1 | 18095 | NK-3 transcription factor, locus 1 (Drosophila) | 1.99 | 7.60E-02 | 1.39E-01 | 0.13 | 9.0E-01 | 198 |
| 104304 r | Polr3k | 67005 | polymerase (RNA) III (DNA directed) polypept | 1.98 | 7.76E-02 | 1.40E-01 | -3.04 | 1.1E-02 | 199 |
| 95336 at | Hif3a | 53417 | hypoxia inducible factor 3, alpha subunit | 1.98 | 7.83E-02 | 1.41E-01 | 0.29 | 7.8E-01 | 200 |

### Experimental validation of candidate genes for pluripotency and self-renewal from gene list generated by Differentiation-Test.

We hypothesized that the Differentiation-Test reported list would include uncharacterized critical regulators of pluripotency and self-renewal. Self-renewal regulators should have a lower expression in differentiated cells and therefore should be detectable in the cell mixture of 4-day EBs. We used short hairpin RNA (shRNA) to further study two transcription regulators detected by the Differentiation-Test, namely, *Smarcad1*and *Pias2*. They ranked 10 and 55 respectively among all transcription regulators (Table 3.2). The other top-ranking regulators were not picked for experimental validation because they had known regulatory roles in ES cell differentiation. Upon 2 days of *Smarcad1* shRNA induction, ES cells started to take on a flattened morphology; large percentages of cells lost Alkaline Phosphatase (AP) staining (Figure 3.7A).

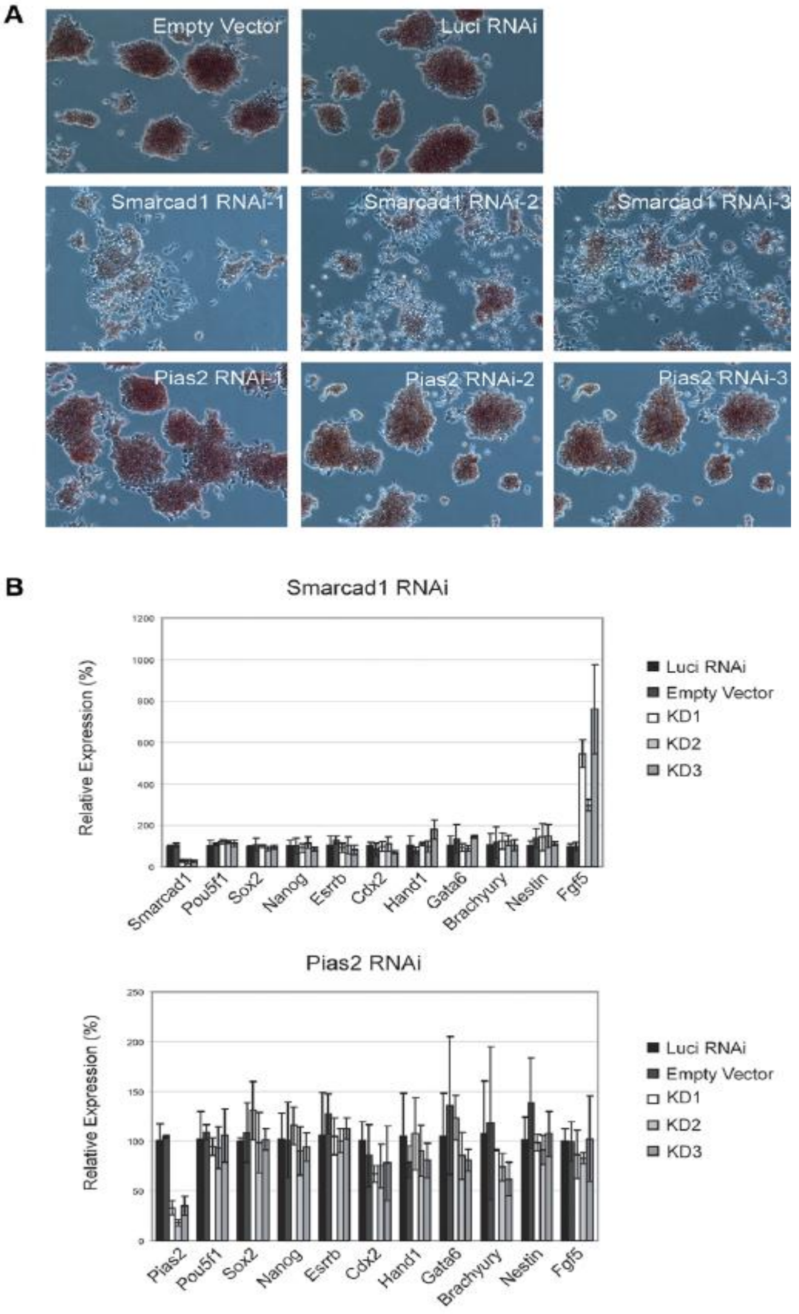


**Figure 3.7 Depletion of candidate genes by RNAi for two days.** Three shRNA constructs are used to target different regions of respective transcripts. (A) Two days after puromycin selection, typical colony morphology of ES cells with positive alkaline phosphatase (AP) staining (red) was maintained after *Pias2* knockdown. Flattened fibroblast-like cells were formed after *Smarcad1* depletion. In control empty vector or *Luc* shRNA transfected cells, normal undifferentiated phenotype with distinct ES cell colonies was maintained. (B) Quantitative real-time PCR analysis of gene expression in two days knockdown ES cells. The levels of the transcripts were normalized against control empty vector transfection. Data are presented as the mean ± SEM and derived from independent experiments.

Quantitative real time polymerase chain reaction (qPCR) analysis showed that the knockdown of *Smarcad1* induced the expression of *Fgf5*, a growth factor involved in multiple differentiation processes including differentiation to the neuronal lineage (Reuss et al., 2003)(Figure 3.7B). On the other hand, neither mock shRNA nor shRNA knockdown of *Pias2* induced ES cell differentiation (Figure 3.7).

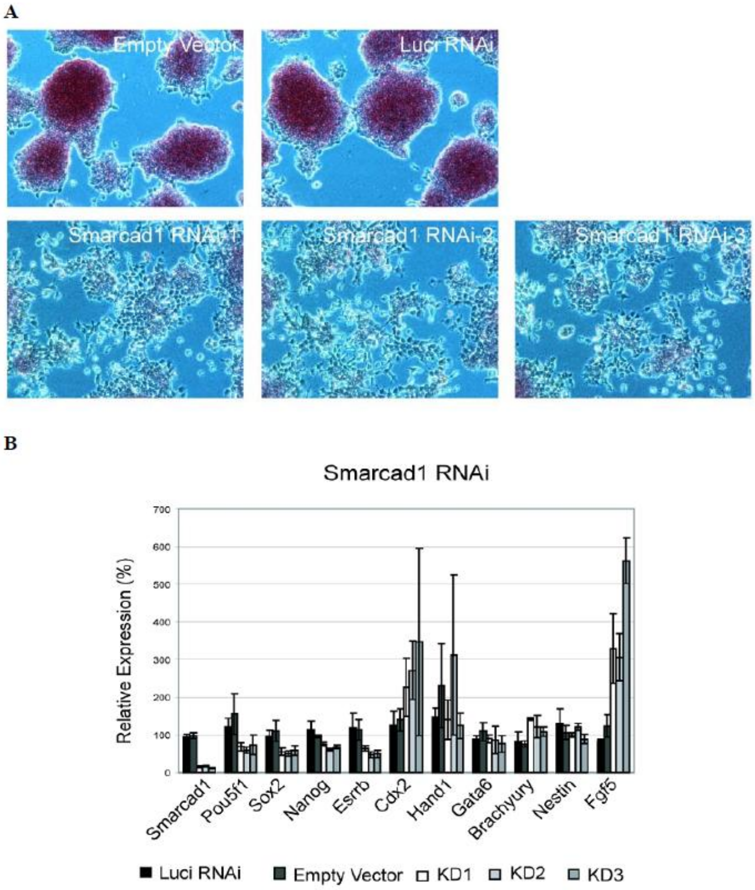


**Figure 3.8 Depletion of candidate genes by RNAi for four days.** (A) Four days after pruomycin selection, *Smarcad1* knockdown cells became more flattened fibroblast-like and completely lost the AP positive colony compared with the cells of two days knockdown. (B) Quantitative real-time PCR analysis of gene expression in four days knockdown ES cells. The levels of the transcripts were normalized against control empty vector transfection. Data are presented as the mean ± SEM and derived from independent experiments.

At 4 days of shRNA induction, we observed further loss of AP staining (Figure 3.8A), reduction in pluripotency markers such as *Pou5f1*, *Sox2*, and *Nanog*, as well as induction of multiple differentiation marker genes including *Fgf5*, *Cdx2*, and *Hand1*, confirming that the cells depleted of *Smarcad1* lost the ability to maintain their stemness state (Figure 3.8B). Multiple shRNA constructs targeting different regions of the target genes gave the same results. These results demonstrate the ability of the Differentiation-Test to identify novel self-renewal regulators.

## A gene regulatory network during differentiation.

A regulatory network of early differentiation genes might reveal the critical events that underlie the earliest differentiation of ES cells. Using the genes identified by the Differentiation-Test, we constructed a gene regulatory network (GRN) that demonstrates the transition of ES cells to 4-day EBs (see Methods).

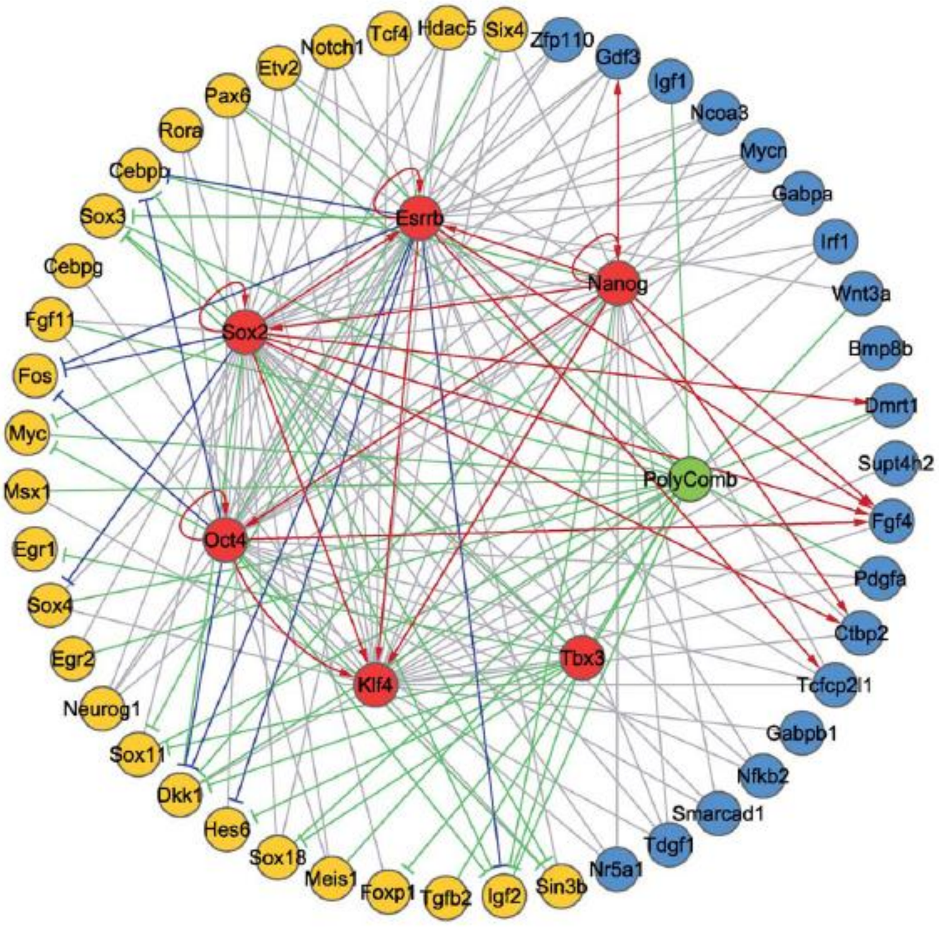


**Figure 3.9 A regulatory network in differentiating ES cells.** Modules and regulatory relationships. Yellow and blue nodes represent genes that are up- and down-regulated in differentiated cells. All blue and yellow nodes are collectively termed as pluripotency and differentiation modules, respectively. Edges (plain edges, activators ↑ and repressors ⊤ ) represent evidence of regulatory relationships. Plain edges: the regulatory relationship is supported by the binding of the regulator to the target gene (ChIP-seq or ChIP-chip data). Activators: the regulatory relationship is supported by both the binding of the regulator to the target gene (ChIP-seq or ChIP-chip data) and down-regulation of the target gene expression when the regulator is knocked down (RNAi microarray data). Repressors: the regulatory relationship is supported by both the binding of the regulator to the target gene (ChIP-seq or ChIP-chip data) and up-regulation of the target gene expression when the regulator is knocked down (RNAi microarray data).

Nodes of this GRN were top-ranked transcription factors and signal transduction genes detected by the Differentiation-Test in 4-day EBs (Figure 3.9). Regulatory relationships among these nodes were taken from published results of ChIP-chip experiments (Boyer et al., 2006b; Jiang et al., 2008; Loh et al., 2006), ChIP-seq experiments (Chen et al., 2008), and RNAi followed by microarray experiments (Ivanova et al., 2006). Comparing the mean expression value of a gene in Oct4 expressing cells (Oct4+) and Oct4 non-expressing cells (Oct4-) (Zhou et al., 2007a), we separated the differentiation regulators into two modules: the upregulated module during differentiation (termed the differentiation module, yellow nodes, Figure 3.9) and the downregulated module (termed the pluripotency module, blue and red nodes, Figure 3.9).

### Systematic over-representation of RBP-J binding sites in the upstream regions of the differentiation module.

The differentiation module of the transcription network for early differentiation is enriched with a number of canonical downstream targets of the Notch signaling pathway (yellow nodes, Figure 3.9). This makes it attempting to hypothesize that Notch signaling is a pathway that triggers the activation of the differentiation module at the early differentiation stage of mosue ES cells. To investigate this hypothesis, we analyzed the binding site distribution of the transcription factor RBP-J, a key mediator of Notch signaling (Tanigaki et al., 2002), in the upstream regions of all regulatory protein genes in Figure 3.9.

Tuning the threshold δ enabled a comprehensive view of the likelihood of RBP-J binding to the genes of the two modules, and minimized the bias of using a predetermined threshold to call putative binding sites. With a wide spectrum of δ, average upstream binding affinity and average motif count in the upstreams of the differentiation module are all consistently larger than their counterparts of the pluripotency module (Figure 3.10A, Figure 3.11). In particular, with a threshold of 500, there are on average 14.2 putative binding sites on an upstream sequence in the differentiation module, as compared to 8.7 putative sites on an upstream sequence in the pluripotency module. Both the scale and the consistency of the overrepresentation of the RBP-J motif in the upstreams of the differentiation module provide supporting data for Notch's potential role of triggering differentiation of mouse ES cells. These data are consistent with recent reports that Notch signaling promotes neural lineage entry of mouse ES cells (Lowell et al., 2006) and it is required for undifferentiated human ES cells to form the progeny of all three embryonic germ layers (Yu et al., 2008).

To test if the RBP-J is among one of the most potent regulators for the differentiation module, we used the PRIMA software (Elkon et al., 2003) to test all 332 non-redundant mammalian DNA binding motifs available in TRANSFAC v10.2 (Figure 3.10B). Four motifs were found to be enriched in the upstream sequences of the differentiation module genes as compared to those of the pluripotency module genes ($p$-value $\leqslant 0.05$). In particular, the RBP-J motif exhibited the second smallest p-value (0.022) and the largest enrichment factor (2.0) among the 332 motifs.

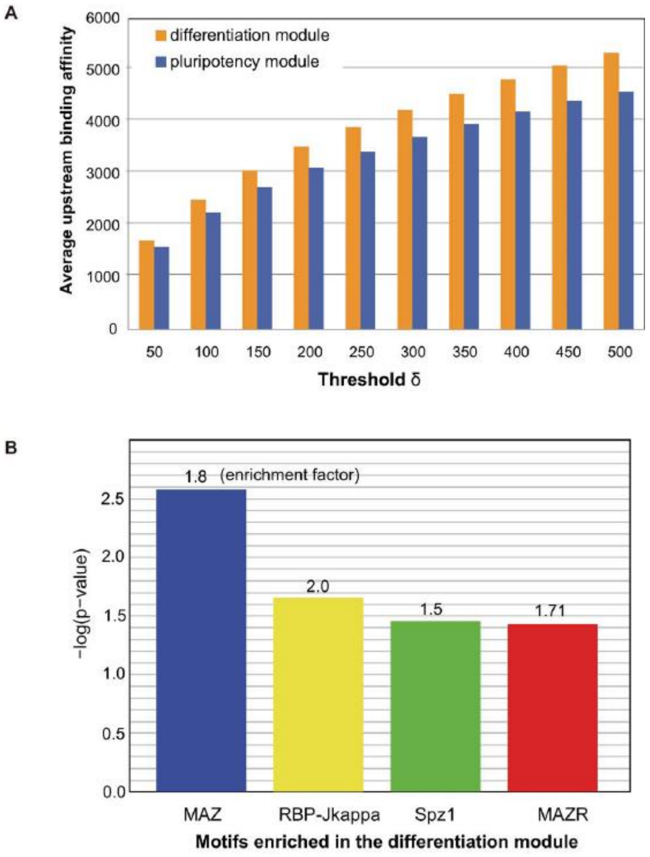


**Figure 3.10 Enrichment of the RBP-J motif in the upstreams of the differentiation module.** (A) Average upstream binding affinity of RBP-J both shows enhanced signals in the upstream sequences of the differentiation module genes as compared to that of the pluripotency module genes. (B) Testing of all 332 non-redundant mammalian DNA binding motifs available in TRANSFAC v10.2, four motifs were found to be enriched in the upstream sequences of the differentiation module genes as compared to that of the pluripotency module genes (p-value ≤0.05). In particular, the RBP-J motif exhibited the second smallest p-value (0.028) and the largest enrichment factor (2.0) among the 332 motifs.

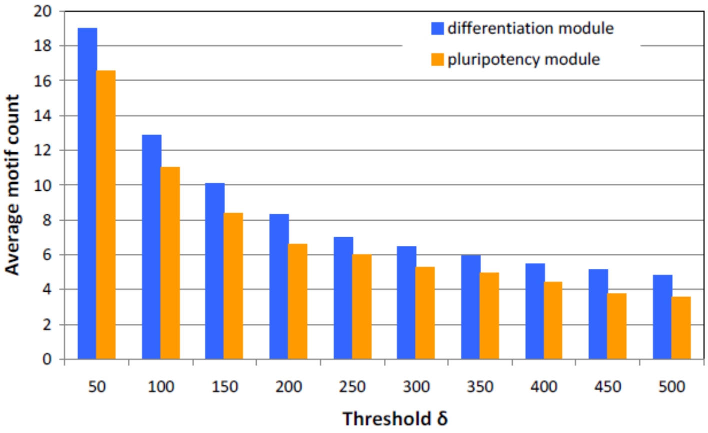
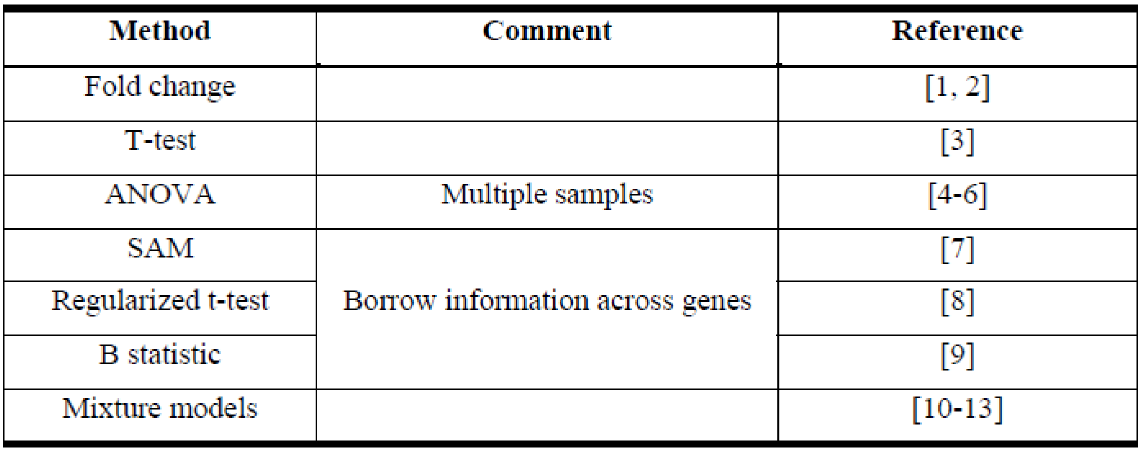


**Figure 3.11 Average motif counts.** Average motif counts of RBP-J in the upstreams of the differentiation module are consistently larger than the counts in the upstreams of the pluripotency module.

## 3.5 DISCUSSION

If high-throughput measurements of gene expression at the single-cell level were available, currently available statistical tools (Table 3) would be applicable to the search for differentially expressed genes during differentiation. However, microarrays typically do not measure gene expression from a single cell but can only measure the average signal from a population of cells. Such data demand new gene expression models from the single-cell level to the cell-mixture level.

**Table 3.3 Two sample comparison methods.** All these methods require gene expression measurements from individual cell types.



| Method | Comment | Reference |
|---|---|---|
| Fold change | | [1, 2] |
| T-test | | [3] |
| ANOVA | Multiple samples | [4-6] |
| SAM | Borrow information across genes | [7] |
| Regularized t-test | | [8] |
| B statistic | | [9] |
| Mixture models | | [10-13] |

The Differentiation-Test method makes a number of abstractions to the differentiation process. Most remarkably, the method assumes that the differentiation process starts from a relatively homogeneous initial cell mixture and progresses into a more heterogeneous cell mixture with identifiable events of divergence of expression levels of certain genes during the process. There are at least two sources contributing to the heterogeneity of gene expression in a cell mixture, including the unsynchronized cell cycle stages and the cell type difference. The first source of heterogeneity is assumed to persist over time, and therefore it is adjusted for by the ratio of variances across time points. Statistically, when the initial cell mixture is not purely homogeneous, Equation (5) would have a non-zero first term in the summation. In such a scenario, the DT statistic still reflects the contrast of variation across time and the null distribution can be approximated by an F distribution with the same degrees of freedom. Therefore, the Differentiation-Test does not require the initial cell mixture to be absolutely homogenous but does require the heterogeneity of the cell mixture to increase over time.

The same set of core regulatory proteins and protein complexes interact and regulate the genes in both the pluripotency module and the differentiation module (Figure 3.9). The complex interactions of these regulatory proteins suggest that their pivotal roles in ES cells may not be sufficiently reflected in a binary description as “activators’’ or ‘‘repressors’’, whereas they may serve to strike a balance between the multiple extrinsic signals that the cells receive, filter intrinsic noise of the system, and collectively predispose the ES cells to pro- or anti-differentiation states. The implications of such complex interactions to data modeling and interpretation are twofold. First, a predictive model for cell fate decision might require modeling the regulators as continuous rather than Boolean variables. A case in point is the observation that the feedback loop of Oct4-Sox2-Nanog is capable of translating continuous differentiation signals into an irreversible bistable switch (Chickarmane et al., 2006). Second, gene knockout data should be interpreted with caution given that a regulator may not merely activate or repress gene expression but may also buffer variability in transcription by minimizing stochastic extrinsic and intrinsic signals that create noise in gene expression (Chi and Bernstein, 2009). A case in point is the deletion experiment of the Polycomb complex protein Suz12 (Pasini et al., 2007). Suz12(-/-) ES cells are viable and exhibit defective differentiation, which seems to contradict the role of the Polycomb group as a repressor complex that suppresses the expression of lineage-specific differentiation genes in ES cells (Boyer et al., 2006b). However Suz12 (-/-) ES cells exhibit a global loss of H3K27 trimethylation (H3K27me3) (Pasini et al., 2007), which may have lost a buffering mechanism that renders the intrinsic signal for pluripotency unrestrictedly amplified. More experiments, such as a series of knockdowns of Suz12 into different concentrations, may produce data to further investigate such questions.

The new gene expression and RNA knockdown data suggest that Smarcad1 is a chromatin modeling factor that contributes to maintaining the pluripotency of ES cells. Smarcad1 is structurally classified into the SWI2/SNF2 superfamily of DNA-dependent ATPases that are catalytic subunits of chromatin-remodeling complexes. Although the importance of other members of the SWR1-like subfamily in chromatin remodeling (EP400, INOC1, and SRCAP) has already been elucidated, little was known about the biological function of Smarcad1 in transcriptional regulation. Homozygous mutation of *Smarcad1* gives rise to a number of phenotypes including prenatal-perinatal lethality (Schoor et al., 1999), confirming the importance of *Smarcad1* in regulating early development. Smarcad1 preferentially binds to transcription start sites in embryonic carcinoma cells (Okazaki et al., 2008), which suggests that *Smarcad1* is a gene specific transcription regulator rather than a ubiquitous chromatin modeling factor. These data and our observations collectively suggest that Smarcad1 might be an overlooked sequence-specific transcription regulator important for both ES cells and early development.

# CHAPTER IV: Coactivators p300/CBP regulate self-renewal of mouse embryonic stem cells by mediating long-range chromatin structure

Fang, F., Xu, Y.F., Chen, X., Chew K.K., Chia, N.Y., Ng, H.H., and Matsudaira, P.

**My contribution to this project:**

I was the main driver of the project, leading every aspects of it, including the conceptuation of the project, the experimental design, executing the experiments, analyzing the data and writing the manuscript.

## 4.1 SUMMARY FOR CHPATER IV

p300 and CBP are two highly homologous transcription coactivators that are essential for transcriptional activation and coordinate a variety of cellular processes, including embryogenesis and development. p300-deficient mouse embryonic stem (ES) cells are viable but are severely compromised in the ability to differentiate. However, the underlying molecular mechanism has not been clearly addressed. In this study, we found that p300 and CBP play redundant roles in maintaining the undifferentiated state of mouse ES cells. They are recruited by master regulators (Oct4, Sox2, Nanog, etc) through protein interaction to activate cell-specific gene expression. Furthermore, using the chromatin conformation capture (3C) technique, we found that p300/CBP are involved in the formation of long-range chromatin looping structure specific to the pluripotent state of ES cells. Characterization of the interacting DNA elements revealed that some contain enhancer activities which were dependent on p300 and CBP. In conclusion, our work, for the first time, characterizes coactivators p300/CBP in ES cells as self-renewal regulators through mediating nuclear architecture, which promote extensive crosstalk among multiple enhancers and promoters to activate specific gene transcription.

## 4.2 INTRODUCTION

Embryonic stem (ES) cells are isolated from the inner cell mass of mammalian preimplantation embryo at the blastocyst stage (Evans and Kaufman, 1981; Keller, 2005). These cells are pluripotent in that they can self-renew continuously while retaining the capacity to differentiate into multiple lineages. Expression of protein-coding genes in ES cells is spatially and temporally regulated in a highly orchestrated and precise pattern by an ES cell-specific transcriptional network. A few essential sequence-specific transcription factors have been characterized, such as Oct4, Sox2 and Nanog (Chambers and Smith, 2004; Loh et al., 2006). They are indispensable for the maintenance of ES cell identity and more strikingly, introducing them into somatic cells is able to reprogram these differentiated cells to pluripotent state (Takahashi and Yamanaka, 2006). Over the last few years, intensive efforts have been directed at identifying transcription factors and their binding targets in order to decipher the secrets of transcriptional network in ES cells, however, besides transcription factors, another critical insight for understanding transcriptional control mechanisms was provided by coactivators, which are multiple intermediary proteins that are recruited by transcription factors and enhance specific gene transcription by countering the repressive effects of local chromatin.

The transcriptional coactivators p300 (Ep300) and CREB-binding protein (CBP) are two highly homologous genes, which are capable of interacting with a large variety of transcription factors playing central roles in a wide range of cellular processes including proliferation, differentiation and apoptosis (Chan and La Thangue, 2001; Goodman and Smolik, 2000). Studies in mice have shed light on the critical roles that p300 and CBP play in embryogenesis. Homozygous p300 and CBP knockouts, as well as CBP/p300 double heterozygotes are embryonically lethal (Yao et al., 1998). Studies with heterozygous and chimeric mice demonstrated requirements for p300 and CBP for tissue and organ development as well as normal adult stem cell self-renewal and differentiation (Kawasaki et al., 1998; Kung et al., 2000; Oike et al., 1999). In the study of ES cells, *p300* null ES cells exhibit normal self-renewal capacity, however, embryoid body (EB) induced by *p300* null ES cells has shown significantly abnormal expression pattern of germ layer markers (Zhong and Jin, 2009). Genome-wide mapping of p300 binding sites in mouse ES cells has uncovered that p300, as an enhancer binding protein, co-occured with Nanog, Oct4 and Sox2 cluster quite often, suggesting that p300 may be recruited by ES specific transcription factors to facilitate the communication of distant regulatory elements with proximal elements (Chen et al., 2008). Despite clear evidence for the participation of p300 and CBP in the transcriptional regulation of ES cells, the mechanisms of how these coactivators assist transcription factors and basal transcripitional machinery in up-regulation of gene expression in the context of ES specific chromatin structure is poorly understood. In addition, whether an extra copy of CBP is able to replace p300 when the function of endogenous p300 is lost *in vivo* (or *vice versa*) is unknown, leaving the issue of functional redundancy between these two homologous proteins unresolved.

In this study, we have characterized the function of p300 and CBP in mouse ES cells and found that they are playing redundant roles in maintaining the undifferentiated state of ES cells. Based on the analysis of genome mapping data and biochemistry assays, we further demonstrated that these coactivators were recruited to specific genomic loci by master regulators Nanog, Oct through protein-protein interactions. Domain mapping studies have identified that KIX and HAT domains are the functional domains for p300 and CBP in ES cells to connect transcription factors and activate gene expression. More importantly, using chromatin conformation capture (3C) technique, we found that loci co-occupied by p300, Nanog and Oct4 form long-range intragenic and intergenic looping interactions that are evolutionary conserved in both mouse and human. The observed *in vivo* chromatin conformation is specific to the pluripotent state as it was abolished in differentiated cells. Through ChIP-3C and RNA interference (RNAi) studies, the presence of p300 and CBP was found to be crucial for the formation of such higher-order chromatin structures. Characterization of the interacting DNA elements revealed that some contain enhancer activities *in vitro* and *in vivo* that is dependent on p300 and CBP. Our work, for the first time, characterizes coactivators p300 and CBP in ES cells as self-renewal regulator as well as bridging proteins for nuclear architecture which promote extensive crosstalk among multiple enhancers and promoters.

## 4.3 MATERIALS AND METHODS

**Cell culture and transfection.** E14 mouse ES cells, cultured under feeder-free conditions on surfaces coated with 0.1% gelatin, were maintained in Dulbecco's modified Eagle's medium (DMEM; GIBCO), supplemented with

15% heat-inactivated fetal bovine serum (FBS; GIBCO), 0.055 mM ß-mercaptoethanol (GIBCO), 2mM L-glutamine (GIBCO), 0.1 mM minimal essential medium with nonessential amino acids (GIBCO), and 1,000 U/ml of leukemia inhibitory factor (LIF) (Chemicon). Transfection of shRNA and overexpression plasmids was performed using Lipofectamine 2000 (Invitrogen) according to manufacturer's instructions. Briefly, 2.0µg plasmid DNA was transfected into ES cells on 60 mm plates for RNA and protein extraction. Detection of alkaline phosphatase, which is indicative of the nondifferentiated state of ES cells, was carried out using a commercial ES cell characterization kit (Chemicon). For RNAi–ChIP assays, 12µg plasmid DNA was transfected into ES cells on 150 mm plates. Puromycin (Sigma) selection was introduced 1 day after transfection at 1.0 µg ml$^{-1}$, and maintained for 2–4 days before harvesting. In differentiation experiments, cells were treated with 1 µM of retinoic acid (RA) for 4 days. The human ES cell-line (H1, WiCell) was cultured feeder-free on Matrigel (BD). Condition medium used for culturing human ES cells contained 20% KO serum replacement, 1mM L-glutamine, 1% non-essential amino acids and 0.1mM 2-mercaptoethanol and an additional 8ng.ml$^{-1}$ of basic fibroblast growth factor (Invitrogen) supplemented to the hESC unconditioned medium. Medium was changed daily. The human ES cells were subcultured with 1mg.ml$^{-1}$ collagenase IV (Gibco) every 5–7days. HEK293T (293) cells were cultured in DMEM supplemented with 10% FBS.

**RNA extraction, reverse transcription and quantitative real-time PCR.** Total RNA was extracted using TRIzol Reagent (Invitrogen) and purified with the RNAeasy Mini Kit (Qiagen). Reverse transcription was performed using

SuperScript II Kit (Invitrogen). DNA contamination was removed by Dnase (Ambion) treatment, and the RNA was further purified by an RNAeasy column (Qiagen). Quantitative PCR analyses were performed in real time using an ABI PRISM 7900 sequence detection system and SYBR green master mix, as previously described. For all the primers used, each gave a single product of the correct size. In all controls lacking reverse transcriptase, no signal was detected. Each RNAi experiment was repeated at least three times with different batches of ES cells. The sequences targeted by shRNA and the primers for gene expression are in Table 1.

**Chromatin Immunoprecipitation (ChIP) and Real-Time PCR.** ChIP assays were carried out as described previously (Loh et al., 2006). Briefly, cells were cross-linked with 1% formaldehyde for 10 min at room temperature, followed by the addition of 0.2 M glycine to inactivate the formaldehyde. Cells were then lysed to obtain chromatin extracts, which were sonicated to obtain DNA fragments with an average size of 300-500 bp. The resulting chromatin extracts were immunoprecipitated using anti-Nanog (Cosmo Bio #RCAB00022PF), anti-p300 (sc-585, Santa Cruz), anti-CBP (sc-583, Santa Cruz) or anti-GST (sc-459, Santa Cruz) polyclonal antibodies immobilized on Protein-G beads. For all ChIP experiments, real-time PCR analyses were performed in technical duplicates using the ABI PRISM 7900 Sequence Detection System and SYBR Green Master Mix as described previously (Loh et al., 2006). Relative occupancy values (fold enrichment) were calculated by determining the apparent immunoprecipitation efficiency (ratios of the amount of immunoprecipitated DNA to that of the input sample) and normalized to the level observed at a control region, which was defined as 1.0. All ChIP experiments were repeated at least three times. For all the primers used, each gave a single product of the right size, as confirmed by agarose gel electrophoresis and dissociation curve analysis. Primer sequences are Table 1.

**Co-immunoprecipitation** – Transfected cells were lysed in cell lysis buffer (50 mM Tris HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 1% NP40, 10% glycerol with protease inhibitor cocktail) for 1 h. Whole cell extracts were collected and precleared. Beads coated with Oct4 (sc-8628, Santa Cruz) or Nanog (sc-7392, Santa Cruz) or p300 (sc-585, Santa Cruz**)** or CBP (sc-583, Santa Cruz**)** antibody were incubated with the precleared whole cell extracts at 4 $^{o}$C for overnight. The beads were washed with cell lysis buffer 4 times. Finally, the beads were boiled in 2x sample buffer for 10 min. The eluents were analyzed by either protein staining or Western blot.

**GST pulldown assay** – Full-length nanog and CBP and various deletion fragments were cloned into pET42b (Novagen). The plasmids were transformed into BL21 *E coli*. The proteins were expressed and purified with GSH-sepharose beads (Amersham) followed by Ni-NTA beads (Qiagen). The purified proteins were bound to GSH beads and incubated with recombinant CBP or nanog proteins for 2 h in 4 $^{o}$C. The beads were washed 6 times with cell lysis buffer. The eluents were analyzed by Western blot.

**Chromosome Conformation Capture (3C) assay and ChIP-3C assay.** The 3C assay was performed as described previously(Miele et al., 2006) with some modifications. Briefly, cells were crosslinked with 1% formaldehyde for 5 min at room temperature. The cells were then subjected to cell lysis. Nuclei were pelleted and resuspended in 1x NEB3 restriction buffer for overnight *Bgl* II digestion at 37$^{o}$C with shaking. The enzyme was inactivated with SDS (1.3% final concentration) and shaking for 15 minutes at 65$^{o}$C. 1x ligation buffer and TritonX-100 (1% final concentration) were then added to the nuclei and incubated for 1 hour at 37$^{o}$C. An 800 µl ligation reaction was prepared with 400 Weiss Units T4 DNA ligase (NEB), 8 µl 10 mg/ml BSA, and 8 µl 100 mM ATP. The sample was incubated for 4 hours at 16$^{o}$C and 30 minutes at room temperature. The percentage of digestion and ligation of *Bgl*II fragment was analyzed. DNA was then purified and subjected to PCR amplification of chimeric products using Hot-Star polymerase (Qiagen). BAC clones of Children's Hospital Oakland Research Institute (CHORI) were used to prepare positive control template for the loci of interest. For mouse *Dppa3-Nanog-Slc2a3* cluster, we used the BAC clone RP24-73P7; for mouse *Tcf3* locus, we used the BAC clones RP23-295I11 and RP24-313P9; for human *DPPA3-NANOG-SLC2A14* loci, we used the BAC clone RP11-277J24; and for human *TCF7L1* locus, we used the BAC clones RP11-312D1. To assure that DNA templates prepared from 3C analyses are working and to standardize the cross-linking frequency in all cell samples, we chose a primer pair that targets two nearby restriction fragments for the *GAPDH* (RP11-72G18) and *Ndufa4* (RP23-230A2) loci to be used as a control for the 3C analyses as they are constitutively expressed genes in human cells and mouse cells respectively, and they are also located on the same chromosome with the loci of interest. ChIP-3C assays were performed as essentially described previously with slight modifications (Horike et al., 2005; Murrell et al., 2004). Briefly, antibody-specific immunoprecipitated chromatin was obtained as described above for ChIP assays. Chromatin still bound to the antibody-Protein-A-Sepharose beads was digested with restriction enzyme, ligated with T4 DNA ligase, eluted, and de-crosslinked. After purification, the ChIP-3C material was detected for long range interaction with primers from the p300, Nanog and Oct4 co-binding regions. Primer sequences used for ChIP, 3C, and ChIP-3C assays are available in Table1.

**Luciferase reporter assays.** The p300, Nanog and Oct4 co-binding fragments of about 500 bp were cloned downstream of a *luciferase* gene driven by the *Oct4* minimal promoter as described previously (Chew et al., 2005). The constructs were co-transfected with either *p300/CBP* or empty vector construct into murine ES cells and the selection was performed with puromycin treatment. Luciferase activity was determined 72 hours after transfection using the dual-luciferase reporter assay system (Promega).

**BAC clones and BAC recombineering using *galK* positive/negative selection.** The bacterial artificial chromosome (BAC) clone RP11-277J24 was obtained from the Children's Hospital of Oakland Research Institute repository (http://www.chori.org/). BAC-containing bacterial stocks were propagated in LB medium supplemented with chloramphenicol. Modification of BACs was performed according to the procedure described previously (Warming et al., 2005). To introduce galK at the desired position, the galK cassette with 50 bp arms homologous to RP11-277J24 was PCR-amplified using 2 ng pGalK and 25 pmol primers and the following PCR conditions: 94 ℃ for 45 s, 58 ℃ for 45 s, 72 ℃ for 90 s, for 30 cycles. PCR primers for Del2, Del3 and Del4 are available in Table 1. Templates were removed from the PCR products by DpnI digestion and gel purification. For positive selection, 50 ml Luria–Bertani

(LB) medium supplemented with 12.5 μg chloramphenicol ml$^{-1}$ was inoculated with SW102 bacteria containing pHB5 (SW102-pHB5); the bacteria were grown at 32 ℃ until an A600 of 0.6 was reached. The culture was heat-shocked at 42 ℃ for 15 min, then cooled briefly in an ice/water bath slurry and pelleted at 4500 g at 0 ℃ for 5 min. Bacteria were washed twice with 20 ml double-distilled H2O (ddH$_2$O) and were finally resuspended in ddH$_2$O. Subsequently, 25 μl of the electrocompetent SW102-pHB5 bacteria were transformed with 150 ng PCR product in a 0.2 cm cuvette using a Bio-Rad Gene Pulser Pulse Controller (Bio-Rad) at 25 μF, 1.75 kV and 200 Ω. Bacteria were recovered in 1 ml LB medium for 1 h at 32 ℃ and then washed three times in 1 ml 1× M9 salts. Bacteria were resuspended in 500 μl 1× M9 salts before plating serial dilutions onto M63 plates supplemented with 0.2 % galactose, 1 mg D-biotin l−1, 45 mg L-leucine l−1 and 12.5 μg chloramphenicol ml−1. Plates were incubated for 3 days at 32 ℃. Gal+ colonies were streaked sequentially twice onto Gal indicator plates (MacConkey agar; BD Biosciences) supplemented with 0.2% galactose and 12.5 μg chloramphenicol ml−1, and incubated overnight at 32 °C. Using the same recombination procedure, galK was replaced by two complementary oligos to the. Again, PCR fragments with arms complementary to 50bp homologous arms used for deletion were amplified in order to replace galK. The conditions for PCR amplification are the same outlined as above. The sequences of complementary are available in supplementary table S3. Again, the plasmid template was removed by DpnI digestion and gel purification. As described above, bacteria were heat-induced and transformed with 300 ng PCR product. The bacteria were recovered in 10 ml LB medium for 4.5 h at 32

℃, and were washed as described above. Serial dilutions were plated on M63 medium supplemented with 0.2% glycerol, 1 mg D-biotin ml−1, 45 mg L-leucine ml−1, 12.5 µg chloramphenicol ml$^{-1}$ and 0.2% 2-deoxygalactose (DOG; Sigma-Aldrich) to select for bacteria in which the galK gene had been removed from the BAC. Plates were incubated for 3 days at 32 ℃. Phosphorylation of DOG by GalK into 2-deoxygalactose-1-phosphate is toxic, resulting in the suppression of bacteria that failed to replace galK. This led to the exclusive growth of clones that contained the desired recombinant BAC. BAC clones with deletion were analysed by PCR, sequencing and BAC DNA restriction analysis. For this, ClaI, PmeI, XhoI triple-digested BAC DNA was separated electrophoretically at 50–100 V for approximately 24 h using a 0.6% agarose gel.

**BAC transfection and generation of stable cell lines.** Stable cell lines were generated by transfection of mouse ES cells with engineered BAC plasmid using lipofectamine (Invitrogen) and selection in 1 mg/mL neomycin (Gibco) selection. Surviving ES cell clones were submitted to a second round of selection with a higher (2 mg/ml) neomycin concentration. ES cell clones survived for the second round were cultured for at least 5 passages and analyzed by RT-PCR to check the integration of complete engineered BAC clones. The correct clones were then extracted for RNA and analyzed for gene expression by RT-PCR.

**Table 4.1: Sequences of primers used in this study.**

## Target Sequence for shRNA

| *Gene* | Target Sequence |
|---|---|
| *p300* | GGACTACCCTATCAAGTAA |
| | GTATTGTCCATGACTACAA |
| *CBP* | GGACAACCCTTTAGTCAAA |
| | GCAGGGATGAATACTATCA |
| *Pou5f1* | GAAGGATGTGGTTCGAGTA |
| *Sox2* | GAAGGAGCACCCGGATTAT |
| *Nanog* | TTCTGGGAACGCCTCATCA |
| *Luciferase* | CGTACGCGGAATACTTCGA |

## Sequence of primers for gene expression

| Genes | Direction | Sequence |
|---|---|---|
| *p300* | F | CAAGGAATTGGCTATCCACCGCAGCAGCA |
| | R | TTGGGAGCATGTGCTGCTGTGGGCTCATAG |
| *CBP* | F | ACGACCCTTCCCAACCTCAGACGACAA |
| | R | AAACCTGAAGGCCAAATGATGTCATAGTGT |
| *Pou5f1* | F | TTGGGCTAGAGAAGGATGTGGTT |
| | R | GGAAAAGGGACTGAGTAGAGTGTGG |
| *Sox2* | F | GCACATGAACGGCTGGAGCAACG |
| | R | TGCTGCGAGTAGGACATGCTGTAGG |
| *Nanog* | F | GGCTATCTGGTGAACGCATCTGGAAG |
| | R | AACTGTACGTAAGGCTGCAGAAAGTCCTC |
| *Esrrb* | F | GCCTCAAAGTGGGGATGCTGAAGGAAGGTG |
| | R | GCCAATTCACAGAGAGTGGTCAGGGCCTTG |
| *Bmp2* | F | CCAAGATGAACACAGCTGGTCACAGATAAGGC |
| | R | AGGTGGTCAGCAAGGGGAAAAGGACACTCC |
| *Fgf5* | F | GAGAGTGGTACGTGGCCCTGAACAAGAGAG |
| | R | CTTCAGTCTGTACTTCACTGGGCTGGGACT |
| *Cdx2* | F | CGCAGAACTTTGTCAGTCCTCCGCAGTACC |
| | R | GTATTCGGCGGGGCTGCTGTAGCCCATAGC |
| *Hand1* | F | CCTGCCCAAACGAAAAGGCTCAGGACCCAA |
| | R | CGACCGCCATCCGTCTTTTTGAGTTCAGCC |
| *Nestin* | F | CAGAGAGGGGACCTGGAACATGAAT |
| | R | CCTGGCCACTGATATCAAAGGTGTCT |

## Sequence of primers for Chromatin Immunoprecipitation (ChIP)

| Nanog Cluster | Coordinates | Direction | Primer sequences |
|---|---|---|---|
| 1 | chr1:36930315-36930854 | Sense | CTAGGCAATACACCGGAGAGGCTCTAGTGA |
| | | Antisense | GATCTGGGCGGAAGAATCAGGTCTATCAAT |
| 2 | chr1:59367651-59368153 | Sense | AGAACCTCAGCCAGCACCCGGAGCAACC |
| | | Antisense | TGTCAAAAGGATGAGCCCGAGAACGCGACC |
| 3 | chr1:134298440-134299029 | Sense | TGGCTCGTACTTTTCCTTACAGTCTGA |
| | | Antisense | GGCAGGATATTTGGTATTGTCATCTCTATG |
| 4 | chr1:135510755-135511374 | Sense | CCTACCCAAGACCTTTGAGCAGCATGAGTA |
| | | Antisense | GGACTGGTCTGCCCTTGAAAGTTCCTAATC |
| 5 | chr1:136351426-136351872 | Sense | ATTCAAATACACAGGATGCCAAAGGTAGGG |
| | | Antisense | CAACCAAGACAAGCTCCTTCCGTACCTTA |
| 6 | chr1:182761354-182761953 | Sense | GAACTTTGGTGGTCTTGGGGGATCA |
| | | Antisense | CCTCATATTGTCGTTTGACACTCAACTGGC |
| 7 | chr2:51893883-51894477 | Sense | CCTGCTTCCCTGACCTTCTGTGCTTAAAGT |
| | | Antisense | AATACCGCCCTTCAGCCTTGAGATTCC |
| 8 | chr2:93728412-93728975 | Sense | CCACTCCTCCTTTTGTCATTTCTGATCCT |
| | | Antisense | ATACAGCGTGTGGGTGTTCTTTCAGTTTGT |
| 9 | chr2:162743518-162744121 | Sense | CCCCCAATGCAATTCTGTTATCCGTTCTGA |
| | | Antisense | GGCAGCAAGGTCACTCCAGTGTCACAAG |
| 10 | chr2:172268270-172269054 | Sense | GAGACAGCAGAAAATGTGGGAGATGGTAAA |
| | | Antisense | CGTGGTGAGGAATGGTAGGCAGAATAC |
| 11 | chr2:180638345-180638906 | Sense | GGACCAACATTCCTTCCAGAAGCATTCCA |
| | | Antisense | TATTCCTCCCAACCCCGAATCCTGTTTACC |
| 12 | chr3:34949336-34949827 | Sense | CCGCACGTCTGACCTTGAGTAAGTT |
| | | Antisense | CGGTATTAAAGTCTCCTTCACAAGACCTCC |
| 13 | chr3:88657705-88658241 | Sense | TCTCTTATTAACCAAACAGGGGTGATGA |
| | | Antisense | TAAAAGCCAGAAAATCCCTAACAGACCTTA |
| 14 | chr3:122137322-122137814 | Sense | ACTCAAAGCAATCTACAGATTCAACACAAT |
| | | Antisense | GCTTGACTGAGATTAGTGTTCTTGTGGTT |
| 15 | chr3:133461886-133462445 | Sense | CAGGGTGTTTTCATCTTGAATACTAACTGC |
| | | Antisense | CCTACTGTCCTTCTTAACTGCCCATAGC |
| 16 | chr4:10859358-10859959 | Sense | TGGTTCAAATTCCTTCCTGTCTTTAGCTCC |
| | | Antisense | TGGTGTGCTGGTGTTGCTCATGTAGA |
| 17 | chr4:34152418-34152898 | Sense | AAGAAACAACAGTTACTCCCCAGATGCTC |
| | | Antisense | GTGAATCACCCTGAATCGACTTAGACTCTC |
| 18 | chr4:55496361-55496919 | Sense | GATGGACTATAACCTGCTAACCTTGAATAC |
| | | Antisense | GGGTTTATCAGAATGACCAAGATACA |
| 19 | chr4:55498339-55498936 | Sense | AGGGTGGGAAGGGCTAGGATTGCTT |
| | | Antisense | GAGGTTGCCAAGTTGAAATTGATGAGTGTG |
| 20 | chr5:33852105-33852787 | Sense | CCAGATCGCCGCTGCTCCCATGTAGG |
| | | Antisense | TGTCACTGTCCCACCCCACGGCTACTAAGG |

| | | | |
|---|---|---|---|
| 21 | chr5:65143737-65144291 | Sense | CACCACACCCTGTAATTCTCGCCACC |
| | | Antisense | TGAAGTTCAGGAGGGCCTCTGTTCTAGTCA |
| 22 | chr5:73181345-73181841 | Sense | GCAGGAGAATGTCCACAGGTTCCAGACTAA |
| | | Antisense | GATGGTGCTCTTGCTGCTGAGTCCAAT |
| 23 | chr5:93187720-93188433 | Sense | GCCCATCGGAGTCCCAGATTAGCTTAATTG |
| | | Antisense | GCGACTTCTTTTTGCTCCCGGCTGT |
| 24 | chr6:72716760-72717250 | Sense | GGAAACTTGGATGTGGAGTAGCGGCTCGAA |
| | | Antisense | GCTGTGTGCCTGGCATTTAGAGCGGTGATT |
| 25 | chr6:149088207-149088837 | Sense | GCAGGCTCAGCGGGCAGTCTATTCT |
| | | Antisense | GACCTCTACAGCGGGCTTTATCTCTAGGGA |
| 26 | chr7:63835859-63836547 | Sense | GCTCATGAGGGCTGCCAGATCACAAC |
| | | Antisense | CCCTGGGAAGGACAAAAGGAGAGGCTAAAC |
| 27 | chr7:107027850-107028498 | Sense | GGCCTTGTCCTTGAACTTAAATGTAGTCA |
| | | Antisense | AGGAGACCATACTTTACTCTGCCTACCAAT |
| 28 | chr7:140810693-140811230 | Sense | AACTCCACACACCCACAGCCACAATAC |
| | | Antisense | GCATCCGGTAATTGGTTTCTTTCATAAGCA |
| 29 | chr8:44819420-44819963 | Sense | GGTGAAAGGACAGAGGAAGGTCGAG |
| | | Antisense | ATAAAATCAGGGCTCTTAGGAGGTGAACAC |
| 30 | chr8:46478018-46478574 | Sense | CCCTAACCTTGATTACTTCTCTTAGCACCC |
| | | Antisense | TTTGTTCAAACAGAGGTCACCGGTAG |
| 31 | chr9:10410913-10411450 | Sense | TGATCAGGGAGGGGCACTATTTTAAGGGAT |
| | | Antisense | ACAAGCAGGCAGACACAGTGGTCTAGAGAA |
| 32 | chr9:58077306-58077895 | Sense | CACCTAACTAGTTTACCACCTGTGCTTAAC |
| | | Antisense | GGTCCAGTCTTTTATTTTACATTTCAATCC |
| 33 | chr11:9016833-9017319 | Sense | CCTTGGGTGACCTTAGCCGAGAGTG |
| | | Antisense | GAGACGAGTACCAGCCCTTTGTTATCCAAT |
| 34 | chr11:44590115-44590686 | Sense | GAGCAGCTGTGAATGCAGATTGAAGT |
| | | Antisense | GCTAATTGCTGCAGTGATCTTTAACCTTTC |
| 35 | chr12:42600023-42600471 | Sense | GAAATGCATAGCAGGTTCTTGTGTATTGGG |
| | | Antisense | GCCTAGACCGGAGCAGGATTTTTATCAGTT |
| 36 | chr12:87390672-87391356 | Sense | GTTGCTTTCTTTTGCTGGTGGTATTCAACT |
| | | Antisense | GTTCCGGTCACGTTGTGGGTTCTATACT |
| 37 | chr13:113583221-113583782 | Sense | GTCTTCAATTGAACAAAGGCCCTTACAAGT |
| | | Antisense | CCCCAGCTCTGAATTATGACCCCTTAG |
| 38 | chr14:77023042-77023657 | Sense | ACCCTTCATGTAATTCTTCTCCAACCTAGT |
| | | Antisense | GTCTTCCTGAGGAGTGATGGGATCT |
| 39 | chr16:13662331-13662766 | Sense | GCTCTATTGTTCTGTCTTCCGCGTTGCTCT |
| | | Antisense | CCGGAACTGTAAGGAAGGAAACGGTGAGA |
| 40 | chr17:35111759-35112443 | Sense | GGTCCCTCTCGTCCTAGCCCTTCCTTAATC |
| | | Antisense | CTTCCGTTTCCTCCACTCTGTCATGCTC |
| 41 | chr17:35112829-35113440 | Sense | GAGGGGATTGGGGCTCAGGAGGGGGTTGGG |
| | | Antisense | GGACTGGGGGAAGGGGGCAGGACAATGGC |
| 42 | chr17:36605361-36605907 | Sense | GGGACGACAAGGAGGTCACAGAAGTCAAA |
| | | Antisense | TCTCCAATAGGACCCATCCCAGAAGCATTA |
| 43 | chr17:47034735-47035280 | Sense | CATGGGAAGTCAGGGTGCTGGACTCTT |
| | | Antisense | CTAACAGAGGCCAAAGGGTAAAGGTCGAGG |

| Myc cluster | Coordinates | Direction | Primer sequences |
|---|---|---|---|
| 1 | chr5:151636756-151636970 | Sense | TGCAAGCGATCAAAGATGGAGAAATCAGCA |
| | | Antisense | GGATATCCAGCCTTCTAGCGTGTTCCC |
| 2 | chr7:35511358-35511719 | Sense | TGGAGCTCAGGCTCGGGTGCGTGGCAGTGC |
| | | Antisense | GACGAGGGCGAGGGTACGGTGGGGTC |
| 3 | chr5:74374842-74375138 | Sense | CGGCGGCCAGAACGGAGGGGTAGAGCAGTT |
| | | Antisense | ATGCCGGGACGTACCTGCGCTTGTGTCCGA |
| 4 | chr4:133182740-33182836 | Sense | TGCCCTGTAAATCAGCCCCGTGCTC |
| | | Antisense | CTCTGACCCGACCATCACAGCCCTCACCGA |
| 5 | chr11:116487757-116487846 | Sense | CCCATCAAACCATTCTCGACATCGC |
| | | Antisense | AACCTCATCTTCCCTTTTCCACAACTTACC |
| 6 | chr15:98700232-98700381 | Sense | AGGAAAGGCTGTGACTCTGCAGTTCTACGC |
| | | Antisense | TGGCCAGAATCCTAGCTTTACCACC |
| 7 | chr9:120767948-120768063 | Sense | AAAGTCAGGTCTCAGGTGTTCCCTACTGGC |
| | | Antisense | CAGAGTTGGCTTCAGAGCGCTTAAGA |
| 8 | chr2:158052777-158052916 | Sense | TAGACCCAGCAGACCAGCCGCCCACCGAC |
| | | Antisense | CATTTGTTCTTCCGCCAGCGCCCATCGTGA |
| 9 | chr7:143105875-143105895 | Sense | AAAAACCATGCAGAGGACGATCACACGGAT |
| | | Antisense | GCCTCATTTTGGATTACTTCGGTGGG |
| 10 | chr18:33919698-33919877 | Sense | GGGAGGCGGTTGTGGCGTGGGATCTAGG |
| | | Antisense | GGGTCTTCGCTTTATTCCCGCCCAGTTCTG |
| 11 | chr3:31207484-31207548 | Sense | CTCTGTCCTCTTGGCTATCTGCTTACCTTC |
| | | Antisense | ACTGGTGAAAGGGAGGAAATGACGACAGAA |
| 12 | chr17:12752839-12752916 | Sense | CAGGAGTGGAAGGCTCTGCTCAAATTCGTC |
| | | Antisense | TCCACTTTTCCCAGCATGAGAAGATCAAGG |

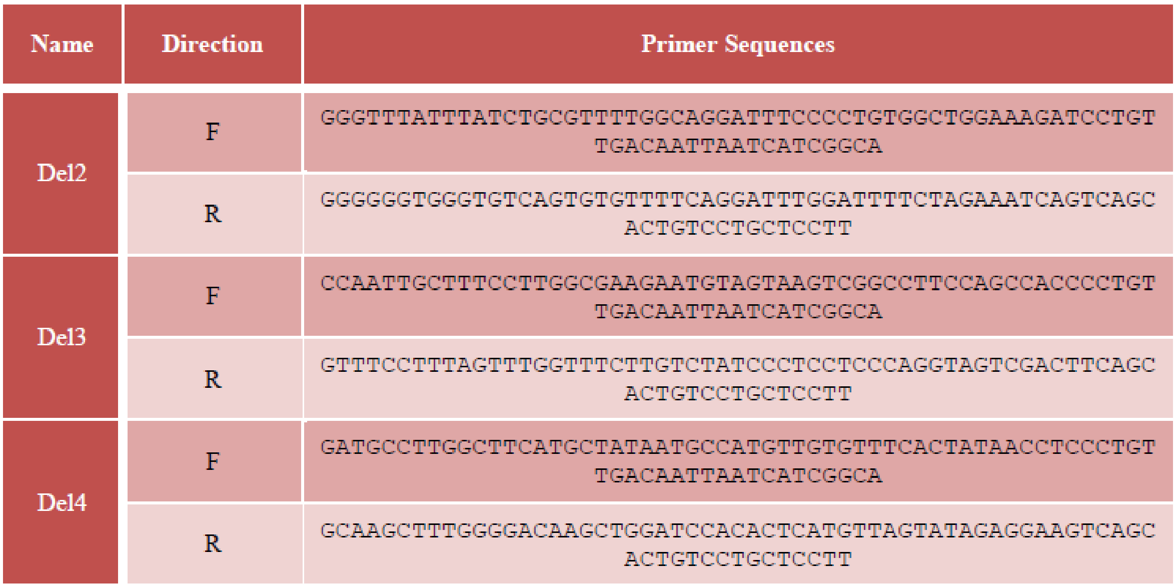
## Sequence of primers for luciferase reporter assays

| Name of Loci | Direction | Primer Sequences |
|---|---|---|
| *Tcf3-i* | F | GGCGTCAGTTCTTCCTAACAACCAATCAG |
| | R | GGTGACCTCCATCCACATTGCATTAAGGTG |
| *Tcf3-ii* | F | ATTAACCTCCCATTTGCCAGGCAGCTTCTC |
| | R | CAACTGTGGTGATGCTCATGTGCCCTGTCT |
| *Tcf3-iii* | F | GAAACTTGGATGTGGAGTAGCGGCTCGAAG |
| | R | CTGTGTGCCTGGCATTTAGAGCGGTGATTA |
| *Nanog-i* | F | AAGTAGATCAACAGGGTCAAGGTGCTGTAA |
| | R | CTTATTAAGATGCAAATTCTCACACGAGGC |
| *Nanog-ii* | F | CGGAAAAACACAGAAGAAGCAAGCGACAC |
| | R | TCACTGCCAAAACAGCCCTCTCCTTAGC |
| *Nanog-iii* | F | ACCGGTGATACGTTGGCCTTCTAGTCTGAA |
| | R | TGGGGTGCTCATTCCAAGCTAGGATGTTAG |
| *Nanog-iv* | F | GCCCCAATCTGGTCTCCTGCTGTTAGCCT |
| | R | GACTGATAGAAATGTAAAGCACGGGGTCTG |
| *Nanog-v* | F | GGTCTATTAAATCCCATGCCCTGACCAATG |
| | R | GTGCTCATGCCTGTAATCCCCTTGCCAG |
| *Nanog-vi* | F | GTCTCATGGTGTGGTTGTAAAAGCGACTGG |
| | R | ATCTTGAAATGCCCTTCCTACGAGTCATC |
| *Nanog-vii* | F | GCAGCATAATGGAGATGAAGGCCGACT |
| | R | GGTACAGAAATGAAACCCAGGCCCTAATCA |

## Sequence of primers for Chromatin Conformation Capture (3C) assays

| Name | Sequence |
|---|---|
| Nanog-Fragment A | CCTGAAATCCACGCTGACCTGGCCTTGA |
| Nanog-Fragment B | GGACCAGGAGAAGAGACACTCGTATGAA |
| Nanog-Fragment C | CCAGCAGTGTGCATTATCGAATCTCCAC |
| Nanog-Fragment D | GCCTCAGTCAAGGTTTGTCCAATCTTCT |
| Nanog-Fragment E | CAGGCAGGGTTAGCAGTGAAAATCTACA |
| Nanog-Fragment F | TAACTGGACCCTCTGACTGGCTGCTCTTGT |
| Nanog-Fragment G | GAAGATGATAGATATCTCCTTGCCTCTCAC |
| Nanog-Fragment H | CAGGAAAGACCAAGGAAGCCTACCTTTA |
| Nanog-Fragment I | TAAATAGGTAGGCTGGGCTTCGGGCTGT |
| Nanog-Fragment J | GGCAAAGACCCTCCTTTAAGGAGAAAGG |
| Nanog-Fragment K | GTGTGATGGAACCAGTTAATCCTGACAAC |
| Nanog-Fragment L | GCTTCAGCAGCCTCTACAGCAAGTGGTA |
| Nanog-Fragment M | CAAGCCTTCATATGAAAGGGTGTTGACCAG |
| Nanog-Fragment N | CAAAGAAGCTCTCAGCCAAGTACTCCTC |
| Nanog-Fragment O | GCCCTGGCAAGGCAGGAGAGAAATTGTA |
| Tcf3-Fragment A | CCTGACCAACCATTGTCGTTGCACAGAG |
| Tcf3-Fragment B | GAATAACAAACTGACCCCGCCCGCTGGT |
| Tcf3-Fragment C | GGTTGAGTCAATCGCTCTGTAAGTTCAA |
| Tcf3-Fragment D | CCCATCCTCTCAAGGGTTGTGGTCAAAT |
| Tcf3-Fragment E | CTGCCATCGTGTTGGTTTGCCTCCTTTTAG |
| Tcf3-Fragment F | GGGACAGCGCGTCATGAGTTTTGTGTCTTC |
| Tcf3-Fragment G | CTCCTTGACTTCCTAGCCCTATGTGTGTT |
| Tcf3-Fragment H | CTTCCATCCATCAGAACGCAACCCCTAA |
| Tcf3-Fragment I | TACACGAACTGTGGAGGCCATCCTGCTC |
| NANOG -Fragment A | GCGTATGTATGGGATACGCCTCACAGTTCG |
| NANOG -Fragment B | CTGGCACTCAGGTGAACCCAATAACCTTGA |
| NANOG -Fragment C | GGGTTTGGGAATGAGCAAGTGGGGATGTGT |
| NANOG -Fragment D | GTGTTCAACTCTCAATTCTACCCCTGCAGT |
| NANOG -Fragment E | CCTGGGTTCATCCTGATAAAGTCTCTCCCT |
| NANOG -Fragment F | CTTCACCGGCTTCCTCATTACCTTCTTGGC |
| NANOG -Fragment G | GGCGCAATTCAATGCTGATTGTCAACCTGT |
| NANOG -Fragment H | TAGTGCTGGGGCGGTTAGAATGCAAACATT |
| TCF7L1-Fragment A | ATGAAAGTCAGGGAGATCCATAGGCCCAAG |
| TCF7L1-Fragment B | CTACTGACCTAGTGGTAAGTTTGCCCAGGA |
| TCF7L1-Fragment C | AGCGACACCTTTGCACAGAAACCCACCGA |
| TCF7L1-Fragment D | ACACCCATGTCTCAAACAGAGATCCCTGC |
| TCF7L1-Fragment E | AGGGAGCCATTGCAAGACAGTGGTAGTTGA |
| TCF7L1-Fragment F | GCCTGTTGTAATGCTCGGGAGCATTTGTTG |
| TCF7L1-Fragment G | CTAAGCTCCCGGAAACCAAGCCAATTTTCG |
| Ndufa4-P1 | CTCAGGTCTCCACTATCTGTGCTCAAAG |
| Ndufa4-P2 | GGGAGGAAGATGGAAAACCTCAACTCTT |

Sequence of primers for BAC modification



| Name | Direction | Primer Sequences |
|---|---|---|
| Del2 | F | GGGTTTATTTATCTGCGTTTTGGCAGGATTTCCCCTGTGGCTGGAAAGATCCTGT<br>TGACAATTAATCATCGGCA |
|  | R | GGGGGGTGGGTGTCAGTGTGTTTTCAGGATTTGGATTTTCTAGAAATCAGTCAGC<br>ACTGTCCTGCTCCTT |
| Del3 | F | CCAATTGCTTTCCTTGGCGAAGAATGTAGTAAGTCGGCCTTCCAGCCACCCCTGT<br>TGACAATTAATCATCGGCA |
|  | R | GTTTCCTTTAGTTTGGTTTCTTGTCTATCCCTCCTCCCAGGTAGTCGACTTCAGC<br>ACTGTCCTGCTCCTT |
| Del4 | F | GATGCCTTGGCTTCATGCTATAATGCCATGTTGTGTTTCACTATAACCTCCCTGT<br>TGACAATTAATCATCGGCA |
|  | R | GCAAGCTTTGGGGACAAGCTGGATCCACACTCATGTTAGTATAGAGGAAGTCAGC<br>ACTGTCCTGCTCCTT |

## 4.4 RESULTS

**p300 and CBP play redundant roles in maintaining the undifferentiated state of ES cells.**

To assess the functional roles of p300 and its closely related gene, CBP in ES cells, we depleted endogenous p300 and CBP to about 40%, respectively, by RNAi (Figure 4.1). Two short-hairpin RNA (shRNA) constructs targeting different regions of the transcript were used to ensure that the effects were specific. Alkaline phosphatase (AP) staining and the expression of pluripotency marker genes ware analyzed to evaluate whether the ES cells underwent differentiation. Consistent with the previous report (Zhong and Jin, 2009), no morphology change of ES cells was observed due to the knockdown of p300. Typical colony morphology of ES cells with positive alkaline phosphatase staining (red) was maintained in p300 knockdown cells as in control cells transfected by luciferase shRNA (Figure 4.1A). Similarly, depletion of CBP has no obvious effect on ES cells morphology neither (Figure 4.1A). In addition, pluripotency markers Pou5f1, Nanog were expressed at comparable levels in p300 or CBP depleted ES cells and control ES cells (Figure 4.1B, C). This data indicates that either p300 or CBP is not required for the self-renewal of ES cells.

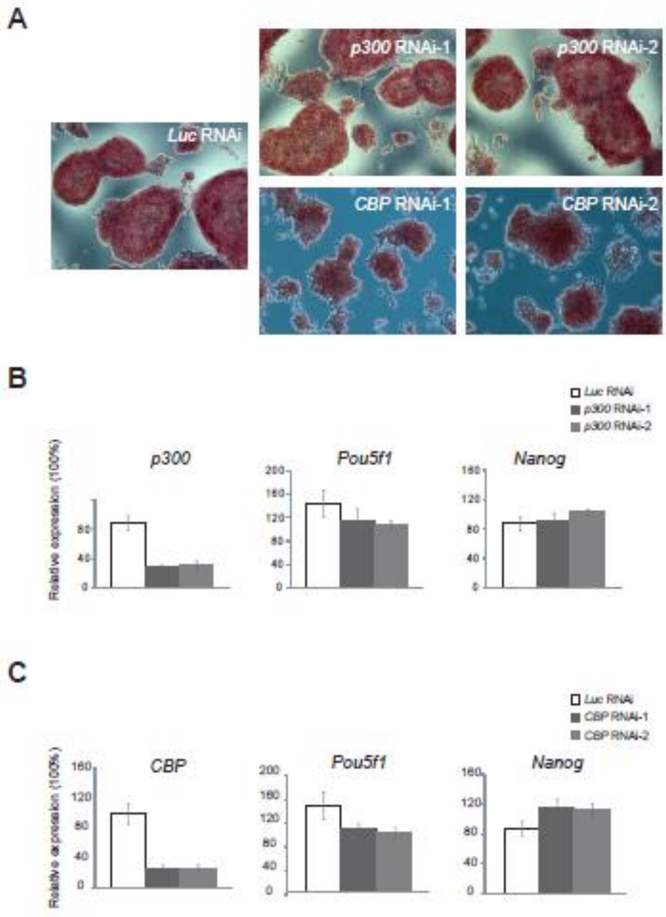


**Figure 4.1. p300 and CBP are dispensable for the maintenance of ES cells.** (A) Alkaline phosphatase (AP) staining (red) was performed on control cells (Luc RNAi), p300 (p300 RNAi-1 and p300 RNAi-2) and CBP (CBP RNAi-1 and CBP RNAi-2) knockdown cells two days after puromycin selection. (B) Quantitative real-time PCR analysis of expression of p300, Pou5f1 and Nanog on control and p300 knockdown cells. (C) Quantitative real-time PCR analysis of expression of CBP, Pou5f1 and Nanog on control and CBP knockdown cells.

Since p300 share very similar amino acid sequences with CBP, we hypothesized that *CBP* play a redundant role with *p300* in ES cells. To test this hypothesis, simultaneous depletion of *p300* and *CBP* by constructs expressing shRNAs targeting both two different transcripts was performed. Strikingly, double knockdown *p300* and *CBP* at the same time led to cell differentiation and consequently disrupted ES cells self-renewal. Alkaline phosphatase (AP) staining of pluripotent ES cells (red color) was reduced dramatically in the double knockdown cells, indicative of differentiation (Figure 4.2A).

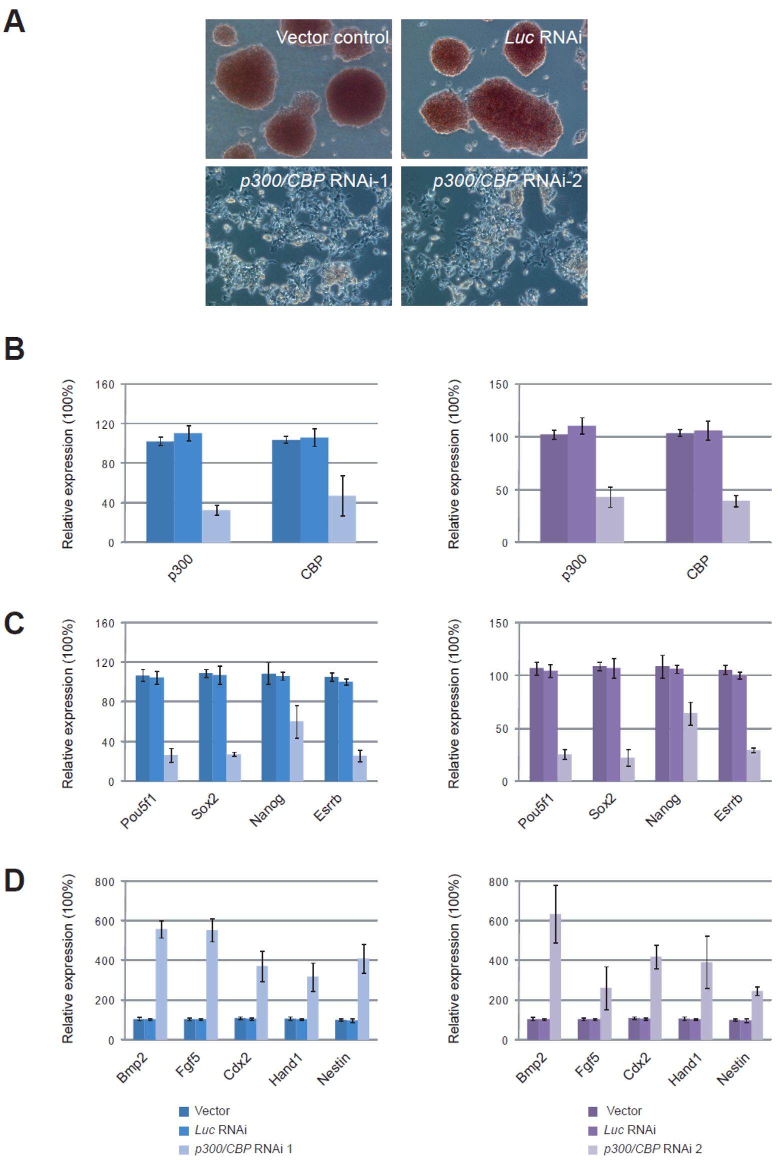


**Figure 4.2. *p300* and *CBP* are required and playing redundant roles for the maintenance of ES cells.** (A) Concurrent knockdown of *p300* and *CBP* led to ES cells differentiation. Differentiated cells with negative alkaline phosphatase staining were formed after knockdown using two sets of shRNA constructs targeting *p300* and *CBP*. Typical colony morphology of ES cells with positive alkaline phosphatase staining (red) was maintained in empty vector and *luciferase* shRNA-transfected cells. (B-D) Quantitative real-time PCR analysis of  expression of (B) *p300* and *CBP*, (C) ES cell-associated genes and (D) lineage-specific marker genes after knockdown using two shRNA constructs targeting different regions of the respective transcripts. The levels of the transcripts were normalized against control empty vector transfection. Data are presented as the mean ± SEM and derived from 3 independent experiments.

We further examined the expression of marker genes in double knockdown cells. To our expectation, the expression of self-renewal associated genes, *Pou5f1*, *Sox2*, *Nanog* and *Esrrb* was reduced in response to *p300/CBP* double knockdown (Figure 4.2C), while a strong induction of differentiation related genes were observed (Figure 4.2D), including mesoderm marker *Bmp2*, ectoderm markers *Fgf5* and *Nestin* as well as trophectoderm markers *Cdx2* and *Hand1*, suggesting that the resulting cells from double knockdown were likely to be composed of multiple differentiated cells.

To further confirm the specificity of the knockdown experiments, RNAi-resistant cDNA encoding p300 or CBP was co-transfected with *p300/CBP* shRNA. Interestingly, expression of either RNAi-resistant p300 or CBP to certain dosage was able to rescue the differentiation phenotype induced by double knockdown (Figure 4.3A). The expression of self-renewal and lineage marker genes was recovered to the comparable level to control ES cell (Figure 4.3B-D).

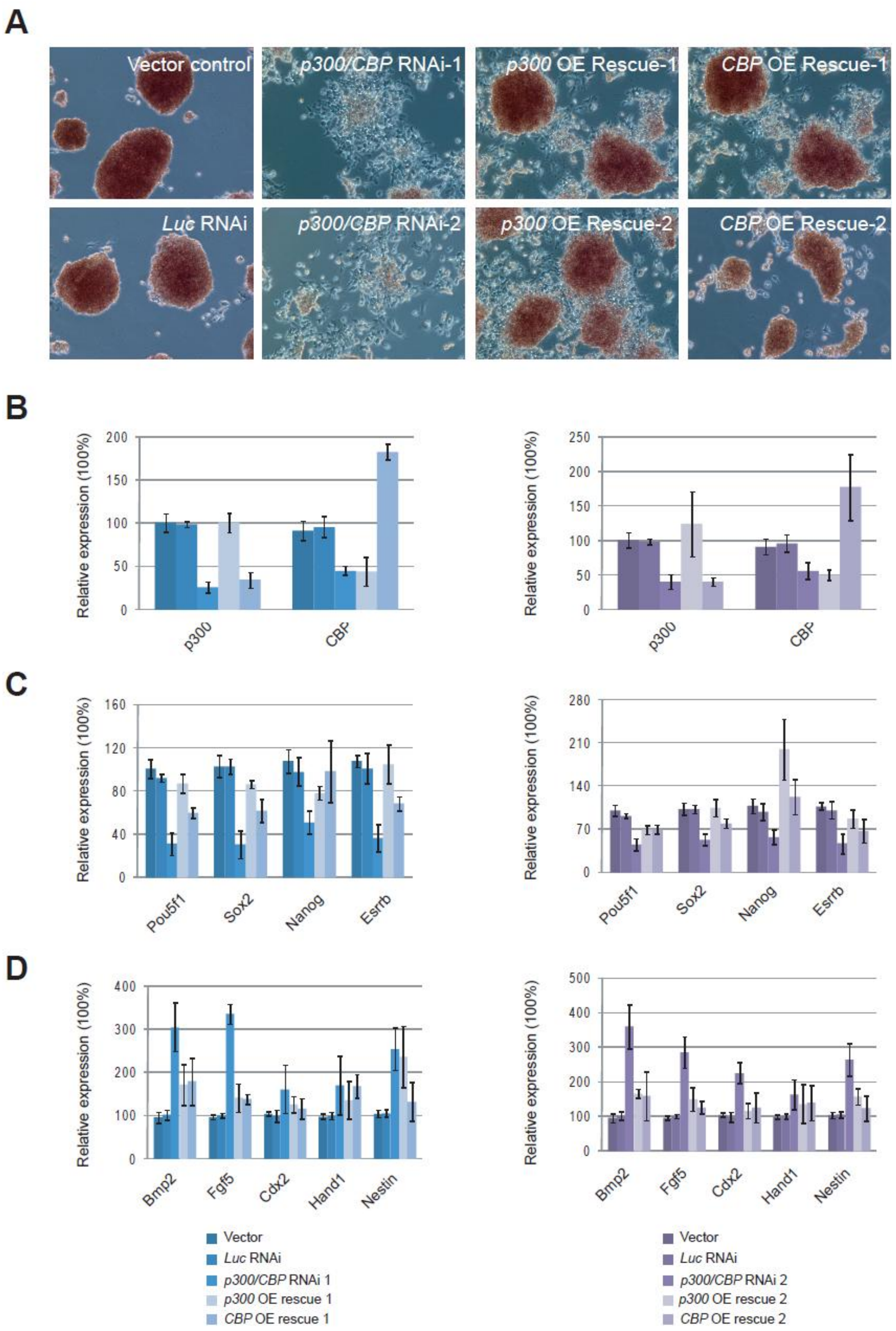


**Figure 4.3. Over-expression of *p300* or *CBP* is able to rescue the double knockdown effect.** (A) Rescue of concurrent knockdown phenotype by co-expression of RNAi-resistant *p300* or *CBP*. RNAi-resistant *p300* or *CBP* expression constructs were co-transfected with corresponding *p300/CBP* RNAi targeting different regions into ES cells respectively. Typical colony morphology of ES cells with positive alkaline phosphatase staining (red) was restored. (B-D) Quantitative real-time PCR analysis of expression of (B) *p300* and *CBP*, (C) ES cell-associated genes and (D) lineage-specific marker genes after p300 or CBP rescue. The levels of the transcripts were normalized

against control empty vector transfection. Data are presented as the mean ± SEM and derived from 3 independent experiments.

All these data strongly suggest that p300 and CBP are functionally redundant in the maintenance of self-renewal of ES cells. A reduction in the dose of one of them can be compensated by the other to cover the function deficit.

**p300 and CBP are recruited by Nanog, Oct4 and Sox2 through direct protein-protein interaction.**

Identifying the binding targets of transcription factors and cofactors is helpful for us to understand their regulatory mechanism through their downstream targets. Genome wide mapping study of p300 in mouse ES cells by ChIP-Seq has found that its binding sites are associated with Nanog, Oct4 and Sox2 binding sites. Furthermore, enriched motif for p300 that generated from its ChIP-Seq data highly resembles sox-oct composite element (Chen et al., 2008). To validate the co-occupancy of p300 with Nanog, Oct4 and Sox2 in the ES cell genome, we performed p300 and CBP ChIP-qPCR on 44 sites randomly chosen from genomic sites that are bound by Nanog-Oct4-Sox2 cluster as well as 12 sites bound by Myc cluster. The qPCR results are consistent with the ChIP-Seq data showing the binding preference of p300 to Nanog-Oct4-Sox2 cluster rather than Myc cluster and CBP has shown similar binding bias (Figure 4.4A). These data suggest that Nanog, Oct4 and Sox2 are recruiting both p300 and CBP as a coactivator complex to specific genomic sites. To test this hypothesis, we depleted Nanog, Oct4 or Sox2 by RNAi and checked for p300 and CBP binding. Our ChIP result showed that the binding intensity of p300 as well as was reduced upon Nanog, Oct4 or Sox2 depletion (Figure 4.4B).

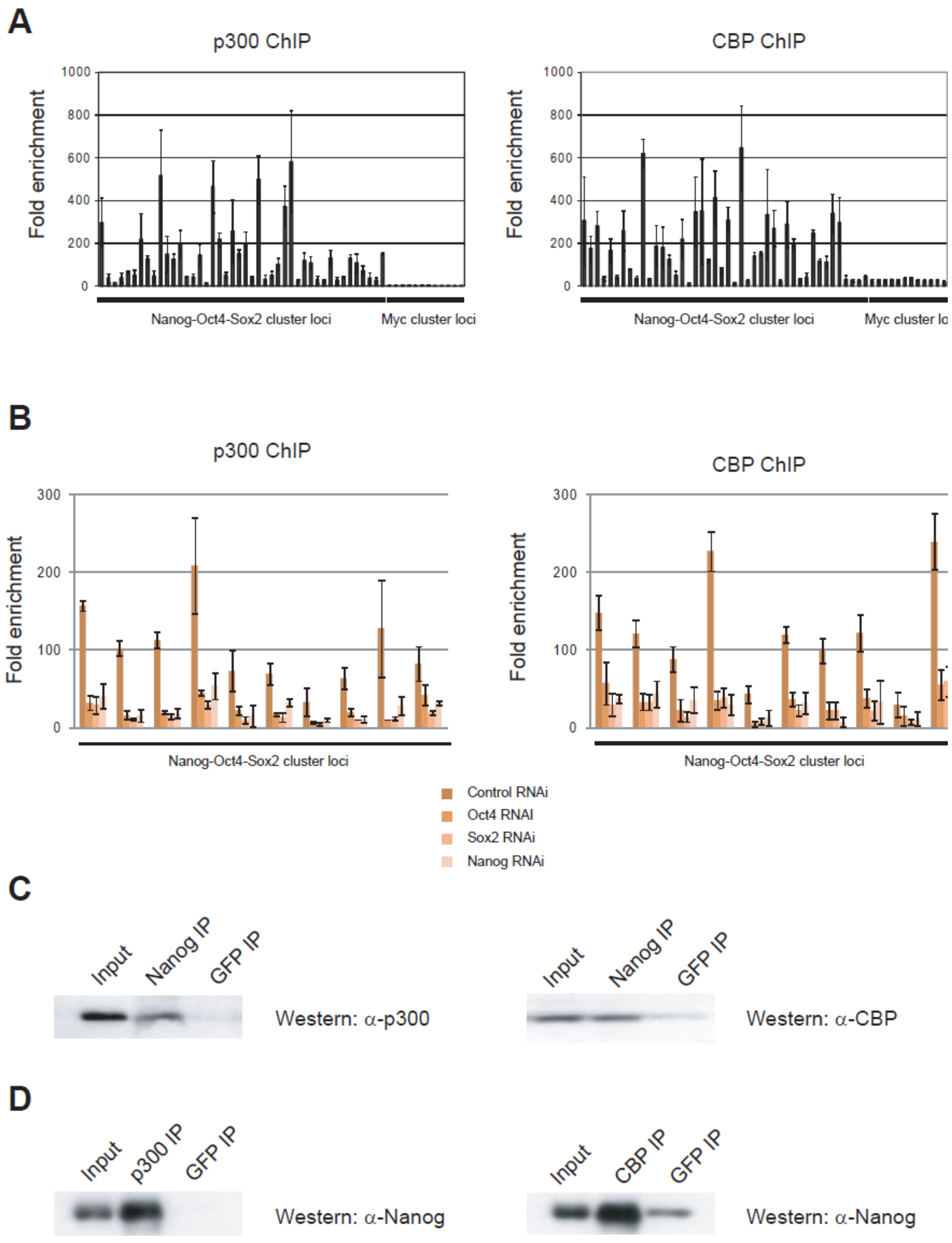


**Figure 4.4. p300 and CBP are recruited to Nanog-Oct4 -Sox2 cluster loci in mouse genome.** (A) p300 and CBP binds to Nang-Oct4-Sox2 cluster loci, but not Myc cluster loci. ChIP assays were performed using anti-p300 or anti-CBP antibody with extracts from ES cells. Fold enrichment is the relative abundance of DNA fragments detected by real-time PCR at the amplified

region over a control amplified region. Data are presented as the mean ± SEM. (B) Recruitment of p300 and CBP is dependent on Oct4, Sox2 and Nanog. ChIP assays were performed using anti-p300 or anti-CBP antibody with extracts from ES cells transfected with control RNAi construct, *Oct4* RNAi construct, *Sox2* RNAi construct or *Nanog* RNAi construct. The level of p300 was not altered after RNAi depletion of these TFs (data not shown). (C) p300 and CBP interact with Nanog. Co-IP using nuclear extracts of ES cells was performed using anti-Nanog antibody. Western blotting was performed with anti-p300 or anti-CBP antibody. Control IP was performed using an anti-greenfluorescent protein (GFP) antibody. (D) Reverse co-IP using the ES cell lysates was performed using anti-p300 or anti-CBP antibody. Western blotting was carried out with anti-Nanog antibody. Control IP was performed using an anti-GFP antibody.

To investigate the possible protein-protein interaction between p300/CBP and master regulators in ES cells, coimmunoprecipitation experiments were performed using ES cell nuclear extracts. p300 and CBP were found to coprecipitate with Nanog (Figure 4.4C), whereas the reciprocal Co-IP also showed that Nanog was able to coprecipitate with p300/CBP (Figure 4.4D).

p300 and CBP are large nuclear proteins with eight distinct functional domains (N terminal, CH1, KIX, Bromo, CH2, HAT, CH3, glutamine-rich) (Blobel, 2000; Kraus et al., 1999), which mediate their interactions with numerous nuclear proteins and allow p300 and CBP to serve as scaffolds to assemble large regulatory complexes or to manipulate chromatin structure through histone modification to activate transcription. To map Nanog-interactive elements on CBP, GST fusion proteins were generated with contiguous segments of murine CBP that collectively span the entire protein (Figure 4.5A). These were used in pull-down experiments with purified Nanog protein; proteins retained on the beads were immunoblotted with an antibody targeting Nanog. Only the fragment containing residues 451–721 of CBP, which is corresponding to KIX domain, retained Nanog proteins beyond background level (Figure 4.5B). On the other hand, to localize the p300/ CBP interaction domain within Nanog, we expressed and purified recombinant Nanog and fragments of Nanog as GST-fusion proteins (Figure 4.5C). These proteins were immobilized onto GSH-Sepharose beads and incubated with purified CBP protein. Homeobox domain of Nanog as well as fragments with homeobox domain could interact with p300 (Figure 4.5D).

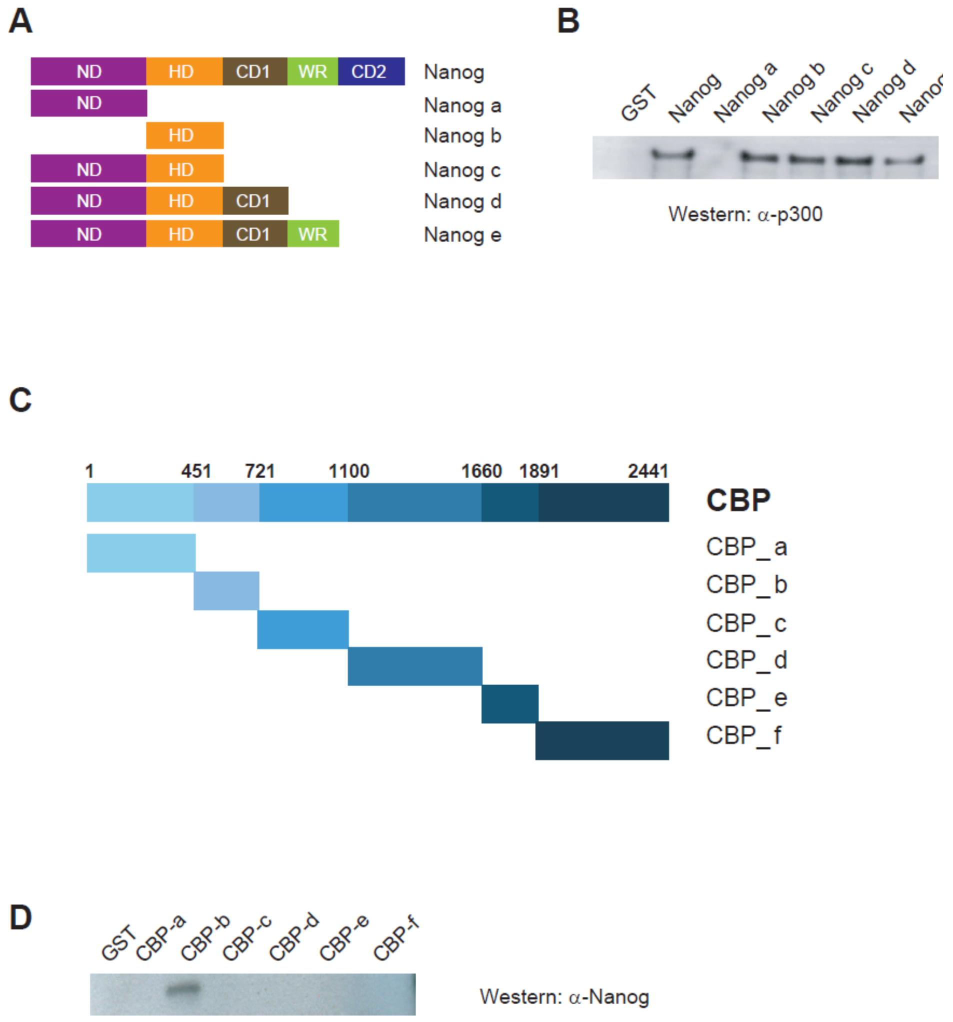
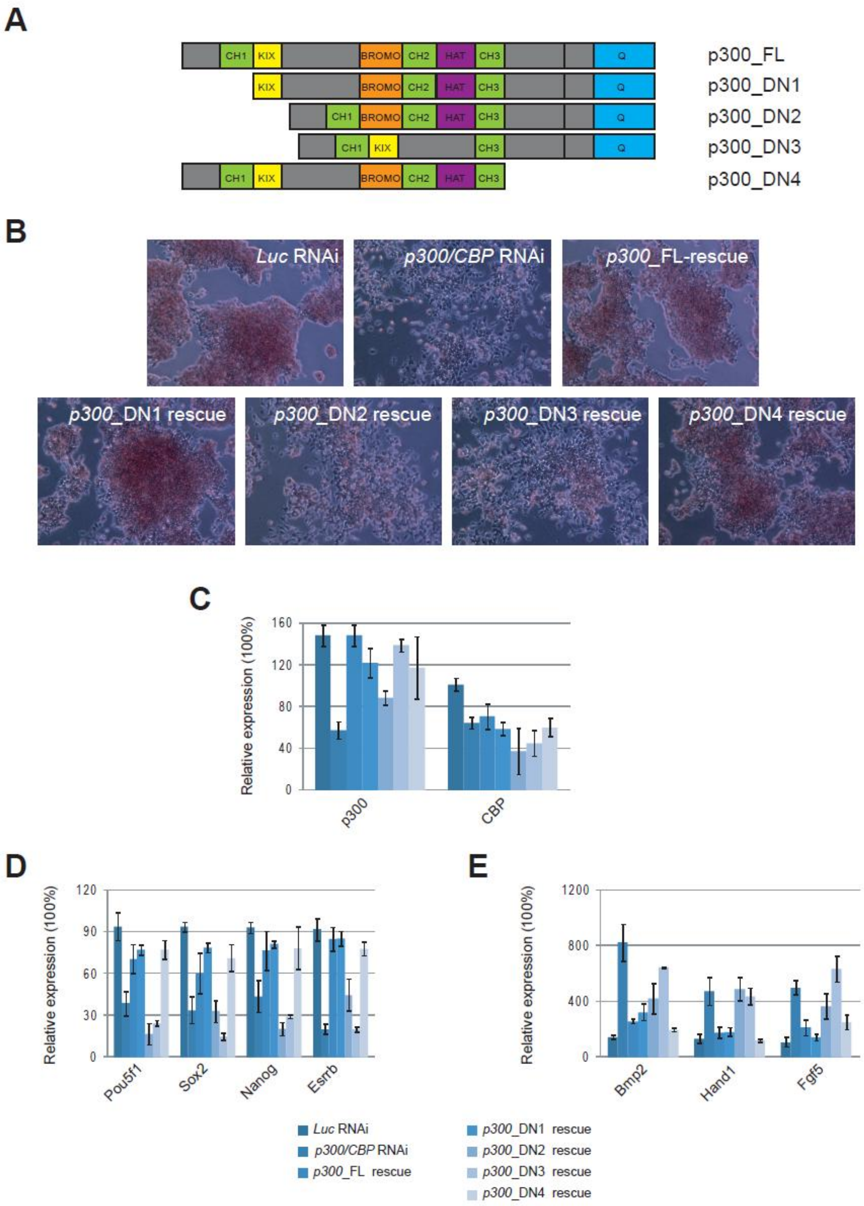


**Figure 4.5**. **Mapping the interaction domains of p300/CBP and Nanog.** (A) Schematic diagram of wild type and deletion forms of Nanog protein. ND, N-terminal domain; HD, homeobox domain; CD1, C-terminal domain 1;WR, tryptophan repeat domain; CD2, C-terminal domain 2. (B) GST pull down was carried out using GST-tagged Nanog proteins and purified p300 protein. Western blot was performed with anti-p300 antibody. GST served as negative control. (C) Portions of CBP expressed as GST fusion proteins in this study.

(D) Pull-downs with the indicated CBP amino acids fused to GST and purified Nanog protein. Western blot was performed with anti-Nanog antibody. GST served as negative control.

**KIX and HAT domains of p300 /CBP are critical for the self-renewal of ES cells.**

To further dissect the functional domain of p300 and CBP in mouse ES cells, we created a series of RNAi-immune expression constructs, each expressing a specific mutant p300 or CBP (Figure 4.6A, F).



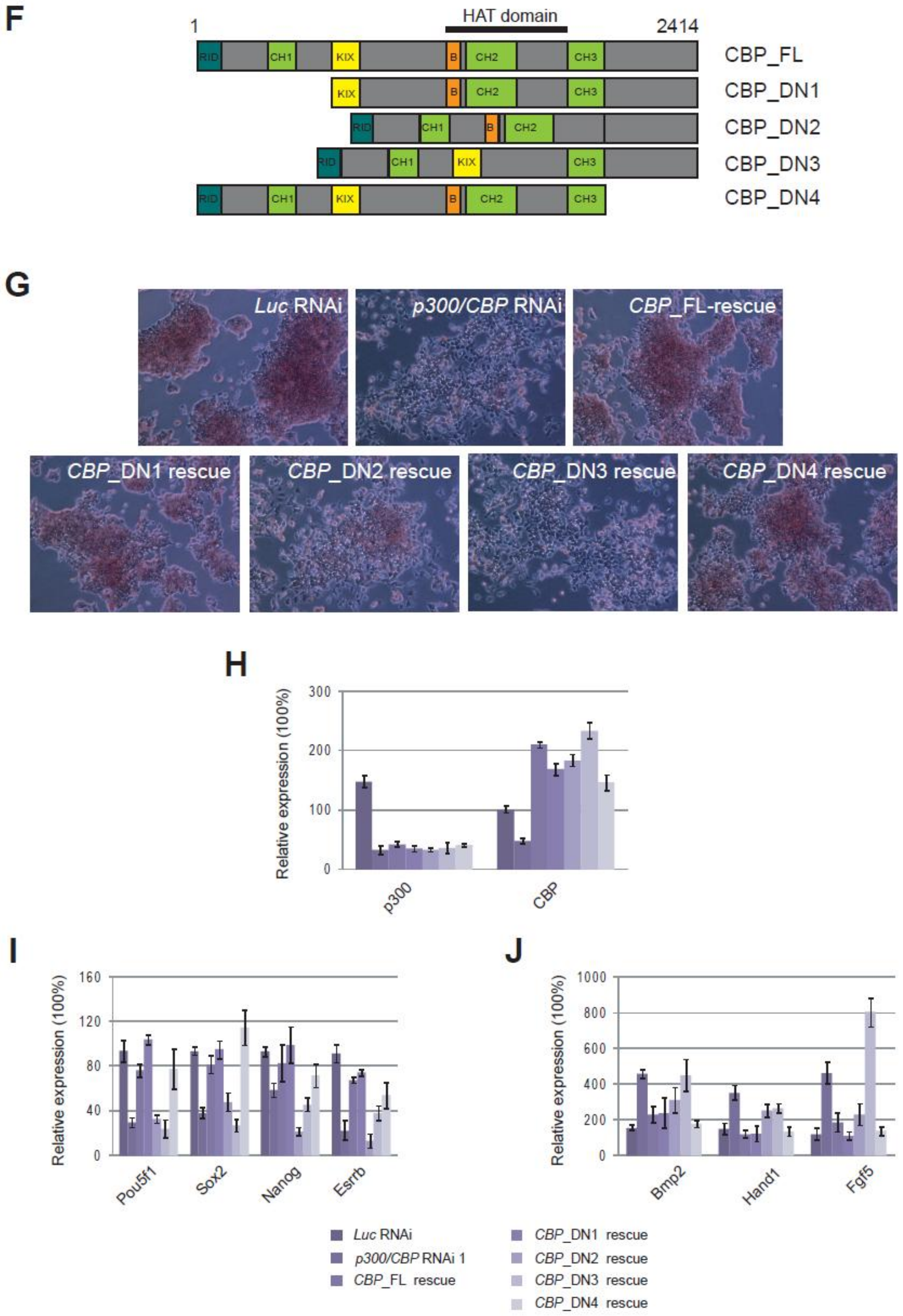


**Figure 4.6. KIX and Histone acetylation (HAT) domain of p300 and CBP are important for their function in the maintenance of ES cells.** (A) Panel of truncated p300 cDNA constructs used for rescue. The deletion of domain is indicated for each construct. Nuclear receptor interaction domain (RID), cysteine/histidine-rich domains (CH1, CH2, CH3), KIX, bromodomain (Br), IRF3-binding domain (IBiD) and HAT. The deletion of domain is indicated for each construct. (B) Rescue of concurrent p300/CBP knockdown phenotype by co-expression of RNAi-resistant truncated p300 constructs. Undifferentiated ES colonies with positive alkaline phosphatase staining were maintained after rescue using truncated p300 constructs p300_DN1 and p300_DN4. (C-E) Quantitative real-time PCR analysis of expression of (C) *p300* and *CBP*, (D) ES cell-associated genes and (E) lineage-specific marker

genes after rescue by overexpression of dominant negative p300 or CBP constructs. The levels of the transcripts were normalized against control empty vector transfection. Data are presented as the mean ± SEM and derived from 3 independent experiments. (F) Panel of truncated CBP cDNA constructs used for rescue. The deletion of domain is indicated for each construct. (G) Rescue of concurrent p300/CBP knockdown phenotype by co-expression of RNAi-resistant truncated CBP constructs. Undifferentiated ES colonies with positive alkaline phosphatase staining were maintained after rescue using truncated CBPconstructs CBP_DN1 and CBP_DN4. (H-J) Quantitative real-time PCR analysis of expression of (H) *p300* and *CBP*, (I) ES cell-associated genes and (J) lineage-specific marker genes after rescue by overexpression of dominant negative p300 or CBP constructs. The levels of the transcripts were normalized against control empty vector transfection. Data are presented as the mean ± SEM and derived from 3 independent experiments.

The mutant p300 or CBP constructs were then co-transfected with *p300/CBP* shRNA to check whether they can rescue the double knockdown effect as well as the constructs encoding wide type p300 and CBP. Interestingly, except the mutant with deleted KIX or HAT domain, all the other mutant p300 or CBP constructs are able to rescue the double knockdown cells, resulting in a normal morphology (Figure 4.6B, G) and gene expression profile as ES cells transfected with control shRNA (Figure 4.6C- E; H-J). These results suggest that the KIX and HAT domains are the functional domains of p300 and CBP in mouse ES cells.

**p300/CBP mediates intragenic looping interactions among p300 and Nanog binding loci in the *Tcf3* locus.**

From ChIP-Seq dataset, we identified over 3,000 high confidence p300 binding loci(Chen et al., 2008). Interestingly, several p300 binding loci colocalize with Nanog and Oct4 and can be found in proximity within a single gene or between genes (Figure 4.7A). This raises the question of whether there are physiological reasons for having multiple binding loci and if so, the molecular mechanisms involved in their functions. We hypothesize that these binding loci specifically interact with each other in ES cells by looping out the intervening regions, and thus may represent a unique chromatin structure in ES cells. To test our hypothesis, we study the loci that are densely bound by p300, Nanog and Oct4.

We first confirmed the ChIP-Seq data by performing ChIP-qPCR analysis using primers targeting the putative binding loci in *Tcf3*. *Tcf3* was chosen due to the presence of three ChIP-Seq identified binding loci of Nanog, Oct4 and p300 that are located at the 5′ end, middle, and 3′ end of the gene locus that is about 150 kb long (Figure 4.7D). Indeed, Nanog and p300 bound to all of the 3 binding loci (Figure 4.7B, C). We observed enrichment with large variance, which could be due to differential efficiency in the crosslinking of protein to the various binding loci as Nanog or p300 may not be binding to all the sites in similar strength or that not all the sites are bound by Nanog and p300 directly. The results point toward an intriguing possibility that these binding loci are interacting with each other by looping and that the interaction is tethered by p300.

To determine whether the binding loci are interacting with each other by looping out the intervening regions, we exploited the recently developed chromosome conformation capture (3C) assay. The 3C assay, which was first developed in yeast(Dekker et al., 2002), and later in mammalian cells(Spilianakis and Flavell, 2004; Tolhuis et al., 2002), is a powerful technique to analyze the overall spatial organization of chromosomes. Similar to ChIP assay, formaldehyde is used to preserve chromatin interactions in living cells. After fixation, the DNA is then subjected to restriction digestion to create cohesive ends for efficient ligation. This assay relies on the ability of DNA restriction fragments in close juxtaposition to each other due to close linear distance, long range looping interactions or interchromosomal interactions *in vivo* to be ligated to form chimeric DNA fragments, which can be amplified using specific primer pairs. It has been used in several recent studies to detect interactions between 2 proposed DNA fragments located far away from each other(Murrell et al., 2004; Spilianakis and Flavell, 2004; Tolhuis et al., 2002)' (Ling et al., 2006; Lomvardas et al., 2006; Spilianakis et al., 2005).

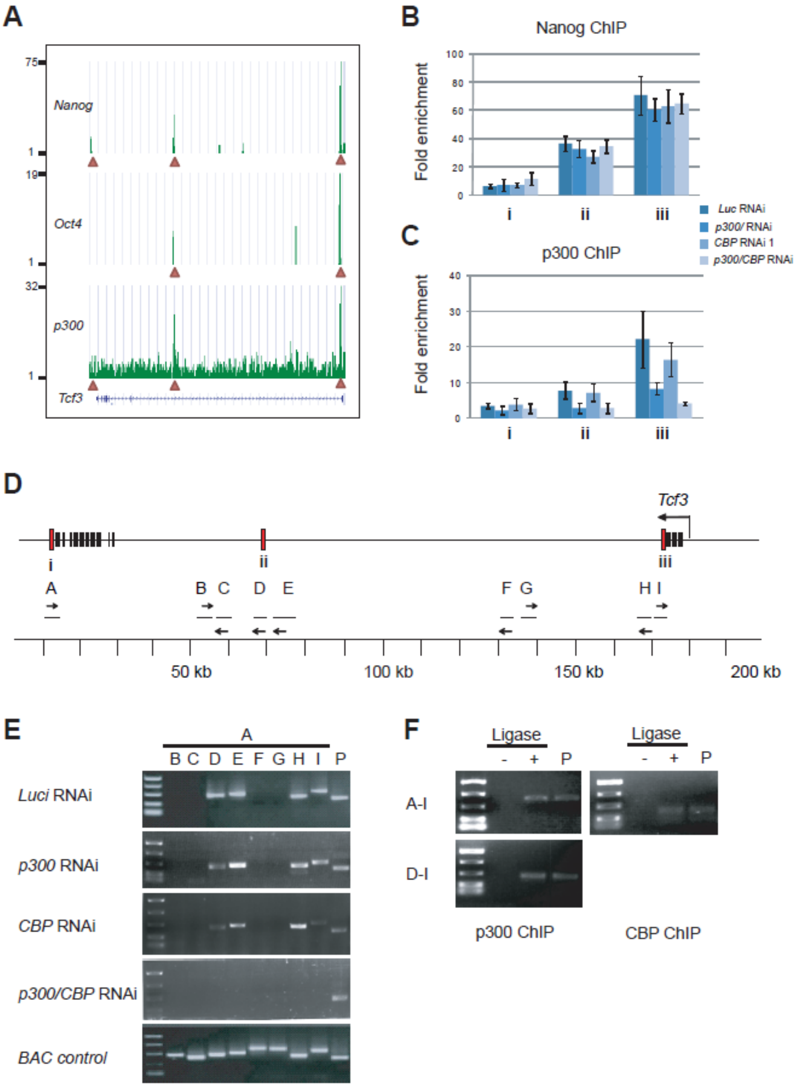


**Figure 4.7. p300 and CBP mediate intragenic looping interactions among colocalization loci** (A) ChIP-seq binding profiles for Nanog, Oct4 and p300 at the *Tcf3* gene loci are shown. (B-C) ChIP assays were performed using anti-Nanog (B) or anti-p300 (C) antibody with extracts from ES cells transfected with control RNAi construct, *p300* RNAi construct, *CBP* RNAi construct or *p300/CBP* RNAi construct. (D) Schematic representation of the murine Tcf3 locus. Dark boxes represent exons and red boxes represent ChIP-Seq identified Nanog binding loci, named by the Roman numerals indicated below them. Relevant BglII restriction fragments are indicated by short green horizontal bars. Primers are named by Roman alphabets and their orientations are indicated by arrows. (E) 3C analyses of Tcf3 locus on murine ES cells transfected with control *Luc* RNAi construct (top panel), p300 RNAi construct(second panel), CBP RNAi (third panel) and p300/CBP RNAi (fourth panel). Bottom panel, PCR controls done using BAC DNA harboring the Tcf3 locus. (F) ChIP-3C analyses done on murine ES cells using p300 antibody (Left panel) and CBP antibody (right panel). Presence or absence of amplicons is detected in a 1.5% agarose gel using the primer combinations indicated on the left of each panel. Lane P is a BAC control and the leftmost lanes represent PCR markers to show sizes of the amplicons. DNA samples used in the PCR reactions were prepared with or without ligase added in the 3C assays to show that presence of amplicons is ligase-dependent.

We first started the 3C analysis using an invariant primer (primer B, Figure 4.7D), targeting the restriction fragment next to the Nanog and p300 binding locus at the 3′ end of *Tcf3*, together with one of a series of primers complementary to different restriction fragments along this gene. Successful amplification of a PCR product from a primer pair signifies the detection of a chimeric fragment (Figure 4.7E). The identities of all the PCR products were verified by sequencing. As recommended by Dekker(Dekker, 2006), we showed that detection of chimeric DNA fragments was ligation dependent in all 3C assays. Furthermore, all primer combinations were verified to be able to give rise to their respective amplicons and the sizes of the PCR products obtained from our 3C samples were identical to that obtained from the BAC controls (Figure 4.7E, lower panel). As an internal control for our 3C assays, we used a primer pair targeting two *Bgl*II restriction fragments, from the unrelated *Ndufa4* locus, that were separated by 7.9 kb apart. Due to the proximity of the restriction fragments, we would expect them to be ligated to form a chimeric DNA fragment in all our 3C assays as formaldehyde is able to capture random collisions between two nearby chromatin loci.

The results indicated that there are interactions between region A with regions D, E, H, and I, which are regions close to Nanog and p300 binding loci. No PCR products were detected when the invariant primer A was used together with any other primers complementary to the regions (B, C, F, and G) in between the Nanog and p300 binding loci (Figure 4.7E,top panel). In addition, the looping interactions appeared to be ES cell-specific as amplicons corresponding to the chimeric DNA fragments were not detectable after subjecting ES cells to retinoic acid (RA)-induced differentiation, indicating the abolishment of looping interactions (Figure 4.8).

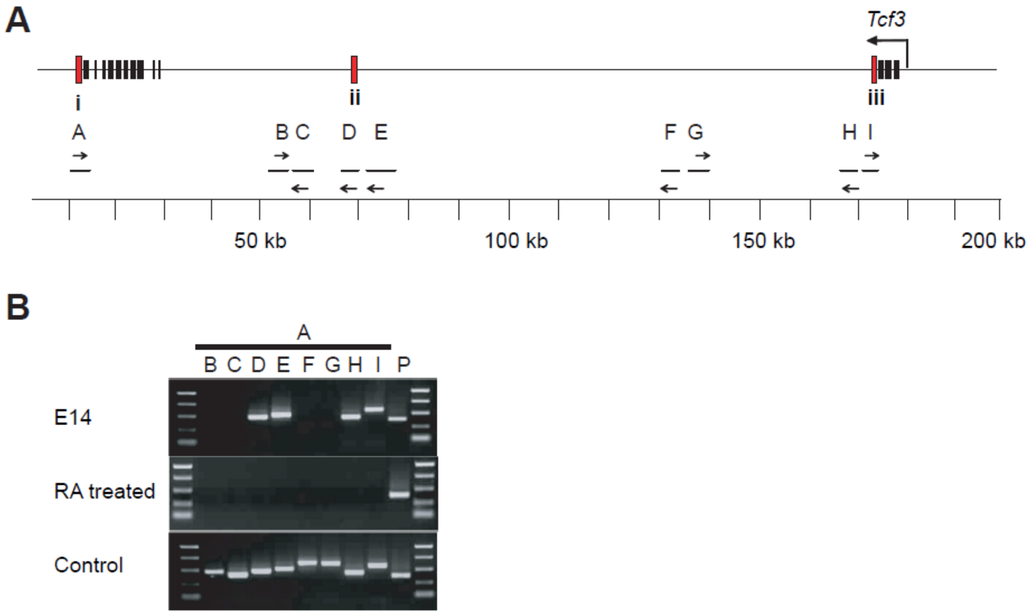


**Figure 4.8. Intragenic looping interactions are specific to the pluripotent state.** (A) Schematic representation of the mouse *Tcf3* locus. (B) 3C analyses of *Tcf3*. Top panel, 3C analyses on mouse E14 ES cells. Second panel, 3C analyses on RA treated ES cells. Bottom panel, PCR controls done using BAC DNA harboring the *Tcf3* locus.

Since the chromosomal loops are formed between fragments bound by p300/CBP, our data strongly suggest that p300 and CBP are involved in the formation of the loops.To further confirm the dependence on p300/CBP, 3C analysis was performed using ChIP-enriched DNA samples (ChIP-3C), using either p300 or CBP as the antibody. Interactions between the binding fragments on the 5′ end and 3’ end of *Tcf3* (Figure 4.7F) were recapitulated. However, other interactions were not successfully verified by ChIP-3C in our hands, possibly due to the low enrichment of certain DNA fragments and/or impaired digestion and ligation efficiencies when both were done on beads.

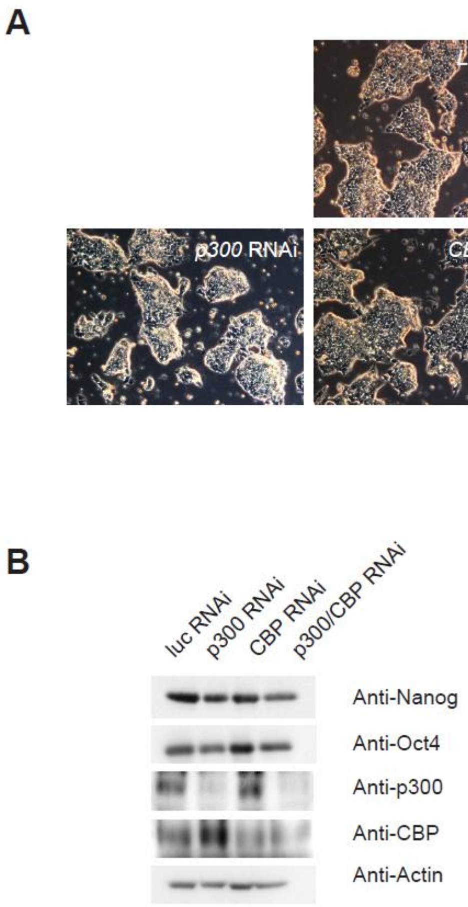


**Figure 4.9. RNAi samples for 3C assays.** (A) Morphology of *Luc*, *p300*, *CBP* and *p300/CBP* knockdown cells harvested for 3C assays. (B) Western blot of RNAi samples for 3C assays using anti- Nanog, Oct4, p300, CBP and Actin antibody.

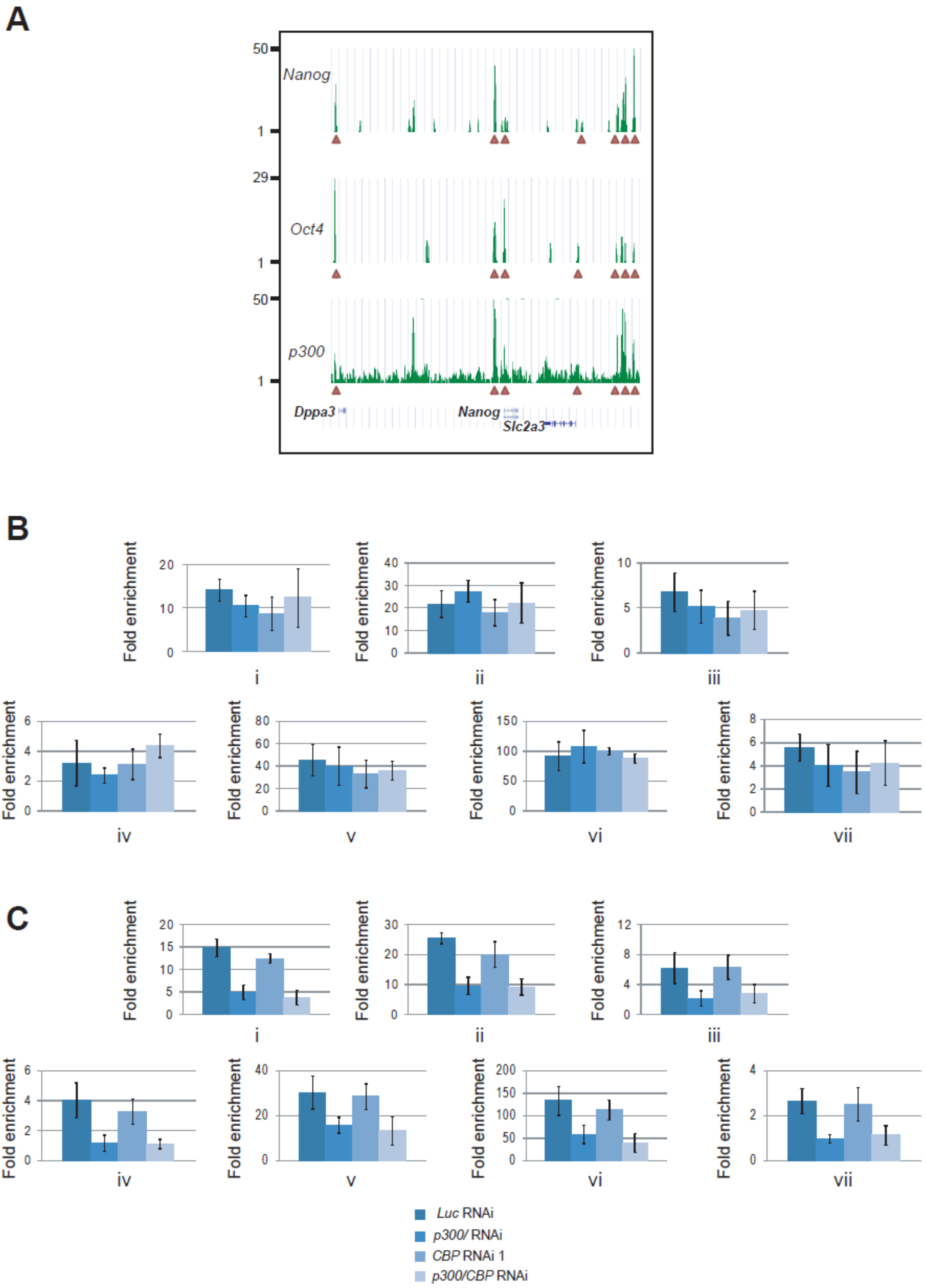
To further determine whether p300 or CBP is required for the formation of the chromosomal loops, 3C analysis was performed on ES cells transfected with a p300 or CBP or p300/CBP double knockdown construct. A luciferase knockdown construct was also introduced into the cells as control. To exclude the effect of comprehensive differentiation, knockdown cells were harvested before morphological changes begin to appear, the protein levels of Nanog, Oct4 and Sox2 were not changed as well (Figure 4.9). The looping structure retained after knocking down either p300 or CBP. However, depletion of both p300 and CBP led to the abolishment of chromosomal loops, while the control knockdown had no effects (Figure 4.7E upper panel).

Altogether, the data support a model whereby the 3′ end of *Tcf3* forms loop structures with two other loci, one located about 50 kb away in the middle of intron 3 and the other located about 160 kb away at the 5′ end of the *Tcf3* gene. The long range communications between these loci correlate with the presence of Nanog and p300 binding loci and these structures are dependent on p300 and CBP.

.

### p300/CBP mediates intergenic looping interactions among Nanog binding loci in the *Dppa3-Nanog-Slc2a3* locus.

Having shown the presence of long range intragenic chromosomal looping, we next extended our investigation further by looking for the presence of intergenic chromosomal looping. The *Dppa3-Nanog-Slc2a3* gene cluster was chosen due to the presence of several ChIP-Seq identified Nanog and p300 binding loci within the cluster containing three genes preferentially expressed in mouse ES cells (Figure 4.10A).



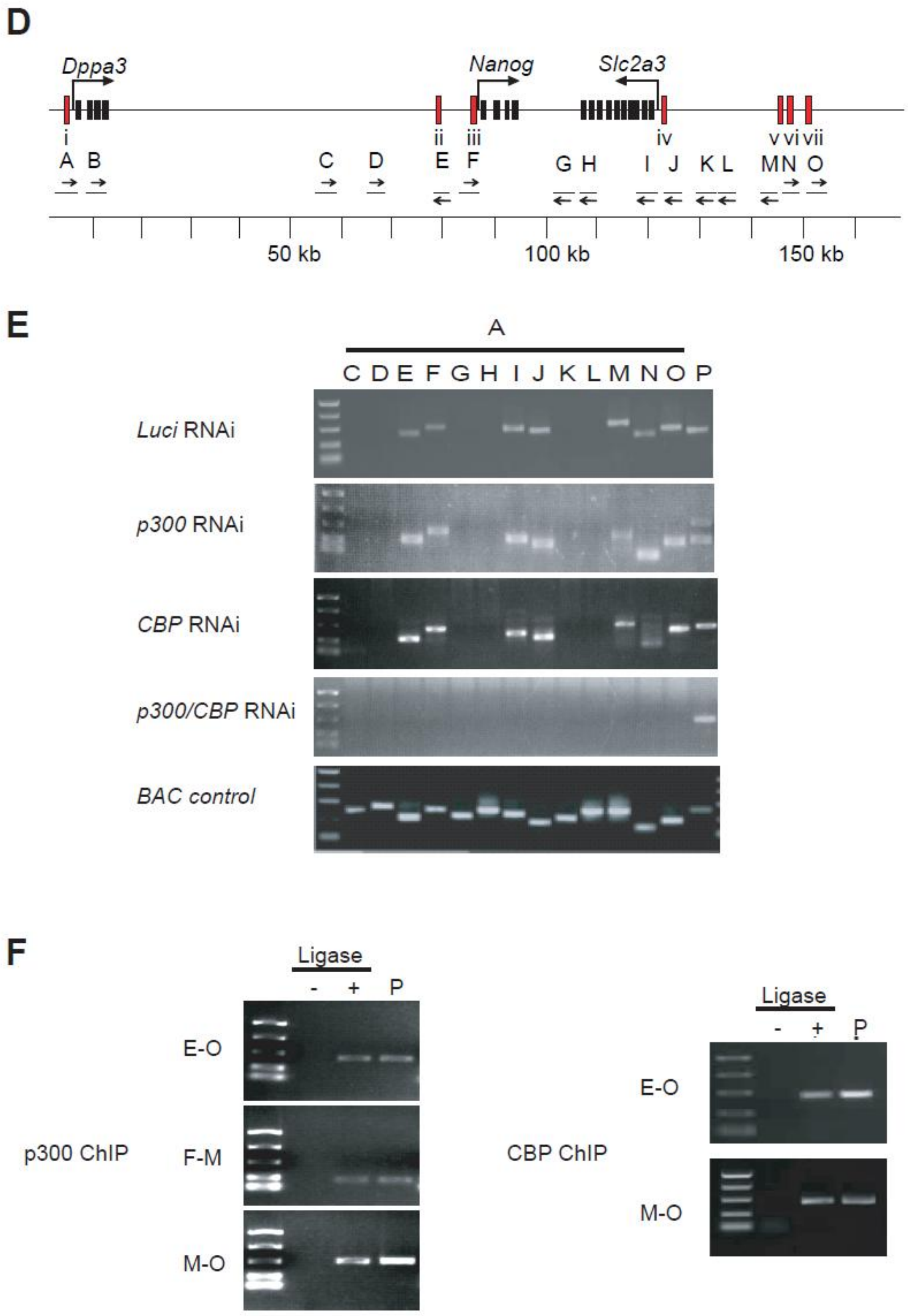


**Figure 4.10. p300 and CBP mediate intergenic looping interactions among colocalization loci**. (**A**) ChIP-seq binding profiles for Nanog, Oct4 and p300 at the *Dppa3-Nanog-Slc2a3* gene loci are shown. (**B-C**) ChIP assays were performed using anti-p300 (**B**) or anti-CBP(**C**) antibody with extracts from ES cells transfected with control RNAi construct, *p300* RNAi construct, *CBP* RNAi construct or *p300/CBP* RNAi construct. (**D**) Schematic representation of the murine *Dppa3-Nanog-Slc2a3* loci. Dark boxes represent exons and red boxes represent ChIP-Seq identified Nanog binding loci, named by the Roman

numerals indicated below them. Relevant BglII restriction fragments are indicated by short green horizontal bars. Primers are named by Roman alphabets and their orientations are indicated by arrows. (**E**) 3C analyses of *Dppa3-Nanog-Slc2a3* loci on murine ES cells transfected with control *Luc* RNAi construct (top panel), p300 RNAi construct(second panel), CBP RNAi (third panel) and p300/CBP RNAi (fourth panel). Bottom panel, PCR controls done using BAC DNA harboring the *Dppa3-Nanog-Slc2a3* loci. (**F**) ChIP-3C analyses done on murine ES cells using p300 antibody (Left panel) and CBP antibody (right panel). Presence or absence of amplicons is detected in a 1.5% agarose gel using the primer combinations indicated on the left of each panel. Lane P is a BAC control and the leftmost lanes represent PCR markers to show sizes of the amplicons. DNA samples used in the PCR reactions were prepared with or without ligase added in the 3C assays to show that presence of amplicons is ligase-dependent.

We found that the four fragments bound by Nanog and p300, one located 2.1 kb upstream of the transcriptional start site (TSS) of *Dppa3*, one located 5.1 kb upstream of the TSS of *Nanog*, one located 2.5 kb upstream of the TSS of *Slc2a3*, and the last one located 30 kb upstream of the TSS of *Slc2a3*, can interact with each other by forming chromosome loops, excluding the intervening regions (Figure 4.10E). Again, such long-range interaction is ES cell-specific as the loops were abolished upon RA-induced differentiation (Figure 4.11) and are dependent on the presence of either p300 or CBP as proved by ChIP-3C and 3C on RNAi samples (Figure 4.10E, F).

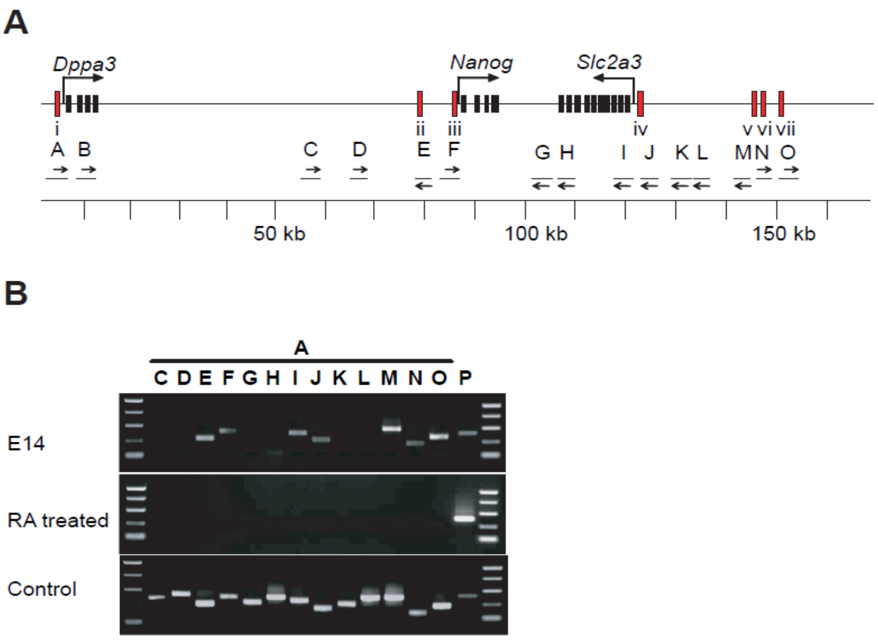


**Figure 4.11. Intergenic looping interactions are specific to the pluripotent state.** (A) Schematic representation of the mouse *Dppa3-Nanog-Slc2a3* loci. (B) 3C analyses of *Dppa3-Nanog-Slc2a3* loci. Top panel, 3C analyses on mouse E14 ES cells. Second panel, 3C analyses on RA treated ES cells. Bottom panel, PCR controls done using BAC DNA harboring the *Dppa3-Nanog-Slc2a3* loci.

**Chromatin looping structure is evolutionarily conserved in human ES cells.**

To access whether similar intragenic looping interactions occur in human ES cells, interacting loci in the *TCF7L1* gene, the ortholog of the murine *Tcf3* gene, were predicted by converting the position of the relevant genomic locus from the mouse assembly to that from the human assembly through the Convert function in the UCSC Genome Browser website (Figure 4.12A). When ChIP and 3C assays were then performed on human ES cells, the predicted interacting loci were confirmed to be bound by Nanog and p300 *in vivo* (Figure 4.12B) and that the regions encompassing the binding loci formed chromosomal loops with each other (Figure 4.12C), indicating that the intragenic loop structures are evolutionarily conserved in mouse and human. Furthermore, this was also verified for intergenic loci. (Figure 4.12D-F). Additionally, we also found that these interactions are ES cell-specific as the looping interactions are not present in HEK293T cells, a differentiated human cell line (Figure 4.12C, F, bottom panel). Interestingly, the human DPPA3-NANOG-SLC2A14 locus is located on human chromosome 12p, a region clustered with pluripotency genes, which has a distinctively central nuclear localization in ES cells but peripheral nuclear localization in differentiated cells27, probably correlating with its unique looping structure described here.

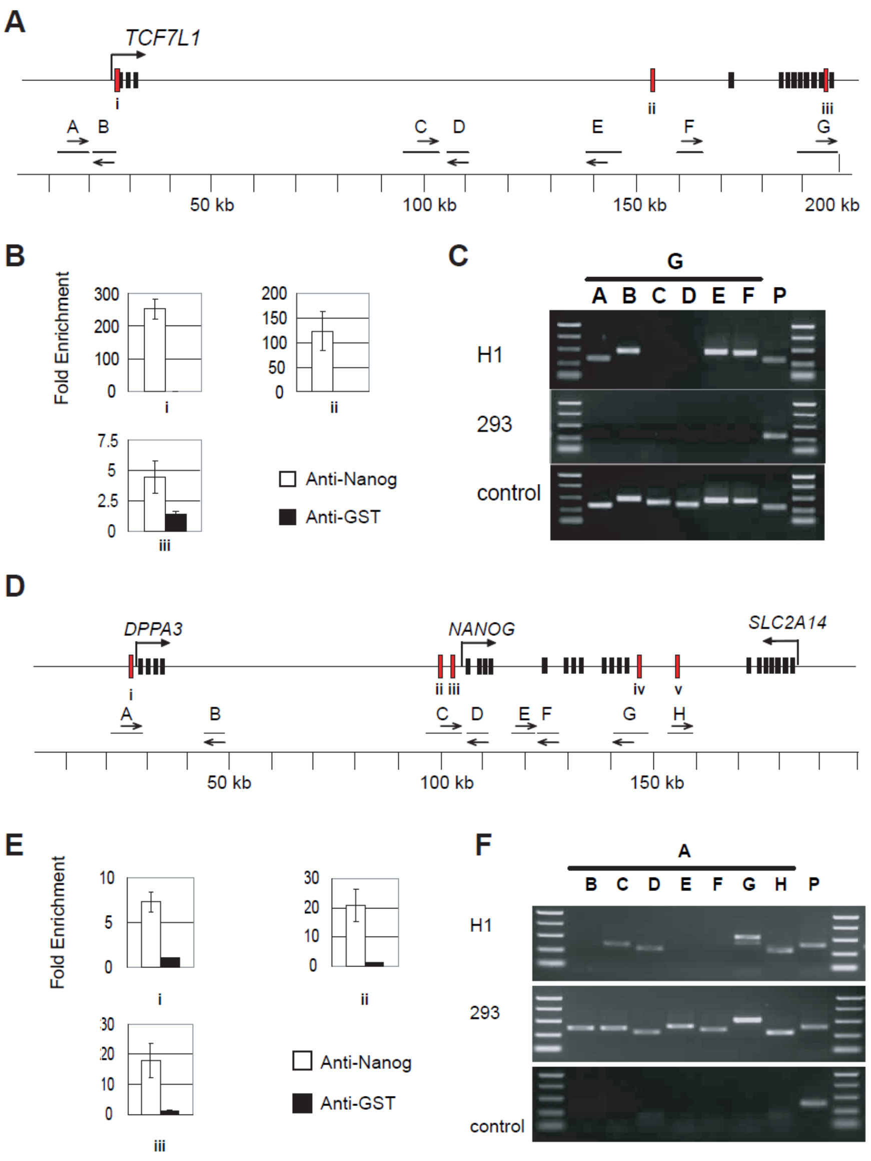


**Figure 4.12. The intragenic and intergenic looping interactions are conserved in human ES cells.** (A, D) Schematic representation of the human (A) *TCF7L1* locus and (D) *DPPA3-NANOG-SLC2A14* loci. Labels are as in (Figure 8, 10D). (B, E) ChIP analyses on Nanog binding loci shown in (B) *TCF7L1* locus and (E) *DPPA3-NANOG-SLC2A14* loci using Nanog antibody (anti-Nanog, white) and GST antibody (anti-GST, black) as a mock ChIP control. Labels are as in (Figure 8, 10E). (C, F) 3C analyses on the (C) *TCF7L1* locus and (F) *DPPA3-NANOG-SLC2A14* loci. Presence or absence of amplicons is detected in a 1.5% agarose gel using the variable primers indicated on top of each lane with the invariant primer G. Top panels, 3C analyses done on human ES cells. Middle panels, 3C analyses done on HEK293T cells (293 cells). Bottom panels (BAC control), 3C analyses done using BAC DNA harboring the relevant genomic regions to show that the all possible ligation products can be amplified using the indicated primer combinations.

### DNA fragments involved in looping interactions have enhancer activities.

Our 3C and ChIP-3C data begs the question of whether such higher order chromatin structures have functional relevance. We hypothesized that the chromosomal loops may help to bring cis-regulatory elements such as enhancer regions in contact with promoters at the active gene loci. To test this, we sought to characterize the genomic DNA associated with the anchor of the loop. Since looping formation has been previously shown to mediate long range communications between enhancers and promoters(Tolhuis et al., 2002), we embarked on testing whether regions around the binding loci on the active loci have enhancer activities. As enhancers are defined as positively-regulating elements that are position-independent, we cloned fragments of about 400-500bp encompassing each binding locus on the active gene loci downstream of a luciferase reporter driven by a *Pou5f1* minimal promoter (Figure 4.13A). The reporters were co-transfected with either p300/CBP double knock down construct or empty vector construct into murine ES cells, which were then assayed for luciferase activity. Interestingly, we found that fragments encompassing the Nanog promoter, the fourth Nanog/p300 binding locus on *Slc2a3*, and the second and third Nanog/p300 binding loci on *Tcf3* have enhancer activities, as shown by the higher luciferase activity relative to a luciferase reporter containing the *Pou5f1* minimal promoter only (Figure 4.13B). The data also confirmed our previous identification of a functional enhancer about 5 kb upstream of Nanog28. Importantly, the enhancer activities are p300/CBP dependent as knocking down p300 and CBP resulted in a significant decrease in the enhancer activities (Figure 4.13B).

To further investigate the functions of chromatin looping interactions in transcription activation *in vivo*, we took advantage of bacterial artificial chromosome (BAC), which encompasses the sequences that are involved in the formation chromatin looping. Human BAC clone RP11-277J24, which contains the genomic fragment from human chromosome12 spanning the *DPPA3-NANOG* loci, was modified by *galK* positive and counterselection system.

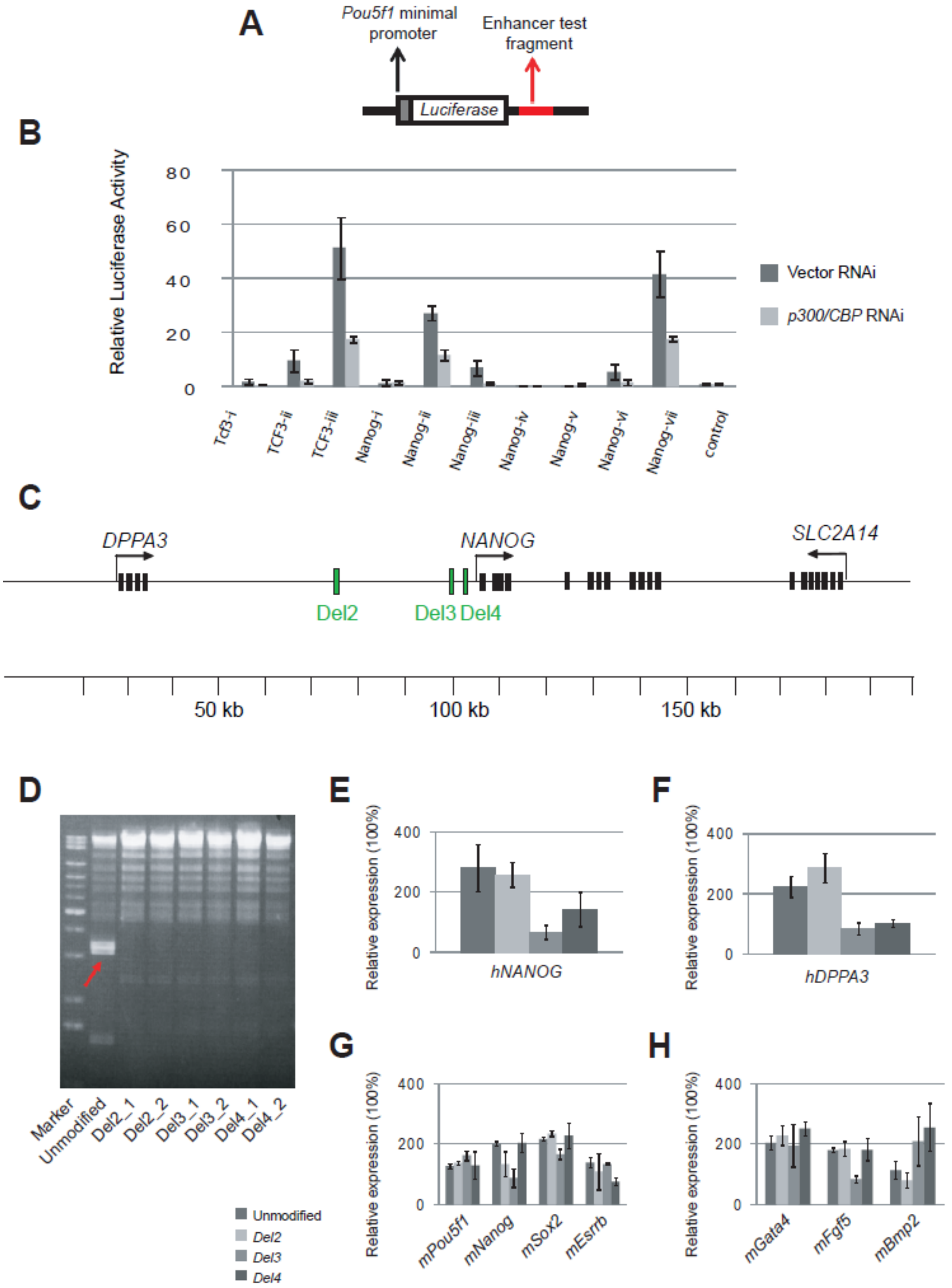


**Figure 4.13. Characterization of the DNA fragments involved in looping interactions.** (A) Reporter constructs used to assay for enhancer activity are shown. Genomic fragment of approximately 300 bp (in red) was inserted downstream (Luc-Nanog Enh) of a luciferase gene driven by Oct4 minimal promoter. (B) Luciferase reporter assay. Murine ES cells were co-transfected with a luciferase reporter plasmid and either a p300/CBP knockdown construct (p300/CBP RNAi) or an empty vector (Vector RNAi) and then subjected to luciferase reporter assay analysis. Y-axis represents the fold enrichment of luciferase activity, calculated relative to a luciferase reporter containing the Oct4 minimal promoter only and normalized over an internal transfection control. X-axis represents the Nanog and p300 binding fragments (labeled as in Fig 5, 6) cloned into the luciferase reporter plasmid. Error bars represent

standard deviations obtained from 3 repeats. (C) Schematic representation of the human DPPA3-NANOG-SLC2A14 loci. Dark boxes represent exons and green boxes represent deleted regions, as indicated below them. (D) ClaI, PmeI, XhoI triple digestion of BAC DNA. Digested BAC DNA is detected in a 1.5% agarose gel. (E-G) Quantitative real-time PCR analysis of expression of (E) human NANOG gene (hNANOG),(F) mouse ES cell-associated genes and (G) mouse lineage-specific marker genes in stable mouse ES cell lines with the insertion of specific modified BAC clones. The levels of the transcripts were normalized against control empty vector transfection. Data are presented as the mean ± SEM and derived from 3 independent experiments.

Three deletions have been introduced to RP11-277J24 specifically. Del1 is located in the middle of *DPPA3* and *NANOG* gene and not involved in chromatin looping based on our results, it was used as a control modification. Del2 and Del3 are upstream of *NANOG* gene and bound by both Nanog and p300. They have been shown to interact with the other fragments and form chromatin loops (Figure 4.13C). To assess whether the BACs were correctly modified, we performed detailed restriction mapping of the original BAC, RP11-277J24 as well as the three modified BACs, Del2, Del3 and Del4, using digests with ClaI, PmeI and XhoI. Two clones for each modification had the desired deletion (Figure 4.13D). Then we transfected the three modified human BAC clones into murine ES cells and establish stable cell lines with entire BAC clones. The expression of mouse genes was not much changed in the BAC tranfected stable cell lines (Figure 4.13F, G). However, the expression of human *NANOG* and *DPPA3* expression was significantly reduced in the cells transfected with Del2 and Del3 compared with control cells that transfected with Del1 (Figure 4.13E). These results demonstrated that deletion of interacting fragments that are involved in chromatin looping does affect the gene expression of related genes.

## 4.5 DISCUSSION

Self-renewal and pluripotency are the key characteristics through which the proliferation and function of ES cells are maintained. Our data demonstrate that two paralogous proteins which act as coactivators in the cells, p300 and CBP, are required for optimal support of ES cell maintenance, as ES cell lost its identity if lacking both p300 and CBP. In addition, p300 and CBP are shown to play overlapping functions in ES cells, probably due to the high protein homology shared by them. Introducing either p300 or CBP into p300/CBP double knock down ES cells to certain dosage is able to retain self-renewal capacity of ES cells. Gene knockout experiments have shown that mice with homozygous p300 or CBP mutations are lethal, even with the normal expression level of the other paralogue, while double heterozygotes p300-/CBP- are also lethal, suggesting either that p300 and CBP have nonoverlapping functions, such that both coactivators are required, or that the total level of CBP and p300 is critical for normal development (Kung et al., 2000; Oike et al., 1999; Yao et al., 1998). Our data, to some extent, support the later notion that the total level of p300 and CBP is essential for ES cell self-renewal, manipulating the overall dosage of p300/CBP is able to control the ES cell identity. However, it is also possible that these two paralogues may play distinct roles for maintaining the pluripotent state of ES cells. The earlier study of p300$^{-/-}$ ES cells has shown that these cells, although having normal expression of CBP and self-renewal capacity, their ability of differentiation is significantly disturbed(Zhong and Jin, 2009). Study of p300 and CBP in hematopoietic stem cells (HSC) has demonstrated their essential but distinct roles in maintaining normal hematopoiesis. CBP, but not p300 is critical, in maintaining the total pool of mouse HSC through self-renewal; whereas p300, but not CBP, appears to contribute to hematopoietic differentiation (Rebel et al., 2002). In our case of ES cells, to better understand whether p300 and CBP play distinct roles to regulate ES cells differentiation, further studies of pluripotency in CBP$^{-/-}$ ES cells or by manipulating the dosage of p300/CBP are remained to be investigated.

As coactivators, p300 and CBP are found to regulate transcription either through endogenous histone acetyltransferase (HAT) activity or through association with transcription factors as well as general transcriptional machinery as adaptor proteins. Domain mapping studies enabled us to discover the functional domains of p300 and CBP, thus understand their molecular mechanisms within the cells. Our domain mapping study has found that KIX domain is the interaction domain of p300/CBP with Nanog. The KIX domain is one of several domains p300/CBP that bind several transcriptional regulators; it is highly conserved in evolution, with 90% identity in human CBP and p300 (Radhakrishnan et al., 1997). It has been shown to have significant functions in mouse development, especially in haematopoiesis system, as in mice homozygous for point mutations in the KIX domain of p300 designed to disrupt the binding surface for the transcription factors c-Myb and CREB7–9, multilineage defects occur in haematopoiesis, including anaemia, B-cell deficiency, thymic hypoplasia, megakaryocytosis and thrombocytosis (Kasper et al., 2002). It will be interesting to see whether mutation of KIX would have any effect on ES cell self-renewal and pluripotency as it disrupts the interaction between p300/CBP with master regulators in ES cells. Further characterization of functional domains of p300/CBP in ES cells has identified that besides KIX domain, HAT domain is indispensable for p300/CBP fulfilling their function as well, suggesting that p300/CBP may regulate transcription through their histone acetyltransferase activity. p300 and CBP are found to be able to acetylate all four core histone *in vitro*, interestingly, they are found to mediate acetylation of histone H3 on lysine 56 recently(Das et al., 2009), which is a newly identified histone modification  marker that overlap strongly with the binding of the master regulators of ES cells(Xie et al., 2009). Altogether, these results strongly indicate that master regulators recruit p300/CBP through contacting KIX domain of them and the intrinsic HAT activity of p300/CBP acetylate histone H3K56 on the local chromatin, resulting in the co-localization of binding sites and H3K56 markers.

Besides the study of transcriptional machinery, understanding the nuclear architecture and higher-order chromatin organization in ES cells is of interest to ES cell research due to the unusual chromatin structure that has been found in ES cells3. For example, chromatin in ES cells has been shown to be marked bivalently by activating and repressive histone modifications at many developmental or lineage specific genes, which is suggested to be a mechanism to silence gene expression while keeping the genes poised for activation during differentiation (Azuara et al., 2006; Bernstein et al., 2006; Giadrossi et al., 2007). In addition, ES cell chromatin has elevated levels of histone modifications associated with gene activity (Lee et al., 2004; Meshorer et al., 2006), and more dynamic interactions with chromatin proteins (Meshorer et al., 2006), pointing towards a unique chromatin state that is relevant to how genes are regulated in ES cells and probably how pluripotency is maintained.

In order to understand complex gene regulation, one key question is how distant regulatory DNA elements communicate with each other. Indeed, recent studies on higher-order chromatin organization in several gene loci have yielded accumulating information on the impact of cell type specific chromatin organizations on gene expression and cell function (Horike et al., 2005; Murrell et al., 2004; Spilianakis and Flavell, 2004; Tolhuis et al., 2002). Several models have been proposed to explain the correct spatial organization of gene expression (Kellum and Schedl, 1991; West et al., 2002). The most widely accepted model of long-range regulatory interactions is the looping model, which proposes that distant enhancers and promoters are in physical contact, while the intervening regions are looped out. The first evidence to support this model is from the study of the chicken beta-globin gene cluster (Choi and Engel, 1988). In this case, all sequences necessary for the efficient transcription of one of the genes in the cluster were found in close proximity, while the inactive regions were pushed aside (Choi and Engel, 1988). The genetic study on the regulation of the homeotic Abdominal-B (Abd-B) gene in Drosophila, as one of the best studied systems, indicates that proper targeting of the Abd-B promoter is most likely to be a result of cooperation among a number of different elements, including promoter targeting sequence (PTS)-like sequences, boundaries, upstream tethering elements, polycomb and trithorax response elements (PREs/TREs), and any other unidentified regulatory units (Kellum and Schedl, 1991). In addition, long-range chromosomal structures within Nanog locus in ES cells have been reported previously by examining the DNaseI hypersensitive sites (HS) using high-throughput quantitative chromatin profiling (QCP) approach. Nevertheless, the study of looping formation has been focused on cis-regulatory elements; information derived from trans-acting factors has only started to be appreciated.

In our study, Nanog and p300 binding loci on *Tcf3* (*TCF7L1*) and *Dppa3-Nanog-Slc2a3* (*DPPA3-NANOG-SLC2A14*) cluster are proven to be involved in long range intragenic and intergenic looping interactions respectively. The interactions are not detectable in differentiated murine ES cells and in a differentiated human cell line. p300 and CBP are further confirmed to be involved in the looping interactions through ChIP-3C and depletion of p300/CBP through RNAi results in abolishment of the interactions. Furthermore, interacting loci are found to bear *in vitro* and *in vivo* enhancer activities which are reduced significantly upon p300/CBP depletion. Our current report on pluripotency-associated looping interactions among enhancers and promoters highlights a new dimension in studying and understanding three-dimensional chromatin architectures. In previous studies, such as those involving the *α*- and *β-globin* loci (Carter et al., 2002; Palstra et al., 2003; Spilianakis and Flavell, 2004; Tolhuis et al., 2002), *$T_H2$* cytokine locus (Spilianakis and Flavell, 2004), and the *PSA* locus (Wang et al., 2005), looping interactions are predicted when *cis*-regulatory elements and the respective genes have been characterized previously, including enhancers, LCRs, and presence of DNase I hypersensitivity sites. Here, we report that looping interactions can be predicted by examining binding locations that are densely occupied by cell type specific transcription factors and coactivators, along individual chromosomes. Such predictions may not only be applied to other cell types in the study of long-range interactions, but also allow for the identification of novel *cis*-regulatory elements and the trans-acting factor that is mediating their functions and organizations.

Furthermore, we found that p300 and CBP are more likely to act as the molecular tethers to stabilize the large protein complex and maintain the higher order structure. Depletion of both p300 and CBP would disrupt the looping structure; even the protein level and binding intensity of Nanog and Oct4 were unchanged. So the model that we are suggesting for this long range looping regulation is that master regulators, such as Nanog, Oct4, recognizes their specific *cis*-regulatory elements and bind to specific loci within a gene or gene cluster; They recruit coactivators p300 and CBP and together with other transcription factors and cofactors, they form a large nucleoprotein complex and draw DNA into a looping structure to bring the self-renewal associated genes, such as *Tcf3*, *Dppa3*, *Nanog*, and *Slc2a3*, together for efficient transcriptional regulation, possibly in a transcription factory (Fraser, 2006). They might be sharing *cis*-regulating elements such as enhancers, activator and coactivator complexes, chromatin-remodeling complexes, and/or the transcription machinery for efficient expression. p300 and CBP are acting as molecular tether connecting different transcription factors and stabilizing this higher-order structure (Figure 4.14). Our finding is consistent with and extends the earlier studies of long range chromatin structure within *Nanog* locus in mouse ES cells. In the earlier study, they have shown that Oct4, as one of the master regulators in ES cells, is essential to maintain the higher order chromatin structure, as the looping structure is abolished in Oct4-

depleted cells. Based on our model, depletion of Oct4 would reduce the recruitment of p300/CBP, thus abolish the looping that tethered by p300/CBP.

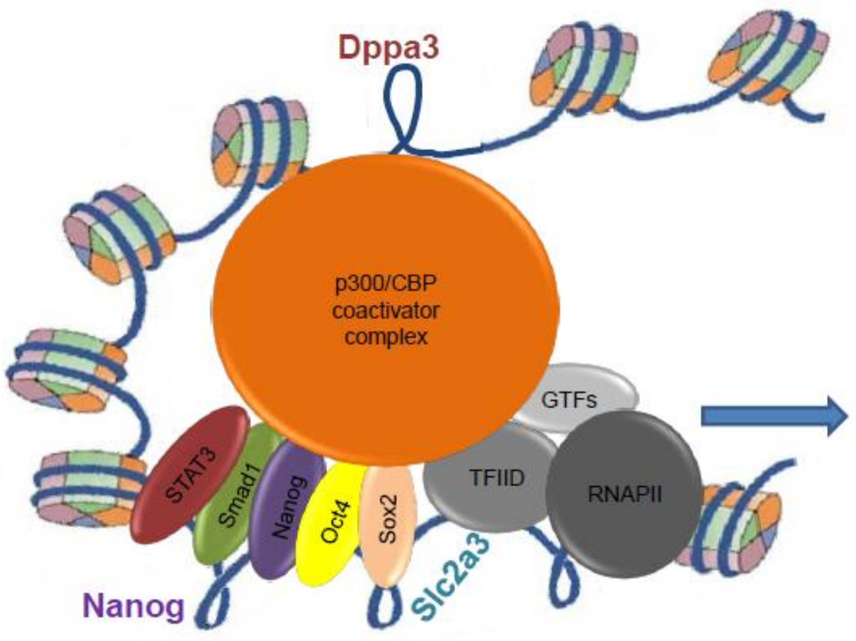


**Figure 4.14. Model showing the three-dimensional organization of Dppa3-Nanog-Slc2a3 loci.** Core regulators (eg., Nanog, Oct4, Sox2) recognize their specific cis-regulatory elements and bind to specific loci within the gene cluster of Dppa3-Nanog-Slc2a . They recruit coactivator complex p300/CBP, and together with other transcription factors and cofactors, they form a large nucleoprotein complex and draw DNA into a loop structure. p300/CBP act as the molecular tether to stabilize the protein complex and also by their intrinsic HAT activity, they may acetylate local chromatin to facilitate transcription response.

In conclusion, this is the first study of *in vivo* higher-order chromatin organization that is unique to pluripotent cells based on the binding sites of transcription factors and coactivators, thereby adding a new player to the list of unusual findings regarding the chromatin structure in ES cells. Another unusual observation is that several enhancers seem to be interacting to coordinate gene expression. Such interactions may represent an important yet poorly understood mechanism whereby crosstalk among enhancers and promoters contribute to proper gene regulation and pluripotency. In view of the recently identified hyperdynamic plasticity of ES cell chromatin (Meshorer et al., 2006), p300 and CBP may have a novel role in maintaining self-renewal and pluripotency by coordinating the chromatin domains into functionally distinct active and repressed domains that can be regulated properly and not subjected to random or unspecific effects from neighboring loci.

# CHAPTER V: Conclusion and Perspectives

Although Oct4, Sox2 and Nanog have long been established as master regulators and forming the core regulatory circuit in the transcriptional network in ES cells and their binding properties have been uncovered by ChIP combined sequencing technology, the large portion of the regulatory mechanism in ES cells is still missing and a few important questions are remained for answers. What are the real essential targets of master regulators? What are the functional partners for master regulators? Why master regulators bind to both activated and repressive loci? Are they activators or repressors? How can they differentiate their functions at different loci? How can they recognize their binding sites?

This thesis work has greatly contributed to dissecting the transcriptional network and extensively expanded our knowledge of the network in mouse ES cells by introducing novel self-renewal and pluripotency associated transcription factors into the known core regulatory circuit. Understanding the molecular function of the novel factors and their interplay with the established master regulators in the network should illuminate fundamental properties of ES cells and shed light on the understanding of the process of directed differentiation and cellular reprogramming, thus ultimately lead to precise manipulation and realization of the full clinical therapeutic benefits of these unique cells.

Furthermore, my research has combined computational and statistical tools with system biology to understand cellular differentiation process. Although microarray experiments are able to provide unprecedented quantities of genome-wide data on gene-expression patterns and has been extensively developed and applied in many biological contexts, the management and analysis of the millions of data points that result from these experiments is still under development. The interpretation of the results from the analysis and its implication on biological questions are of more importance to biologists. Chapter III of my research has opened the possibility of identifying the early differentiation markers during ES cell differentiation based on statistical analysis of gene expression data. Besides *Smacard1*, which we have experimentally validated in our research, other candidate genes in the list derived from our statistical analysis, have also been identified and validated as critical regulators governing ES cell maintenance by other groups (Guo and Smith, 2010; Walker et al., 2010; Zhang et al., 2010). Their work further support the robust and reliability of our statistical algorithm. This work highlights the importance and necessity of combining novel and sophistical computational tools with genome-wide biological studies in advancing and understanding early cellular differentiation. As the price for microarray and sequencing continues dropping, a comprehensive study of total RNA transcripts may further developed and become popular (Wang et al., 2009). Application of in-depth bioinformatics analysis to these genome-wide data will provide more thorough insights into the dynamics of gene expression in a biological system (Pepke et al., 2009).

In addition to the functional study of novel transcription factors, I looked into coactivators that facilitate the functions of transcription factors and further linked coactivator regulation to higher-order chromatin structure. This is the first study of *in vivo* higher-order chromatin organization that is unique to pluripotent cells based on the binding sites of transcription factors and coactivators, adding a new content to the list of unusual findings regarding the chromatin structure in ES cells as well as a new layer to the ES cell specific transcriptional network. However, the higher-order looping structures are not only presented at activated loci, how they are stabilized and mediated at repressive loci by transcription factors and cofactors remain to be elucidated.

Future work is needed in the following specific directions:

1. In order to construct a more comprehensive transcriptional network in ES cells, putative novel transcription regulators indicated in our statistical analysis need to be identified and functionally characterized. On the other hand, to get a better understanding of cellular reprogramming, our statistical model can be applied to analyze the dynamics of time point gene expression during reprogramming process as well. It will be helpful to identify novel critical regulators of reprogramming and deepen our understanding of transcriptional network governing pluripotency from a reverse perspective.

2. It would be interesting to study the dynamics of gene expression profile with the dynamics of binding affinity and binding profiles by essential transcription factors during cellular differentiation or reprogramming. A statistical model capturing these two dynamics would be instructive to understand how a specific transcription factor is functioning *in vivo* to control the cell identity.

3. Besides p300 and CBP, it is important to study other cofactors, including coactivators and corepressors, as embryos display severe defects with deletion of several cofactors. A proposed model is that although transcription factors bind to both activated and repressive loci, it is coactivators and corepressors recruited by transcription factors that mediate gene activation or repression, rather than the transcription factors. However, how coactivators or corepressors are recruited at specific loci would be another interesting question and it is possible that it may be mediated by their protein binding partners of transcription factors. This idea is supported by the study of Eset gene, which is shown to interact with Oct4 and restrict extraembryonic trophoblast lineage potential (Yuan et al., 2009).

4. Chimera formation experiments by injecting an engineered ES cells with deletion of regulatory sequence mediating long-range looping structure should be extremely helpful to understand the *in vivo* function of higher-order chromatin structure during embryogenesis and development.

5. It would also be interesting to study the higher-order chromatin structure at repressive loci. Combining the binding profile data of transcription factors, particularly transcriptional repressors as well as the cis-regulatory sequence, would shed light on how specific lineage is restricted and how we can manipulate it for directed differentiation.

# BIBLIOGRAPHY

Aubert, J., H. Dunstan, et al. (2002). "Functional gene screening in embryonic stem cells implicates Wnt antagonism in neural differentiation." Nat Biotechnol 20(12): 1240-1245.

Avilion, A. A., S. K. Nicolis, et al. (2003). "Multipotent cell lineages in early mouse development depend on SOX2 function." Genes Dev 17(1): 126-140.

Azuara, V., P. Perry, et al. (2006). "Chromatin signatures of pluripotent cell lines." Nat Cell Biol 8(5): 532-538.

Bernstein, B. E., T. S. Mikkelsen, et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." Cell 125(2): 315-326.

Berstine, E. G., M. L. Hooper, et al. (1973). "Alkaline phosphatase activity in mouse teratoma." Proc Natl Acad Sci U S A 70(12): 3899-3903.

Blobel, G. A. (2000). "CREB-binding protein and p300: molecular integrators of hematopoietic transcription." Blood 95(3): 745-755.

Boiani, M. and H. R. Scholer (2005). "Regulatory networks in embryo-derived pluripotent stem cells." Nat Rev Mol Cell Biol 6(11): 872-884.

Boyer, L. A., T. I. Lee, et al. (2005). "Core transcriptional regulatory circuitry in human embryonic stem cells." Cell 122(6): 947-956.

Boyer, L. A., C. Logie, et al. (2000). "Functional delineation of three groups of the ATP-dependent family of chromatin remodeling enzymes." J Biol Chem 275(25): 18864-18870.

Boyer, L. A., K. Plath, et al. (2006). "Polycomb complexes repress developmental regulators in murine embryonic stem cells." Nature 441(7091): 349-353.

Bradley, A., M. Evans, et al. (1984). "Formation of germ-line chimaeras from embryo-derived teratocarcinoma cell lines." Nature 309(5965): 255-256.

Brandenberger, R., H. Wei, et al. (2004). "Transcriptome characterization elucidates signaling networks that control human ES cell growth and differentiation." Nat Biotechnol 22(6): 707-716.

Briggs, R. and T. J. King (1952). "Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs." Proc Natl Acad Sci U S A 38(5): 455-463.

Brons, I. G., L. E. Smithers, et al. (2007). "Derivation of pluripotent epiblast stem cells from mammalian embryos." Nature 448(7150): 191-195.

Bultman, S., T. Gebuhr, et al. (2000). "A Brg1 null mutation in the mouse reveals functional differences among mammalian SWI/SNF complexes." Mol Cell 6(6): 1287-1295.

Burdon, T., A. Smith, et al. (2002). "Signalling, cell cycle and pluripotency in embryonic stem cells." Trends Cell Biol 12(9): 432-438.

Burdon, T., C. Stracey, et al. (1999). "Suppression of SHP-2 and ERK signalling promotes self-renewal of mouse embryonic stem cells." Dev Biol 210(1): 30-43.

Card, D. A., P. B. Hebbar, et al. (2008). "Oct4/Sox2-regulated miR-302 targets cyclin D1 in human embryonic stem cells." Mol Cell Biol 28(20): 6426-6438.

Carter, D., L. Chakalova, et al. (2002). "Long-range chromatin regulatory interactions in vivo." 32(4): 623-626.

Cartwright, P., C. McLean, et al. (2005). "LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism." Development 132(5): 885-896.

Casella G, B. R. (2002). Statistical Inference: Duxbury.

Chambers, I., D. Colby, et al. (2003). "Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells." Cell 113(5): 643-655.

Chambers, I., J. Silva, et al. (2007). "Nanog safeguards pluripotency and mediates germline development." Nature 450(7173): 1230-1234.

Chambers, I. and A. Smith (2004). "Self-renewal of teratocarcinoma and embryonic stem cells." Oncogene 23(43): 7150-7160.

Chan, H. M. and N. B. La Thangue (2001). "p300/CBP proteins: HATs for transcriptional bridges and scaffolds." J Cell Sci 114(Pt 13): 2363-2373.

Chen, X., H. Xu, et al. (2008). "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." Cell 133(6): 1106-1117.

Chew, J. L., Y. H. Loh, et al. (2005). "Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells." Mol Cell Biol 25(14): 6031-6046.

Chi, A. S. and B. E. Bernstein (2009). "Developmental biology. Pluripotent chromatin state." Science 323(5911): 220-221.

Chickarmane, V., C. Troein, et al. (2006). "Transcriptional dynamics of the embryonic stem cell switch." PLoS Comput Biol 2(9): e123.

Choi, O. R. and J. D. Engel (1988). "Developmental regulation of beta-globin gene switching." Cell 55(1): 17-26.

Chou, Y. F., H. H. Chen, et al. (2008). "The growth factor environment defines distinct pluripotent ground states in novel blastocyst-derived stem cells." Cell 135(3): 449-461.

Clodfelder-Miller, B., P. De Sarno, et al. (2005). "Physiological and pathological changes in glucose regulate brain Akt and glycogen synthase kinase-3." J Biol Chem 280(48): 39723-39731.

Conlon, E. M., X. S. Liu, et al. (2003). "Integrating regulatory motif discovery and genome-wide expression analysis." Proc Natl Acad Sci U S A 100(6): 3339-3344.

Cope, L. M., R. A. Irizarry, et al. (2004). "A benchmark for Affymetrix GeneChip expression measures." Bioinformatics 20(3): 323-331.

Creyghton, M. P., S. Markoulaki, et al. (2008). "H2AZ is enriched at polycomb complex target genes in ES cells and is necessary for lineage commitment." Cell 135(4): 649-661.

Dalton, S. (2009). "Exposing hidden dimensions of embryonic stem cell cycle control." Cell Stem Cell 4(1): 9-10.

Das, C., M. S. Lucia, et al. (2009). "CBP/p300-mediated acetylation of histone H3 on lysine 56." Nature 459(7243): 113-117.

Davis, R. L., H. Weintraub, et al. (1987). "Expression of a single transfected cDNA converts fibroblasts to myoblasts." Cell 51(6): 987-1000.

Dejosez, M., J. S. Krumenacker, et al. (2008). "Ronin is essential for embryogenesis and the pluripotency of mouse embryonic stem cells." Cell 133(7): 1162-1174.

Dekker, J. (2006). "The three 'C' s of chromosome conformation capture: controls, controls, controls." Nat Methods 3(1): 17-21.

Dekker, J., K. Rippe, et al. (2002). "Capturing chromosome conformation." Science 295(5558): 1306-1311.

Dietrich, J. E. and T. Hiiragi (2007). "Stochastic patterning in the mouse pre-implantation embryo." Development 134(23): 4219-4231.

Dimos, J. T., K. T. Rodolfa, et al. (2008). "Induced pluripotent stem cells generated from patients with ALS can be differentiated into motor neurons." Science 321(5893): 1218-1221.

Doble, B. W. and J. R. Woodgett (2003). "GSK-3: tricks of the trade for a multi-tasking kinase." J Cell Sci 116(Pt 7): 1175-1186.

Doetschman, T. C., H. Eistetter, et al. (1985). "The in vitro development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium." J Embryol Exp Morphol 87: 27-45.

Durbin, B. P., J. S. Hardin, et al. (2002). "A variance-stabilizing transformation for gene-expression microarray data." Bioinformatics 18 Suppl 1: S105-110.

Ebert, A. D., J. Yu, et al. (2009). "Induced pluripotent stem cells from a spinal muscular atrophy patient." Nature 457(7227): 277-280.

Elkon, R., C. Linhart, et al. (2003). "Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells." Genome Res 13(5): 773-780.

Elling, U., C. Klasen, et al. (2006). "Murine inner cell mass-derived lineages depend on Sall4 function." Proc Natl Acad Sci U S A 103(44): 16319-16324.

Evans, M. J. and M. H. Kaufman (1981). "Establishment in culture of pluripotential cells from mouse embryos." Nature 292(5819): 154-156.

Fair, J. H., B. A. Cairns, et al. (2003). "Induction of hepatic differentiation in embryonic stem cells by co-culture with embryonic cardiac mesoderm." Surgery 134(2): 189-196.

Fazzio, T. G., J. T. Huff, et al. (2008). "An RNAi screen of chromatin proteins identifies Tip60-p400 as a regulator of embryonic stem cell identity." Cell 134(1): 162-174.

Filipczyk, A. A., A. L. Laslett, et al. (2007). "Differentiation is coupled to changes in the cell cycle regulatory apparatus of human embryonic stem cells." Stem Cell Res 1(1): 45-60.

Finch, B. W. and B. Ephrussi (1967). "Retention of multiple developmental potentialities by cells of a mouse testicular teratocarcinoma during prolonged culture in vitro and their extinction upon hybridization with cells of permanent lines." Proc Natl Acad Sci U S A 57(3): 615-621.

Fraser, P. (2006). "Transcriptional control thrown for a loop." Curr Opin Genet Dev 16(5): 490-495.

Galan-Caridad, J. M., S. Harel, et al. (2007). "Zfx controls the self-renewal of embryonic and hematopoietic stem cells." Cell 129(2): 345-357.

Gangaraju, V. K. and H. Lin (2009). "MicroRNAs: key regulators of stem cells." Nat Rev Mol Cell Biol 10(2): 116-125.

Gaspar-Maia, A., A. Alajem, et al. (2009). "Chd1 regulates open chromatin and pluripotency of embryonic stem cells." Nature 460(7257): 863-868.

Geijsen, N., M. Horoschak, et al. (2004). "Derivation of embryonic germ cells and male gametes from embryonic stem cells." Nature 427(6970): 148-154.

Giadrossi, S., M. Dvorkina, et al. (2007). "Chromatin organization and differentiation in embryonic stem cell models." Curr Opin Genet Dev 17(2): 132-138.

Goodman, R. H. and S. Smolik (2000). "CBP/p300 in cell growth, transformation, and development." Genes Dev 14(13): 1553-1577.

Grossman, C. E., Y. Qian, et al. (2004). "ZNF143 mediates basal and tissue-specific expression of human transaldolase." J Biol Chem 279(13): 12190-12205.

Gu, P., D. LeMenuet, et al. (2005). "Orphan nuclear receptor GCNF is required for the repression of pluripotency genes during retinoic acid-induced embryonic stem cell differentiation." Mol Cell Biol 25(19): 8507-8519.

Guo, G. and A. Smith (2010). "A genome-wide screen in EpiSCs identifies Nr5a nuclear receptors as potent inducers of ground state pluripotency." Development 137(19): 3185-3192.

Haegele, L., B. Ingold, et al. (2003). "Wnt signalling inhibits neural differentiation of embryonic stem cells by controlling bone morphogenetic protein expression." Mol Cell Neurosci 24(3): 696-708.

Herrera, R. E., F. Chen, et al. (1996). "Increased histone H1 phosphorylation and relaxed chromatin structure in Rb-deficient fibroblasts." Proc Natl Acad Sci U S A 93(21): 11510-11515.

Ho, L., J. L. Ronan, et al. (2009). "An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency." Proc Natl Acad Sci U S A 106(13): 5181-5186.

Horike, S., S. Cai, et al. (2005). "Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome." Nat Genet 37(1): 31-40.

Houbaviy, H. B., M. F. Murray, et al. (2003). "Embryonic stem cell-specific MicroRNAs." Dev Cell 5(2): 351-358.

Huang, S., A. A. Yeo, et al. (2004). "At what scale should microarray data be analyzed?" Am J Pharmacogenomics 4(2): 129-139.

Huber, W., A. von Heydebreck, et al. (2003). "Parameter estimation for the calibration and variance stabilization of microarray data." Statistical Applications in Genetics and Molecular Biology 2(1): 3.

Irizarry, R. A., S. L. Ooi, et al. (2003). "Use of mixture models in a microarray-based screening procedure for detecting differentially represented yeast mutants." Stat Appl Genet Mol Biol 2: Article1.

Ishiguchi, H., H. Izumi, et al. (2004). "ZNF143 activates gene expression in response to DNA damage and binds to cisplatin-modified DNA." Int J Cancer 111(6): 900-909.

Ivanova, N., R. Dobrin, et al. (2006). "Dissecting self-renewal in stem cells with RNA interference." Nature 442(7102): 533-538.

Jaenisch, R. and R. Young (2008). "Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming." Cell 132(4): 567-582.

Jensen, S. T. and J. S. Liu (2004). "BioOptimizer: a Bayesian scoring function approach to motif discovery." Bioinformatics 20(10): 1557-1564.

Jiang, J., Y. S. Chan, et al. (2008). "A core Klf circuitry regulates self-renewal of embryonic stem cells." Nat Cell Biol 10(3): 353-360.

Kaji, K., I. M. Caballero, et al. (2006). "The NuRD component Mbd3 is required for pluripotency of embryonic stem cells." Nat Cell Biol 8(3): 285-292.

Kanellopoulou, C., S. A. Muljo, et al. (2005). "Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing." Genes Dev 19(4): 489-501.

Kang, L., J. Wang, et al. (2009). "iPS cells can support full-term development of tetraploid blastocyst-complemented embryos." Cell Stem Cell 5(2): 135-138.

Kasper, L. H., F. Boussouar, et al. (2002). "A transcription-factor-binding surface of coactivator p300 is required for haematopoiesis." Nature 419(6908): 738-743.

Kawasaki, H., R. Eckner, et al. (1998). "Distinct roles of the co-activators p300 and CBP in retinoic-acid-induced F9-cell differentiation." Nature 393(6682): 284-289.

Keirstead, H. S., G. Nistor, et al. (2005). "Human embryonic stem cell-derived oligodendrocyte progenitor cell transplants remyelinate and restore locomotion after spinal cord injury." J Neurosci 25(19): 4694-4705.

Keller, G. (2005). "Embryonic stem cell differentiation: emergence of a new era in biology and medicine." Genes Dev 19(10): 1129-1155.

Kellum, R. and P. Schedl (1991). "A position-effect assay for boundaries of higher order chromosomal domains." Cell 64(5): 941-950.

Kielman, M. F., M. Rindapaa, et al. (2002). "Apc modulates embryonic stem-cell differentiation by controlling the dosage of beta-catenin signaling." Nat Genet 32(4): 594-605.

King, T. J. and R. Briggs (1955). "Changes in the Nuclei of Differentiating Gastrula Cells, as Demonstrated by Nuclear Transplantation." Proc Natl Acad Sci U S A 41(5): 321-325.

Kloosterman, W. P. and R. H. Plasterk (2006). "The diverse functions of microRNAs in animal development and disease." Dev Cell 11(4): 441-450.

Knoepfler, P. S. (2008). "Why myc? An unexpected ingredient in the stem cell cocktail." Cell Stem Cell 2(1): 18-21.

Kopp, J. L., B. D. Ormsbee, et al. (2008). "Small increases in the level of Sox2 trigger the differentiation of mouse embryonic stem cells." Stem Cells 26(4): 903-911.

Kraus, W. L., E. T. Manning, et al. (1999). "Biochemical analysis of distinct activation functions in p300 that enhance transcription initiation with chromatin templates." Mol Cell Biol 19(12): 8123-8135.

Kunath, T., M. K. Saba-El-Leil, et al. (2007). "FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment." Development 134(16): 2895-2902.

Kung, A. L., V. I. Rebel, et al. (2000). "Gene dose-dependent control of hematopoiesis and hematologic tumor suppression by CBP." Genes Dev 14(3): 272-277.

Kuroda, T., M. Tada, et al. (2005). "Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression." Mol Cell Biol 25(6): 2475-2485.

Laiosa, C. V., M. Stadtfeld, et al. (2006). "Determinants of lymphoid-myeloid lineage diversification." Annu Rev Immunol 24: 705-738.

Lamba, D. A., J. Gust, et al. (2009). "Transplantation of human embryonic stem cell-derived photoreceptors restores some visual function in Crx-deficient mice." Cell Stem Cell 4(1): 73-79.

Lee, J. H., S. R. Hart, et al. (2004). "Histone deacetylase activity is required for embryonic stem cell differentiation." Genesis 38(1): 32-38.

Lee, T. I., R. G. Jenner, et al. (2006). "Control of developmental regulators by Polycomb in human embryonic stem cells." Cell 125(2): 301-313.

Lessard, J., J. I. Wu, et al. (2007). "An essential switch in subunit composition of a chromatin remodeling complex during neural development." Neuron 55(2): 201-215.

Li, C. and W. H. Wong (2003). DNA-Chip Analyzer (dChip). The Analysis of Gene Expression Data: An Overview of Methods and Software, Springer.

Liang, J., M. Wan, et al. (2008). "Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells." Nat Cell Biol 10(6): 731-739.

Lim, L. S., Y. H. Loh, et al. (2007). "Zic3 is required for maintenance of pluripotency in embryonic stem cells." Mol Biol Cell 18(4): 1348-1358.

Lin, T., C. Chao, et al. (2005). "p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression." Nat Cell Biol 7(2): 165-171.

Ling, J. Q., T. Li, et al. (2006). "CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1." Science 312(5771): 269-272.

Loebel, D. A., C. M. Watson, et al. (2003). "Lineage choice and differentiation in mouse embryos and embryonic stem cells." Dev Biol 264(1): 1-14.

Loh, Y. H., Q. Wu, et al. (2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." Nat Genet 38(4): 431-440.

Loh, Y. H., W. Zhang, et al. (2007). "Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells." Genes Dev 21(20): 2545-2557.

Lomvardas, S., G. Barnea, et al. (2006). "Interchromosomal interactions and olfactory receptor choice." Cell 126(2): 403-413.

Lowell, S., A. Benchoua, et al. (2006). "Notch promotes neural lineage entry by pluripotent embryonic stem cells." PLoS Biol 4(5): e121.

Marson, A., S. S. Levine, et al. (2008). "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells." Cell 134(3): 521-533.

Martin, G. R. (1981). "Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells." Proc Natl Acad Sci U S A 78(12): 7634-7638.

Maruyama, M., T. Ichisaka, et al. (2005). "Differential roles for Sox15 and Sox2 in transcriptional control in mouse embryonic stem cells." J Biol Chem 280(26): 24371-24379.

Masui, S., Y. Nakatake, et al. (2007). "Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells." Nat Cell Biol 9(6): 625-635.

Matsuda, T., T. Nakamura, et al. (1999). "STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells." Embo J 18(15): 4261-4269.

McLaren, A. (2000). "Cloning: pathways to a pluripotent future." Science 288(5472): 1775-1780.

Meshorer, E., D. Yellajoshula, et al. (2006). "Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells." Dev Cell 10(1): 105-116.

Miele, A., N. Gheldof, et al. (2006). "Mapping chromatin interactions by chromosome conformation capture." Curr Protoc Mol Biol Chapter 21: Unit 21 11.

Miller, R. A. and F. H. Ruddle (1976). "Pluripotent teratocarcinoma-thymus somatic cell hybrids." Cell 9(1): 45-55.

Miller, R. A. and F. H. Ruddle (1977). "Properties of teratocarcinoma-thymus somatic cell hybrids." Somatic Cell Genet 3(3): 247-261.

Mitsui, K., Y. Tokuzawa, et al. (2003). "The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells." Cell 113(5): 631-642.

Murchison, E. P., J. F. Partridge, et al. (2005). "Characterization of Dicer-deficient murine embryonic stem cells." Proc Natl Acad Sci U S A 102(34): 12135-12140.

Murrell, A., S. Heeson, et al. (2004). "Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops." Nat Genet 36(8): 889-893.

Myslinski, E., M. A. Gerard, et al. (2006). "A genome scale location analysis of human Staf/ZNF143-binding sites suggests a widespread role for human Staf/ZNF143 in mammalian promoters." J Biol Chem 281(52): 39953-39962.

Ng, H. H., F. Robert, et al. (2003). "Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity." Mol Cell 11(3): 709-719.

Nichols, J., B. Zevnik, et al. (1998). "Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4." Cell 95(3): 379-391.

Niwa, H., T. Burdon, et al. (1998). "Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3." Genes Dev 12(13): 2048-2060.

Niwa, H., J. Miyazaki, et al. (2000). "Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells." Nat Genet 24(4): 372-376.

Niwa, H., Y. Toyooka, et al. (2005). "Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation." Cell 123(5): 917-929.

Ogawa, S., Y. Tagawa, et al. (2005). "Crucial roles of mesodermal cell lineages in a murine embryonic stem cell-derived in vitro liver organogenesis system." Stem Cells 23(7): 903-913.

Ohtsuka, S. and S. Dalton (2008). "Molecular and biological properties of pluripotent embryonic stem cells." Gene Ther 15(2): 74-81.

Oike, Y., N. Takakura, et al. (1999). "Mice homozygous for a truncated form of CREB-binding protein exhibit defects in hematopoiesis and vasculo-angiogenesis." Blood 93(9): 2771-2779.

Okazaki, N., S. Ikeda, et al. (2008). "The novel protein complex with SMARCAD1/KIAA1122 binds to the vicinity of TSS." J Mol Biol 382(2): 257-265.

Orford, K. W. and D. T. Scadden (2008). "Deconstructing stem cell self-renewal: genetic insights into cell-cycle regulation." Nat Rev Genet 9(2): 115-128.

Palmieri, S. L., W. Peter, et al. (1994). "Oct-4 transcription factor is differentially expressed in the mouse embryo during establishment of the first two extraembryonic cell lineages involved in implantation." Dev Biol 166(1): 259-267.

Palstra, R. J., B. Tolhuis, et al. (2003). "The beta-globin nuclear compartment in development and erythroid differentiation." Nat Genet 35(2): 190-194.

Pan, G., J. Li, et al. (2006). "A negative feedback loop of transcription factors that controls stem cell pluripotency and self-renewal." Faseb J 20(10): 1730-1732.

Papaioannou, V. E., M. W. McBurney, et al. (1975). "Fate of teratocarcinoma cells injected into early mouse embryos." Nature 258(5530): 70-73.

Park, I. H., N. Arora, et al. (2008). "Disease-specific induced pluripotent stem cells." Cell 134(5): 877-886.

Pasini, D., A. P. Bracken, et al. (2007). "The polycomb group protein Suz12 is required for embryonic stem cell differentiation." Mol Cell Biol 27(10): 3769-3779.

Pepke, S., B. Wold, et al. (2009). "Computation for ChIP-seq and RNA-seq studies." Nat Methods 6(11 Suppl): S22-32.

Pereira, L., F. Yi, et al. (2006). "Repression of Nanog gene transcription by Tcf3 limits embryonic stem cell self-renewal." Mol Cell Biol 26(20): 7479-7491.

Radhakrishnan, I., G. C. Perez-Alvarado, et al. (1997). "Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions." Cell 91(6): 741-752.

Rajan, P., D. M. Panchision, et al. (2003). "BMPs signal alternately through a SMAD or FRAP-STAT pathway to regulate fate choice in CNS stem cells." J Cell Biol 161(5): 911-921.

Ramalho-Santos, M., S. Yoon, et al. (2002). ""Stemness": transcriptional profiling of embryonic and adult stem cells." Science 298(5593): 597-600.

Raz, R., C. K. Lee, et al. (1999). "Essential role of STAT3 for embryonic stem cell pluripotency." Proc Natl Acad Sci U S A 96(6): 2846-2851.

Rebel, V. I., A. L. Kung, et al. (2002). "Distinct roles for CREB-binding protein and p300 in hematopoietic stem cell self-renewal." Proc Natl Acad Sci U S A 99(23): 14789-14794.

Reubinoff, B. E., M. F. Pera, et al. (2000). "Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro." Nat Biotechnol 18(4): 399-404.

Reuss, B., R. Dono, et al. (2003). "Functions of fibroblast growth factor (FGF)-2 and FGF-5 in astroglial differentiation and blood-brain barrier permeability: evidence from mouse mutants." J Neurosci 23(16): 6404-6412.

Reynolds, A., D. Leake, et al. (2004). "Rational siRNA design for RNA interference." Nat Biotechnol 22(3): 326-330.

Robbins, J., J. Gulick, et al. (1990). "Mouse embryonic stem cells express the cardiac myosin heavy chain genes during development in vitro." J Biol Chem 265(20): 11905-11909.

Rocke, D. M. and B. Durbin (2001). "A model for measurement error for gene expression arrays." J Comput Biol 8(6): 557-569.

Rodda, D. J., J. L. Chew, et al. (2005). "Transcriptional regulation of nanog by OCT4 and SOX2." J Biol Chem 280(26): 24731-24737.

Rosner, M. H., M. A. Vigano, et al. (1990). "A POU-domain transcription factor in early stem cells and germ cells of the mammalian embryo." Nature 345(6277): 686-692.

Sakaki-Yumoto, M., C. Kobayashi, et al. (2006). "The murine homolog of SALL4, a causative gene in Okihiro syndrome, is essential for embryonic stem cell proliferation, and cooperates with Sall1 in anorectal, heart, brain and kidney development." Development 133(15): 3005-3013.

Sato, N., L. Meijer, et al. (2004). "Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor." Nat Med 10(1): 55-63.

Schaub, M., A. Krol, et al. (1999). "Flexible zinc finger requirement for binding of the transcriptional activator staf to U6 small nuclear RNA and tRNA(Sec) promoters." J Biol Chem 274(34): 24241-24249.

Schaub, M., E. Myslinski, et al. (1999). "Maximization of selenocysteine tRNA and U6 small nuclear RNA transcriptional activation achieved by flexible utilization of a Staf zinc finger." J Biol Chem 274(35): 25042-25050.

Schaub, M., E. Myslinski, et al. (1997). "Staf, a promiscuous activator for enhanced transcription by RNA polymerases II and III." Embo J 16(1): 173-181.

Schmitt, T. M., R. F. de Pooter, et al. (2004). "Induction of T cell development and establishment of T cell competence from embryonic stem cells differentiated in vitro." Nat Immunol 5(4): 410-417.

Scholer, H. R., S. Ruppert, et al. (1990). "New type of POU domain in germ line-specific protein Oct-4." Nature 344(6265): 435-439.

Schoor, M., K. Schuster-Gossler, et al. (1999). "Skeletal dysplasias, growth retardation, reduced postnatal survival, and impaired fertility in mice lacking the SNF2/SWI2 family member ETL1." Mech Dev 85(1-2): 73-83.

Schuster, C., A. Krol, et al. (1998). "Two distinct domains in Staf to selectively activate small nuclear RNA-type and mRNA promoters." Mol Cell Biol 18(5): 2650-2658.

Schuster, C., E. Myslinski, et al. (1995). "Staf, a novel zinc finger protein that activates the RNA polymerase III promoter of the selenocysteine tRNA gene." Embo J 14(15): 3777-3787.

Sharp, J., J. Frame, et al. (2010). "Human embryonic stem cell-derived oligodendrocyte progenitor cell transplants improve recovery after cervical spinal cord injury." Stem Cells 28(1): 152-163.

Smith, L. J. and S. Benchimol (1988). "Introduction of new genetic material into human myeloid leukemic blast stem cells by retroviral infection." Mol Cell Biol 8(2): 974-977.

Spilianakis, C. G. and R. A. Flavell (2004). "Long-range intrachromosomal interactions in the T helper type 2 cytokine locus." Nat Immunol 5(10): 1017-1027.

Spilianakis, C. G., M. D. Lalioti, et al. (2005). "Interchromosomal associations between alternatively expressed loci." Nature 435(7042): 637-645.

Stavridis, M. P., J. S. Lunn, et al. (2007). "A discrete period of FGF-induced Erk1/2 signalling is required for vertebrate neural specification." Development 134(16): 2889-2894.

Stead, E., J. White, et al. (2002). "Pluripotent cell division cycles are driven by ectopic Cdk2, cyclin A/E and E2F activities." Oncogene 21(54): 8320-8333.

Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proc Natl Acad Sci U S A 100(16): 9440-9445.

Suh, M. R., Y. Lee, et al. (2004). "Human embryonic stem cells express a unique set of microRNAs." Dev Biol 270(2): 488-498.

Surani, M. A., K. Hayashi, et al. (2007). "Genetic and epigenetic regulators of pluripotency." Cell 128(4): 747-762.

Suzuki, A., A. Raya, et al. (2006). "Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells." Proc Natl Acad Sci U S A 103(27): 10294-10299.

Takahashi, K. and S. Yamanaka (2006). "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." Cell 126(4): 663-676.

Tanigaki, K., H. Han, et al. (2002). "Notch-RBP-J signaling is involved in cell fate determination of marginal zone B cells." Nat Immunol 3(5): 443-450.

Tay, Y. M., W. L. Tam, et al. (2008). "MicroRNA-134 modulates the differentiation of mouse embryonic stem cells, where it causes post-transcriptional attenuation of Nanog and LRH1." Stem Cells 26(1): 17-29.

Thomas, K. R. and M. R. Capecchi (1986). "Introduction of homologous DNA sequences into mammalian cells induces mutations in the cognate gene." Nature 324(6092): 34-38.

Thomson, J. A., J. Itskovitz-Eldor, et al. (1998). "Embryonic stem cell lines derived from human blastocysts." Science 282(5391): 1145-1147.

Thomson, J. A. and V. S. Marshall (1998). "Primate embryonic stem cells." Curr Top Dev Biol 38: 133-165.

Tolhuis, B., R. J. Palstra, et al. (2002). "Looping and interaction between hypersensitive sites in the active beta-globin locus." Mol Cell 10(6): 1453-1465.

Ui-Tei, K., Y. Naito, et al. (2004). "Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference." Nucleic Acids Res 32(3): 936-948.

Viswanathan, S. R., G. Q. Daley, et al. (2008). "Selective blockade of microRNA processing by Lin28." Science 320(5872): 97-100.

Walker, E., W. Y. Chang, et al. (2010). "Polycomb-like 2 associates with PRC2 and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation." Cell Stem Cell 6(2): 153-166.

Wang, G. G., C. D. Allis, et al. (2007). "Chromatin remodeling and cancer, Part II: ATP-dependent chromatin remodeling." Trends Mol Med 13(9): 373-380.

Wang, J., S. Rao, et al. (2006). "A protein interaction network for pluripotency of embryonic stem cells." Nature 444(7117): 364-368.

Wang, J. and A. Wynshaw-Boris (2004). "The canonical Wnt pathway in early mammalian embryogenesis and stem cell maintenance/differentiation." Curr Opin Genet Dev 14(5): 533-539.

Wang, Q., J. S. Carroll, et al. (2005). "Spatial and temporal recruitment of androgen receptor and its coactivators involves chromosomal looping and polymerase tracking." Mol Cell 19(5): 631-642.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet 10(1): 57-63.

Warming, S., N. Costantino, et al. (2005). "Simple and highly efficient BAC recombineering using galK selection." Nucleic Acids Res 33(4): e36.

Wells, J. M. and D. A. Melton (1999). "Vertebrate endoderm development." Annu Rev Cell Dev Biol 15: 393-410.

West, A. G., M. Gaszner, et al. (2002). "Insulators: many functions, many mechanisms." Genes Dev 16(3): 271-288.

White, J. and S. Dalton (2005). "Cell cycle control of embryonic stem cells." Stem Cell Rev 1(2): 131-138.

Williams, R. L., D. J. Hilton, et al. (1988). "Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells." Nature 336(6200): 684-687.

Wilmut, I., A. E. Schnieke, et al. (1997). "Viable offspring derived from fetal and adult mammalian cells." Nature 385(6619): 810-813.

Wu, D. and W. Pan (2010). "GSK3: a multifaceted kinase in Wnt signaling." Trends Biochem Sci 35(3): 161-168.

Wu, Q., X. Chen, et al. (2006). "Sall4 interacts with Nanog and co-occupies Nanog genomic sites in embryonic stem cells." J Biol Chem 281(34): 24090-24094.

Xie, H., M. Ye, et al. (2004). "Stepwise reprogramming of B cells into macrophages." Cell 117(5): 663-676.

Xie, W., C. Song, et al. (2009). "Histone h3 lysine 56 acetylation is linked to the core transcriptional network in human embryonic stem cells." Mol Cell 33(4): 417-427.

Yan, Z., Z. Wang, et al. (2008). "BAF250B-associated SWI/SNF chromatin-remodeling complex is required to maintain undifferentiated mouse embryonic stem cells." Stem Cells 26(5): 1155-1165.

Yang, P., S. A. Arnold, et al. (2008). "Ciliary neurotrophic factor mediates dopamine D2 receptor-induced CNS neurogenesis in adult mice." J Neurosci 28(9): 2231-2241.

Yao, T. P., S. P. Oh, et al. (1998). "Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300." Cell 93(3): 361-372.

Ying, Q. L., J. Nichols, et al. (2003). "BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3." Cell 115(3): 281-292.

Ying, Q. L., J. Wray, et al. (2008). "The ground state of embryonic stem cell self-renewal." Nature 453(7194): 519-523.

Yoshikawa, T., Y. Piao, et al. (2006). "High-throughput screen for genes predominantly expressed in the ICM of mouse blastocysts by whole mount in situ hybridization." Gene Expr Patterns 6(2): 213-224.

Yu, X., J. Zou, et al. (2008). "Notch signaling activation in human embryonic stem cells is required for embryonic, but not trophoblastic, lineage commitment." Cell Stem Cell 2(5): 461-471.

Yuan, P., J. Han, et al. (2009). "Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells." Genes Dev 23(21): 2507-2520.

Zhang, J., W. L. Tam, et al. (2006). "Sall4 modulates embryonic stem cell pluripotency and early embryonic development by the transcriptional regulation of Pou5f1." Nat Cell Biol 8(10): 1114-1123.

Zhang, X., I. Neganova, et al. (2009). "A role for NANOG in G1 to S transition in human embryonic stem cells through direct binding of CDK6 and CDC25A." J Cell Biol 184(1): 67-82.

Zhang, Z., A. Jones, et al. (2010). "PRC2 Complexes with JARID2, and esPRC2p48 in ES Cells to Modulate ES Cell Pluripotency and Somatic Cell Reprogramming." Stem Cells.

Zhao, R. and G. Q. Daley (2008). "From fibroblasts to iPS cells: induced pluripotency by defined factors." J Cell Biochem 105(4): 949-955.

Zhao, X. Y., W. Li, et al. (2009). "iPS cells produce viable mice through tetraploid complementation." Nature 461(7260): 86-90.

Zhong, X. and Y. Jin (2009). "Critical roles of coactivator p300 in mouse embryonic stem cell differentiation and Nanog expression." J Biol Chem 284(14): 9168-9175.

Zhou, Q., J. Brown, et al. (2008). "In vivo reprogramming of adult pancreatic exocrine cells to beta-cells." Nature 455(7213): 627-632.

Zhou, Q., H. Chipperfield, et al. (2007). "A gene regulatory network in mouse embryonic stem cells." Proceedings of the National Academy of Sciences 104(42): 16438.

Zhou, Q. J., L. X. Xiang, et al. (2007). "In vitro differentiation of hepatic progenitor cells from mouse embryonic stem cells induced by sodium butyrate." J Cell Biochem 100(1): 29-42.

# APPENDIX I: Integration of external signaling pathways with the core transcriptional network through transcription factor colocalization hotspots in embryonic stem cells

This chapter is published as: Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, Loh YH, Yeo HC, Yeo ZX, Narang V, Govindarajan KR, Leong B, Shahab A, Ruan Y, Bourque G, Sung WK, Clarke ND, Wei CL, Ng HH. *Cell.* 2008 Jun 13;133(6):1106-17.

.

**My contribution to this project:**

The profiling of transcription factors in mouse ES cells was lead by Chen Xi and Ng Huck Hui. I was responsible for leading the validation analysis ChIP-Seq results. I worked closely with Chen Xi to validate the binding peak identified by ChIP-Seq using ChIP-qPCR and RNAi-ChIP for the 13 transcription factors. In addition, I also independently validate the predicted binding motif for Tcfcp2l1 dataset..

# ABSTRACT

Transcription factors and their specific interactions with targets are crucial in specifying gene expression programmes. To gain insights into the transcriptional regulatory networks in embryonic stem cells, we use chromatin immunoprecipitation coupled to ultra-high-throughput DNA sequencing (ChIP-seq) to map the locations of thirteen sequence specific transcription factors (Nanog, Oct4, STAT3, Smad1, Sox2, Zfx, c-Myc, n-Myc, Klf4, Esrrb, Tcfcp2l1, E2f1 and CTCF) and two transcription regulators (p300 and Suz12). These factors are known to play different roles in ES cell biology as components of the LIF and BMP signaling pathways, self-renewal regulators and key reprogramming factors. Our study provides new insights into the integration of the signaling pathways to the ES cell-specific transcription circuitries. Intriguingly, we find specific genomic regions extensively targeted by different transcription factors. Collectively, the comprehensive mapping of transcription factor binding sites identifies new features of the transcriptional regulatory networks that define ES cell identity.

# INTRODUCTION

Embryonic stem (ES) cells are derived from early preimplantation embryos and they can be maintained for extended periods in culture through self-renewing division (Evans and Kaufman, 1981; Martin, 1981). These cells are pluripotent as they retain the ability to differentiate into many, and perhaps all, cell lineages. The ability to generate transgenic mouse ES cells through homologous recombination has revolutionized biological research through the creation of genetically altered animals (Thomas and Capecchi, 1986). In addition, human ES cells can potentially serve as an inexhaustible source of cells for the derivation of clinically useful cells for regenerative medicine and cell-based therapy.

Mouse ES cells were first isolated in 1981 from mouse blastocysts (Evans and Kaufman, 1981; Martin, 1981). Maintenance of the self-renewing state of mouse ES cells requires the cytokine leukemia inhibitory factor (LIF). The binding of LIF to its receptor activates STAT3 through phosphorylation (Matsuda et al., 1999; Niwa et al., 1998; Raz et al., 1999). LIF alone is however not sufficient to maintain ES cells as their maintenance requires the presence of fetal calf serum. Bone morphogenetic proteins (BMPs) appear to be a key serum-derived factor that acts in conjunction with LIF to enhance self-renewal and pluripotency of mouse ES cells (Ying et al., 2003). The binding of BMP4 to its receptors triggers the phosphorylation of Smad1 and activates the expression of members of the *Id* (inhibitor of differentiation) gene family. As ES cells overexpressing *Ids* can self-renew in the absence of BMP4, it is proposed that induction of *Id* expression is the critical contribution of the BMP/Smad pathway. Hence, the LIF and BMP signaling pathways play a central role in the maintenance of pluripotent stem cell phenotype.

Besides these signaling pathways, which sense the presence of extrinsic growth factors in the environment, intrinsic factors such as transcription factors (TFs) are also essential for specifying the undifferentiated state of ES cells. Oct4, encoded by *Pou5f1*, is a POU domain-containing transcription factor known to be essential to ES cells and early embryonic development (Boiani and Schöler, 2005; Nichols et al., 1998; Smith, 2001). Oct4 interacts with Sox2 (an HMG-containing transcription factor) and genome wide mapping of OCT4 and SOX2 sites in human ES cells show that they co-target multiple genes (Boyer et al., 2005). The cis-regulatory element in which the Sox2-Oct4 complex is bound consists of neighboring sox (CATTGTA) and oct (ATGCAAAT) elements (Loh et al., 2006). Recent works indicate that Oct4 and Sox2, along with c-Myc and Klf4, are sufficient to reprogram fibroblasts to induced pluripotent stem cells (iPS) which are functionally similar to ES cells (Maherali et al., 2007; Okita et al., 2007; Takahashi and Yamanaka, 2006; Wernig et al., 2007). Hence, these transcription factors can exert a dominant role in reconstructing the transcriptional regulatory network of ES cells. A third well studied transcription factor in ES cells is Nanog. Nanog is a homeodomain-containing transcription factor that can sustain pluripotency in ES cells even in the absence of LIF (Chambers et al., 2003; Mitsui et al., 2003). Other transcriptional regulators are required as well to maintain ES cells. Recent work has begun to identify new components of the transcriptional regulatory network required for the maintenance of pluripotency. Through genetic studies, Esrrb and Zfx have been shown to regulate self-renewal of ES cells (Ivanova et al., 2006; Loh et al., 2006; (Galan-Caridad et al., 2007).

Despite the critical roles of transcriptional regulators in the maintenance of mouse ES cells, detailed knowledge of their *in vivo* targets is lacking. The targets of the downstream effectors of key signaling pathways are poorly studied and the targets of many of the transcription factors in ES cells have not been defined. How the different transcriptional circuitries are integrated is also not clear. Elucidation of the transcriptional regulatory networks that are operating in embryonic stem cells is fundamental to understand the molecular nature of pluripotency, self-renewal and reprogramming.

In this study, we use chromatin immunoprecipitation coupled to massively parallel ultrahigh throughput short tag based sequencing (ChIP-seq) (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007) to map the *in vivo* binding loci for thirteen sequence-specific transcription factors and two transcription co-regulators in living mouse ES cells. Intriguingly, these transcription factors are wired to the ES cell genome in two major ways. The first cluster includes Nanog, Oct4, Sox2, Smad1 and STAT3. The second cluster consists of c-Myc, n-Myc, Zfx and E2f1. The co-activator p300 is predominantly recruited to dense binding loci involving proteins found in the first type of cluster. Our analysis also reveals that highly dense binding loci involving these factors have characteristic features of enhanceosomes. ES cell-specific gene expression is associated with binding of many of the factors studied. Based on these associations between binding and expression, we have constructed a transcriptional regulatory network model that integrates the two key signaling pathways with the intrinsic factors in ES cells.

# MATERIALS AND METHODS

**Cell culture and transfection.** E14 mouse ES cells, cultured under feeder-free conditions were maintained in Dulbecco's Modified Eagle-Medium (DMEM, GIBCO), with 15 % heat-inactivated ES qualified fetal bovine serum (FBS, GIBCO), 0.055 mM β-mercaptoethanol (GIBCO), 2mM L-glutamine, 0.1 mM MEM non-essential amino acid, 5,000 units/ml penicillin/streptomycin and 1,000 units/ml of LIF (Chemicon). 293T cells were cultured in DMEM with 10 % FBS and maintained at 37 $^{o}$C with 5 % $CO_2$. For serum-free cell cultures, feeder-free E14 mouse ES cells were plated onto gelatin-coated plates in ESGRO Complete Basal Medium (Chemicon) supplemented with 10 ng/ml LIF (Chemicon) and 50 ng/ml BMP4 (Sigma). Cells were passaged every 2–3 days using accutase (Chemicon). Dissociated cells were pelleted and the cell pellet was resuspended and replated directly.

*Oct4*, *Sox2* and *Nanog* shRNA constructs were designed as described previously (Loh et al. 2006). Transfection of shRNA was performed using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. For the ChIP assay, 35 µg of plasmids were transfected into ES cells on 150-mm plates. Puromycin (Sigma) selection was introduced 1 d after transfection at 1.0 µg/ml, the cells were cross-linked and harvested 48 h after transfection.

**Luciferase assay.** For the 25 Nanog-Oct4-Sox2 cluster fragments and 8 Myc cluster fragments tested for enhancer activity, the fragments (about 300bp) were amplified from genomic DNA and cloned into BamHI and SalI sites of pGL3-*Pou5f1* pp vector (an *Pou5f1* minimal promoter driving *luciferase*) and sequence-verified. E14 mouse ES cells or 293T cells were transfected with these reporter constructs by Lipofectamine 2000 (Invitrogen) following the manufacturer's protocol. A *Renilla* luciferase plasmid (pRL-SV40 from Promega) was co-transfected as an internal control. Cells were harvested 36 h after transfection and the luciferase activities of the cell lysates were measured using Dual-luciferase Reporter Assay System (Promega).

**ChIP assay.** ChIP assay was carried out as described previously (Loh et al. 2006). Briefly, cells were cross-linked with 1% formaldehyde for 10 min at room temperature and formaldehyde was then inactivated by the addition of 125 mM glycine. Chromatin extracts containing DNA fragments with an average size of 500 bp were immunoprecipitated using antibodies shown in Table S1. Quantitative PCR analyses were performed in real time using the ABI PRISM 7900 sequence detection system and SYBR green master mix. Threshold cycles (Ct) were determined for both immunoprecipitated DNA and known amount of DNA from input sample for different primer pairs. Relative occupancy values (also known as fold enrichments) were calculated by determining the immunoprecipitation efficiency (ratios of the amount of immunoprecipitated DNA to that of the input sample) and were normalized to the level observed at a control region, which was defined as 1.0. All ChIP experiments were repeated at least three times independently. For all the primers used, each gave a single product of the right size, as confirmed by agarose gel electrophoresis and dissociation curve analysis.

**Computational analyses.** To identify the MTL, a list of genomic sites co-bound by any of the 13 TFs was generated. Two binding regions were clustered if their centers were 100 bp apart at most. This clustering procedure was done iteratively to form the largest possible clusters, forming what we call

MTL. ChIP-seq data sets for p300 and Suz12 were also generated to determine where these coregulators are recruited with respect to the TFs profiled. Distances from one coregulator site to the nearest TFBSs were then calculated. Pairs of sites within 50 bp of one another were considered to belong to the same group. We computed the Pearson correlation coefficient for each pair of such colocalization vectors and used it as a similarity measure to cluster these factors. To associate binding site information with gene expression, we computed an association score for each pair of gene and TF based on the relative distance to the TSS of the gene. We then performed k-mean clustering on an association matrix to group the genes with similar TF association. Gene groups by this method were then analyzed with a previously published RA-induced differentiation data set (Ivanova et al., 2006).

Two published sets of gene-expression experiments were used in combination with the ChIP-seq data reported here to obtain a set of genes that are enriched in direct transcriptional targets (Ivanova et al., 2006; Zhou et al., 2007). For a given TF, we scored and ranked each gene based on the number and ‘‘intensity’’ of ChIP-seq-defined binding sites. For a given expression change ranking and a given TF-binding ranking, we used responder analysis to determine the significance of association between binding and expression, as well as to define gene sets that are at least 2-fold enriched in direct targets. Regulatory targets were inferred from the intersection of top-ranked bound genes and top-ranked differentially expressed genes.

# RESULTS

## Mapping of *in vivo* binding sites of 13 transcription factors by ChIP-seq approach

Whole genome binding sites of thirteen sequence-specific transcription factors (TFs) were profiled in mouse ES cells by the ChIP-seq approach (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007). Nanog, Oct4, Sox2, Esrrb and Zfx are known regulators of pluripotency and/or self-renewal. Smad1 and STAT3 are key components of the signaling pathways mediated by BMP and LIF, respectively. Tcfcp2l1 has uncharacterized DNA binding property and function, and is preferentially up-regulated in ES cells. E2F1 is best known for its role in regulating cell cycle progression and has also been shown to associate extensively with promoter regions (Bieda et al., 2006). Klf4 and Myc TFs are reprogramming factors also implicated in maintenance of undifferentiated state of ES cells (Cartwright et al., 2005; Jiang et al., 2008). CTCF is required for transcriptional insulation (Kim et al., 2007). Through mapping the binding sites of these 13 TFs, we seek to investigate the binding behavior of these factors and uncover novel insights into how they are wired in the ES cell genome. Here, chromatin immunoprecipitation (ChIP) with specific antibodies against these TFs was used to enrich the DNA fragments bound by these transcription factors followed by direct ultra high throughput sequencing using Solexa Genome Analyzer platform. Genomic regions defined by multiple overlapping DNA fragments derived from the ChIP enrichments were considered as putative binding sites. We used Monte Carlo simulations to determine the minimal number of overlapping ChIP fragment reads required to distinguish true binding from non-specific randomly expected overlaps. Regions with overlapping ChIP DNA counts higher than the threshold were further filtered by removing peaks that were also found in the negative control (anti-GFP ChIP) liberary.

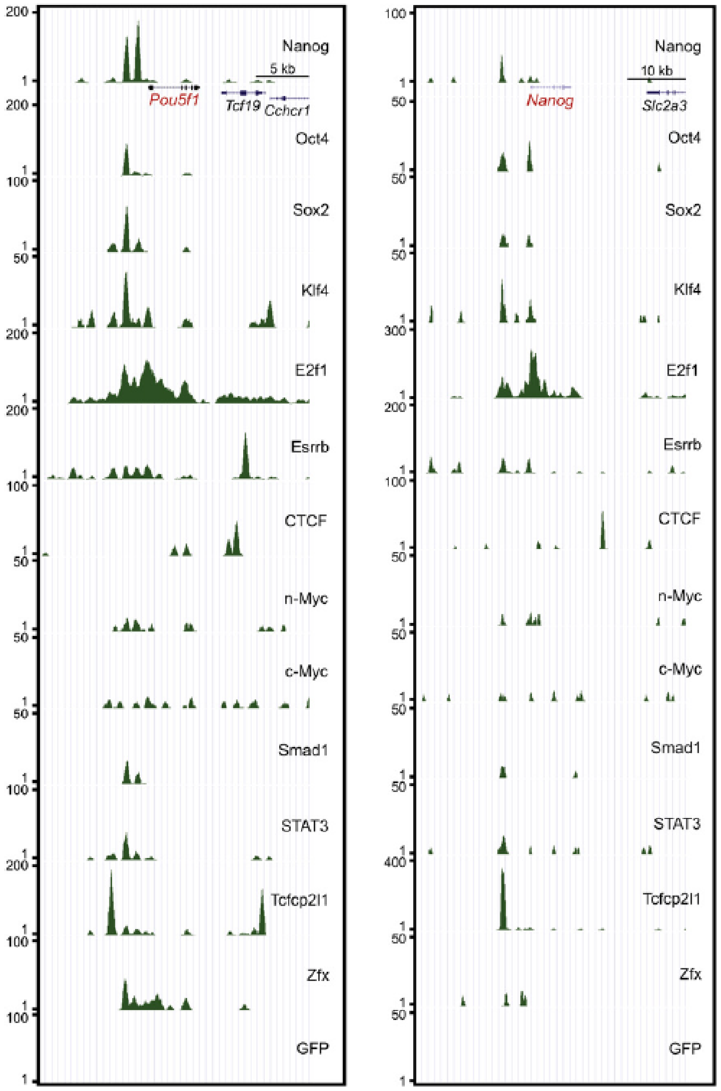


**Figure 1. Genome wide mapping of thirteen factors in ES cells using ChIP-seq technology.** TFBS profiles for the sequence-specific transcription factors and mock ChIP control at the *Pou5f1* and *Nanog* gene loci are shown.

ChIP-qPCR validations were carried out on randomly selected sites with different “intensities” (i.e., ChIP tag counts within the defined overlap region) to further refine the threshold used. Based on the ChIP-qPCR analyses, we determined that the specificity of binding site determination was greater than 95% for the majority of the libraries. We identified between 1,126 to 39,609 transcription factor binding sites (TFBS) for the 13 factors.  As examples, the binding profiles for all 13 factors at the *Pou5f1 (Oct4)* and *Nanog* gene loci are shown in Figure 1.

## Motif analyses of transcription factor binding sites

To determine the *in vivo* sequence specificity of these TFs, we derived the consensus sequence motifs using *de novo* motif discovery algorithm (as described in Loh et al, 2006). Sequences (± 100 bp) from the top 500 binding peaks were selected from each factor, repeats were masked and the program Weeder (Pavesi et al., 2001) was used to find over-represented sequences. Because of the high resolution in defining the binding sites offered by the high sequence depth coverage, over-represented motifs could be uncovered from all of the thirteen factors except E2f1 (Figure 2). Consistent with our previous study, we obtained a sox-oct composite element consisting of a Sox2 binding site consensus (CATTGTT) and canonical Oct4-binding sequence (ATGCAAAT) adjacent to one another from both the Oct4 and Sox2 datasets (Loh et al., 2006). The presence of a common motif suggests that the Sox2 and Oct4 heterodimer is the functional binding unit. Interestingly, the *de novo*-predicted matrices for Nanog and Smad1 bound sequences resemble the sox-oct joint motif. This reflects the frequent co-binding of Nanog and Smad1 with Sox2 and Oct4. It is noteworthy that the Nanog motif reported previously (Loh et al., 2006) can be found using another motif discovery algorithm NestedMICA (Figure S9). The binding consensus sequences identified for Klf4, Esrrb, CTCF, c-Myc, n-Myc, STAT3 and Zfx are closely related to the binding sequences reported previously (Ehret et al., 2001; Galan-Caridad et al., 2007; Kaczynski et al., 2003; Kim et al., 2007; Pettersson et al., 1996; Zeller et al., 2006). Hence, we showed that sequence motifs can be identified from the *in vivo* bound sites.

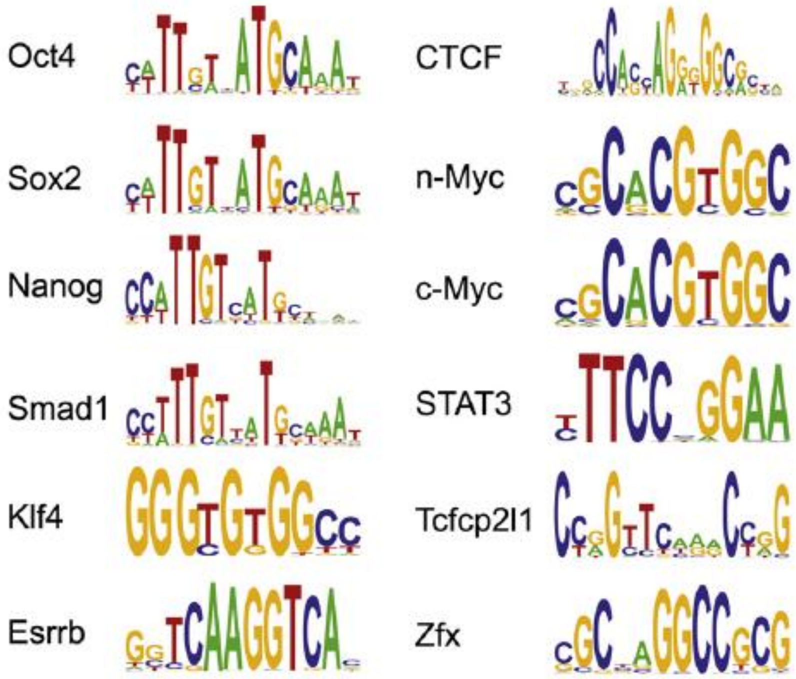


**Figure 2. Identification of enriched motifs using a *de novo* approach.** Matrices predicted by *de novo* motif discovery algorithm Weeder.

### A subset of multiple transcription factor binding loci as ES cell enhanceosomes

Upon close examination of the binding profiles from these thirteen TFs, we found that a subset of binding sites was bound by many of these TFs. To investigate their biological relevance, we first determined the significance of such enrichments of TF binding sites (see Supplemental Experimental Procedures in Supplemental Data). Peak sites within 100 bp were iteratively clustered to define Multiple Transcription factor binding Loci (MTL) (see Table S8). The number of these MTL, plotted as a function of the number of different TFs in the MTL, is shown in Figure 3A. Loci bound by four or more TFs are highly significant (p<0.001, Figure 3A), and there is a total of 3,583 such MTL. Of these, 1,440 loci (40.2 %) were found in the intergenic regions and the remaining loci were spread between promoter regions (1,334 loci, 37.2 %) and within gene regions (809 loci, 22.6%). Less than 20% of the clusters with seven or more TFs are found at promoter regions (yellow columns, Figure 3B) compared with 40% of clusters that have fewer than 5 TFs. Hence, the co-occurrence of TFBS within the MTL is not mainly due to their occurrence at promoters.

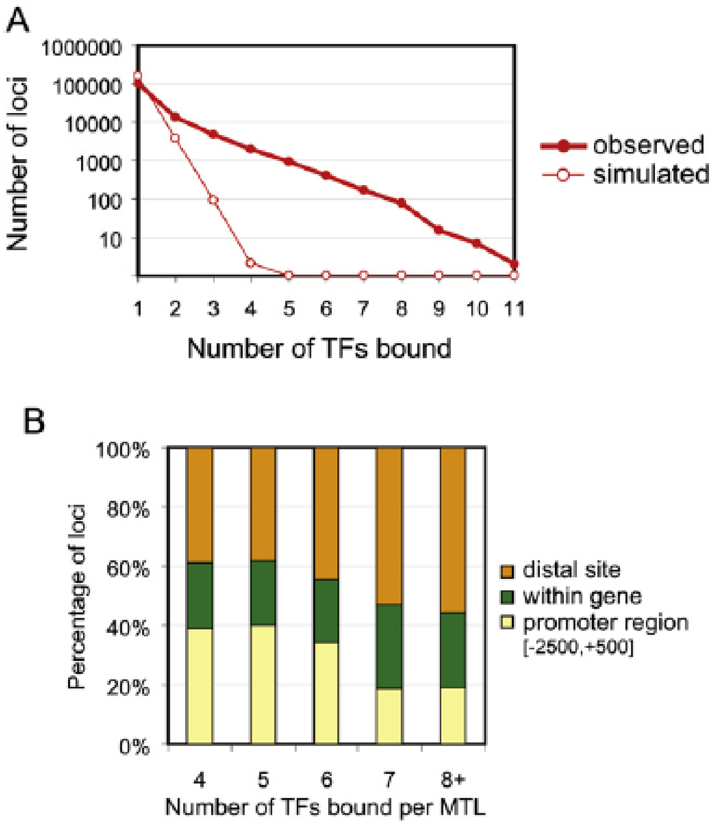


**Figure 3. Multiple Transcription factor binding Loci (MTL).** (A) Plot of the number of TFs bound per co-bound loci. The distribution of randomly occurring co-bound loci is obtained by simulation. (B) Distribution of clusters with different number of co-bound TFs. Promoter regions are defined as sequences 2,500 bp upstream and 500 bp downstream of TSS (Heintzman et al., 2007).

To further dissect the composition of the MTL, we examined the co-occupancy of different factors found in the 3,583 MTL. Among the 13 TFs, Nanog, Sox2, Oct4, Smad1 and STAT3 (blue box, Figure 4A) tend to co-occur quite often, as do members of a second, distinct group comprised of n-Myc, c-Myc, E2f1 and Zfx (green box, Figure 4A). In addition to these two high-level groupings of TFs, we find it useful to define four groups of MTLs based on the presence or absence of binding sites for (i) Oct4, Sox2 or Nanog and (ii) c-Myc or n-Myc. The Nanog-Oct4-Sox2 clusters (binding sites for Nanog, Oct4 or Sox2, but not n-Myc or c-Myc) constitute 43.4% of the 3,583 MTL (orange sector, Figure 4B). The Myc-specific clusters (n-Myc or c-Myc, but not Nanog, Oct4 or Sox2) make up 32.9% of the MTL (light blue sector, Figure 4B).

Consistent with the pair-wise co-occurrence shown in Figure 4A, 87.4% of Smad1 and 56.8% of STAT3 binding sites within MTLs were associated with the Nanog-Oct4-Sox2 specific MTL (orange sector, Figure 4C). This indicates that Smad1 and STAT3 share many common target sites with Nanog, Oct4 and Sox2 and reflects a point of convergence of the two key signaling pathways (Smad1 and STAT3) with the core circuitry defined by Nanog, Oct4 and Sox2 (Boyer et al., 2005). This is consistent with previous study showing the link between Nanog and the LIF pathway (Chambers et al., 2003). 56.9% of Esrrb and 41.9% of Klf4 binding sites within MTLs were found in the

Nanog-Oct4-Sox2 specific MTL. Indeed, Esrrb has been shown to reside in the same complex as Nanog (Wang et al., 2006). In contrast, the co-occurrence of Zfx, CTCF and E2f1 were skewed towards the Myc-specific cluster (light blue sector, Figure 4C).

As the majority of the Nanog-Oct4-Sox2 specific MTL are found outside of promoter regions (91.2%), we assayed genomic sequences from this MTL cluster types for enhancer activity. 25 genomic fragments from the Nanog-Oct4-Sox2 cluster and 8 genomic fragments from the Myc cluster were cloned downstream of a luciferase reporter. The genomic fragment was placed 2 kb away from the minimal *Pou5f1* promoter used to drive the luciferase gene. These constructs were transfected into ES cells and 293T cells and luciferase activity was measured. Remarkably, all 25 constructs with genomic fragments spanning Nanog-Oct4-Sox2 clusters showed robust ES cell-specific enhancer activity (Figure 4D). 21 of the constructs were even more active than a *Nanog* enhancer positive control, which we had characterized previously. In contrast, the control constructs with genomic fragments from the Myc cluster were either not active or showed very weak ES cell-specific enhancer activity.

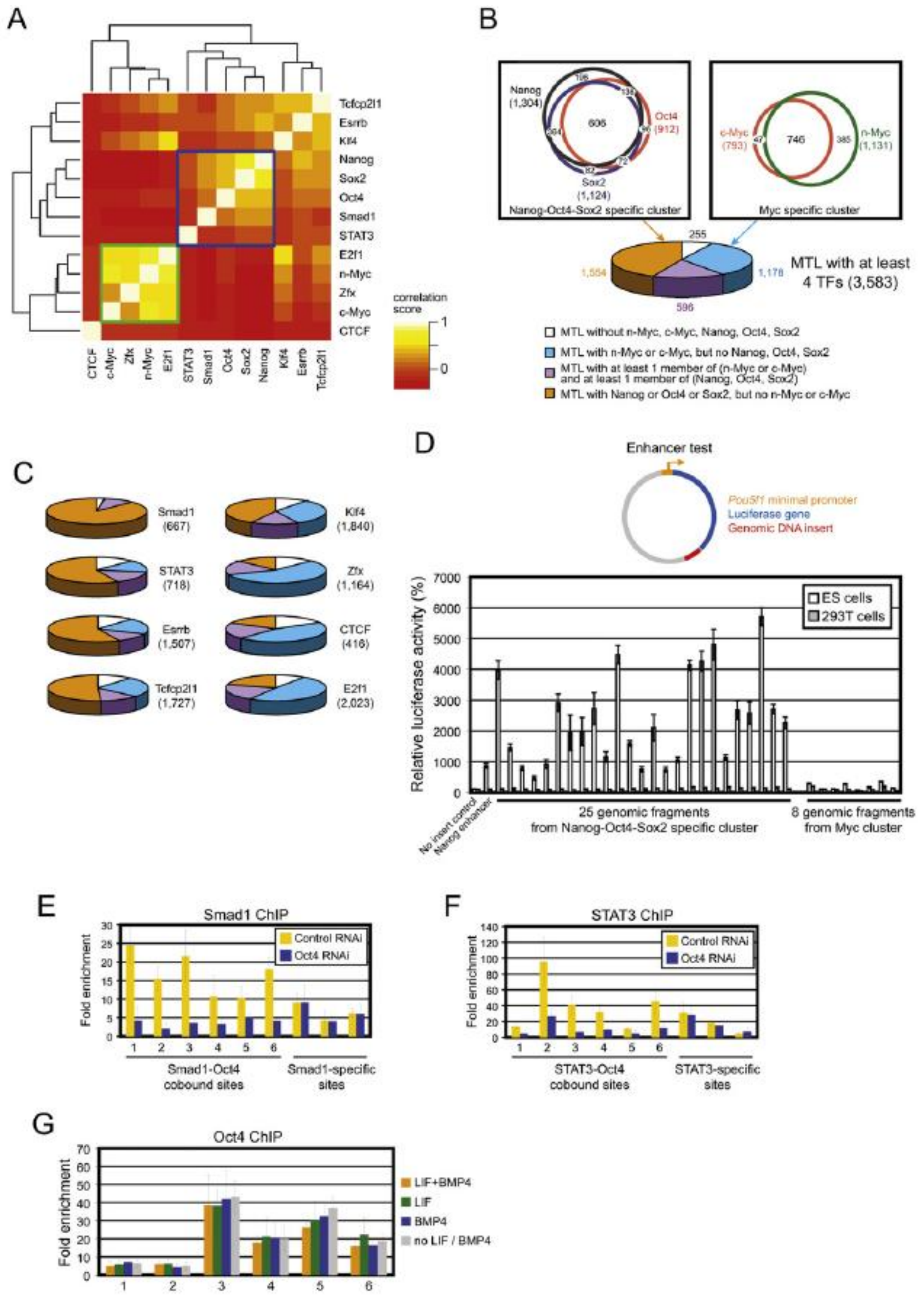


**Figure 4. MTL associated with Nanog, Oct4, Sox2, Smad1 and STAT3 as ES cell enhanceosomes.** (A) Co-occurrence of TF groups within MTL. Colors in the heat map reflect the co-localization frequency of each pair of transcription factors in MTL (yellow means more frequently co-localized, red less). Transcription factors have been clustered along both axes based on the similarity in their co-localization with other factors. (B) Dissection of the transcription factor makeup within MTL. Two major clusters exist within the

3,583 MTL. The first group (orange sector) consists of Oct4, Nanog or Sox2, The second group (light blue sector) consists of n-Myc or c-Myc. The purple sector is a mixture of the first two groups (orange and light blue sectors). (C) The occurrence of the other transcription factors (Smad1, STAT3, Esrrb, Tcfcp2l1, Klf4, Zfx, CTCF and E2f1) within the 3,583 MTL. The color legend is the same as B. (D) Genomic fragments associated with the Nanog-Oct4-Sox2 cluster show enhancer activity. To test for enhancer activity, genomic fragment of approximately 300 bp (shown in red) was cloned downstream of a luciferase reporter (shown in blue) driven by minimal *Pou5f1* promoter (shown in orange). These reporter constructs were transfected into ES cells or 293T cells to determine ES cell-specific enhancer activity. The loci tested for enhancer activity and primers for cloning these genomic fragments are listed in Table S9. Data are presented as the mean ± SEM. (E) Smad1 occupancy is dependent on Oct4. ChIP assays were performed using anti-Smad1 antibody with extracts from ES cells transfected with control RNAi construct (yellow bar) or *Oct4* RNAi construct (blue bar). Coordinates and q-PCR primers of Smad1-Oct4 co-bound sites and Smad1 specific sites are listed in Table S10. Fold enrichment is the relative abundance of DNA fragments detected by q-PCR at the amplified region over a control amplified region. (F) STAT3 occupancy is dependent on Oct4. ChIP assays were performed using anti-STAT3 antibody with extracts from ES cells transfected with control RNAi construct (yellow bar) or *Oct4* RNAi construct (blue bar). (G) Oct4 occupancy is not dependent on LIF and BMP pathways. ChIP assays were performed using anti-Oct4 antibody with extracts from ES cells treated with LIF+BMP4 (orange bar), LIF alone (green bar), BMP4 alone (blue bar) or no LIF and BMP4 (grey bar).

Combinatorial binding of transcription factors to enhancers can impart transcriptional synergy (Struhl, 2001). To address the relationships between Oct4, Smad1 and STAT3, we perturbed the binding of these factors through RNAi or growth factor withdrawal. Depletion of Oct4 led to reduction in Smad1 and STAT3 binding (Figure 4E, F). The alteration of Smad1 and STAT3 binding occurs specifically on Oct4, Smad1, STAT3 co-bound sites and was not due to reduction in Smad1 and STAT3 levels (data not shown). Perturbation of the two signaling pathways however did not affect the binding of Oct4 (Figure 4G). This indicates that Oct4 is pivotal in stabilizing the nucleoprotein complex and establishes a hierarchy of regulatory interactions between Oct4, STAT3 and Smad1. The mechanism for Oct4 dependent

STAT3 and Smad1 binding is not clear. It is possible that Oct4 may interact with STAT3 or Smad1 to facilitate their interactions with chromatin.

In summary, through the global binding sites of transcription factor profiling, we uncovered over three thousand genomic regions densely bound by TFs. The Nanog-Oct4-Sox2 cluster exhibits features of enhanceosomes by enhancing transcription from a distance and shows extensive co-occupancy with Smad1 and STAT3. Importantly, we showed that Oct4 is required for the binding of Smad1 and STAT3, suggesting that Oct4 plays a pivotal role in stabilizing the TF complex.

### p300 is recruited to the Nanog-Oct4-Sox2 cluster

To further assign functionality to the MTL, we determined the locations of transcriptional co-activator p300 using ChIP-seq. p300 is a histone acetyltransferase commonly found at enhancer regions (Heintzman et al., 2007; Ogryzko et al., 1996). Genome-wide mapping of chromatin regulator like p300 has the potential to reveal the DNA binding factor(s) responsible for recruiting the regulator to specific sites in the genome (Birney et al., 2007). We also profiled the locations of another chromatin regulator Suz12, to serve as a control .

Strikingly, p300 was found to co-occur with Nanog-Oct4-Sox2 cluster type (Figure 5A). Most p300 binding sites are associated with three to six other transcription factors, up to as many as nine in one case (Figure 5B). The composition of these p300-containing clusters is highly diverse, but typically they include one or more of the factors Nanog, Oct4 or Sox2 followed, at lower probability, by Smad1, Esrrb, Klf4, Tcfcp2l1, and STAT3 (Figure 5B). In contrast to p300, Suz12 did not show strong association with any of the 13 TFs (data not shown). Using the *de novo* motif discovery algorithm Weeder, we were able to uncover an enriched motif from p300-enriched sequences that resembles the sox-oct composite element (Figure 5C). The association of p300 with Oct4 binding sites was validated for 12 sites using ChIP-qPCR.

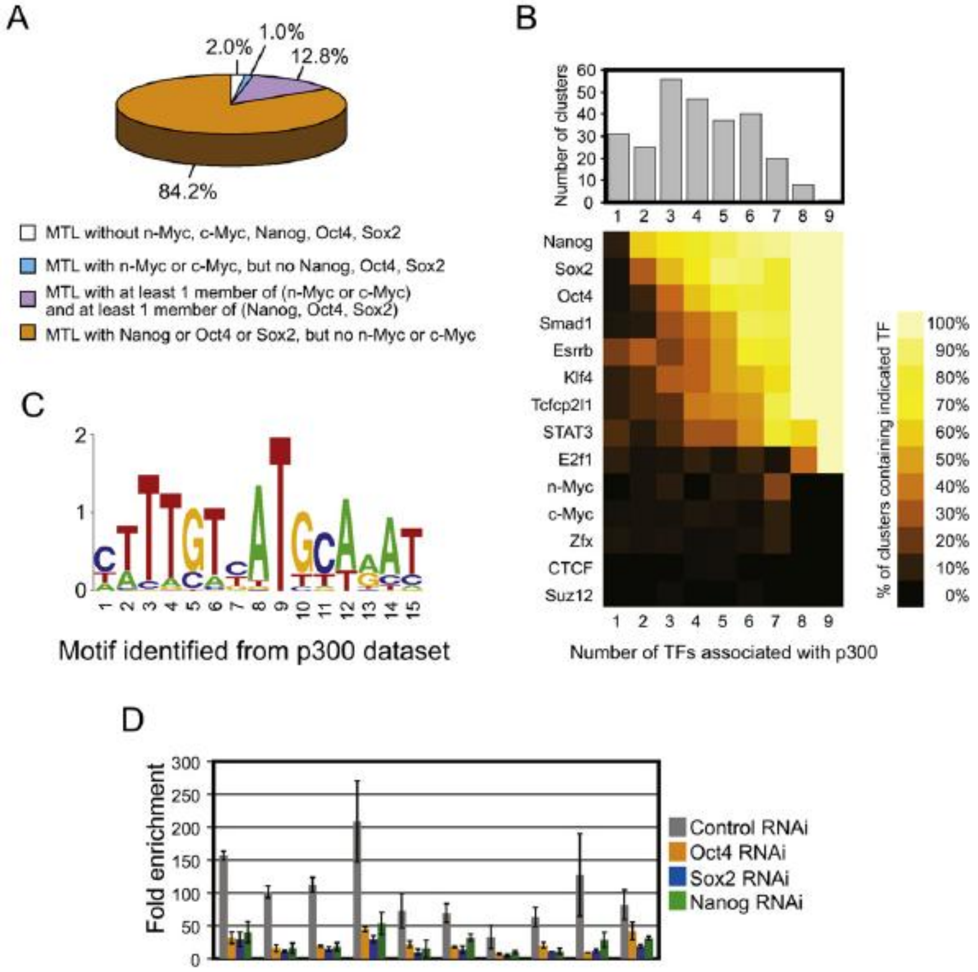


**Figure 5. p300 is recruited to the Nanog-Oct4-Sox2 cluster.** (A) p300 is associated with the Nanog-Oct4-Sox2 cluster, but not the Myc cluster. Pie-chart showing the occurrence of p300 in different MTL types. Color legend is the same as Figure 4B. (B) Size distribution and composition of binding site groups containing p300. (top) Histogram showing the number of binding site groups of different sizes. Size, here, refers to the number of non-p300 transcription factors that have binding sites in the same group. (bottom)

Composition of p300-containing binding site groups for different group sizes. Composition is expressed in terms of the percentage of p300-containing groups that contain the indicated transcription factor. For example, Nanog, Sox2 and Oct4 are each found in 70% or more of the p300 containing clusters that have five other factors bound, while Smad1, Esrrb, Klf4, Tcfcp2l1 and STAT3 are each found at a frequency of around 30-50%. (C) Motif predicted by *de novo* motif discovery algorithm Weeder. (D) Recruitment of p300 is dependent on Oct4, Sox2 and Nanog. ChIP assays were performed using anti-p300 antibody with extracts from ES cells transfected with control RNAi construct (gray bar), *Oct4* RNAi construct (orange bar), *Sox2* RNAi construct (blue bar) or *Nanog* RNAi construct (green bar). Coordinates of loci and q-PCR primers are listed in Table S10. Fold enrichment is the relative abundance of DNA fragments detected by real-time PCR at the amplified region over a control amplified region. The level of p300 was not altered after RNAi depletion of these TFs (data not shown). Data are presented as the mean ±SEM.

These data suggest that Oct4, Sox2 and Nanog are recruiting p300 to the genomic sites. To test this hypothesis, we depleted Oct4, Sox2 or Nanog by RNAi and checked for p300 binding. Our ChIP result showed that p300 binding was reduced by Oct4, Sox2 or Nanog depletion (Figure 5D). Previous work has shown that c-Myc interacts with p300 and mediates the recruitment of p300 to *hTERT* promoter (Faiola et al., 2005). In ES cells, we did not observe global recruitment of p300 to Myc sites. Depletion of c-Myc by RNAi however did not affect p300 recruitment to these sites (data not shown). The data suggests that p300 could be a general factor being recruited to enhancers (Heintzman et al., 2007) and we conclude that p300 recruitment is promoted by Oct4, Sox2 and Nanog.

### Combinatorial binding of transcription factors is associated with ES cell-specific expression

Next we sought to establish the correlation between TF occupancies and gene expression. A commonly employed approach for assigning target genes to a TF is to associate TF binding sites with genes based on proximity. However, the relevant threshold for proximity could be different for different TFs. For that reason, we developed a novel approach to cluster genes based on TF binding data (see Supplemental Experimental Procedures in Supplemental Data). For each pair of transcription factor and gene, we assigned an association score based on the genomic location of the binding site that is closest to the transcription start site (TSS). This association score is based on the distribution of nearest site-to-TSS distances in the genome, and is thus different for, and characteristic of, each TF. A higher score implies higher chance of the gene being the target of the TF. This avoids an arbitrary threshold. Based on the association scores for all TFs, we performed k-means clustering to define five classes of genes that are associated with similar set of transcription factors (Figure 6A).

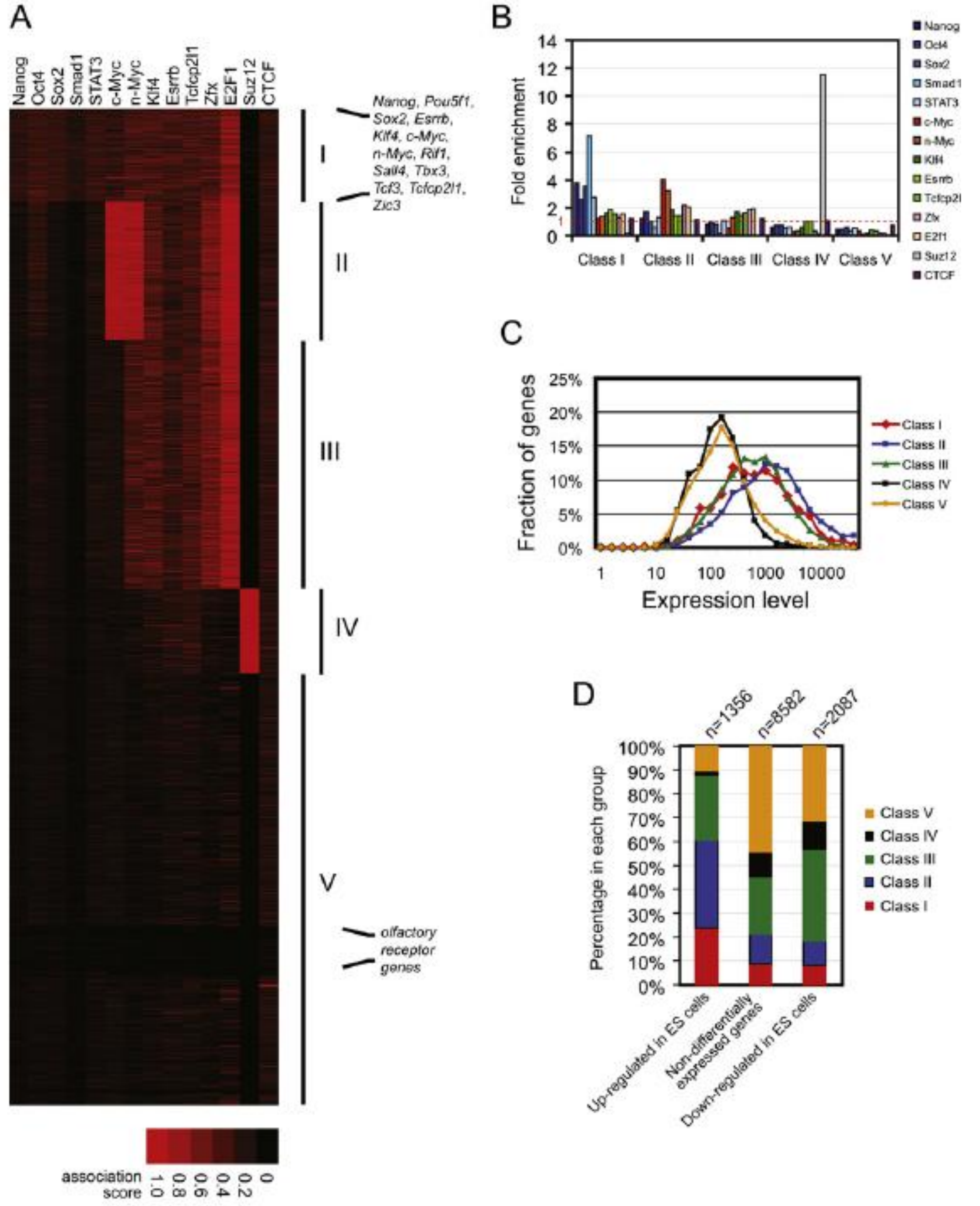


**Figure 6. Association between TF binding and gene expression in ES cells.** (A) Heatmap showing five classes of genes obtained from k-means clustering based on TF-gene association score. In this analysis, we included a Suz12 ChIP-seq dataset to explore the potential association of Suz12 and the other thirteen TFs. (B) Enrichment of transcription factors in the five classes. The Y axis represents the ratio of average TF-gene association score for the group to the average association score for all genes. (C) Histogram of the levels of gene expression for genes found in each of the five classes. (D) Proportion of different classes of genes found in differentially (up- or down-regulated in ES cells) and non-differentially expressed genes in published expression dataset (Ivanova et al., 2006).

Class I genes are enriched in binding sites for Nanog, Oct4, Sox2, Smad1 and STAT3 (Figure 6B). Class II genes are bound heavily by c-Myc and n-Myc. Class III genes show enrichment (more than 1 fold) in binding by n-Myc, Klf4, Esrrb, Tcfcp2l1, Zfx and E2f1. Class IV is highly enriched in Suz12 bound genes while class V genes are deficient in all the transcription factors. In total, 48% of genes are deficient in transcription factor binding by the thirteen transcription factors (class IV and class V). We note that E2f1 and Suz12 localization is essentially mutually exclusive, suggesting that polycomb repressor complexes inhibit the binding of E2f1 to its target sites.

To further characterize the gene expression profiles of each class, we used a microarray dataset that interrogated the transcriptome dynamics of retinoic acid (RA)-induced differentiation (Ivanova et al., 2006). The genes in this dataset were divided into three categories (see Supplemental Experimental Procedures in Supplemental Data). They are genes up-regulated in ES cells, non-differentially expressed genes and genes down-regulated in ES cells. Class I genes constitute less than 10% of the non-differentially expressed genes and genes down-regulated in ES cells (compare the red columns in Figure 6D). This compares to 24% of the genes in the up-regulated category. The percentage of class II genes is only 12% among non-differentially expressed genes, but 36% in the up-regulated set (compare the blue columns in Figure 6D). Hence, class I and class II genes are 2.7 (p=8.14E-52) and 2.9 (p=1.28E-91) fold enriched, respectively, in genes up-regulated in ES cells. In contrast, class IV and class V genes are underrepresented in this set. Class III is slightly enriched in genes that are down-regulated in ES cells, but not enriched in genes that are preferentially up-regulated in ES cells. As a validation, we compared the five classes with another independent microarray dataset (Zhou et. al. 2007) and similar results were obtained (Figure S14). In summary, our global analysis showed that 60% of genes up-regulated in ES cells are from class I and class II. Most importantly, the result demonstrates that gene clustering based on TF occupancies has the potential to predict ES cell-specific gene expression. This suggests that the TF binding patterns of these two groups are relevant in specifying ES cell-specific expression. In summary, we demonstrate that combinatorial binding patterns of TFs have greater predictive power for ES cell-specific expression.

### Constructing a regulatory network defining ES cell-specific expression

The self-renewing state of undifferentiated ES cells is characterized by the expression of genes specifically up-regulated in this cell-type. We sought to construct a regulatory network that specifies ES cell-specific expression using binding sites of transcriptional regulators under the undifferentiated state. In order to infer regulatory interactions, we made use of published expression profiling data that compared undifferentiated with differentiating ES cells. The rationale is that nine (Nanog, Oct4, Sox2, Klf4, n-Myc, c-Myc, Esrrb, Zfx, Tcfcp2l1) out of the thirteen TFs we studied are known to be coordinately up-regulated in ES cells and their levels are down-regulated upon differentiation or in differentiated cell-types (Fortunel et al., 2003; Ivanova et al., 2006; Ivanova et al., 2002; Ramalho-Santos et al., 2002). Two sets of published experiments were used to define genes that are differentially expressed during differentiation (Ivanova et al., 2006; Zhou et al., 2007b). The use of two independently generated datasets minimizes biases in gene expression differences that are due to different ways of differentiating ES cells.

The regulatory interaction between a transcription factor and its target gene is first defined for individual transcription factor by intersecting the rank-ordered bound genes (based on the total number of sequence tags associated with binding site peaks) and rank-ordered differentially expressed genes (see Supplementary Data for method). The thresholds for defining top-ranked genes in the two lists were determined empirically by maximizing the number of genes in the intersection subject to two constraints: there had to be at least twice as many genes in the intersection as the number expected by chance, and the null model (that there are no genes in excess) had to be rejected with $p < 10^{-3}$. This method allows us to make use of the unique features of our binding datasets (signal intensity and unbiased survey) and avoid the use of a single cutoff for all datasets.

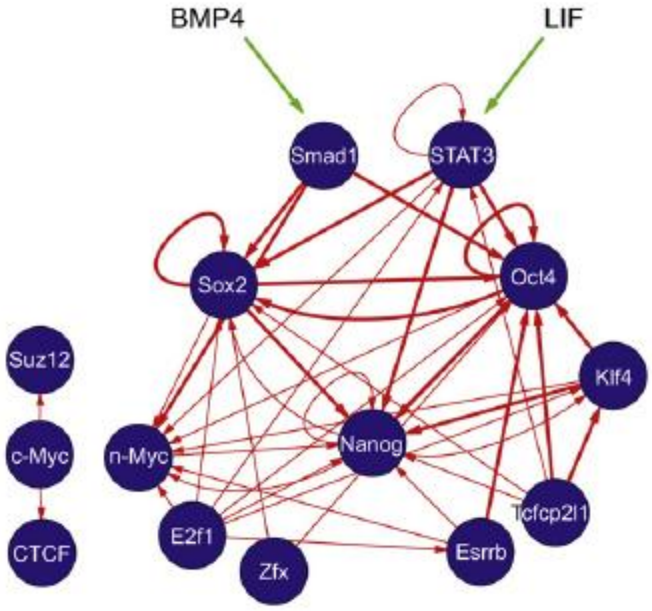


**Figure 7. Transcriptional regulatory network in ES cells.** Network of regulatory interactions inferred from ChIP-seq binding assays and from gene expression changes during differentiation. Nodes are ChIP-seq assayed transcription factors. Arrows point from the transcription factor to the target gene. Two sets of published experiments were used to define genes that are differentially expressed during differentiation (Ivanova et al., 2006; Zhou et al., 2007). Thick arrows represent interactions inferred from binding data and both expression experiments, while thin arrows represent interactions inferred from binding data and only one of the expression experiments. Regulatory targets were inferred from the intersection of top-ranked bound genes and top-ranked differentially expressed genes. Thresholds for defining top-ranked genes in the two lists were determined empirically by maximizing the number of genes in the intersection, subject to two constraints: the p-value for the enrichment of genes in the intersection had to be 0.001 or better, and there had to be at least twice as many genes in the intersection as expected. All regulatory interactions in this network involve higher level expression in ES cells and lower level expression during differentiation. There were no interactions among the factors in this network when regulation in the opposite direction was evaluated. The network was drawn using Cytoscape.

A network model based on the thirteen transcription factors as depicted in Figure 7 reveals both anticipated and novel aspects of the relationships between these transcription factors. Consistent with previous studies, this model shows regulatory feedback loops for Oct4, Sox2 and Nanog (Boyer et al., 2005; Chew et al., 2005; Loh et al., 2006). An interesting feature of this network is the inter-connectivity among eleven of the thirteen transcription factors being profiled.

## DISCUSSION

### The repertoire of binding sites in mammalian genome revealed by global mapping of transcription factor binding sites

Ultra high throughput sequencing technology through massively parallel short read sequencing has recently been developed for mapping transcription factor binding sites and histone modification profiles in mammalian cells (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007). In this study, we performed the first large scale mapping study of multiple TFs in mammalian cells.

Genome-wide mapping studies reveal abundant binding sites for different TFs in mammalian cells (Bieda et al., 2006; Birney et al., 2007; Cawley et al., 2004). It is thus a challenging task to identify the biologically relevant sites among the large repertoire of binding sites. It is also important to note that ChIP experiments may capture indirect TF-DNA interactions through protein-protein interaction. While advances in mapping technologies allow for comprehensive and unbiased disclosure of the repertoire of binding sites, it is difficult to determine which sites are functional regulatory elements that influence transcription. It is also possible that a sizeable fraction of these binding sites are non-functional and are the consequence of biological noise (Struhl, 2007). The strength of this study lies in the concurrent survey of the locations of multiple TFs in a single cell-type. Our data show that there are genomic regions extensively co-occupied by TFs (transcription factor colocalization hotspots) and they could represent functional important sites.

**ES cell-specific enhanceosomes**

An enhanceosome is a nucleoprotein complex composed of distinct sets of transcription factors bound directly or indirectly to enhancer DNA (Thanos and Maniatis, 1995). The density of transcription factors occurring on this short segment of DNA is high compared to more “modular” enhancers that have less dense binding clusters occurring over a longer segment of genomic DNA (Arnosti and Kulkarni, 2005). The virus-inducible enhancer of the interferon-β gene (*IFN- β*) is a prototypical enhanceosome. This 55 base pairs enhancer is bound by the p50 and p65 subunits of NF-kB, ATF-2, IRF-3, IRF-7, c-Jun, and the architectural transcription factor HMGA. An atomic model for the complex of eight of these factors on the DNA has been constructed based on three crystal structures (Panne et al., 2007). The basis for cooperativity is unlikely to be mediated through protein-protein interactions as these structures reveal limited contact between the transcription factors. It is proposed that the binding of these eight transcription factors on a composite DNA interface creates a continuous surface for recruiting co-activators such as p300 (Merika et al., 1998; Wathelet et al., 1998).

Our genome-wide mapping study reveals genomic regions with features of enhanceosomes. First, the binding sites are densely clustered within relatively compact genomic segments. It is of interest to note that the densest binding locus we identified is the distal enhancer of *Pou5f1*. This region (Chew et al., 2005) was bound by eleven transcription factors. Second, we showed that 25 of these genomic regions act as enhancers when placed downstream of the luciferase reporter. Third, they are associated with the H3K4me3 mark which is one of the signatures of active genomic regions. Fourth, our p300 ChIP-seq analysis revealed on a global scale the recruitment of this co-activator to the Nanog-Oct4-Sox2 cluster, but not the Myc cluster. Importantly, we showed that the recruitment of p300 is dependent on Oct4 and Sox2.

In higher eukaryotes, transcriptional enhancers play important roles in integrating multiple signaling pathways to achieve activation of specific genes. By profiling multiple transcription factor binding sites on the whole genome scale, we discovered extensive co-localization of multiple transcription factors on selective sites in the ES cell genome.

**Wiring of the ES cell genome**

LIF has long been known to be essential for the derivation or maintenance of mouse ES cells (Smith et al., 1988; Williams et al., 1988). Beside LIF, other factors in fetal calf serum (FCS) could be essential for self-renewal of mouse ES cells. Smith and co-workers have identified bone morphogenetic proteins as growth factors that work in conjunction with LIF to promote self-renewal (Ying et al., 2003). Addition of BMP4 to chemically defined media leads to the phosphorylation of Smad1 in ES cells. As constitutive expression of the *Id* genes bypasses the BMP4 or FCS requirement for maintenance of ES cells, the *Id* genes are implicated as downstream targets of the BMP / Smad signaling pathway (Ying et al., 2003). ES cells can be passaged without differentiation with LIF and BMP4, indicating that the pathways induced by these two ligands are sufficient to maintain stem cells. Importantly, we showed here that the binding of STAT3 and Smad1 to genomic sites is dependent on the LIF and BMP pathways respectively, confirming the importance of these transcription factors as effectors of the signal transduction pathways that maintain pluripotency in ES cells (Figure S3L, M). Until the present study, the role of transcriptional regulatory proteins downstream of these signaling pathways has not been well defined in the context of ES cell transcriptional regulatory networks. STAT3 had been shown to bind to the *Nanog* enhancer (Suzuki et al., 2006), but there were no known target of Smad1.

Consistent with a previous study implicating *Id* genes as downstream targets of BMP4 pathway (Ying et al., 2003), we identified a MTL (bound by Smad1, Oct4, Sox2, Nanog, Klf4, E2f1, Esrrb and Tcfcp2l1) at 1.5 kb upstream of the TSS of *Id3*. Strikingly, the majority (97.3%, 649/667) of Smad1 at the MTL is associated with Nanog, Oct4 or Sox2. STAT3 (72.5%, 521/718) is also predominantly localized with Nanog, Oct4 or Sox2 within the MTL.

The multiple transcription factor binding site maps provide us with the opportunity to examine the mode of targeting genes by these factors on a global scale. E2f1 binds to approximately 50% of all genes, almost all of which fall into what we call classes I, II and III (Figure 6A). Genes in these three classes (I, II and III) are expressed at higher levels in ES cells than are the other classes (Figure 6C), and class I and class II genes are enriched in genes that are expressed at higher levels in ES cells than in differentiating cells. Roughly 50% of all genes, those in classes IV and V, are not enriched in transcription factor binding (Figure 6A). These transcription factor deficient genes are not expressed or are expressed at a low level. A fraction of these genes are bound by Suz12, suggesting that Suz12 plays a role in preventing transcription factor occupancy and in silencing these genes (Boyer et al., 2006a; Lee et al., 2006). However, a larger fraction of the transcription factor deficient genes are not bound by Suz12. It is possible that the chromatin structure of these genes is not permissive to facilitate the binding of transcription factors.

In summary, the genome-wide maps of transcription factors and co-regulators demarcate different gene compartments in the ES cell genome. The densely co-occupied sites represent key regions of potential functional importance and will assist in the identification of new regulators of self-renewal, pluripotency and reprogramming. We demonstrate that the two key signaling pathways are integrated to the Oct4, Sox2 and Nanog circuitries through Smad1 and STAT3. Our data also provide a framework for modeling gene expression and understanding the transcriptional regulatory networks in pluripotent cells.

# APPENDIX II: A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data

This chapter is published as: He X, Chen CC, Hong F, Fang F, Sinha S, Ng HH, Zhong S.

*PLoS One*. 2009 Dec 1;4(12):e8155.

**My contribution to this project:**

It is a collaboration project with a bioinformatic group at UIUC. I was responsible for leading the experimental analysis for this project. I have validated the predicated Nanog binding motif by Electrophoretic Mobility Shift Assays and compared its binding affinity with published Nanog binding motif. I also worked with Sheng Zhong to construct most of the manuscript figures as well as the writing of manuscript text.

## ABSTRACT

How regulatory DNA sequences control gene expression, in a quantitative manner, through the combinatorial interactions with transcription factors (TFs) is not well understood. We present a computational method to address this question, relying on the established biophysical principles. This method, STAP (sequence to affinity prediction), takes into account all combinations and configurations of strong and weak binding sites to analyze large scale transcription factor (TF)-DNA binding data to discover cooperative interactions among TFs, infer sequence rules of interaction and predict TF target genes in new conditions with no TF-DNA binding data. The distinctions between STAP and other statistical approaches for analyzing cis-regulatory sequences include the utility of physical principles and the treatment of the DNA binding data as quantitative representation of binding strengths. Applying this method to the ChIP-seq data of 12 TFs in mouse embryonic stem (ES) cells, we found that the strength of TF-DNA binding could be significantly modulated by cooperative interactions among TFs with adjacent binding sites. However, further analysis on five putatively interacting TF pairs suggests that such interactions may be relatively insensitive to the distance and orientation of binding sites. Testing a set of putative Nanog motifs, STAP showed that a novel Nanog motif could better explain the ChIP-seq data than previously published ones. We then experimentally tested and verified the new Nanog motif. A series of comparisons showed that STAP has more predictive power than several state-of-the-art methods for cis-regulatory sequence analysis. In conclusion, STAP is an effective method to analyze binding site

arrangements, TF cooperativity, and TF target genes from genome-wide TF-DNA binding data.

## INTRODUCTION

The spatial-temporal patterns of gene expression are controlled by cis-regulatory sequences (Davidson, 2006), through the binding of transcription factors (TFs) to specific sites in these sequences. Numerous studies point out that the final transcriptional "read-out" is determined, not by an individual TF, but by the combinatorial interactions of multiple TFs with DNA. Most notably, in developmental genes, multiple binding sites of different TFs are often located close to each other in genomes, forming so called cis-regulatory modules (CRMs), and work together to generate very precise expression patterns (Howard and Davidson, 2004).

Sequence-specific binding of TF molecules to DNA has been well studied, both in theory (Berg and von Hippel, 1988) and in practice (Stormo, 2000). In contrast, the interactions between TF molecules that enhance or inhibit their DNA binding affinities or transcriptional effects are not well understood. It is not clear, at a quantitative level, how important such interactions are, and in most systems the identities of interacting TFs remain unknown.  In cases where multiple TF molecules do interact, it is unknown how the spatial organization of their binding sites affects DNA binding. Some studies suggest that binding sites must be arranged in specific ways, following "grammar-like rules" (Beer and Tavazoie, 2004; Struhl, 2001) in order for them to interact properly; others provide evidence of a flexible organization of regulatory sequences (Arnosti and Kulkarni, 2005; Brown et al., 2007). Knowledge of the importance of TF interactions will be a central piece of our picture of gene regulation.

Genomewide DNA-binding data from chromatin immunoprecipitation followed by either genome tiling array analysis (ChIP-chip) or sequencing (ChIP-seq), provide an opportunity to address the above-mentioned problems quantitatively (Barski and Zhao, 2009; Bulyk, 2006). DNA-binding by TFs is an inevitable step in transcriptional regulation, thus modeling the combinatorial TF-DNA interactions will serve as a bridge to understanding the complex transcriptional process. Focusing on ChIP-based data, instead of gene expression data, enables a simplification of the task at hand. Gene expression is often accomplished through an intricate process involving not only TF-DNA interactions, but also chromatin remodeling, epigenetic modifications, communications between multiple enhancers, etc (Blackwood and Kadonaga, 1998). For this reason, several studies have argued for studying combinatorial interactions among TFs using ChIP-based technologies (Yu et al., 2008; Zhang et al., 2007).

A number of computational methods have been proposed to study the combinatorial aspect of gene regulation. Typically, these methods attempt to extract information from the statistical patterns in DNA sequences, e.g., the occurrence of sequence motifs. For example, some researchers detect possible interactions between pairs of TFs based on frequent co-occurrence of their motifs (Smith et al., 2005a; Zhou et al., 2007a). Various techniques from statistics and machine learning, such as Bayesian networks (Beer and Tavazoie, 2004), multivariate regression (Smith et al., 2005b), decision trees (Jin et al., 2006), SVMs and artificial neural networks (Zhou and Liu, 2008), have been developed to extract important features (e.g., motifs and their combinations) from sequences, using either gene expression data or ChIP-chip data. However, these statistical methods do not reflect underlying physical principles. As such, it is not clear to what extent their underlying assumptions, e.g., additivity of different features, are valid. Additionally, important sequence features, e.g. the interactions among adjacent binding sites, are often not represented in these approaches. In many cases, parts of the results, e.g. the physical meanings of the model parameters, are not easy to interpret.

By directly modeling the underlying process, a biophysics-based approach can overcome many limitations of the statistical methods mentioned above. Shea and Ackers (Shea and Ackers, 1985) and Buchler et al. (Buchler et al., 2003) pioneered the use of thermodynamic principles in the study of regulatory mechanisms. A number of recent studies applied these principles to model expression data on promoters/enhancers (Gertz et al., 2009; Janssens et al., 2006; Segal et al., 2008; Zinzen and Papatsenko, 2007; Zinzen et al., 2006) or TF-DNA binding data from ChIP-chip experiments (Foat et al., 2006; Roider et al., 2007). However, these methods have certain limitations when considered in the context of our study. Importantly, the possibility of multiple transcription factors interacting with DNA and at the same time influencing each other has not been adequately addressed in most of these methods (Foat et al., 2006; Janssens et al., 2006; Roider et al., 2007; Segal et al., 2008; Zinzen and Papatsenko, 2007). Often the physical interactions were not explicitly modeled (Janssens et al., 2006; Segal et al., 2008) or overly simplified (Zinzen and Papatsenko, 2007; Zinzen et al., 2006). Finally, the studies are often limited to individual regulatory sequences (Janssens et al., 2006; Zinzen and Papatsenko, 2007; Zinzen et al., 2006), or theoretical simulations (Zinzen and Papatsenko, 2007), or artificial promoters, which are by design far simpler than natural systems (Gertz et al., 2009).

We developed a novel method, called STAP (Sequence To Affinity Prediction), to analyze large scale TF-DNA binding data. The heart of this method is a thermodynamic model adapted from earlier theoretical studies (Buchler et al., 2003; Shea and Ackers, 1985). The key novel feature of our model is the explicit treatment of cooperative interactions among different TF molecules. In addition, our specially developed computational techniques based on dynamic programming will enable the model to be efficiently applied to complex sequences and ChIP-seq data. This combined biophysical and computational approach distinguishes our method from previous methods that rely on statistical patterns of DNA sequences or inadequate modeling of physical details. Another main feature of our method is the utility of ChIP-seq data not only as binary indicators of TF binding regions, as been done by most existing studies, but also as quantitative measurements of the binding strengths of a region. Thus, more information from ChIP-seq data will be utilized by this new model.

STAP was applied to analyze the ChIP-seq data of 12 TFs in mouse embryonic stem (ES) cells (Chen et al., 2008a) . A novel DNA binding motif of NANOG was identified and subsequently experimentally validated. Consistent to previous observations, we found that the TFs were often "co-localized", in the sense that multiple TFs would bind to the same loci. We also identified several instances of cooperative interaction between TFs contributing to enhancing their DNA-binding, although such instances were in the minority among all instances of co-localization. Finally, the analysis suggested that the precise arrangement of binding sites is not critical for cooperative interactions between TFs.

## MATERIALS & METHODS

**Biophysical model of TF-DNA interaction.** Given a sequence $S$, our goal is to predict its binding intensity with the experimental TF, denoted as TFexp. For a single binding site $S_i$, its binding affinity to the TF is given by (Berg and von Hippel, 1988):

$$q_i = [TF]K(S_i) = [TF]K(S_{\max})e^{-\Delta E(S_i)} (1)$$

where $[TF]$is the TF concentration, $K(.)$ is the equlibrium constant of a site, $S_{\max}$ denotes the consensus sequence of this TF, and $\Delta E(S_i)$ is the mismatch energy of $S_i$ in the unit of $\beta = 1/kT$. Note that $[TF]K(S_{\max})$ can be considered as a single TF-specific constant, denoted as $R$. Supposing there are a total of $n$ binding sites, a state $\sigma$ is represented by an $n$-bit vector, where $\sigma_i$ represents whether the $i$-th site is occupied by its corresponding TF (equal to 1) or not (0). The sequence is thus viewed as being in a mixture of $2^n$ states. The probability of a state $\sigma$, denoted as $P(\sigma)$, is determined by its Boltzman weight, $W(\sigma)$, given by (Buchler et al., 2003):

$$W(\sigma) = \prod_{i=1}^{n} q_i^{\sigma_i} \prod_{i<j} \omega(i,j)^{\sigma_i \sigma_j} \quad (2)$$

where $\omega(i,j)$ denotes the interaction between the two sites $i$ and $j$ when both are occupied. Note that the interaction may depend on the arrangement of the binding sites. Our default model of interaction is a simple binary model: the bound factor at position $i$, $f$, and the bound factor at position $j$, $f'$, can interact with constant $\omega_{f,f'}$ if the distance of their binding sites is less than $d_{\max}$. Basically, the above equation states that the weight of a particular state has two components: one from the binding of TF to each individual site; and the other from cooperative interactions among bound TFs. In theory, any two bound TF molecules can form interactions; in reality, however, this is quite unlikely to be true. So we make the assumption that only two adjacent bound TF molecules can interact with each other. We assume that the binding affinity of the whole sequence to TFexp (denoted as index $k$) is proportional to the expected number of bound molecules of $k$, averaging over all states:

$$\overline{N_k} = \sum_{\sigma} N_k(\sigma) P(\sigma) = \frac{\sum_{\sigma} N_k(\sigma) W(\sigma)}{\sum_{\sigma} W(\sigma)} \quad (3)$$

where $N_k(\sigma)$ is the number of bound molecules of $k$ in the state $\sigma$ (a simple counting).

Because the number of states is exponential to the number of sites in a sequence, the computation of the above quantity is expensive. We developed a dynamic programming algorithm that computes it efficiently. In the first step, we compute the denominator $Z = \sum_{\sigma} W(\sigma)$. Let $\sigma[i]$ be one configuration up to the site $i$, where $i$ is bound by its cognate TF $f_i$. We could decompose the configuation $\sigma[i]$: supposing the nearest site to $i$ that is occupied in this configuration is $j$ ($j < i$, $j = 0$ if no site is occupied before $i$), then we have:

$$W(\sigma[i]) = W(\sigma[j])\omega(i,j)q(i) \text{ (4)}$$

We use $Z(i)$ to denote the total statistical weight of all configurations up to $i$, where the site $i$ is occupied, i.e., $Z(i) = \sum_{\sigma[i]} W(\sigma[i])$. Summing over all $\sigma[i]$ in the above equation and plugging in the expression of $Z(j)$ lead to the following recurrence:

$$Z(i) = q(i)\left[\sum_{j\in\Phi(i)} \omega(i,j)Z(j) + 1\right] \text{(5)}$$

where $\Phi(i)$ is the set of sites before $i$ that do not overlap with $i$. In order to compute $Z$, we note that the last bound site in any configuration could be $1, 2, \cdots, n$ or no bound site. So we have: $Z = 1 + \sum_{i=1}^{n} Z(i)$.

Next we compute the numerator $Y_k = \sum_{\sigma} W(\sigma)N_k(\sigma)$. We define the variable $Y_k(i) = \sum_{\sigma[i]} W(\sigma[i])N_k(\sigma[i])$. For any specific configuration $\sigma[i]$, we have:

$$W(\sigma[i])N_k(\sigma[i]) = \left[W(\sigma[j])q(i)\omega(i,j)\right]\left[N_k(\sigma[j]) + I(f_i,k)\right] \quad (6)$$

where $I(f_i,k)$ is the indicator variable of whether $f_i$ is equal to $k$. Summing over all $\sigma[i]$ and plugging in the expressions of $Z(j)$ and $Y_k(j)$, we have the following recurrence:

$$Y_k(i) = q(i)\left\{\sum_{j\in\Phi(i)} \omega(i,j)\left[Y_k(j) + I(f_i,k)Z(j)\right] + I(f_i,k)\right\} (7)$$

The last bound site could be $1,2,\cdots,n$ (if no site is bound, no contribution to $Y_k$), so we have: $Y_k = \sum_{i=1}^{n} Y_k(i)$.

**Model fitting in ChIP-seq data.** In the model, the free parameters are $R_f$ for each factor $f$, and $\omega_{f,f'}$ for any two factors $f$ and $f'$. The mismatch energy of any site is related to the commonly used PWM matching score (Berg and von Hippel, 1988; Stormo, 2000). Given the data mapping sequence to binding intensities, we use the simplex and BFGS algorithms (Press et al., 1992; Segal et al., 2008) to train the parameter values that maximize the Pearson correlation between the observed binding intensities and the predicted values.

Our program takes as input ChIP-seq data of one experiment (i.e. of one TF) and a set of TF motifs (including TFexp), learns a TF-binding model that can be used to predict the binding affinities of any new sequences, and predict a set of interactions between other TFs and TFexp. Specifically, as the first step, we identify the genomic loci with the highest tag counts and extract the surrounding sequences, defined as 250 bp upstream/downstream of the peaks. We also randomly choose sequences which do not show significant binding. This collection of pairs of sequences and binding intensities will be used for training data. In the next step, we perform a step-wise learning of the model: starting from TFexp, if adding a motif significantly improves the predictability of the model (reflected via the Pearson correlation between predicted and observed binding intensities), it will be added to a motif set. Once a motif is added, we will check the existing motifs in the set: if removing it does not significantly drop the model's predictability, it will be removed. This process is repeated until it converges to a stable set of motifs. At this stage, the program outputs the final set of motifs, ranked by their contributions to binding, and the model parameters.

**Models of cooperative interactions.** We denote the cooperative interaction between two bound factors, $\omega(d)$, where $d$ is the distance between the two sites. It may also depend on the orientations of the two sites (in the same direction or not). Let $d_{\max}$ be the maximum distance where two bound factors can interact. We consider several forms of the function $\omega(d)$. Under the Binary function, the interaction term is equal to a constant, $\omega$, if $d$ is less than $d_{\max}$; and 1.0 otherwise (no interaction, corresponding to free energy at 0). The orientation bias (one orientation will be favored over the other) is modeled by multiplying a constant to $\omega$ if two sites are at different strands. The Linear function is defined by:

$$\omega(d) = \begin{cases} \omega & d \le d_0 \\ 1 + (d_{\max} - d)\cdot(\omega - 1)/(d_{\max} - d_0) & d_0 < d \le d_{\max} \\ 1 & d > d_{\max} \end{cases} \quad (8)$$

The orientation bias is modeled similarly. To derive the Periodic function, we assume that the free energy of interaction consists of a constant plus a term corresponding to the energetic cost of DNA looping. Following (Saiz et al., 2005), the effective interaction between $A$ and $B$ is given by:

$$\Delta G = \Delta G_{A-B} - \Delta G' \sin(2\pi \frac{d}{T} + \varphi) \, (9)$$

where $T$ is the period, $\varphi$ is the phase parameter and $\Delta G_{A-B}$ and $\Delta G'$ are constants. The interaction weight is $\exp(-\Delta G / RT)$ when $d$ is less than $d_{\max}$ and 1.0 otherwise. Also note that $\varphi$ can in fact take two values, depending on whether the two sites are in the same orientation.

**Learning the Interaction Model between Two TFs.** In studying the effect of binding site arrangement on TF interaction, we adopt a different model fitting procedure. Suppose we want to study the interaction of the factors A and B. We estimate a single set of parameters: RA, RB and the relevant interaction parameters (depends on how we model their interaction) from the binding data of both factors. The objective function is the average correlation coefficients between predictions and observations in the two sets of sequences. Also we vary the interaction parameters to observe their effects on the predictability of the model, as shown in the text, instead of estimating single optimal values. We note that such procedure is not applicable to fitting a ‘‘global’’ model of a large number of TFs (e.g. all 12 TFs in the mouse ESC dataset). In that case, the number of possible interactions is probably too large (66 in the ESC case) to be reliably estimated. Our software, however, does support estimating the global model when the number of factors is small (less than four, for instance).

**Testing TF co-localization.** We took the ChIP-seq data from (Chen et al., 2008b) and followed their procedure to identify peaks that are bound by a TF. Our goal is to test if a factor, A, co-localizes with another factor, B. This translates to the hypothesis that A sites which are adjacent to some B sites (250 bp in our experiments) are enriched among all sites of A. We estimated the expectation of this number as well as the expected number of A sites that are not adjacent to some B sites, assuming the distribution of B sites follows a Poisson distribution whose rate is the genome-wide density of B peaks. These expected numbers are compared with the observed numbers of peaks via Pearson's $\chi$2 test.

**Expression of Nanog protein.** Recombinant proteins of the Nanog (GST tagged) were used for the gel shift assays. The full length Nanog protein was cloned into the pET42b (Novagen) vector. The proteins were expressed and purified with GSH-sepharose beads (Amersham). Eluents were dialyzed against a dialysis buffer (10 mM Tris–HCl, pH 7.4, 100 mM NaCl, 10 mM ZnCl2 and 10% glycerol) at 4 ℃ for 6 h. Proteins were stored at -80 ℃.Concentrations of proteins were verified with the Biorad protein measurement assay.

**Electrophoretic Mobility Shift Assay (EMSA).** DNA oligonucleotides (Proligo) labeled with biotin at the 5' end of the sense strands were annealed with the antisense strands in the annealing buffer (10 mM Tris-HCl , pH 8.0, 50 mM NaCl, 1 mM EDTA) and purified with agarose gel DNA extraction kit (Qiagen). DNA concentrations were determined by the NanoDrop ND-1000 spectrophotometer. The gel shift assays were performed using a LightShift Chemiluminescent EMSA kit (Pierce Biotechnologies). 100 ng of protein were added to a 5ul reaction mixture (final) containing 1 ug of poly(dI-dC) (Amersham), 1 ng of biotinlabeled oligonucleotide in the binding buffer (12 mM HEPES, pH 7.9, 12% glycerol, 60mM KCl, 0.25 mM EDTA, 1 mM

DTT). Binding reaction mixtures were incubated for 20 min at room temperature. Binding reaction mixtures were resolved on pre-run 6%native polyacrylamide gels in 0.5X Tris-buffered EDTA. Gels were transferred to Biodyne B nylon membranes (Pierce Biotechnologies) using Western blot techniques and detected using chemiluminescence.

## RESULTS

### ChIP-seq data can be quantitatively reproduced.

We hypothesized that ChIP-seq data quantitatively reflect the binding strength between the TF and the respective genomic binding regions. The rationale is that the binding strength is proportional to the proportion of cells that have this genomic locus bound by this TF (Buchler et al., 2003), and therefore proportional to the counts of overlapping ChIP-seq tags. To verify this hypothesis, we randomly picked 28 Nanog ChIP-seq detected binding regions from (Chen et al., 2008a) and repeated the ChIP experiments in E14 mouse ES cells. We used real-time qPCR to quantify the ChIP precipitated DNA on the 28 pre-selected regions. The ChIP-seq and ChIP-qPCR signals exhibited a strong correlation (*r*2= 0.656). We performed the same experiment on 11 SUZ12 binding regions from ChIP-seq data and similarly found a strong correlation (*r*2= 0.792). These data suggest that the counts of overlapping ChIP-seq tags are quantitatively reproducible by independent experiments. Thus it becomes possible to model and utilize the quantitative nature of ChIP-seq data for investigating the biophysical rules of protein-protein and protein-DNA interaction.

### Transcription factors are extensively co-localized

We studied ChIP-seq data on 12 TFs active in embryonic stems cells (Chen et al., 2008a): cMyc, CTCF, E2f1, Esrrb, Klf4, Nanog, nMyc, Oct4, Sox2, STAT3, Tcfcp2l1 and Zfx. Combinatorial gene regulation leads to a statistical tendency of multiple factors to bind to proximally located sites, a phenomenon we call TF "co-localization". We developed a statistical test for co-localization of TF pairs and found extensive evidence for this phenomenon. The majority (121) of all 132 possible pairs show significant co-localization (P-value < 0.01, Pearson's $\chi 2$ test). Our results are broadly consistent with the results of Chen et al. (Chen et al., 2008a), which also revealed extensive co-localization of TFs (though no statistical tests were provided). In summary, both analyses strongly indicate a combinatorial mode of action by multiple factors.

### A biophysical model of TF binding to DNA sequences

A possible explanation for TF co-localization is that DNAbinding of one factor helps recruit another factor to its binding site, through favorable TF-TF interaction. (Note that the binding sites in this paper refer to 10–20 bp regions actually occupied by TFs, while other papers may refer to putatively larger regions identified in ChIP-chip or ChIP-seq experiments – these will be called TF bound regions in our paper). Thus, when co-localized, both factors may access the DNA with higher affinity than their individual binding sites alone would allow. We adapted the biophysical model from (Buchler et al., 2003) that incorporates such cooperative binding, for the purpose of analyzing TF-DNA binding data. Given a transcription factor (called ''TFexp''), our goal is to predict the binding affinity of TFexp to any sequence. The basic assumption is that many putative binding sites, including the sites of TFexp and of other factors, not just the single best match, may contribute to interaction of this sequence to TFexp. Indeed, the importance of weak binding sites and cooperative interactions has been supported by a number of recent studies (Gertz et al., 2009; Roider et al., 2007; Segal et al., 2008; Tanay, 2006). Under this picture: binding sites of TFexp directly attract TFexp, and sites of other factors may interact cooperatively with TFexp, thus indirectly recruiting TFexp. The cooperative interactions may occur among adjacent binding sites of the same TF (self-cooperativity) or of different TFs (heterotypic cooperativity). Thermodynamically, each binding site of a sequence may be occupied or not, thus a sequence with n sites exists in 2n states, where each state represents the occupancy status of all sites (Figure 1). The probability of a state depends on interactions of TFs with their binding sites, as well as TF-TF interactions, as quantified by Equation (2) in Methods. Following earlier work on ChIP-chip data analysis (Foat et al., 2006; Roider et al., 2007), we assume that the binding affinity of TFexp to this sequence is proportional to the average number of TFexp molecules occupying their sites, over all states weighted by their probabilities (Figure 1). Note that the number of states is exponential to the number of binding sites, thus it is computationally difficult to calculate the binding affinities of complex sequences by the brute-force method. We developed a dynamic programming algorithm to carry out the computation efficiently. The details of the model and the algorithm can be found in Methods.

When analyzing the genome-wide binding data of some TF (hereafter called the primary factor), the goal is to learn the TFs (called cooperative factors) that interact with this factor, as well as the relevant model parameters. The STAP model is fit by maximizing the Pearson's correlation coefficient between the predicted binding affinities and the overlapping ChIP-seq counts (or ChIP-chip intensities). To search for interacting factors, we iterate the motifs in a motif collection, such as the JASPAR database (Bryne et al., 2008) . Each motif in this collection is tested by whether adding this motif to the STAP model with only the primary factor will significantly improve the Pearson's correlation coefficient. The significance of this improvement is assessed by using a large number of randomized motifs as negative controls. After all cooperative factors are learned, and STAP re-trains the model parameters. The STAP model is designed for analyzing ChIP data from a single TF; a variation of STAP is developed for simultaneously analyze ChIP data from several TFs (see "Exploring the effects of binding sites arrangement").

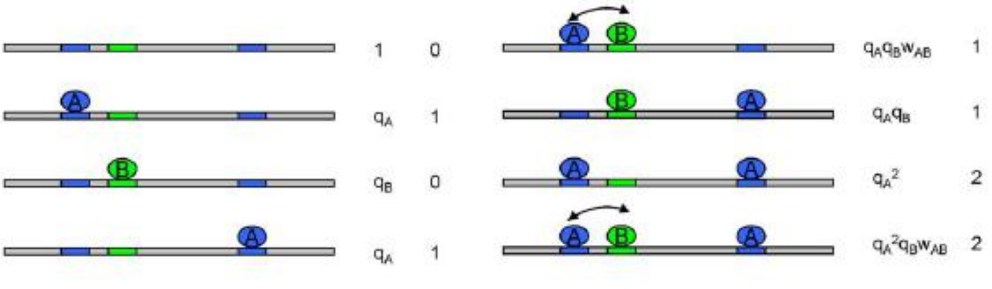


**Figure 1. Models of cooperative DNA binding.** The sequence contains three binding sites, two for factor A, and one for factor B. All possible eight configurations of the sequences, in terms of binding site occupancy, are shown. The arrow connecting two adjacent bound molecules indicates cooperative interaction. For each configuration, the first column represents the weight, i.e., un-normalized probability, and the second column represents the number of bound molecules of A. The parameters in the weight terms are: qA (qB) – strength of factor A (B) binding to DNA; wAB – strength of the

interaction between A and B. The binding affinity of the sequence to A is the average of the second column, weighted by the first column.

### ChIP-seq data reveals a novel characterization of Nanog binding specificity

Our method needs to use motifs of TFs, representing binding specificities, to identify putative binding sites in target sequences (though it is theoretically possible to learn novel motifs under our framework, similar to (Foat et al., 2006)). So at the first step, we identified the motifs of the 12 TFs. For each factor, we ran the MEME program (Bailey and Elkan, 1994) on the top 100 regions (ranked by tag counts) detected in the ChIP-seq experiments. These motifs are by and large similar to those reported in the original ChIP-seq paper (Chen et al., 2008b). However, we noted that the motifs of Oct4, Sox2 and Nanog, learned by (Chen et al., 2008b) were remarkably similar to each other. We hypothesized that this similarity was due to co-localization of the factors, which resulted in similar collections of genomic regions being used for enrichment-based motif finding. To test this hypotheses, we used sequences bound exclusively by each of these three factors and performed MEME analysis again (NestedMICA (Down and Hubbard, 2005) and Gibbs sampler (Thompson et al., 2007) gave similar results). The resulting Oct4 and Sox2 motifs are similar to the corresponding parts of the previously identified Oct4-Sox2 joint motif, while the Nanog motif is different (Figure 2A, Nanog1). We also note that several other DNA binding profiles of Nanog were reported from previous studies [(Chen et al., 2008b; Loh et al., 2006; Mitsui et al., 2003), but they do not resemble each other. Inspired by the importance of Nanog as an essential regulator in ES cell proliferation and self-renewal

(Mitsui et al., 2003), we set out to characterize the binding specificity of Nanog using a combination of computational and experimental approaches.

Even though STAP was not designed for *de novo* motif finding, it is applicable to compare multiple motifs of the same factor. By setting these motifs as alternative inputs and comparing the model fit to genome-wide binding data, the best motif can be recognized. We applied this strategy to the new Nanog motif as well as two previously published ones (Nanog2 (Mitsui et al., 2003) and Nanog3 (Loh et al., 2006), Figure 2A) to test if the new motif better explains the ChIP-seq data. The new Nanog motif resulted in a higher correlation than the other two in the sequences bound only by Nanog, but not Oct4 and Sox2 (Figure 2B, Nanog-only), providing initial support to the novel Nanog motif. In a second test, we utilized STAP's capability of analyzing cases where multiple factors are bound. As discussed before, the enrichment of Oct4 and Sox2 binding sites in the Nanog-bound sequences tend to confuse the motif discovery tools. This obstacle was resolved by setting Oct4 and Sox2 as cooperative factors, and varying the candidate primary motif. In this way, the difference of results was attributed to the different Nanog motifs, with the effects of Oct4 and Sox2 sites automatically disentangled. Again, the new Nanog motif provided a significantly better fit to the ChIP-seq counts of the Nanog bound sequences than the other motifs (Figure 2B, Nanog-500). In addition, the fitting of observations with the new Nanog motif is highly significant under a test using randomized motifs.

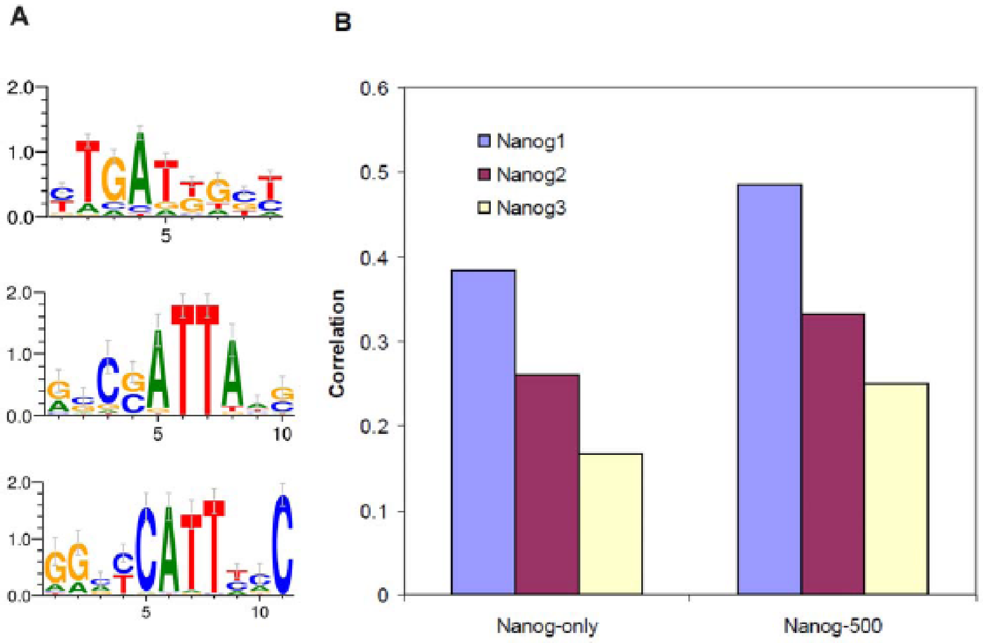
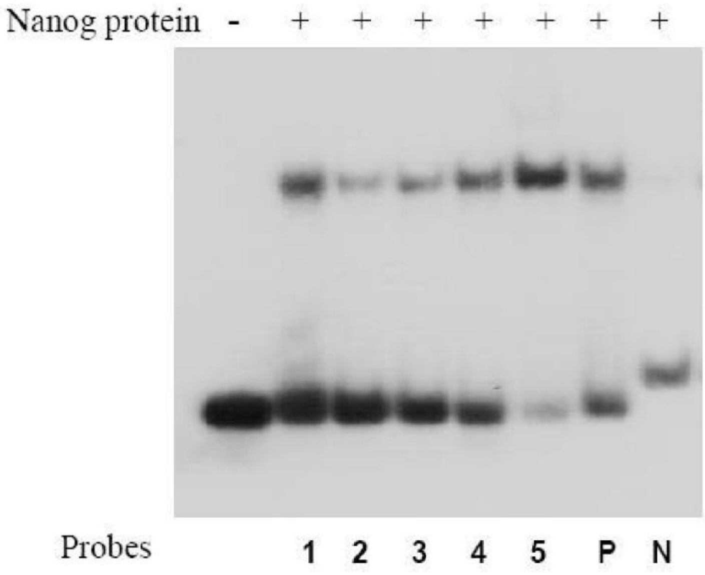


**Figure 2. Comparison of three versions of the Nanog motif.** (A) Nanog1 – the motif learned from the sequences bound by Nanog, but not Oct4 and Sox2, in the ChIP-seq data; Nanog2 – the motif in (Mitsui et al., 2003); Nanog3 – the motif in (Loh et al., 2006). (B) Performance of models using three different versions of the motif, measured by the correlation between model predictions and observations. The models are applied to two different sets of data. Nanog only: the sequences bound by Nanog, but not Oct4 and Sox2; Nanog-500: the 500 sequences with strongest binding to Nanog.

### Experimental tests of the novel Nanog motif

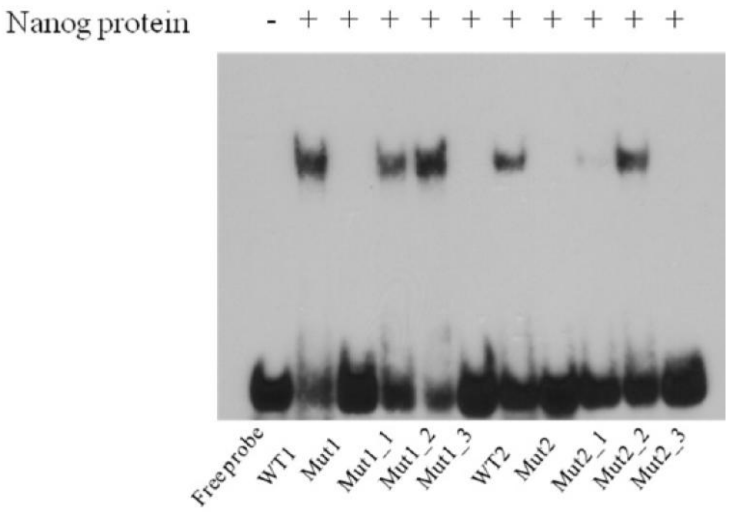
We used electrophoretic mobility shift assay (EMSA) to test the binding of Nanog to the DNA sequences that match the novel Nanog motif. First, from the the Nanog ChIP-seq positive regions, we randomly selected five sequences that match to the new Nanog motif but do not match the Oct4-Sox2 joint motif (Table S-EMSA). EMSA produced the same band from these five sequences, which also match the band produced from a positive control region that is known to interact with Nanog. In contrast, a randomly selected negative control sequence produced a completely different band (Figure 3).



| Probe | Coordinate | strand | Sequence |
|---|---|---|---|
| 1 | chr13_3712191_3712231 | + | TCCTGCAACCAGCCCTTGATGGCCCTCCTTGATGGCCCGC |
| 2 | chr19_21852503_21852543 | + | GGATTCCTTTCAGCTCTGATGGGTTTCTTTCAGCTATTGA |
| 3 | chr4_41045395_41045435 | - | AAGGCTTAGTCCTTGATGGGTTCTTTGTCATCCCAATCAA |
| 4 | chr6_112851211_112851251 | + | TCACTTAATTCCTCCTTGATTGCTTTTCAAAAGCAATGTA |
| 5 | chr5_142668044_142668084 | - | TGTGATTTATCCCTGATGGCCCATTAGTCCGGATGGTTTG |

**Figure 3. EMSA experiments of five genomic regions with high similarities to the new Nanog motif.** Probes 1 to 5 correspond to the genomic regions 1 to 5 in the Table. Probes P and N are positive and negative control probes, respectively. Negative control region: chr12:122668133–122668172 (mm8). Positive control region: chr18: 46513245–46513285 (mm8).

Second, we performed a series of point mutations to a wild type sequence that matches the new Nanog motif. EMSA was again used to test the binding affinities of the mutated sequences. Since the “TGA” from position 2 to position 5 is the most conserved part of the new motif, we focused the point mutations to these three positions. First, mutating the “TGA” core of the motif completely abolished the binding. Second, except the “G to A” mutation on position 3, all the rest six point mutations to the “TGA” core severely reduced the binding or completely abolished the binding. All these results were reproduced by at least two independent EMSA experiments. We also tested the potential difference between the DNA binding domain (DBD) of Nanog and the whole Nanog protein. Nanog DBD and Nanog generated the same binding specificities in all of EMSA experiments. In summary, both the EMSA data on the five wildtype sequences and the point mutation data are consistent with the notion that Nanog bind to the the novel Nanog DNA motif.



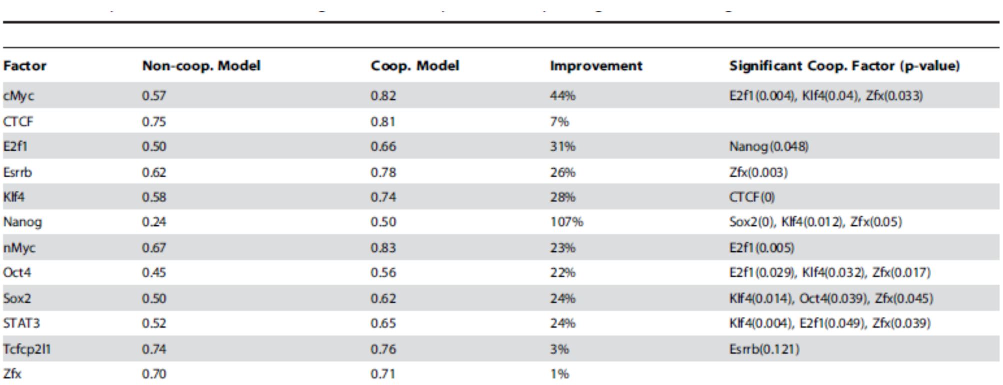
| | |
|---|---|
| CAACCAGCCCTTGATGGCCCTCCTTGATGG | WT1 |
| CAACCAGCCCTcacTGGCCCTCCTTGATGG | Mut1 |
| CAACCAGCCCTcGATGGCCCTCCTTGATGG | Mut1_1 |
| CAACCAGCCCTTaATGGCCCTCCTTGATGG | Mut1_2 |
| CAACCAGCCCTTGcTGGCCCTCCTTGATGG | Mut1_3 |
| CCTTTCAGCTCTGATGGGTTTCTTTCAGCT | WT2 |
| CCTTTCAGCTCcacTGGGTTTCTTTCAGCT | Mut2 |
| CCTTTCAGCTCcGATGGGTTTCTTTCAGCT | Mut2_1 |
| CCTTTCAGCTCTaATGGGTTTCTTTCAGCT | Mut2_2 |
| CCTTTCAGCTCTGcTGGGTTTCTTTCAGCT | Mut2_3 |

**Figure 4. Mutation results do not depend on the wild-type binding sites.** A subset of mutations was repeated on two independent wild-type sequences. EMSA results of these mutated sequences are shown. The two independent wild-type sequences in the mutagenesis analysis generated similar results.

**Cooperativity among TFs is frequently associated with DNA binding**

We next assessed the importance of cooperative interactions among TFs for DNA binding. For each ChIP-seq experiment, we compared the effectiveness of a “non-cooperative model” (that disables cooperative interactions) and a “cooperative model” that allows these interactions. To account for different complexities of the compared models, we used 10-fold cross validation on 500 sequences from each experiment, and measured performance as the average correlation coefficient between predictions and observations. For most factors, incorporating cooperative TF interactions substantially improved the predictive ability of the model (Table 1). In a different test, we trained the model on the strongest bound 500 sequences and used the next strongest 500 sequences as independent test data. Again, we found that the cooperative model explains the test data significantly better than the non-cooperative model. These results suggest that cooperative interactions are an important part of the process of TF-DNA binding.

**Table 1. Cooperative interactions among factors are important in explaining TF-DNA binding data.** In non-cooperative (non-coop.) model, only the motif of TFexp is used for fitting the data and no cooperativity is allows. In cooperative (coop.) model, both the motif of TFexp and the motifs of significant cooperative factors are used, and the cooperative interactions among factors, including the homotypic interaction, are allowed. The performance of a model is measured by the Pearson correlation between model predictions and observations in an independent testing data (not used for training the models). Significance of a cooperative factor is determined through comparison with a large number of randomized motifs. Only the factors with p value less than or equal to 0.05 are shown.



| Factor | Non-coop. Model | Coop. Model | Improvement | Significant Coop. Factor (p-value) |
|---|---|---|---|---|
| cMyc | 0.57 | 0.82 | 44% | E2f1(0.004), Klf4(0.04), Zfx(0.033) |
| CTCF | 0.75 | 0.81 | 7% | |
| E2f1 | 0.50 | 0.66 | 31% | Nanog(0.048) |
| Esrrb | 0.62 | 0.78 | 26% | Zfx(0.003) |
| Klf4 | 0.58 | 0.74 | 28% | CTCF(0) |
| Nanog | 0.24 | 0.50 | 107% | Sox2(0), Klf4(0.012), Zfx(0.05) |
| nMyc | 0.67 | 0.83 | 23% | E2f1(0.005) |
| Oct4 | 0.45 | 0.56 | 22% | E2f1(0.029), Klf4(0.032), Zfx(0.017) |
| Sox2 | 0.50 | 0.62 | 24% | Klf4(0.014), Oct4(0.039), Zfx(0.045) |
| STAT3 | 0.52 | 0.65 | 24% | Klf4(0.004), E2f1(0.049), Zfx(0.039) |
| Tcfcp2l1 | 0.74 | 0.76 | 3% | Esrrb(0.121) |
| Zfx | 0.70 | 0.71 | 1% | |

Many of the specific predictions of cooperative interaction listed in Table 1 are either known or consistent with evidence from the literature. Oct4 is a cooperative factor of Sox2, and both Oct4 and Sox2 are found to stimulate DNA-binding by Nanog. These results are consistent with the observation that Oct4, Sox2 and Nanog work together to control gene regulation in ES cells (Boyer et al., 2005). Similarly, the specific interaction between Esrrb and Nanog has been reported earlier in a study of protein-protein interactions among TFs in ES cells (Wang et al., 2006). We also found that Klf4 is cooperative with a number of other factors, including Oct4, Sox2, Nanog and STAT3. Interestingly, Klf4 has been recently found to be a key factor for maintaining self-renewal of ES cells (Jiang et al., 2008), through mechanisms that are not yet clear. Our results suggest that the cooperative interaction between Klf4 and other key TFs may underlie the function of Klf4.

We repeated the above analysis using motifs from the JASPAR database (Bryne et al., 2008), in addition to the motifs in this dataset. In the most interesting result from this analysis, we found the GABPA factor to cooperate with Oct4. (The correlation coefficients for the non-cooperative and cooperative models were 0.45 and 0.54 respectively.) GABPA expression is known to be induced in undifferentiated ES cells and its expression decreases during differentiation (Hailesellasse Sene et al., 2007). Moreover, GABPA has been shown to regulate the expression of Oct4 in mouse ES cells (Kinoshita et al., 2007). Thus, it would be interesting to test experimentally how GABPA is related to the function of Oct4. This is an example where our method can be utilized to automatically discover biologically plausible hypothesis from existing resources of DNA binding and motif data.

**STAP Improves Prediction of TF Targets over Existing Methods.**

An intended application of STAP is to use the learned binding model to predict affinities of unseen sequences to a set of TFs. An initial support to this application came from the results above showing incorporating cooperative interactions were more predictive than simple models without interactions (Table 1). We then compared STAP with the existing methods that are also capable of predicting TF target sequences. Two popular programs were chosen for this purpose, Cluster-Buster (Frith et al., 2003) and Stubb (Sinha, 2006). Both programs take a set of TF motifs as input, and predict if some binding site clusters appear in a test sequence. To use these programs to predict the targets of some TF, it was necessary to obtain the relevant motifs (in addition to the motif of this TF). Neither program provides such capabilities, and therefore we used another program Clover for this purpose (Frith et al., 2004). In summary, the executed procedure of applying these two programs was: first learn all overrepresented motifs using Clover from TF-bound sequences in the training data, and then classify all sequences in the test data using Cluster-Buster or Stubb (the same training and testing data as used in the previous section). We evaluated the classification performance with the standard ROC curves, which quantifies the tradeoff of specificity and sensitivity as the classification threshold varies.

Clover identified a number of overrepresented motifs from the collection of 12 motifs of the 12 assayed TFs. These results were similar to STAP's predictions in some aspects: both predicted few interacting factors for CTCF, E2f1 and Esrrb, and some pairs were predicted by both including Nanog-Sox2 and Tcfcp2l1-Esrrb. But Clover and STAP generated quite different results on other factors.We noticed that Clover results were largely parallel to the co-localization results in (Chen et al., 2008b), with Oct4, Sox2, Nanog and Esrrb forming a cluster of mutually interacting factors. Clover effectively identified motifs whose presence in the training sequences could not be explained by chance alone, regardless of whether these motifs actually facilitate binding of the primary factor. We comment on these different ways of defining ''interacting'' factors in Discussion. For now, this motif set was simply applied to predict TF targets by Cluster-Buster and Stubb. In almost all cases, STAP better classified the sequences in the testing data than the other two programs (see Figure 5 for the Oct4 result).

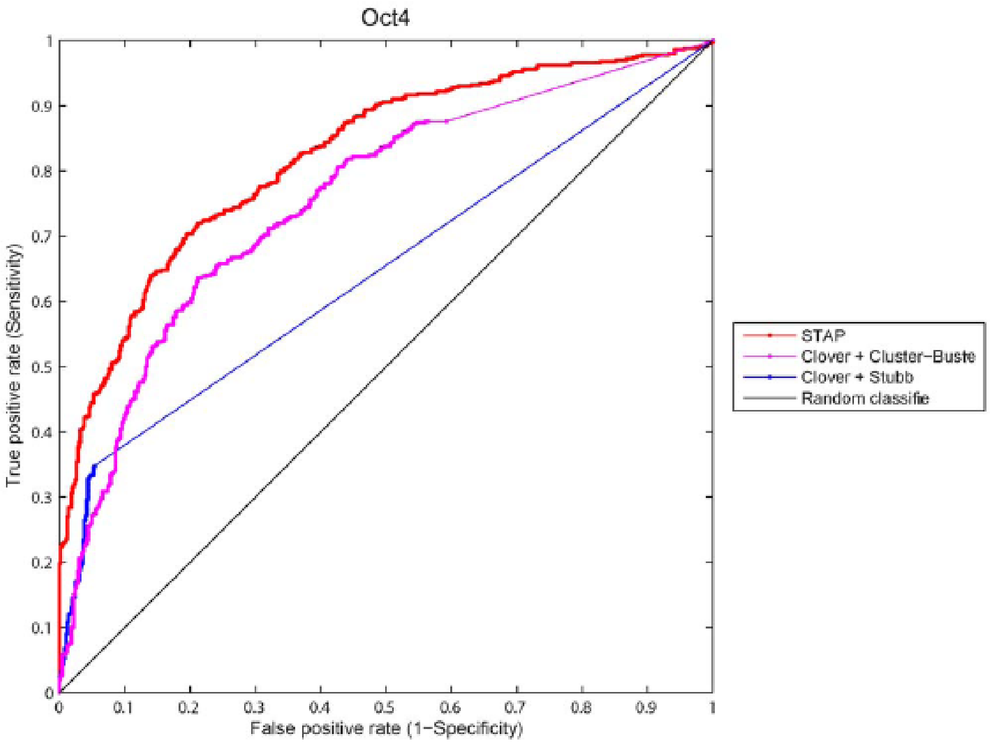


**Figure 5. ROC curves comparing the performance of three methods for classification of Oct4 target sequences in the ChIP-seq data of Oct4.** For evaluation of Cluster-Buster and Stubb, the Clover program is run first on the training data to extract a set of overrepresented motifs, which will be used as inputs of Cluster-Buster and Stubb.

### Exploring the effect of binding site arrangement

How binding sites are arranged in a regulatory sequence is an important, but poorly understood aspect of combinatorial gene regulation. Our biophysical model includes a component that describes how the strength of interaction between bound TF molecules depends on the arrangement (distance and relative orientation) of their respective binding sites. By varying the details of this component, we tested if the data supports a particular mode of TF interaction over others. Specifically, we compared three different models of cooperative interaction between two bound TF molecules. In each, we assume that there is a maximum distance *d*max between the two bound factors, beyond which there is no interaction. Under the “Binary” model, which is also our default model used in previous studies, the strength of interaction is constant within the range of 0 to *d*ma*x*. Under the “Linear” model, the interaction is stronger when the two cooperative sites are closer. For both Binary and Linear models, there may also be an orientation bias: the interaction when two factors bind in the same direction may be different from that when they bind in the opposite direction. The extent to which one orientation is favored over the other is encoded by a bias parameter. Finally, under the “Periodic” model, the strength of interaction is a periodic function of the distance. This periodicity has been reported in a few cases before and often corresponds to the helical period of DNA molecules (Makeev et al., 2003; Saiz et al., 2005). In all cases, a particular model is evaluated by the Pearson correlation between predictions and observations in an independent testing dataset, which is distinct from the one used for training the model parameters.

We focused on two TF pairs: Oct4-Sox2 and Nanog-Esrrb. Both interactions have been suggested before by earlier work (Chen et al., 2008a; Wang et al., 2006). First, for the Binary model of cooperative interaction, we vary the the *d*max parameter and for each value of *d*max, we optimize the orientation bias parameter and compare this optimized model with the one without bias. We found small orientation bias in most cases, in the range of 0.4-1.0 *kBT*, in terms of the free energy that penalizes one orientation over the other. For comparison, the energy of TF interaction falls in the range of 2.0-4.0 *kBT*. As further evidence of the lack of orientation bias, the performance of the models which optimized the bias parameter is very close to the one without bias (Figure 6A, B). The differences in terms of correlation are less than 1% in most cases. In contrast, the parameter *d*max plays a much larger role (Figure 6A, B). We found that most TF interactions occur in the range of 200 bp, but for Oct4-Sox2 pair, the majority of interaction seems to happen within 60 bp (Figure 6A). Next, we found that the Linear models did not improve the predictability (the differences between Linear model and Binary model are less than 0.5% for both pairs), suggesting that interaction between the two factors does not decrease significantly with distance, i.e. the interaction is tolerable to distance change. Finally, for the Periodic model, we vary the periodicity from 10.0 to 12.0 bp, and for each of these values, we also vary the amplitude parameter, which is a measure of the strength of periodicity, i.e. how greatly the interaction changes within a period (see Methods).

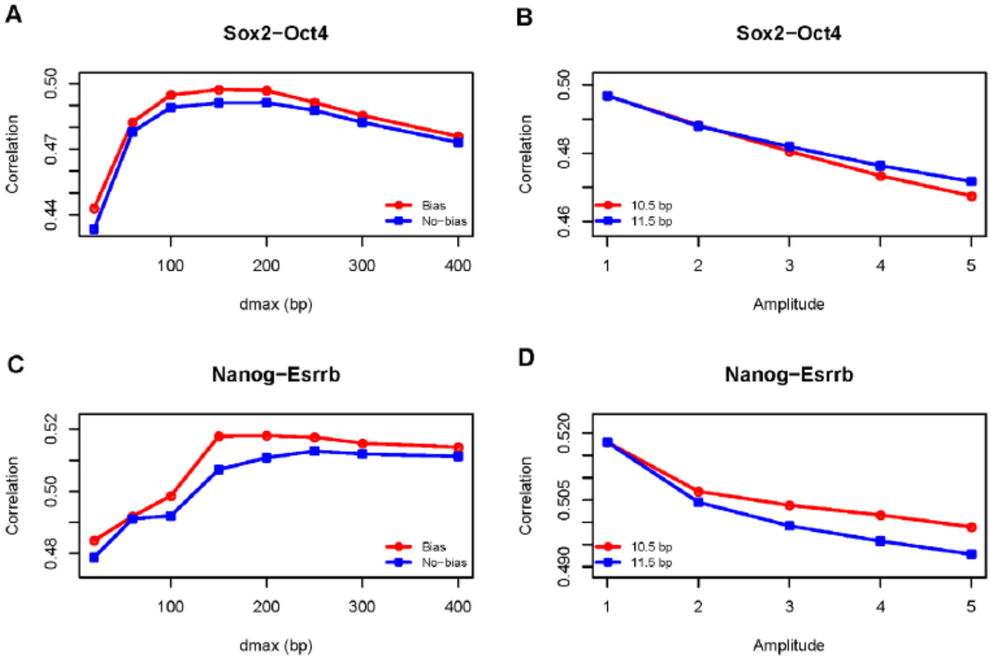


**Figure 6. The effect of binding site arrangement on TF interactions.** (A,C) Under the Binary model of interaction, the relationship between model performances, measured by correlation between predictions and observations, and the distance parameter (maximum distance, measured in bp, where two

factors can interact along DNA sequence). For each value of the distance parameter, two models are compared: one in which the orientation bias parameter is optimized, and the other not allowing the bias. (B,D) Under the Periodic model of interaction, the relationship between model performances and the amplitude parameter (the change of the interaction strength within a period). Only two values of periodicity are shown.

Similar to the results from the Linear model, we found that the correlations under this more complex model is no better than the simpler Binary model. In fact, the performance of the Periodic model always decreases when the amplitude parameter is increased under all values of periodicity we tested, suggesting that the interactions are not periodic for both pairs (Figure 6C, D, only two values of periodicity are shown). All these results: lack of orientation bias, tolerance to distance and lack of periodicity, together indicate that binding site interactions do not follow strict rules; rather, a flexible organization, within a certain distance, seems to be sufficient for enabling TF interactions.

## DISCUSSION

In this work, we adapted the theoretical models pioneered by Shea-Ackers (Shea and Ackers, 1985)and formulated by Buchler et al. (Buchler et al., 2003)to the analysis of large-scale TF binding data. Different from these previous works, we explicitly expressed the expected number of TFs bound by a given regulatory sequence, and thus derived a variation of the Shea-Ackers model suitable for analysis of genome-wide binding data. We developed a dynamic programming algorithm that efficiently computes the binding affinity of any sequence. We provided software, STAP, to automatically learn the best models from the binding data. Through extensive evaluations, we demonstrated that this is an effective computational framework to extract information from and extrapolate over TF-DNA binding data.

STAP was applied to several important analysis tasks, including comparison of TF binding profiles, identification of TF interactions, studying the effect of binding site arrangement (regulatory grammar) and prediction of TF target sequences. These tasks are commonly encountered in analysis of genome-wide data, and we believe STAP offers key benefits over existing methods. First, STAP was applied to compare several putative Nanog motifs. Such functionality can be useful, for example, when one needs to compare outputs from multiple motif-finding programs or from different experiments. Furthermore, when multiple factors access the same target regions, STAP is able to disentangle the effects of confounding factors. This was demonstrated in the analysis of Nanog-bound sequences, which are often bound by Oct4 and Sox2 as well. Second, we took advantage of the new method to predict TF-TF interactions. Similar analyses were done previously by first predicting the binding sites of the pair of motifs, and then analyzing the co-occurrence pattern of two types of sites (Smith et al., 2005a; Zhou et al., 2007a). Co-occurrence based analysis does not utilize the measured TFbinding intensities, sacrificing a significant amount of available information. Co-occurrence based analysis also requires the explicit annotation of binding sites, a task known for its inaccuracy. Weak binding sites were shown to contribute significantly to TF binding (Roider et al., 2007; Segal et al., 2008), making a binary demarcation of sites and nonsites more problematic. Thirdly, STAP was applied to test different regulatory rules for binding site arrangement. This task has been gaining attention from the community (Arnosti and Kulkarni, 2005; Brown et al., 2007), but a computational tool for addressing this challenge has been missing so far. Finally, we demonstrated that STAP is able to make more accurate predictions of TF targets in new sequences than other state-of-the-art programs. This capability enables the study of the evolution of TF binding across species despite that the binding data are often available in only one species. We also found that limiting to sequences with conserved affinities would improve the identification of functional TF targets.

The recent work by Segal et al. (Segal et al., 2008)also uses the thermodynamic model to predict the functional properties (expression patterns) of DNA sequences, and it is worthwhile to point out the similarity and the difference between the two papers. Both Segal et al. and this work rely on the same thermodynamic framework of Buchler et al. (Buchler et al., 2003)to model TF-DNA interactions as well as cooperative DNA binding by multiple TFs. In the algorithmic side, both use dynamic In this work, we adapted the theoretical models pioneered by Shea-Ackers (Shea and Ackers, 1985) and formulated by Buchler et al. (Buchler et al., 2003) to the analysis of large-scale TF binding data. Different from these previous works, we explicitly expressed the expected number of TFs bound by a given regulatory sequence, and thus derived a variation of the Shea-Ackers model suitable for analysis of genome-wide binding data. We developed a dynamic programming algorithm that efficiently computes the binding affinity of any sequence. We provided software, STAP, to automatically learn the best models from the binding data. Through extensive evaluations, we demonstrated that this is an effective computational framework to extract information from and extrapolate over TF-DNA binding data. STAP was applied to several important analysis tasks, including comparison of TF binding profiles, identification of TF interactions, studying the effect of binding site arrangement (regulatory grammar) and prediction of TF target sequences. These tasks are commonly encountered in analysis of genome-wide data, and we believe STAP offers key benefits over existing methods. First, STAP was applied to compare several putative Nanog motifs. Such functionality can be useful, for example, when one needs to compare outputs from multiple motif-finding programs or from different experiments. Furthermore, when multiple factors access the same target regions, STAP is able to disentangle the effects of confounding factors. This was demonstrated in the analysis of Nanog-bound sequences, which are often bound by Oct4 and Sox2 as well. Second, we took advantage of the new method to predict TF-TF interactions. Similar analyses were done previously by first predicting the binding sites of the pair of motifs, and then analyzing the co-occurrence pattern of two types of sites (Smith et al., 2005a; Zhou et al., 2007a). Co-occurrence based analysis does not utilize the measured TFbinding intensities, sacrificing a significant amount of available information. Co-occurrence based analysis also requires the explicit annotation of binding sites, a task known for its inaccuracy. Weak binding sites were shown to contribute significantly to TF binding (Roider et al., 2007; Segal et al., 2008), making a binary demarcation of sites and nonsites more problematic. Thirdly, STAP was applied to test different regulatory rules for binding site arrangement. This task has been gaining attention from the community (Arnosti and Kulkarni, 2005; Brown et al., 2007), but a computational tool for addressing this challenge has been missing so far. Finally, we demonstrated that STAP is able to make more accurate predictions of TF targets in new sequences than other state-of-the-art programs. This capability enables the study of the evolution of TF binding across species despite that the binding data are often available in only one species. We also found that limiting to sequences with conserved affinities would improve the identification of functional TF targets. The recent work by Segal et al. (Segal et al., 2008) also uses the thermodynamic model to predict the functional properties (expression patterns) of DNA sequences, and it is worthwhile to point out the similarity and the difference between the two papers. Both Segal et al. and this work rely on the same thermodynamic framework of Buchler et al. (Buchler et al., 2003) to model TF-DNA interactions as well as cooperative DNA binding by multiple TFs. In the algorithmic side, both use dynamic programming to optimize the computational task, which is also a familiar technique in statistical mechanics (known as the transfer matrix method), and has been used before for similar calculations involving cis-regulatory sequences (Hermsen et al., 2006; Teif, 2007). These similarities are not surprising as both attempts to capture the same underlying physics. There are two main differences. Segal et al. uses a logistic function as the expression ‘‘readout’’ of any molecular configuration (s in our notation) and predicts the expression of the sequence as the average readout over all configurations. The downside of this approach is that the logistic function has no connection to thermodynamics, and the computation involves expensive sampling. In this work, the relevant quantity we compute has a clear physical interpretation: the average number of TF molecules bound to the sequence. This also enables the derivation of dynamic programming, which is far more efficient than sampling. The other main difference lies in the intended applications of the models. STAP was applied to questions that were not addressed previously, such as the characterization of rules of cooperative interactions and evolution of TF-target relationship.

Combinatorial gene regulation by definition involves the relationship among different transcription factors. However, how such relationships should be defined and inferred is not clear in practice. We believe it is important to distinguish among three types of relationship between a pair of transcription factors: (A) co-localization of two factors as revealed by ChIP experiments; (B) direct binding of two factors to the neighboring DNA sites (co-binding) and (C) cooperative interaction of two factors bound in the neighborhood. Note that these three classes correspond to progressively more specific relationships. Colocalization of two TFs in a ChIP experiment may be due to cobinding, or due to one of the TFs being bound to DNA and recruiting the other TF (without the latter directly binding to DNA). Similarly, when two factors bind to adjacent sites on DNA (co-binding), they may not actually interact with each other, i.e. no cooperative interactions. The different results we obtained from our co-localization analysis, from motif enrichment test using Clover and from our identification of cooperative factors may partly come from these distinctions. This picture of a hierarchy in the relationships of TFs (in the context of DNA binding) suggests that it is important to interpret the results in a way that is appropriate for the type of analysis performed.

We assumed that cooperative interactions are due to proteinprotein interactions, but this may not always be true. For example, the factor B may stimulate DNA-binding of the factor A through chromatin modification that makes DNA more accessible. This point has also been commented before (Hermsen et al., 2006). It is difficult to distinguish different mechanisms of cooperative interactions when only DNA binding data is available. This is important for interpreting the results, as the predictions may not be confirmable through protein-protein interaction assays. In addition, this suggests that the cooperative interactions, as defined by stimulated effects of DNA binding on another factor, may not be symmetric. In the example we cited above, the factor A itself may not modify chromatin structure, thus has no effect on DNA binding affinity of the factor B.

We studied the effect of binding site orientation and relative distance on the cooperative TF interactions. Because the effect is likely to be subtle, we focused on the TF pairs with the strongest signals in the data. We did not found evidence supporting rigid rules, such as the periodicity of distance (in the range of period tested). This may suggest that the interactions may occur indirectly, rather than through physical protein-protein interactions, such as the well known case of lambda repressor (Hochschild and Ptashne, 1986). If a TF modifies the chromatin structure through chemical modifications of histones or remodeling of nucleosomes, the effect of this TF on other TFs will be less specific (as it could affect all binding sites in the neighborhood) and less likely to follow strict rules. We recognize there are several limitations in our methodology: only several forms of cooperative functions were tested while the actual function may be much more complex; and in the thermodynamic model, only immediately adjacent binding sites may interact with each other, an assumption taken for the ease of computation without much theoretical justification. These limitations coupled with the fact that only five TF pairs were tested in a single dataset limit our ability to extrapolate any general regulatory rules. Still, the STAP method is relatively sensitive, as demonstrated by the large effect of dmax and the amplitude parameters we observed (Figure 6), and represents one concrete step towards an important but difficult problem.

## BIBLIOGRAPHY:

Arnosti, D. N. and M. M. Kulkarni (2005). "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?" J Cell Biochem 94(5): 890-898.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol 2: 28-36.

Barski, A., S. Cuddapah, et al. (2007). "High-resolution profiling of histone methylations in the human genome." Cell 129(4): 823-837.

Barski, A. and K. Zhao (2009). "Genomic location analysis by ChIP-Seq." J Cell Biochem 107(1): 11-18.

Beer, M. A. and S. Tavazoie (2004). "Predicting gene expression from sequence." Cell 117(2): 185-198.

Berg, O. G. and P. H. von Hippel (1988). "Selection of DNA binding sites by regulatory proteins." Trends Biochem Sci 13(6): 207-211.

Bieda, M., X. Xu, et al. (2006). "Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome." Genome Res 16(5): 595-605.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature 447(7146): 799-816.

Blackwood, E. M. and J. T. Kadonaga (1998). "Going the distance: a current view of enhancer action." Science 281(5373): 60-63.

Boiani, M. and H. R. Schöler (2005). "Regulatory networks in embryo-derived pluripotent stem cells." Nat Rev Mol Cell Biol 6(11): 872-884.

Boyer, L. A., T. I. Lee, et al. (2005). "Core transcriptional regulatory circuitry in human embryonic stem cells." Cell 122(6): 947-956.

Boyer, L. A., K. Plath, et al. (2006). "Polycomb complexes repress developmental regulators in murine embryonic stem cells." Nature 441(7091): 349-353.

Brown, C. D., D. S. Johnson, et al. (2007). "Functional architecture and evolution of transcriptional elements that drive gene coexpression." Science 317(5844): 1557-1560.

Bryne, J. C., E. Valen, et al. (2008). "JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update." Nucleic Acids Res 36(Database issue): D102-106.

Buchler, N. E., U. Gerland, et al. (2003). "On schemes of combinatorial transcription logic." Proc Natl Acad Sci U S A 100(9): 5136-5141.

Bulyk, M. L. (2006). "DNA microarray technologies for measuring protein-DNA interactions." Curr Opin Biotechnol 17(4): 422-430.

Cawley, S., S. Bekiranov, et al. (2004). "Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs." Cell 116(4): 499-509.

Chambers, I., D. Colby, et al. (2003). "Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells." Cell 113(5): 643-655.

Chen, X., V. B. Vega, et al. (2008). "Transcriptional Regulatory Networks in Embryonic Stem Cells." Cold Spring Harb Symp Quant Biol.

Chen, X., H. Xu, et al. (2008). "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells." Cell 133(6): 1106-1117.

Chew, J. L., Y. H. Loh, et al. (2005). "Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells." Mol Cell Biol 25(14): 6031-6046.

Davidson, E. H. (2006). The Regulatory Genome: Gene Regulatory Networks in Development and Evolution, Academic Press.

Down, T. A. and T. J. Hubbard (2005). "NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence." Nucleic Acids Res 33(5): 1445-1453.

Ehret, G. B., P. Reichenbach, et al. (2001). "DNA binding specificity of different STAT proteins. Comparison of in vitro specificity with natural target sites." J Biol Chem 276(9): 6675-6688.

Evans, M. J. and M. H. Kaufman (1981). "Establishment in culture of pluripotential cells from mouse embryos." Nature 292(5819): 154-156.

Foat, B. C., A. V. Morozov, et al. (2006). "Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE." Bioinformatics 22(14): e141-149.

Fortunel, N. O., H. H. Otu, et al. (2003). "Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature"." Science 302(5644): 393; author reply 393.

Frith, M. C., Y. Fu, et al. (2004). "Detection of functional DNA motifs via statistical over-representation." Nucleic Acids Res 32(4): 1372-1381.

Frith, M. C., M. C. Li, et al. (2003). "Cluster-Buster: Finding dense clusters of motifs in DNA sequences." Nucleic Acids Res 31(13): 3666-3668.

Galan-Caridad, J. M., S. Harel, et al. (2007). "Zfx controls the self-renewal of embryonic and hematopoietic stem cells." Cell 129(2): 345-357.

Gertz, J., E. D. Siggia, et al. (2009). "Analysis of combinatorial cis-regulation in synthetic and genomic promoters." Nature 457(7226): 215-218.

Hailesellasse Sene, K., C. J. Porter, et al. (2007). "Gene function in early mouse embryonic stem cell differentiation." BMC Genomics 8: 85.

Heintzman, N. D., R. K. Stuart, et al. (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nat Genet 39(3): 311-318.

Hermsen, R., S. Tans, et al. (2006). "Transcriptional regulation by competing transcription factor modules." PLoS Comput Biol 2(12): e164.

Hochschild, A. and M. Ptashne (1986). "Cooperative binding of lambda repressors to sites separated by integral turns of the DNA helix." Cell 44(5): 681-687.

Howard, M. L. and E. H. Davidson (2004). "cis-Regulatory control circuits in development." Dev Biol 271(1): 109-118.

Ivanova, N., R. Dobrin, et al. (2006). "Dissecting self-renewal in stem cells with RNA interference." Nature 442(7102): 533-538.

Ivanova, N. B., J. T. Dimos, et al. (2002). "A stem cell molecular signature." Science 298(5593): 601-604.

Janssens, H., S. Hou, et al. (2006). "Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene." Nat Genet 38(10): 1159-1165.

Jiang, J., Y. S. Chan, et al. (2008). "A core Klf circuitry regulates self-renewal of embryonic stem cells." Nat Cell Biol 10(3): 353-360.

Jin, V. X., A. Rabinovich, et al. (2006). "A computational genomics approach to identify cis-regulatory modules from chromatin immunoprecipitation microarray data--a case study using E2F1." Genome Res 16(12): 1585-1595.

Johnson, D. S., A. Mortazavi, et al. (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science 316(5830): 1497-1502.

Kaczynski, J., T. Cook, et al. (2003). "Sp1- and Kruppel-like transcription factors." Genome Biol 4(2): 206.

Kim, T. H., Z. K. Abdullaev, et al. (2007). "Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome." Cell 128(6): 1231-1245.

Kinoshita, K., H. Ura, et al. (2007). "GABPalpha regulates Oct-3/4 expression in mouse embryonic stem cells." Biochem Biophys Res Commun 353(3): 686-691.

Lee, T. I., R. G. Jenner, et al. (2006). "Control of developmental regulators by Polycomb in human embryonic stem cells." Cell 125(2): 301-313.

Loh, Y. H., Q. Wu, et al. (2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." Nat Genet 38(4): 431-440.

Maherali, N., R. Sridharan, et al. (2007). "Directly Reprogrammed Fibroblasts Show Global Epigenetic Remodeling and Widespread Tissue Contribution." Cell Stem Cell 1: 55-70.

Makeev, V. J., A. P. Lifanov, et al. (2003). "Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information." Nucleic Acids Res 31(20): 6016-6026.

Martin, G. R. (1981). "Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells." Proc Natl Acad Sci U S A 78(12): 7634-7638.

Matsuda, T., T. Nakamura, et al. (1999). "STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells." Embo J 18(15): 4261-4269.

Merika, M., A. J. Williams, et al. (1998). "Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription." Mol Cell 1(2): 277-287.

Mikkelsen, T. S., M. Ku, et al. (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature 448(7153): 553-560.

Mitsui, K., Y. Tokuzawa, et al. (2003). "The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells." Cell 113(5): 631-642.

Nichols, J., B. Zevnik, et al. (1998). "Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4." Cell 95(3): 379-391.

Niwa, H., T. Burdon, et al. (1998). "Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3." Genes Dev 12(13): 2048-2060.

Ogryzko, V. V., R. L. Schiltz, et al. (1996). "The transcriptional coactivators p300 and CBP are histone acetyltransferases." Cell 87(5): 953-959.

Okita, K., T. Ichisaka, et al. (2007). "Generation of germline-competent induced pluripotent stem cells." Nature 448(7151): 313-317.

Panne, D., T. Maniatis, et al. (2007). "An atomic model of the interferon-beta enhanceosome." Cell 129(6): 1111-1123.

Pavesi, G., G. Mauri, et al. (2001). "An algorithm for finding signals of unknown length in DNA sequences." Bioinformatics 17 Suppl 1: S207-214.

Pettersson, K., K. Svensson, et al. (1996). "Expression of a novel member of estrogen response element-binding nuclear receptors is restricted to the early stages of chorion formation during mouse embryogenesis." Mech Dev 54(2): 211-223.

Press, W. H., B. P. Flannery, et al. (1992). Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press.

Ramalho-Santos, M., S. Yoon, et al. (2002). ""Stemness": transcriptional profiling of embryonic and adult stem cells." Science 298(5593): 597-600.

Raz, R., C. K. Lee, et al. (1999). "Essential role of STAT3 for embryonic stem cell pluripotency." Proc Natl Acad Sci U S A 96(6): 2846-2851.

Robertson, G., M. Hirst, et al. (2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." Nat Methods 4(8): 651-657.

Roider, H. G., A. Kanhere, et al. (2007). "Predicting transcription factor affinities to DNA from a biophysical model." Bioinformatics 23(2): 134-141.

Saiz, L., J. M. Rubi, et al. (2005). "Inferring the in vivo looping properties of DNA." Proc Natl Acad Sci U S A 102(49): 17642-17645.

Segal, E., T. Raveh-Sadka, et al. (2008). "Predicting expression patterns from regulatory sequence in Drosophila segmentation." Nature 451(7178): 535-540.

Shea, M. A. and G. K. Ackers (1985). "The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation." J Mol Biol 181(2): 211-230.

Sinha, S. (2006). "On counting position weight matrix matches in a sequence, with application to discriminative motif finding." Bioinformatics 22(14): e454-463.

Smith, A. D., P. Sumazin, et al. (2005). "Mining ChIP-chip data for transcription factor and cofactor binding sites." Bioinformatics 21 Suppl 1: i403-412.

Smith, A. D., P. Sumazin, et al. (2005). "Identifying tissue-selective transcription factor binding sites in vertebrate promoters." Proc Natl Acad Sci U S A 102(5): 1560-1565.

Smith, A. G. (2001). "Embryo-derived stem cells: of mice and men." Annu Rev Cell Dev Biol 17: 435-462.

Smith, A. G., J. K. Heath, et al. (1988). "Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides." Nature 336(6200): 688-690.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics 16(1): 16-23.

Struhl, K. (2001). "Gene regulation. A paradigm for precision." Science 293(5532): 1054-1055.

Struhl, K. (2007). "Transcriptional noise and the fidelity of initiation by RNA polymerase II." Nat Struct Mol Biol 14(2): 103-105.

Takahashi, K. and S. Yamanaka (2006). "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." Cell 126(4): 663-676.

Tanay, A. (2006). "Extensive low-affinity transcriptional interactions in the yeast genome." Genome Res 16(8): 962-972.

Teif, V. B. (2007). "General transfer matrix formalism to calculate DNA-protein-drug binding in gene regulation: application to OR operator of phage lambda." Nucleic Acids Res 35(11): e80.

Thanos, D. and T. Maniatis (1995). "Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome." Cell 83(7): 1091-1100.

Thomas, K. R. and M. R. Capecchi (1986). "Introduction of homologous DNA sequences into mammalian cells induces mutations in the cognate gene." Nature 324(6092): 34-38.

Thompson, W. A., L. A. Newberg, et al. (2007). "The Gibbs Centroid Sampler." Nucleic Acids Res 35(Web Server issue): W232-237.

Wang, J., S. Rao, et al. (2006). "A protein interaction network for pluripotency of embryonic stem cells." Nature 444(7117): 364-368.

Wathelet, M. G., C. H. Lin, et al. (1998). "Virus infection induces the assembly of coordinately activated transcription factors on the IFN-beta enhancer in vivo." Mol Cell 1(4): 507-518.

Wernig, M., A. Meissner, et al. (2007). "In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state." Nature 448(7151): 318-324.

Williams, R. L., D. J. Hilton, et al. (1988). "Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells." Nature 336(6200): 684-687.

Ying, Q. L., J. Nichols, et al. (2003). "BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3." Cell 115(3): 281-292.

Yu, X., J. Zou, et al. (2008). "Notch signaling activation in human embryonic stem cells is required for embryonic, but not trophoblastic, lineage commitment." Cell Stem Cell 2(5): 461-471.

Zeller, K. I., X. Zhao, et al. (2006). "Global mapping of c-Myc binding sites and target gene networks in human B cells." Proc Natl Acad Sci U S A 103(47): 17834-17839.

Zhang, Z. D., A. Paccanaro, et al. (2007). "Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions." Genome Res 17(6): 787-797.

Zhou, Q., H. Chipperfield, et al. (2007). "A gene regulatory network in mouse embryonic stem cells." Proc Natl Acad Sci U S A 104(42): 16438-16443.

Zhou, Q. and J. S. Liu (2008). "Extracting sequence features to predict protein-DNA interactions: a comparative study." Nucleic Acids Res 36(12): 4137-4148.

Zinzen, R. P. and D. Papatsenko (2007). "Enhancer responses to similarly distributed antagonistic gradients in development." PLoS Comput Biol 3(5): e84.

Zinzen, R. P., K. Senger, et al. (2006). "Computational models for neurogenic gene expression in the Drosophila embryo." Curr Biol 16(13): 1358-1365.