Epigenetic Regulation of Pluripotent and Multipotent Stem Cell Systems

Stephanie Lauren Battle

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

R. David Hawkins, Chair

Cole Trapnell

C. Anthony Blau

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Epigenetic Regulation of Pluripotent and Multipotent Stem Cell Systems

Stephanie Lauren Battle

Chair of the Supervisory Committee:
Associate Professor R. David Hawkins
Division of Medical Genetics, Department of Medicine
Department of Genome Sciences

The epigenome defines what a cell has the potential to do and allows a cell to elicit a response to external cues. This is exemplified in organisms where all the hundreds of different cell types share one genome but differ in their gene expression and response to environmental stimuli. Stem cells with their inherent ability to self-renew and differentiate, are a useful model for studying how the epigenome confers these attribute in a cell. Here I present research performed in two stem cell systems, pluripotent naïve human embryonic stem cells and multipotent ovarian cancer stem cells. I use different next-generation sequencing techniques to quantify changes in gene expression and measure three epigenetic characteristics: DNA methylation, histone modifications and genome three-dimensional architecture. From the data generated, I am able to identify regions that regulate gene expression, identify proteins that potentially target these regions, and predict the genes targets of regulatory regions. I have shown that the pluripotent naïve human embryonic stem cell line, Elf1, has a more open chromatin structure than primed embryonic stem cells, a known feature of early development. This is observed in their less methylated genome and presence of broad domains of active chromatin modifications. I found that architectural differences between the naïve and primed genome can be explained through the presence of histone modifications. I observed that Elf1 naïve embryonic stem cells are also a good model for development due to their ability to gain DNA methylation at imprinted regions and gain repressive histone modifications

at key genes when pushed forward to a primed-like state. In our multipotent stem cell system, I was able to identify key genes and protein interactions that distinguish ovarian cancer stem cells from the cells in the bulk of the tumor. I was able to link differentially methylated regions to regulatory elements and identify putative gene targets. Many of the interacting proteins and gene targets have previously been shown to be responsible for chemotherapy resistance and quiescence in cancer stem cells. I uncover evidence that ovarian cancer stem cells use pluripotent cell transcription factors at their regulatory elements, creating a surprising and unexpected connection between the two stem cell systems in this study. There is still much to learn about the epigenetic regulatory network in pluripotent and multipotent stem cells and the work presented here can be viewed as a launching pad for future studies.

Table of contents

List of Figures

List of Tables

Acknowledgments

This thesis has been the culmination of my own efforts and the vast village that has supported me over the years. I begin by thanking my academic department, Genome Science of the University of Washington. That also extends to the staff in the Division of Medical Genetics and Molecular Medicine Training Program. Their collective support has been greatly appreciated. A very special thank you goes to the Ford Foundation for their funding and supportive family of Ford Fellows. I would like to thank my committee for their intellectual contributions to my development as a researcher. To David Hawkins, my advisor who has steered me through my graduate research, thank you for your tireless efforts to see me through this process. I would also like to thank my lab mates for their contributions to this work and the memories we've made. Last, but not least, I would like to thank the larger village that has helped mold me into the person I am and continued to support along the way - my husband, family and friends. Thank you for listening, laughing, caring, cheering, encouraging, and loving me through the best and worst of this journey. Now, as I come to close this chapter, a new one begins - thus, I would like to dedicate this thesis to my soon-to-be born child. Perhaps someday you'll read this and know (even better) what you are made of!

Chapter 1:
Introduction

1.1 Naïve vs Primed hESC

1.1.1 mESCS are unlike hESCs

After the derivation of mouse embryonic stem cells (mESCs) in 1981, the derivation of human ESCs (hESCs) came in 1998, over a decade later (commonly used acronyms and terms defined in Table 1.1). Previous attempts in 1994 failed to derive hESCs because human cells were derived in mouse conditions, serum+Lif[3]. However the 1998 study used culture conditions that were originally teased apart in a study performed on Rhesus macaque in 1995 and derived hESCs from donated blastocysts from IVF (*in vitro* fertilization)[4,5]. The hESCs were grown in the same culture condition as mouse, containing fetal calf serum, and were shown to form teratomas containing all three germ layers[5]. Like mESCs, hESCs also had active alkaline phosphatase (AP) activity but lacked the cell surface marker SSEA-1 a characteristic of mESCs and mouse ICM but did express surface markers SSEA-3 and SSEA-4, which were not expressed in mouse[5]. It is worth noting that the male line H1 and the female lines H7 and H9, were derived in the original 1998 paper and are still widely used today. Additional lines HES-1 and HES-2 were derived in 2000[6]

From their derivation, hESCs and mESCs were different. This was apparent from their morphology, growth conditions and other features. The big question in the stem cell field at the time was were these differences due to fundamental, molecular attributes of the species or not? Follow up papers further teased apart the molecular differences. Wnt pathway activation helps to sustain pluripotency in both mESCs and hESCs[7]. Lif supplementation and subsequent STAT3 phosphorylation are not able to support hESC growth in feeder-free conditions unlike in mESCs[8,9] indicating different signalling pathways are required for each cell type. Human ESCs pluripotency depended on SMAD2/3 activation through TGFβ signalling via ligands Activin and Nodal [7]. Inhibition of SMAD2/3 activation via this pathway leads to loss of pluripotency genes *POU5F1*(also known by its protein name OCT4) and *NANOG* in hESCs but leads to no effect in mESCs[7].

Term	Description	Ref
2iL	Naïve mESC growth condition; represents the "ground state" of pluripotency; 2i are the two inhibitors Meki and Gsk3i while “L” stands for Lif supplementation	Silva, 2008
2iL+IF	Elf1 line naïve hESC growth condition	Sperber, 2015
5hmC	5-hydroxymethylcytosine; DNA hydroxymethylation	
5mC	5-methylcytosine; DNA methylation	
AP	Alkaline phosphatase activity is used as readout of pluripotent stem cells	
CTS	Cell-type specific	
DHS	DNase I hypersensitive site; region of DNA that is not bound by TFs or protein, not incorporated into a nucleosome and is accessible for DNase I enzymatic cleavage	
epiblast	The cells of the developing blastocyst that will give rise to three germ layers of the embryo proper	
EpiSC	Mouse epiblast stem cells, derived from postimplantation embryonic cells	Brons, 2007; Tesar, 2007
ESC	Embryonic stem cell; mouse ESCs denoted mESCs; human ESCs denoted hESCs	
Hi-C	Chromatin conformation capture assay that assess genome-wide chromatin interactions	Lieberman-Aiden, 2009
hypermethylation	Genomic region or nucleotide that on average contains more or higher levels of 5mC relative to a comparison region or nucleotide	Lister, 2009
hypomethylation	Genomic region or nucleotide that on average contains less or lower levels of 5mC relative to a comparison region or nucleotide	Lister, 2009
KD	Knockdown	
KO, DKO	Knockout, double knockout	
naïve	ESCs that represent preimplantation embryonic cells	Silva, 2008; Nichols, 2009
NPC	neural progenitor cells; early progenitor cells that can be made from ESC differentiation *in vitro*	
OSN	OCT4, SOX2, NANOG; master regulators of stem cell pluripotency; may refer to either gene expression or protein level	
primed	ESCs that represent postimplantation embryonic cells	Silva, 2008; Nichols, 2009
reset cells	Primed ESCs that have been "pushed" backwards to a naive state, usually through growth media changes	
serum/Lif	mESC growth condition that does not fully represent the naive state however cells can still contribute to the developing mouse preimplantation embryo and generate an adult mouse	Silva, 2008
TF	Transcription factor	

Table 1.1 List of Terms
Definitions of terms and acronyms commonly used throughout this manuscript.

1.1.2 Primed ESCs represent the postimplantation epiblast

In 2007, two research groups derived a new line of mouse ESC from a later stage of embryonic development. ESCs from the postimplantation epiblast, E5.5-E5.75, were isolated and cultured in Activin and FGF2[10,11]. Preimplantation ICM cells could not grown in these conditions but postimplantation grew well[11]. The cells were termed EpiSCs, for epiblast stem cells, due to their cell-type of origin and had a striking resemblance to hESCs[10,11]. EpiSCs in culture were morphologically similar to hESCs with large, flattened colonies compared to mESCs which had smaller, domed colonies[10,11]. EpiSCs are dependent on Activin/Nodal signalling (via TGFβ pathway) and can not be passaged as single cells by trypsin treatment, much like hESCs[10,11]. EpiSCs are very inefficient at producing chimaera animals when injected in preimplantation blastocysts[10,11] likely due to more limited potential since they were derived from later stage of development. This is in contrast to mESCs which were derived from the preimplantation blastocysts and can re-incorporate into the ICM of a developing embryo and can contribute to any cell lineage of the adult organism. However, EpiSCs still prove to be pluripotent cells through their ability to generate teratoma and EBs with differentiation to all three germ layers[10,11]. They express OSN at similar levels to mESCs[10] but do not express Zfp42/Rex1[11], a mESC marker. Thus the stem cell community had evidence to suggests that hESCs reflected a later stage from traditional mESCs. This later stage was termed the "primed" state of pluripotency, representing the postimplantation embryo and the earlier staged mESCs that represented the preimplantation state were termed "naïve"[12,13].

In 2009 additional evidence supported that hESCs did not reflect the preimplantation blastocyst, when a research group compared the gene expression profiles of different hESCs lines (H9, HSF-1,HSF-6) to ICM blastomeres isolated from 5 human embryos[14]. Hierarchical clustering of the microarray gene expression data showed that the hESCs lines were distinct from ICM and the hESC lines clustered by sex and not by time or place (institution) of derivation[14]. They detected 4,821 genes expressed in the ICM and >75% of them were also expressed in hESCs lines, although the difference in expression level was not investigated[14]. This is even more interesting given that these hESC lines (H9, HSF-1, HSF-6) were previously shown to be very distinct in their gene expression. Of the >14, 000 genes expressed between

all 3 hESCs lines, only ~50% were expressed in all 3 and many of those genes were >2-fold differentially expressed between the cell types[15]. Even with their differential gene expression, hESC line are able to form EBs and differentiate into all 3 germ layers[15]. This provides evidence that the drivers of pluripotency are not solely directed by gene expression, but other factors outside of expression.

Functional validation of primed hESCs representing the postimplantation stage of development came later. In 2012, it was shown that EpiSCs could integrate into different regions of the egg cylinder, proliferate, and adopt integration site-specific cellular markers of *in vitro* cultured post-implantation embryos[16]. EpiSCs could not integrate into embryos older than E7.5 possibly due to reduced pluripotency of epiblast cells as the embryo develops[16]. Later, in 2016, primed H9 hESCs, when injected into early (E6.5-E6.75) or late (E7.5-E7.75) gastrulation mouse embryos and were able to form chimeras 70% to 100% of the time[17]. Primed hESCs were able to contribute to many different subregions of the mouse embryo[17]. These experiment provided functional validation that primed ESCs represent postimplantation embryo.

1.1.3 Derivation of naïve hESCs

The gold standard assay for determining if a stem cell is truly naïve is to put it back into the developing embryo and observe if it can contribute to formation of all tissues of the adult organism. For ethical reasons we would not want to perform these experiments on human embryos therefore scientists have to come up with other metrics to measure the naïvety of hESCs. Since the derivation of naïve hESCs is fairly recent, identifying the minimal medium for their growth is still underway, although most published protocols give definitive rationale for why their growth conditions are the most ideal (list of naïve hESC publications and growth conditions summarized in Table 1.2).

The ability to induce a more naïve state in hESC was illustrated through transgene expression. A number of early studies used human iPSCs (induced pluripotent stem cells) manipulation to achieve naviety. Li *et al.* in 2009 created hiPSCs from transgene expression of *OSN* and *LIN28* in fibroblasts[18]. When grown in MEKi, GSK3i, TGFβ receptor inhibitor and human Lif supplement, the hiPSCs adopted a more naïve

mESC-like morphology, hypomethylated endogenous *POU5F1* (OCT4) promoter, and expression of endogenous *OSN*, *ZFP42* (*REX1*), *TDGF2* and *FGF4*[18]. Expression of endogenous genes indicates that transcriptional regulation in the cell was also reprogrammed and cells had adopted lesser reliance on transgene expression. In 2010, Buecker et al. made naïve hiPSCs through reprogramming of fibroblasts with *OSN*, *c-MYC*, and *KLF4* which developed morphologically naïve colonies in the presence of Lif[19]. These cells had active Lif pathways, as shown by STAT3 phosphorylation, but did not gain expression of endogenous genes[19]. ChIP-qPCR of gene promoters showed many pluripotency genes such as *SOX2*, *DNMT3b*, and *SALL4* were bivalently marked (H3K4me3 and H3K27me3 discussed in depth in Section 1.3.6) while *ZFP42*, *POU5F1*, and *NANOG* are H3K27me3 repressed at their promoters[19]. Interestingly, *POU5F1* promoter was hypomethylated, suggesting that some epigenetic reprogramming had occurred however cells were not truly Lif dependent and Lif withdrawal caused changes in cell morphology but not differentiation[19].

These early studies were instrumental in proving the human pluripotent cells could exist in the naïve state like mouse and that slightly different growth conditions might be required to reach the human naïve state. However, complete reprogramming of the epigenome is not achieved in iPSCs. As a cell transitions down Waddington's Epigenetic landscape, chemical modifications are put in place to remind a cell of its new function. Reprogramming attempts to erase these epigenetic memories and move a cell backwards in development. By the time a cell is fully differentiated, like a fibroblast, there may already be too many epigenetic hurdles in the way to fully achieve naivety. Therefore, starting from primed ESCs should be easier to achieve the naïve state as less epigenetic barriers and developmental time exists between the two states. Discussed next are attempts to "push" primed ESCs backwards into the naïve state and also successful derivation of naïve hESCs from human embryos.

In order to dissect the best conditions to push backwards or "reset" a primed ESCs, many studies utilized iPSCs or primed hESCS expressing transgenes containing factors relevant in reprogramming primed to naïve state[18-24] (Table 1.2). Exogenous expression via transgenes can force naïve morphology in primed cells. The induction agent, which is generally doxycycline (DOX), can turn on tetracycline inducible

transgenes. DOX can be easily washed off cells and removed from media and the transgene turns off, thus creating a easy reporter system to screen for compounds that will induce naïve state: use DOX to turn on transgene and push primed ESCs to naïve state, add a series of chemical compounds (mostly inhibitors), remove DOX and observe which chemicals maintain the naïve state. Every paper that has tried to figure out the ideal human naïve growth conditions, uses morphology as the first readout of naivety. Cells must form the rounded, domed colony structure similar to naïve mESCs. This is the first inspection for naïvety as cells in the preimplantation ICM have a mounded, clump-like structure in the blastocyst. Postimplantation the cells arrange themselves into a different shape forming either the cup-like egg cylinder in mouse or flattened embryonic disk in humans and other mammals, a more planar structure[25]. After that a number of different factors may be investigated. Naïve hESCs, when injected into early embryos should be able to contribute to the developing ICM of mouse blastocysts and form embryonic chimeras. If naïve lines are female, they should not have undergone X inactivation and should have two active X chromosomes. Naïve cells should have a faster doubly time than primed cells and should be able to be passaged as single cells after trypsinization. Most studies also compare the transcriptional profiles of naïve cells to human embryo ICM data. When available, the data source of the human embryo transcriptome is reference in text. All ESCs are expected to be able to form EBs *in vitro* and teratomas *in vivo*, but this is not a distinguishing feature of specifically naïve or primed cells.

Hanna *et al.* (2010) started with transient expression of *POU5F1* (OCT4), *SOX2* and *KLF4* and were able to sustain the naïve state in hESCs by growing them in MEKi, GSK3i with added hLif and Forskolin[20]. The cells were sensitive to JAK inhibition and had high level of phosphorylated STAT3 protein showing a reliance on the JAK/STAT pathway[20]. They also exhibited upregulation of genes shown to be upregulated in naïve mESCs compared to primed, including *KLF4*, *KLF2*, *TBX3*, *GBX2*, *LIN28*, and *SOCS3*[20]. In 2013, naïve hESCs were reset to primed using MEKi, GSKi, AMPKi with Lif supplementation, termed 3iL (referred to as 3iL-AMPKi in this text)[26]. These reset hESCs were also dependent on Lif, FGF, PI3K, and Activin signaling, and expressed genes that were seen upregulated in human preimplantation single cell RNA-seq data from Yan et al.: *NANOG* (2x as much as primed cells), *DPPA3*, *KLF4*, *TBX3*. Takashima *et al.* (2014) reset primed H9 hESCs through exogenous transient overexpression of *KLF2* and *NANOG* and

stable growth in MEKi, PKCi and a lower concentration GSK3i than typically used, along with hLif supplementation termed t2i+Gö[21]. These reset H9 hESCs had upregulation of stem cell factors *POU5F1*, *TBX3*, *ZFP42*, *DPPA3* (*STELLA*), *TFCP2L1*, *KLF4*, *GBX2* and *SALL4* compared to primed cells[21]. Reset H9 also expressed endogenous *NANOG* and *KLF2* without transgene expression and cluster with naïve mESCs, but not human ICM, according to their transcriptome[21,27]. Reset t2i+Gö cells are insensitive to FGFR or TGFβR inhibition but did have nuclear localization of TFE3, a feature of naïve cells[21].

Several research groups also undertook the task of deriving new hESCs line from embryos, eliminating any potential epigenetic memory of once being a primed cell. Gafni *et al.* (2013) used iPSCs with an inducible transgene expressing *POU5F1*, *SOX2*, KLF4, and *MYC* to screen for compounds that could stabilize the naïve state[22]. They determined the optimal growth condition to be MEKi, GSKi, p38i, JNKi with Lif, FGF2, and TGFβ supplementation and optional ROCKi and PKCi inclusion[22]. They named this medium naïve human stem cell medium or NHSM. Blastomeres from human blastocysts were plated on mouse feeders in NHSM in order to derive the naïve lines LIS1, LIS2, WIS1, and WIS2[22]. The naïve lines passed the usual tests of naivety (mentioned previously) and could be passaged at least up to p33 (WIS1) with normal karyotype[22]. Naïve NHSM hESCs were transcriptionally more similar to naïve mESCs and human ICM[28] and had upregulation of key genes like *NANOG*, *TEAD4*, *CD44* and DUSP family genes but not *KLF4*, *TBX3*, *PRDM14* or *STAT3* as observed in other naïve hESCs[22]. NHSM hESCs had downregulation of *DNMT3a*, *DNMT3b*, and *DNMT3L* which is in contrary to what was observed in human embryo single cell transcriptomes[22,29].

A few months later in 2014, another naïve hESC line was derived. Ware *et al.* first tested their naïve conditions by resetting H1, H9, HUES1, HUES2, mEpiSC5, and mEpiSC7 primed ESCs. They found that efficient resetting occurred after treating primed cells with an HDACi (histone deacetylase inhibitor)[30,31]. Mouse lines could be reset to 2iL after HDACi treatment but human lines grew best in 2i with FGF supplementation (2iF)[31]. Ware *et al.* started with an 8-cell embryo, allowed it to develop *in vitro*, plated the blastomeres from blastocysts in 2iF, and mechanically passage cells to pick naïve looking colonies[31]. After a few passages in naïve media, they established the line Elf1 and converted them to a different

growth condition contain MEKi, GSK3i, FGFRi with hLif supplementation (also called 3iL in the published work)[31]. Naïve Elf1 cells were stable in culture for up to 60 passages in 2iL or 3iL, have the expected naïve morphology and shorter doubling time in culture compared to primed hESCs[31]. Elf1 cells grown in primed conditions containing Activin and FGF (AF) were able to reset back to naïve growth conditions without HDACi treatment suggesting that the derivation protocol is important in maintain pluripotent flexibility of hESCs[31]. Naïve Elf1 cells are sensitive to STAT3i, an indication of their dependence on LIF/STAT pathway. Microarray transcriptomic analysis data suggests Elf1 cells are more similar to naïve mESCs and diapause mESCs than primed cells[31]. Naïve mESCs (2iL) and naïve Elf1 grown in an updated culture conditions, 2iL+IF (described in next paragraph), have higher mitochondrial respiration and lower glycolytic activity than primed EpiSCs and hESCs[2,32]. This correlates with previous data which showed in a comparison between naïve 2iL mESCs, primed hESCs, and EpiSCs that the mitochondrial respiratory pathway is repressed in primed ESCs as evident by lower expression of mitochondrial complex IV cytochrome c oxidase family genes and by the widespread cellular death that occurs when glycolysis is inhibited[32]. In spite of their reliance on glycolysis, primed mitochondria look more mature having an elongated shape with more developed cristae while naïve mitochondria are more rounded and have less cristae[32].

Later that same year another group published an additional naïve hESC with yet another set of growth conditions. Theunissen *et al.* used primed hESCs (WIBR2 and WIBR3) expressing inducible transgene of *KLF2* and *NANOG* to screen for compounds ideal to maintain naivety without transgene expression[23]. The hESCs also contain a GFP reporter, controlled by the *POU5F1* distal enhancer, a naïve-specific enhancer element (also mentioned in the histone H3K4me1 section below). From this assay and subsequent refining of growth conditions they settled on the 5iLA media, MEKi, GSKi, BRAFi, ROCKi, SRCi with hLif and Activin A supplementation[23]. They then thawed human 8-cell embryos or blastocysts in 5iLA media and derived the WIN1 naïve hESC line. Unlike the previous naïve conditions, “5i” naïve cells could not contribute to the ICM of a developing mouse embryo. However this experiment was performed using reset hESCs, not the naïve derived WIN1 line, and grown in a slightly different growth condition 5iLA+FGF[23]. WIN1 cells could also be grown with JNKi in the 6iLA growth conditions.

Hierarchical clustering of WIN1 cells grown in 5iLA or 6iLA conditions, reset 6iLA naïve cells and naïve mESCs shows these cell types are more similar to each other at the transcription level than they are to primed hESCs and mEpiSCs[23]. WIN1 cells were more similar to each other regardless of growth condition, suggesting the addition of the JNKi inhibitor may not be necessary for this cell line. Genes upregulated in this naïve state include *NANOG*, *TFCP2L1*, *KLF5*, *KLF4*, *ZFP42*, *DPPA2*, *DPPA3*, and *DPPA5*. Later in 2016, this research group determined that removing an additional inhibitor from their cocktail would provide slightly better proliferation of naïve cells[33]. The 4iLA media lacks GSK3i but still contains MEKi, BRAFi, ROCKi, SRCi with hLif and Activin A. Reset 4iLA hESCs, although having a slightly flatter morphology, have similar transcriptomic profile to 5iLA[33].

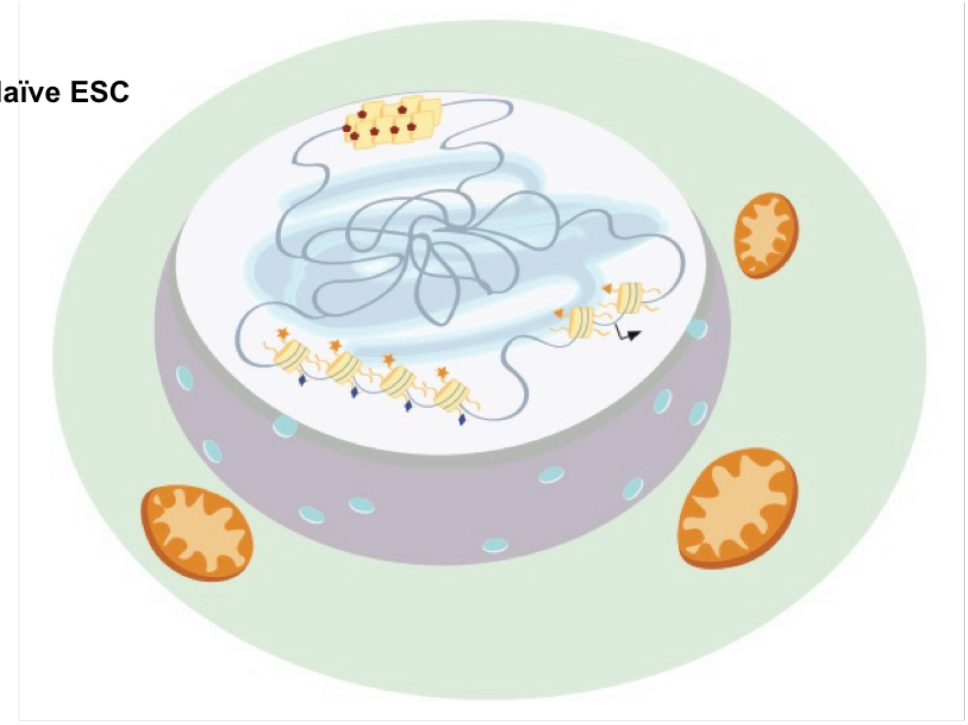
Between 2013 and 2014, the majority of original naïve hESC papers came out. Since then researchers have further refined their culture conditions and continued to derive additional naïve hESC lines. In 2016, Guo *et al.* (the same groups from Takashim 2014) published a derived line using their t2iL+Gö cocktail, but with the addition of ROCKi, and called their new growth media t2iLGöY[34]. They plated the ICM from human blastocysts 6 days postfertilization and manually picked naïve-morphed colonies for ~20 passages until they established the HNES1, HNES2, HNES3, HNES4 lines[34]. HNES1 was a karyotypically normal line, therefore the majority of their characterization focused on this line. The t2iLGöY naïve line upregulated *DPPA3*, *KLF4*, *TFCP2L1*, *KLF17*, *NANOG*, *TET1*, *DNMT3L*, and *TBX3* genes in a manner similar to reset t2i+Gö H9 [34] as compared to primed H9 hESCs .

In addition to the various inhibitor and supplement combinations, it is important to note that many of the naïve iPSCs, reset and derived hESCs were also grown in slightly different basal media. Whichever basal media chosen, all had been previously used to grow stem or progenitor cells. Different growth cocktails may behave differently depending on the basal medium used, in addition to how cells are derived.

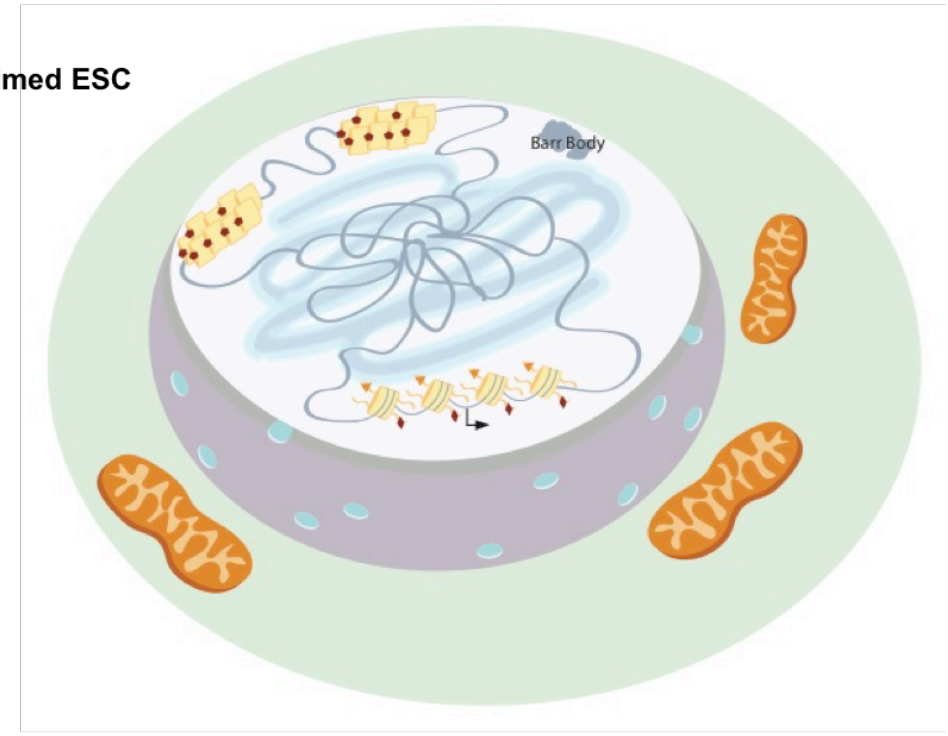
Starting from an 8-cell embryo and allowing it to develop in culture versus starting with a blastocyst may also influence the pluripotency state researchers are able to capture *in vitro*. Oxygen level during derivation and growth may also influence naïve state of hESCs. Lengner et al (2010) were able to derive

new hESCs line from embryos that maintain XaXa phenotype[35]. Cells derived from 8-cell embryo under physiological oxygen level, 5% O_2, maintained two active X chromosome as determined by XIST promoter methylation, lack of XIST expression and absence of XIST immunostained clouds compared to the same lines grown at atmospheric levels of O_2 (20%)[35].

Naïve ESC



Primed ESC



Histone Symbol Key

H3K36me3 H3K27ac H3K4me3

H3K4me1 H3K27me3 H3K9me3

Figure 1.1 Naïve and Primed hESC
Summary figure of naïve and primed ESC chromatin and nuclear organization. Naïve ESCs have a more open chromatin structure, two active X chromosomes, and more active histone modifications such as H3K27ac and H3K4me1 (see also Section 1.3). Naïve cells also have immature mitochrondria, indicated by their round shape and less developed cristae[2]. Primed ESCs begin to have more condensed chromatin regions and more bivalent domains (see also Section 1.3). Primed cells have more mature mitochondria, with an elongated shape and more developed cristae[2].

Condition	Cell Line Established	Derivation Publication	Growth Condition Defined
3iL-AMPKi		Chan *et al.* (2013)	MEKi, GSKi, AMPKi, Lif
		Hanna *et al.* (2010)#,§	transient transgene with OCT4/SOX2/KLF4 then MEKi, GSKi, Lif, Forskolin
NHSM	LIS1, LIS2, WIS1, WIS2	Gafni *et al.* (2013)#	MEKi, GSKi, p38i, JNKi, Lif, FGF2, TGFβ supplementation and optional ROCKi and PKCi
3iL-FGFRi	Elf1	Ware *et al.* (2014)*	MEKi, GSK3i, FGFRi, hLif
2iL+IF		Sperber *et al.* (2015)*	MEKi, GSK3i, hLif
5iLA(F), 6iLA	WIN1	Theunissen *et al.* (2014)§	MEKi, GSK3i, BRAFi, ROCKi, SRCi, hLif, Activin A, FGF(taken out for some of the paper); 6iLA (5iL+JNKi)
4iLA		Theunissen *et al.* (2016)§	MEKi, BRAFi, ROCKi, SRCi, hLif, Activin A
t2iL+Gö		Takashima *et al.* (2014)@	lower concentration of GSK3i, MEKi, hLif, PKCi
t2iLGöY	HNES1, HNES2, HNES3, HNES4	Guo *et al.* (2016)@	lower concentration of GSK3i, MEKi, hLif, PKCi, ROCKi, ascorbic acid

Symbol	Lab
#	Hanna (first author, senior author)
§	Jaenisch
*	Ruohola-Baker, Ware
@	Smith

Table 1.2 Naïve hESCs and Growth Conditions
Summary of naïve hESC publications and growth conditions. Shown are the growth condition shorthand used in this text and how each growth conditions is defined in its publication.

1.1.4 Are Naïve ESCs representative of the human preimplantation blastocyst?

The outstanding question still remains, which naïve hESCs line and growth condition truly represents the ground state of pluripotency in human embryos? For ethical reasons, this cannot be tested directly, but by taking advantage of immunostaining and transcriptomic data from human embryos[27] we can begin to answer this question. Huang *et al.*, addressed this by looking at gene expression modules (groups of genes that are co-expressed or regulated in the same direction) in published transcriptomic data for reset cells grown in 2iL + forskolin[20], 3iL-AMPKi[26], t2iL + Gö[21], the derived lines LIS/WIS grown in NHSM[22], Elf1 grown in 3iL-FGFi[31], and WIN1 grown in 5iLA[23]. Huang *et al.* found that the 5iLA and t2iL + Gö most resembled the 2iL mESCs and post-zygotic genome activation (ZGA; used interchangably with maternal-zygotic transition, MZT, in this text) preimplantation human embryo[36]. However all the naïve hESCs did share a conserved gene network that included genes involved in RNA processing, ribosome biogenesis, and mitochondrial genes[36]. In spite their shared network, the majority of naïve hESCs have cell-type specific expression modules[36], suggesting that these naïve cells mostly capture different pluripotent states. A study published by Blakeley *et al.* reanalyzed the human embryo RNA-seq data from Yan *et al.* [27] and compared it to a subset of naïve hESCs: 3iL-AMPKi, t2iL + Gö, NHSM, and 5iLA. Based on the transcrptome, 3iL-AMPKi and t2iL + Gö look more like their primed counterparts than cells from human ICM[29]. Human ICM transcriptome is enriched for gene in the oxidative phosphorylation pathway compared to 3iL-AMPKi and t2iL + Gö hESCs, while the naïve hESCs were enriched for FGF, Wnt and MAPK signaling pathways[29]. In a gene-by-gene comparison, NHSM and 5iLA naïve hESCs show similar expression levels of *OSN* and *NODAL* but with some dysregulation of other genes[29]. While 5iLA cells showed higher expression of *FGF2*, *FGF4* and *KLF2*, 5iLA along with t2iL + Gö cells showed upregulation of the human ICM-specific gene *KLF17*[29]. NHSM hESCs dysregulated many more genes including *DNMT3L*, *TET1*, *DPPA3*, *DPPA5*, *KLF4*, and *TBX3*[29].

The issue with any molecular biology technique in ESCs is that the results can be highly variable depending on the the lab the cells were grown in. Transcriptomic data can be generated in many different ways giving slightly different results for gene expression and making it hard to compare across different

datasets. This is further exacerbated by the fact that the studies discussed are being compared to very limited human embryo data. Even upon pooling the single cell RNA-seq data, an incomplete profile of human embryonic development exists. Additionally it should be questioned whether a cell that is strongly driven into the naïve state, but can't exit that state, genuinely reflects the human preimplantation ICM (this is examined more in the DNA methylation section).

1.1.5 Interpretation and Opinions

Some may argue that the difference in pluripotency state captured between various hESCs may also be due to their genetic background. To some degree this could be true as some genetic backgrounds may make it more difficult to conceive and it is likely that a researcher would encounter individuals of these genetic background at IVF clinics where most human embryos are taken from. It has been shown that cells respond differently depending on their genetic background when left to differentiate without passaging[35]. Nonetheless, organisms from a variety of genetic backgrounds are able to conceive and produce viable offspring. Development is not organism specific but species specific therefore organisms with different genetic background will go about embryogenesis in roughly the same way. Placing too much emphasis on the genetic background of the embryo exonerates the field from being knowledgeable on what it takes to sustain a hESC in culture.

When discussing the naïve and primed states or preimplantation and postimplantation epiblast, it is important to recognize that their are differences between mouse and human development. Mouse obviously develop from fertilized embryo to live birth faster than humans. This is also seen at the zygotic level as mouse initiate ZGA at the 2-cell stage with human initiate ZGA at the 4- to 8-cell stage. The timing of expression of certain pluripotency markers can vary between mouse and human[29]. For example, *Klf4* and *Esrrb* are expressed throughout mouse preimplantation development while *KLF4* is expressed mostly at the 8-cell stage onward and *ESRRB* at the morula stage onward in human[29]. Some genes are expressed exclusively in a species specific manner such as *Klf2*, which is expressed in the mouse ICM, and *KLF17*, which is expressed in the human epiblast[29]. Luckily, pluripotency is more of an epigenetic feature than a genetic one. Embryos of diverse genetic backgrounds are able to create pluripotent cells

and differentiate. Therefore gene expression must be accompanied by the proper epigenetic program in order to truly define pluripotency.

1.2 DNA methylation

1.2.1 5-methylcytosine

Methylation on cytosine nucleotides (5-methylcytosine, 5mC) is maintained in somatic cells through the activity of DNA methyltransferase enzyme DNMT1, the maintenance methyltransferase, with its affinity for hemimethylated DNA[37] typically present on the newly synthesized strand after DNA replication. The *de novo* DNMTs, DNMT3a and DNMT3b, which have an affinity for unmethylated cytosines were identified roughly a decade later[38,39]. DNMT1, DNMT3a and DNMT3b are all essentially for proper development[38,39]. DNA methylation in mammalian somatic cells occurs mainly at CpGs and is symmetric, present on both strands of the CpG dyad[40]. CpG residues are depleted in mammalian genomes but where the do occur they are generally methylated[41]. The exception to this are regions dense with CpG residues, CpG islands (CGIs), which are typically unmethylated and are found at or near the promoters of many mammalian housekeeping and developmental genes[41]. Methylation at islands may be blocked through TF binding or the histone 3 lysine 4 methylation[42] (histone modifications discussed further in Section 1.3.1). DNA methylation at certain regions of the genome also helps to regulate gene expression. DNA hypermethylation at promoters and TSS and hypomethylation at gene bodies generally correlates with repression while hypermethylation in gene body and hypomethylation at transcription start sites (TSS) and promoters generally correlates with expression [40].

Ground state mESCs grown naïve 2iL conditions differ from serum/Lif mESCs in their global DNA methylation levels. Naïve mESCs, which have been shown to look more like the ICM of the preimplantation embryo based on gene expression, also resemble the ICM at the DNA methylation level[43]. Serum/Lif mESCs have higher 5mC levels and higher hmC (hydroxymethylcytosine, Table 1.1 and discussed in Section 1.2.2) than 2iL mESCs[43,44]. The observed lower mC level in naïve mESCs can be explained due to the *de novo* Dnmts, *Dnmt3a*/*3b*/*3L*, and *Tet1* have lower expression in 2iL than in serum/Lif[43,44]. Naïve mESCs also express *Prdm14* at a much higher level than serum/Lif[44]. KO of *Prdm14*

in mESC causes upregulation of *Dnmt3b*, and to a lesser extent *Dnmt3a1* (an isoform of *Dnmt3a*), resulting in 5mC levels similar to serum/Lif mESCs[44]. This shows that in naïve mESCs Prdm14 is a key regulator of the hypomethylated phenotype. Although 2iL and serum/Lif cultured mESCs have striking differences at the DNA methylation level and at the gene expression level, mESCs can easily be converted via media change between the two states[44,45]. This suggests that at least in mouse, at some levels of pluripotency, ESCs can easily fluctuate or transition between one state to another. The fluidity of the epigenome aids in an ESCs ability to glide along the pluripotency spectrum.

Double knockout *Dnmt3a*/*Dnmt3b*, and not single knockouts, serum grown mESCs lose their DNA methylation over time with passaging. They are resistant to differentiate via Lif removal and instead retain their stem cell characteristics[46,47]. In this particular hypomethylated background, there was an measureable increase in H4K5ac (histone 4, lysine 5, acetylation) that was not observed in *Dnmt1* KO mESCs[46]. Double knockouts also lose DNA methylation at imprinted regions and exogenous expression of the proteins could restore some paternal methylated imprints but not maternal[47]. Double knockouts were also observed having lost methylation at repetitive elements, particularly LINE1 elements, while knocking out all three Dnmts (1/3a/3b) results in additional loss of methylation at major satellite (pericentromeric repeats) regions and intracisternal-A particle retroelements (IAP)[48]. Interesting, *Dnmt1* overexpression can not restore global DNA methylation in a *Dnmt3a*/*Dnmt3b* DKO background[47], emphasizing the nonoverlapping roles of the maintenance and *de novo* methyltransferases.

Despite its known importance in development, DNA methylation is not required for maintenance of pluripotency. Triple knockout *Dnmt1*/*Dnmt3a*/*Dnmt3b* serum/Lif mESCs retain pluripotency markers and self-renewal properties[48]. Cells seem mostly unaffected by the drop in DNA methylation, most dramatically at CpG sites[48]. *Dnmt1* KO mESCs are viable, much like the TKO mESCs, but retain ~20% global methylation level[38,46,49]. These cells, however, are not able to differentiate to other embryonic lineages efficiently[38,46]. However, *Dnmt1* KO mESCs are able to differentiate to the trophoblast lineage through activation of the TF Elf5, which is normally repressed through DNA methylation in embryonic

cells[50]. This suggests a different role for DNA methylation, to regulate cell fate to either the embryonic or extra-embryonic lineages.

DNMT3L shares homology with DNMT3a/3b but lacks the catalytic domain for activity[51]. Knockout of *Dnmt3L* in mESCs has similar effects on DNA methylation as DKO of *Dnmt3a* and *Dnmt3b* although the loss of DNA methylation is not as strong[52]. Dnmt3L interacts with Dnmt3a and Dnmt3b to help direct DNA methylation, particularly at CpA sites[51,52]. Dnmt3L interacts with histone H3 tail however its binding is inhibited by methylation on histone lysine 4[51]. Thus providing an explanation for the anticorrelation between H3K4 methylation and DNA methylation observed in stem and somatic cells. Dnmt3L expression changes during post-implantation development[53]. Its repression in the postimplantation epiblast/blastocyst is mediated through promoter methylation which is directed by Dnmt3a, Dnmt3b, and Dnmt3L itself[53].

DNA methylation likely exhibits the most dynamic changes during early embryogenesis compared to other epigenomic marks. In humans embryos, a similar methylation pattern across development was observed as in mouse. DNA methylation studies on human embryos have found that demethylation occurs rapidly by about the 2-cell stage and is at its lowest level in the genome in the ICM of the preimplantation blastocyst(~30%)[42,54,55]. Much like mouse, sperm are hypermethylated (~60-75%) and upon fertilization the paternal genome is actively demethylated prior to the first cell division, within the first hours of development, until the paternal genome is hypomethylated with respects to the maternal genome[42,54-56]. Human oocytes have intermediate levels of methylation that slightly decreases as oocytes mature (~54% at germinal vesicle/metaphase I stage to ~48% at metaphase II stage)[54,55]. Interestingly, oocytes have more non-CpG methylation than all other cell types investigated (>2% in MII oocytes compared to ~≤1% in sperm, zygote or most somatic cells), and are specifically enriched for CpA methylation (GV/MI ~5.6%)[54,55]. Additionally immature oocytes (GV/MI) express *DNMT1*, *DNMT3a* and *DNMT3b* but not *DNMT3L*[55]. Human preimplantation blastocysts have a similar methylation pattern to oocytes suggesting that the embryo mostly adopts the maternal genomic methylation[55].

Naïve ESCs capture many aspect of the demethylated, preimplantation state of the embryo with some exceptions. Naïve hESCs grown in many different growth conditions show a decrease in global DNA methylation (Figure 1.2). Naïve 4iLA, 5iLA, t2iLGöY and reset t2iL+Gö cells have global methylation levels ~26-35% compared to >75% in primed hESCs[33,34]. Reset t2iL+Gö hESCs additionally showed lower non-CpG methylation compared to primed cells[21]. Naïve derived t2iLGöY HNES1 and reset t2iL+Gö H9 cells have higher *TET1* and *DNMT3L* expression and lower *DNMT3B* compared to primed H9 and primed HNES1[34].

Some naïve hESCs however do not recapitulate the gain of methylation observed in postimplantation development. Naïve hESCs grown in 5iLAF have decreased expression of *DNMT3B* and increased expression of *DNMT3L* and *UHRF1*[57]. *Uhrf1* KO in mESCs was shown to result in an increase in hemimethylated DNA, similar to an *Dnmt1* KO[52]. Naïve hESCs 5iLAF and t2iL+Gö exhibit global hypomethylation levels (~30%) similar to the ICM/blastocyst embryo but unlike the blastocyst, naïve hESCs (5iLAF and reverted t2iL+Gö) methylation pattern shows little to no correlation with oocyte methylation patterns[57]. At regions that are to be stably imprinted later in development, 5iLAF and t2iL+Gö naïve hESCs show loss of DNA methylation and biallelic expression of imprinted genes in 5iLAF hESCs. When shifted to primed hESC growth conditions, 5iLAF naïve hESCs fail to fully regain DNA methylation at imprinted regions and still exhibit biallelic expression[57]. Reset naïve hESCs grown in 5iLA or 4iLA also failed to fully methylate DNA at imprinted regions and retain biallelic expression[33]. Similar investigations have yet to be conducted in other human naïve conditions. Similar to the naïve hESCs, mESCs growth as naïve in 2iL or metastable serum/Lif show allelic DNA methylation at some imprinted loci[44].

Primed hESCs are globally hypermethylated compared to naïve and differentiated cells. The cell accomplishes this through CG and non-CG methylation. H1 primed hESCs methylated the same number of CGs in their genome as differentiated cells but 25% of their total global methylation also occurs at non-CG sites, enriched for CAH motif (often referred to as CHG or CHH, where H is any base other than G

)[40].This non-CpG methylation is not affected by DNMT1 null mutations, but are affected by downregulation/knockout of DNMT3a and DNMT3b[49,58-60]. Most (77%) CG sites in primed hESCs are >70% methylated[40] while non-CG sites (85%) are partially methylated, ~10-40%[40]. In contrast differentiated fibroblasts have >80% methylation at only >50% of their CGs[40]. Unlike CG methylation, non-CG methylation, particularly CHG, is asymmetric[40] suggesting the propagation of non-CG methylation may be due to the activity of the *de novo* DNMTs since DNMT1 has a preference for hemi-methylated DNA and likely creates the symmetry at CpG dyads. However this abundance of non-CG methylation is lost upon differentiation[40]. Compared to somatic cells ESCs (also iPSCs and progenitor cells) have higher amounts of non-CpG methylation[40,58,59,61]. All this indicating a higher globally methylated genome in primed hESCs.

1.2.2 5-hydroxymethylcytosine

DNA demethylation can occur passively or actively. Passive demethylation may occur if DNMTs are not present or active to methylate DNA after each cellular division. Through this process, DNA methylation would decrease gradually with each successive replication, which is thought to be the mechanism for maternal genome methylation loss in the preimplantation embryo. The TET mediated demethylation pathway has been suggested as the mechanism for active demethylation of the paternal genome and is required for proper embryogenesis and differentiation of ESCs in mammals [62,63] (Figure 1.2). TET (ten-eleven-translocation methylcytosine dioxygenase) proteins are conserved throughout metazoan[64]. TET proteins (TET1, TET2 and TET3) bind 5-methylcytosine (5mC) and catalyzes its conversion to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine(5fC), and 5-carboxylcytosine (5caC)[64,65], although TET3 makes 5caC to a lesser extent[65]. 5caC can be bound by thymine-DNA glycosylase (TDG), removed through the base excision repair pathway and replaced with an unmethylated cytosine[66,67].

Tets are expressed in many mouse tissues[68] and are imperative for mouse preimplantation development. Tet2 and Tet3 mRNA can be found in mouse and human oocytes[55,62]. Tet3 in particular is highly expressed in mouse oocytes and zygotes. RNAi knockdown of Tet3 strongly diminishes 5hmC in the parental genomes[62]. Pre-cleavage mouse embryos have a sharp increase in 5hmC by the PN3 stage

(pronuclear stage of development, prior to first cellular division), which peaks prior to the first cellular division[62]. The paternal genome has the largest increase of 5hmC, and subsequently the largest decrease in 5mC compared to the maternal genome[62]. Knockdown of *Tet3* prevents the loss of 5mC in the paternal pronucleus[62]. *Tet1* and *Tet2* are expressed from the 2-cell stage to the preimplantation blastocyst[62,68]. Tet1 protein is enriched in the ICM, and blastomeres with siRNA knockdown of *Tet1* exhibit a preference toward the trophectoderm lineage[68].

Similar results have been observed in serum/Lif mESCs. *Tet1* knockdown in serum/Lif mESCs affect ESC self-renewal and causes a reduction of OSN expression through promoter methylation[68]. Knockdown of *Tet1* and *Tet2* causes an increase in early lineage markers, particularly those for trophectoderm and primitive endoderm lineages[68]. The reduction of ESC maintenance is also observed through Lif withdrawal, which also results in *Tet1* downregulation[68].

Immunofluorescence studies in human zygotes, both parental genome contain 5mC and 5hmC but the paternal genome is more hydroxymethylated than the maternal[63]. Human oocytes contain both modifications but sperm only have detectable 5mC[63] suggesting that the paternal genome becomes a substrate for TET protein very soon after fertilization. Most 5hmC is found in euchromatic regions of the chromosomes, in the T-band regions which are GC and gene rich regions[63]. The asymmetry between maternal and paternal genome hydroxymethylation and methylation is eventually lost through cellular divisions throughout preimplantation development.[63].

Naïve hESCs also show a reduction of DNA methylation compared to primed hESCs, indicative of them representing the preimplantation stage of embryonic development[21,22,31,33]. Human epiblast and naïve hESCs were also shown to have increased expression of *TET1* and TET2 compared to primed hESCs, the level of expression depends on the growth condition[29]. Naïve 5iLAF hESCs have increased expression of *TET1* and *TET2*[57]. Mass spectrometry measurements in reset t2iL+Gö hESCs measured lower levels of 5mC and 5hmC compared to primed cells[21].

When discussing 5mC and 5hmC levels it is important to remember that WGBS cannot distinguish these two modifications[69,70]. The presence of 5hmC in the genome, although small, is significant when interpreting data from a bisulfite reaction. Chemical (TLC, mass spectrometry), antibody (immunostaining, 5hmC antibodies, hMeDIP), enzymatic (β-glucosyltransferase labelling, restriction enzymes: MspI, MspJI50, PvuRts1I, SauUSI, TaqαI) and non-amplification single strand sequencing (SMRT-Seq) 5hmC detection techniques[71] allow us to infer the presence of 5hmC in the genome. Many techniques require an enrichment step that makes direct quantification of 5hmC levels challenging. In 2012 and 2013, two protocols were developed to measure the level of 5hmC directly at nucleotide resolution.

TAB-seq, Tet-assisted bisulfite-sequencing, was developed by Yu *et al.* to directly detect hmC levels. Yu and colleagues protect 5hmC by using β-glucosyltransferase, then use a recombinant Tet1 protein to oxidize 5mC to 5caC. Bisulfite treatment then deaminates 5caC to uracil, which after PCR, is read as a thymine in next-generation sequencing. Thus, 5hmC are read as a C in TAB-seq and 5mC are read as a T. A WGBS or BS-seq library should be generated for side-by-side comparison and confident detection of 5hmC. The accuracy of 5hmC quantification is greatly influenced by efficient enzymatic glucosylation of 5hmC, efficient conversion of 5mC to 5caU and efficient bisulfite conversion. Using TAB-seq primed H1 hESCs were found to have 691,414 confident 5hmC sites while serum grown mESCs had considerably more 5hmC sites at 2,057,636 5hmCs genome-wide which may be explained by the higher expression of *Tet1* and *Tet2* in mESCs[72]. The vast majority of 5hmC sites were found at CpG dinucleotides (>99% for human and >98% for mouse) and were asymmetrically hydroxymethylated[72]. Over 46% of 5hmCs were found at distal regulatory elements in primed hESc and mESCs that were marked as either enhancers, CTCF binding sites and Dnase I Hypersensitive Site (DHS) with local depletion at TF binding sites[72]. Additionally, 5hmC was enriched at low-CpG density promoters in hESCs and mESCs[72].

Another protocol name oxBS-Seq, for oxidative bisulfite sequencing, published by Booth *et al.*, uses potassium perruthenate (KRuO4) to selectively oxidize 5hmC to 5fC which can then be converted to uracil following bisulfite treatment[73]. Thus, 5hmC are read as a thymine in oxBS-seq and 5mC are read as a cytosine. Like TAB-Seq, oxBS-Seq also requires the generation of a traditional bisulfite sequencing

library for accurate detection and quantification of 5hmC levels. There are other products and kits from biological science research companies that allow the detection of 5hmC, using immunoprecipitation and sequencing techniques. Downstream data analysis allows a researcher to infer the nucleotide position of 5hmC.

The two inhibitors, Meki and Gsk3i which important for ground state of pluripotency in mESCs, were shown to direct hypomethylation of the genome using two different pathways. Researchers teased apart these pathways through a series of biochemical assays in mESCs grown in one inhibitor at a time in low Lif conditions so that the effects of Lif wouldn't cloud the effects of the inhibitors. One pathway involves the demethylase JMJD2C whose protein levels were stabilized by MEKi[74]. JMJD2C removes methylation from amino acids like K9 and K27, where methylation can result in a repressive chromatin structure. JMJD2C interacts with Tet1/2 proteins and MEKi was shown to increase 5hmC levels genome wide[74]. The second pathway involves downregulation of *de novo* methylation. MEKi and GSK3i cause a decrease in *DNMT3a/3b* transcription and a decrease in 5mC levels[74]. These inhibitors also decrease Tet1 expression but Tet2 and 5hmC levels are unaffected. MEKi and GSK3i also promote *PRDM14* transcription. PRDM14 interacts with DNMT3s to recruit the methyltransferase G9a. G9a activity on DNMT3s marks them for ubiquitination and degradation. To summarize, the first pathway of ground state hypomethylation involves the downregulation of repressive chromatin structure and upregulation of active DNA demethylation pathways. The second pathway involves downregulation of the *de novo* DNA methyltransferases through transcriptional repression and protein degradation.

1.2.3 Interpretations and Opinions

While DNA methylation has been explored in naïve hESCs, DNA hydroxymethylation has not been investigated at single-base resolution, genome-wide. Here, I will present the first naïve hESC TAB-seq data. I will provide a comparison of 5mC and 5hmC in both naïve and primed hESCs and investigate whether the naïve line Elf1 undergoes appropriate shifts in DNA methylation as it is pushed to the primed state.

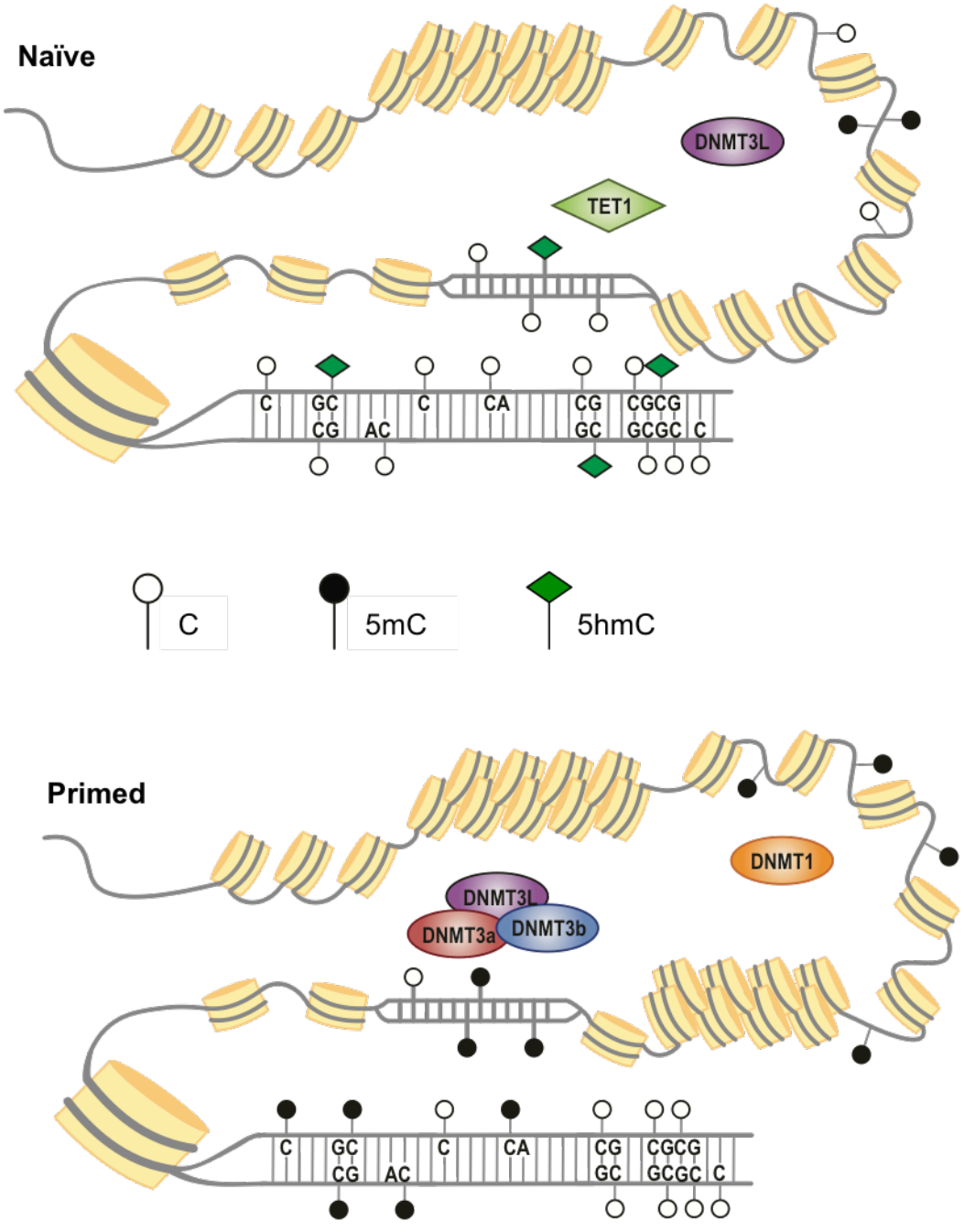


Figure 1.2 DNA Methylation and Hydroxymethylation in Naïve and Primed ESCs
Naive ESCs are hypomethylated compared to primed, similar to how the preimplantation embryonic cells are hypomethylated. The activity of TET proteins keeps 5mC levels low in the naïve epigenome and 5hmC level relatively high compared to primed ESCs. Primed ESCs and postimplantation embryonic cells are hypermethylated relative to other cell types due to the activity of DNMTs.

1.3 Chromatin dynamics

1.3.1 Histone Modifications

DNA is tightly wound and packaged into the nucleus of the cell with the assistance of histones. Chromatin is the collection of nucleosomes, packages of DNA wrapped around a histone core, that make up the nuclear genomic material in a eukaryotic cell. The nucleosome is comprised of a histone core made up of an octamer of 4 histones, H2A, H2B, H3, and H4, that 147 bp of DNA wrap around[75]. The core histones H2A, H2B, and H3 also have variants (such as H2A.Z and H3.1) which may be inserted into the nucleosome to affect DNA binding and gene regulation. Additionally, much like DNA can be chemically modified, modifications of histones are a way to regulate gene expression and chromatin structure. Histone post-translational modifications can occur at many amino acid in the body, globular, domain but most of the modifications studied occur on the tails, the unstructured ends of histones that wave out from the nucleosome. Some commonly studied modifications include acetylation, mono-, di- or tri- methylation, and phosphorylation occurring mostly on lysine (K), arginine (R) and serine (S) residues[75]. If one considers all the modifications and amino acid residues they can occur on, there are well over 100 different post-translational modifications a nucleosome can acquire. These modifications can occur in different combinations on the same nucleosome drastically expanding the vocabulary of the histone code.

Many histone modification studies have been on modifications of histone H3. The tail of H3 interacts with the major groove of DNA[76] making it an ideal candidate for orchestrating DNA-protein interaction or altering DNA-histone association. A number of H3 modifications have been studied and the genomic element they mark have been identified, with little to no variation between species examined. Much of the what we know about histone modifications came from biochemical, molecular biology, and sequencing studies. Chromatin immunoprecipitation, where an antibody is used to pull down nucleosomes or fragments of DNA bound with the histone modification or protein of interest, combined with microarray or next-generation sequencing have provided a genome-wide view of histone and post-translationally modified histone localization. The enzymes that establish these chromatin marks may overlap in function and many have the ability to act on different histone residues and other proteins in the cell. Thus, when interpreting the results of experiments of KO chromatin modifiers it is important to consider that enzymes

may have wide ranging effects on cell phenotype.

1.3.2 Histone dynamics in early embryogenesis

Like DNA methylation, histone occupancy is also very dynamic during the earliest stages of embryogenesis as part of "resetting" the epigenome. Fluorescent studies in mouse embryos provide insight into the dynamic nature of histone occupancy during embryogenesis. Immature oocytes have many histone modifications present in their genome including methylation of H3K9/K4/R17, acetylation of H4 lysine residues (K5, K8, K12, and/or K16), and S1 phosphorylation of either H4/H2A (H4/H2AS1ph, indistinguishable because of the identical amino acid sequence near the S1 residue)[77]. By the time oocytes are mature, they retain many of the same histone modifications, also having H3K4me1/me3[78] but H3R17 methylation is lost from the nucleus[77]. Sperm DNA is packaged with protamines, which are removed and replaced with histones shortly after fertilization in mouse[78,79].

After fertilization, during the pronuclear stages prior to the first cellular division, both the maternal and paternal genomes have acetylation of H4 lysine residues, H4/H2aS1ph, H3K9ac, H3K9me1, H3R17 methylation, H3K27me1/me2/me3[77,80,81]. H3K4me1 is detected in the paternal genome starting PN1 (~5 hours post-fertilization) but doesn't reach maternal levels until about PN3-4 (~<8 hours post-fertilization)[78]. H3K4me3 is detected in the paternal genome around PN4 (~8-10 hours post-fertilization) and is roughly equal to maternal level by PN5 (~12 hours post-fertilization)[78]. H3K9me2/3 is only present on the maternal genome and localizes to centromeric regions until the morula stage where speckles of H3K9me3 can be seen in all nuclei[77,80,82]. H3K27me1/me2/me3 has been observed in the maternal genome starting early after fertilization along with components of the Polycomb complex, Eed and Ezh2[80,81]. The male pronucleus also gains the Polycomb complex proteins Eed and Ezh2 later around PN1 then subsequently begins to gain H3K27 methylation[80,81]. H3K27me3 is observed in cells of morula stage embryos and is dispersed in the ICM of preimplantation blastocysts compared to punctate staining (indicative of the inactive X chromosome) in trophectoderm cells[81]. Lastly, H4R3 methylation is also observed but only becomes present from the 4-cell stage onward and only in interphase nuclei[77]

Histones are detected on the paternal genome prior to H3K4me1 modification and H3K4me1 occurs prior to H3K4me3. These data point to a mechanism where histone methylation, at least of H3K4, occurs on the nucleosome after histones have been incorporated. Histone acetylation, however has been shown to occur in the cytoplasm, prior to incorporation[83]. H3K4me1 incorporates into the paternal genome roughly around the same time as active DNA demethylation is occurring[78,80] suggesting a shared mechanism or co-dependence between these two epigenetic marks in embryogenesis. The persistence of H3K9me2/3 staining in the maternal genome could potentially act as way to protect the maternal genome from the active DNA demethylation pathway.

1.3.3 Broad H3K4me3 domains in mouse embryos

H3K4me3 has been observed to be mostly invariant between the different ESC states and differentiated cells[45] and is generally always found at the promoters of genes[84]. H3K4me3 regions are hypomethylated at levels ~<10% throughout all stages of preimplantation, postimplantation, primed hESCs, and in the gametes (oocyte 10%, sperm 5%)[54]. However H3K4me3 has shown unique genome patterning at the earliest stages of development. Three studies in 2016 identified these unique H3K4me3 chromatin features in mouse embryos. Dahl *et al.*, Lui *et al.*, and Zhang *et al.* individually developed ChIP-seq methods to profile the distribution of H3K4me3 in the genome of developing oocytes and preimplantation embryos.

Mature MII oocytes have broad regions of H3K4me3, many over 10kb that can be distal from TSS and promoter regions[85,86]. These domains develop as oocytes undergo maturation but decline upon ZGA in the late 2-cell stage[85,86]. Due to the antagonistic relationship between DNA methylation and H3K4me3 the broad domains are hypomethylated, overlapping partially methylated domains[85,86] but have sharp increases of DNA methylation outside their boundaries[85]. Dahl *et al.* postulates that the hypermethylation outside the boundaries help to define the domains[85]. Unlike conventional H3K4me3, 75% of the broad domains are not located near a TSS[85]. The more canonical, promoter based H3K4me3 gains rapidly from MII oocytes to the 2-cell stages and remains relatively stable during the rest of development, mostly at promoters with high Cpg density[87]. H3K27me3, however, gradual gains over the course of

preimplantation development and is preferentially deposited at low CpG promoters[87].

H3K4me3 is essential for ZGA as 79.4% of ZGA genes are marked with H3K4me3 in oocytes and sperm[85] providing a mechanism for their rapid transcription during the Maternal-to-Zygotic Transition (MZT). Distal, broad domains of H3K4me3 are only on the maternal genome and disappear upon ZGA in the late 2-cell stage[86]. If transcription is inhibited and ZGA doesn't initiate, these broad domains remain[86] indicating that transcription is required for proper genomic H3K4me3 distribution. At 2-cell stage, H3K4me3 broad domains with ZGA genes gain H3K27ac[85], an indication of their activation. Kmt2b, which is responsible for establishing H3K4me3 in oocytes and is essential for ZGA, has all 8 genes present in broad H3K4me3 domains[85]. The demethylases Kdm2a and Kdm2b, cause developmental delay and inhibit blastocyst formation when KD in mouse embryos[85]. High level of H3K4me3 are retained when Kdm2a/2b are KD in 2-cell stage embryos and there is downregulation of ~1300 ZGA genes[85]. KD of Kdm5b in embryos also impedes blastocyst formation and increases the number of broad domains in the morula stage[87].

Domains carrying both modifications H3K4me3 and H3K27me3 (termed bivalent domains, discussed more in Section 1.3.6) are few and sparse during the majority of preimplantation development[87]. However there is a large increase in the number of bivalent domains as embryos transition from morula to the cells of the blastocysts, ICM and trophectoderm, and even more bivalent domains observed in ESCs and trophoblast stem cells[87].

1.3.4 Enhancers and active chromatin regions

Enhancers are regions of DNA that bind proteins to help regulate gene expression and, unlike promoters, are highly cell type specific. Enhancers can be proximal or distal to a gene promoter and can be as far as megabases away from the gene or genes they regulate[88]. Enhancers not only bind transcription factors (TFs) but also may bind Pol II, coactivators (i.e. p300), or other chromatin structural proteins (i.e. Mediator complex, CTCF) that loop them in close spacial proximity to the promoter(s) they regulate[88,89]. Enhancers tend to be hypomethylated, especially in the cell-type in which they are active, but are also depleted of

CG dinucleotides unlike promoters which commonly overlap CGIs[40,61,90] (Figure 1.3). Enhancers are identifiable by the presence of DNase I hyper sensitive site (DHS), high conservation across mammals (particularly at TF motifs), and enrichment for H3K4me1[61,84,89,91]. Since H3K4me1 and H3K4me3 can often colocalize, enhancers must also have a depletion of H3K4me3[84], ruling out overlap promoter which can also have enhancer function.

Enhancers are identified via functional test or sequencing based methods. The traditional functional test for an enhancers is to clone the DNA region into a reporter plasmid expressing luciferase and look for increased gene expression. ChIP combined with sequencing based methods allows for genome-wide identification of enhancers while 3C technologies (discussed more in Section 1.4) allows for identification of the genes, or other enhancers, they interact with. Global genomic enhancer profiles are highly cell-type specific and these elements exists anywhere from ~20-100 thousand times in the genome depending on cell type[84,89,92]. This is because different cells may express the same genes using different regulatory elements. For example, *POU5F1* is expressed in both naïve and primed ESCs but naïve cells use the distal enhancer (named because it is located further upstream of the promoter/TSS) while primed cells use the enhancer that is proximal to the promoter and TSS[93].

Enhancers regions of chromatin can have overlapping histone modifications that distinguish them as being active, repressed or poised depending on whether H3K4me1 co-occurs with H3K27ac, H3K27me3 or occurs alone[90,91,94]. Serum/Lif mESCs have very few H3K4me1 enhancers regions that also contain overlapping H3K27me3, at only 1.2%[94]. For naïve cells this is to be expected since the genome is depleted of H3K27me3 in general. Active enhancers are co-marked with H3K27ac, which is a general feature of active enhancers in mammal cells not just ESCs[91,94]. Acetyltransferases p300 and CBP acetylate H3K27 and can acetylate all four core histones[95]. Both p300 and CBP are required for embryogenesis as null mutants are embryonic lethal[96]. High expression levels of the nearest neighboring genes (aka most proximal gene) to active enhancers is more positively correlated with active enhancers than it is with TFs, p300 or enhancers lacking H3K27ac, meaning active enhancers are more likely to be associated with expressed genes than other chromatin modifiers or DNA bind proteins[90,91,94]. Enhancers

with only H3K4me1 or co-marked with H3K27me3 are considered “poised” for activation and may gain H3K27ac or retain H3K27me3 as they differentiate[91,94]. However many poised enhancers lose their H3K4me1 modification upon differentiation in a process known as enhancer decommissioning (discussed in Section 1.3.5).

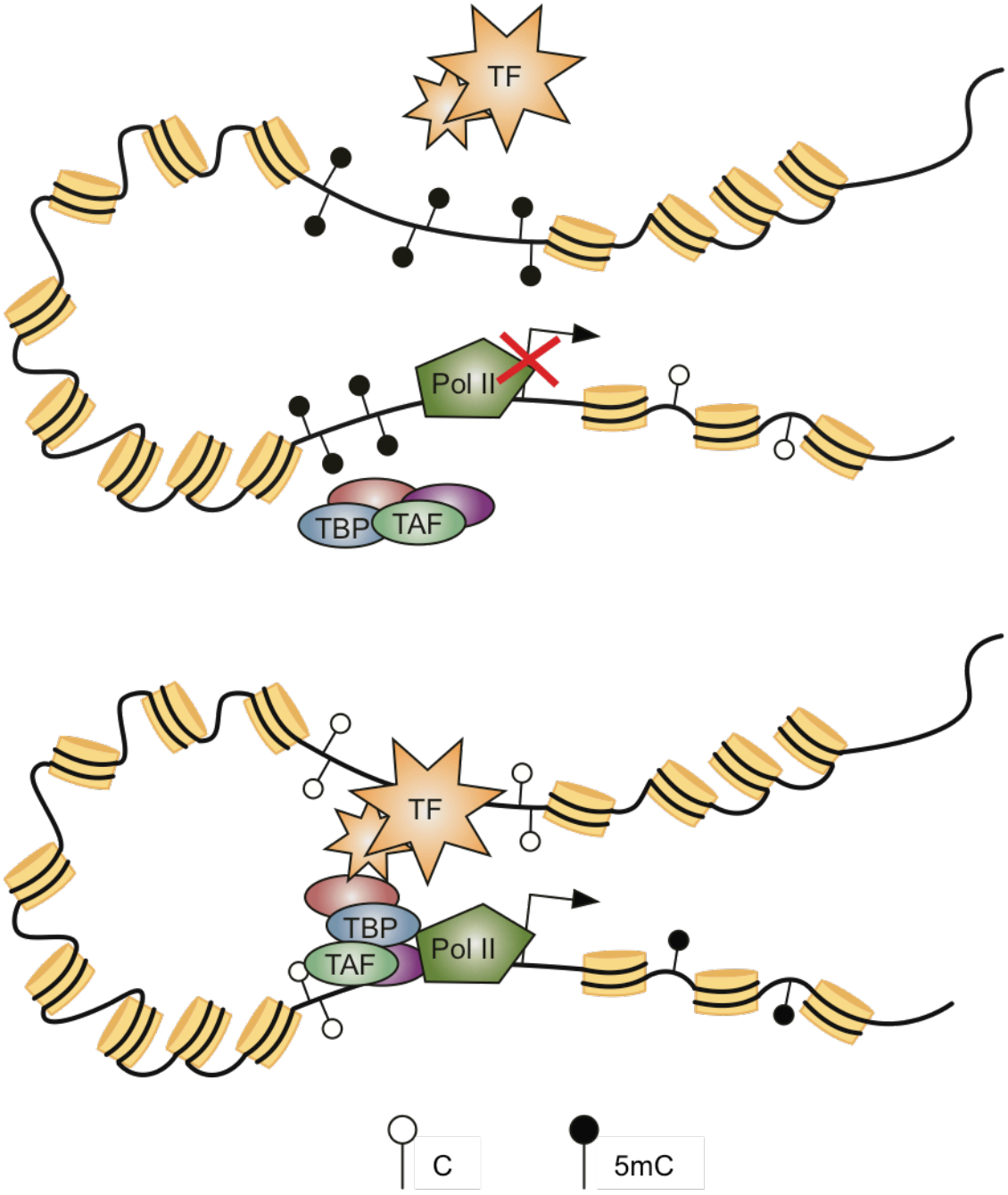


Figure 1.3 TF Binding at Enhancers – Methylation Dependent
DNA methylation at regulatory elements can affect binding of protiens suchs as TF and activators. DNA binding proteins are typically inhibited by the presence of DNA methylation, Hypomethylation at regulatory elements is associated with activation of these elements.

1.3.5 Broad Enhancers and Decommissioning

Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. Whyte *et al.* (2013) originally defined super-enhancers in serum/Lif mESCs as regions on average 8.7kb long with clusters of OSN binding, enrichment of Mediator, Klf4 and Esrrb binding[97]. They are also enriched for H3K27ac and DHS sites[97]. Super-enhancers are cell type specific and in ESCs are near genes important for stem cell identity and function including *OSN*, *Prdm14*, *Sall4*, *Tbx3*, *Klf4*, *Esrrb*, *Zfp42*, *Tet1*, *Tet2*, and the most abundant miRNA in ESCs, miR-290-295 [97]. Genes near super-enhancers are expressed higher and under *Pou5f1* KD, are downregulated sooner than other genes in the cell[97]. The parallel of super-enhancers, stretch enhancers, were defined in differentiated cells by Parker *et al.* (2013), have similar features including enrichment of H3K27ac and cell-type specificity. Stretch enhancers were defined as greater than 3kb and they were found to overlap locus control regions in their parent cell-type[98].

Primed cells undergo further reduction of H3K4me1 in a process termed enhancer decommissioning during differentiation[99]. Enhancer decommissioning is performed by the K4/K9 demethylase Lsd1, which binds at the promoters and enhancers of active and bivalent genes[99,100]. The activity of Lsd1 is inhibited by the presence of histone acetylation[100], presenting a secondary role of H3K27ac as a protective modification to preserve active enhancers in the genome. In mESCs, Lsd1 was found to occupy 97% of enhancers co-occupied by OSN and Mediator complex[99]. Lsd1 directly interacts with members of the NuRD complex, such as HDAC1 and HDAC2, who also bind at these enhancers[99]. During differentiation some stemness enhancers are lost and lineage specific enhancers are gained. Enhancers that are decommissioned lose their H3K4me1 signal, along with loss of p300 and H3K27ac, presumably through the action of HDAC1 or HDAC2 which are co-bound with Lds1 at these regions. However with inhibition of Lsd1, hESCs either die or are resistant to differentiation and retain stem-like phenotypes[99]. For example, Lsd1-inhibited primed hESCs form small, spherical colonies, stain positive for AP activity, and continue to express Sox2 while also expressing the trophoblast lineage gene Cdx2[99]. ESCs treated with Lsd1 inhibitor and forced to differentiate also retain H3K4me1 signal at regions it should be depleted[99].

1.3.6 H3K27me3 and Bivalency

The Polycomb Repressive Complex (PRC) is important for all three forms of H3K27 methylation[101,102]. H3K27me3 is depleted in naïve ESCs [2] but is required for primed pluripotency[103]. The main components of Polycomb Repressive Complex 2 (PRC2), EZH2, EED, SUZ12, and RBAP46/RBBP7 or RBAP48/RBBP4[102] are each important in the function of the PRC2 complex and embryonic development.

Ezh2 contains a SET domain giving it histone methyltransferase activity for PRC2. Ezh2 has high activity for H3K27[81]. The timing for Ezh2 activity is key for proper development. Null embryonic mutants for Ezh2 are embryonic lethal while depletion of maternal, cytoplasmic Ezh2 results in growth retardation but eventually fully developed, fertile adult animal[81]. Maternally depleted Ezh2 is eventually "rescued" by zygotic expression of *Ezh2* and accumulation of H3K27me3 in the genome[81].

The subunit of PCR2 Eed is required for H3K27me1/me2/me3[101]. Eed null mESC still express *Ezh2* but lack detectable protein product[101]. Eed null mouse embryos form post-implantation embryos *in vitro* but with larger epiblast compartment compared to extra-embryonic compartment and have impaired gastrulation[104].

PRC2 subunit Suz12 null embryos start to show defects around E7.5-8.5 but do not survive pass E8.5 and, like the Eed mutants, show defect in gastrulation[105]. KO of Suz12 also results in a undetectable levels of Ezh2 protein and lower levels of Eed protein[105]. Suz12 helps maintain stability of Ezh2 and prevent its degradation. KD of Suz12 lowers the enzymatic activity of Ezh2 and prevents incorporation of another PRC2 subunit RbAp48 but does not prevent Ezh2 from binding Eed. Additionally, Suz12 KO embryos were shown to only be depleted in H3K27me2/me3 but not H3K27me1[105] suggesting that different components of PRC2 may directly influence specific types of H3K27 methylation.

H3K27me3 deposition in the genome is necessary for primed ESCs to differentiate into embryonic and extraembryonic lineages. H3K27me3 regions, identified in primed hESCs, show lower levels of DNA

methylation (<25%) in human blastocyst ICM compared to background genomic levels[54]. Primed hESCs and postimplantation human embryos also have low DNA methylation at these H3K27me3 regions, although slightly higher than ICM (~25-40%). Promoters with stronger H3K27me3 ChIP-seq signal, tend to have lower levels of DNA methylation[54]. Suggesting an antagonistic role between DNA methylation and H3K27me3 and differential mechanisms of gene repression for DNA hypermethylated promoters and Polycomb repressed promoters.

One epigenetic characteristic of primed ESCs is the prevalence of domains >4-5 kb of H3K27me3 with smaller regions of H3K4me3 at TSSs[106-108]. These domains that contain both repressive H3K27me3 and active H3K4me3 marks are termed bivalent domains[106-108] (Figure 1.4). NGS in H9 primed hESCs showed that the active and repressive modifications were not necessary on the same nucleosomes but still very close and within the same region[108]. However, mass spectrometry analysis in mESCs found that H3K4me3 and H3K27me3 can co-occur on the same nucleosome but opposing histone H3 tails[109]. Additionally, the presence of H3K27me3 on one H3 tail of a nucleosome stimulates PRC2 activity on the unmodified adjacent nucleosome[109], presenting a mechanism for spreading the H3K27me3 modification across chromatin.

The genes in bivalent domains are have very low expression levels, due to the presence of the repressive H3K27me3 at their loci[107,108]. Bivalent genes also tend to be TFs and important developmental regulators and the genes are resolved to either active or repressed in differentiated cells[106-108]. In *Eed* mutant mESCs, there is early expression of lineage specific genes due to loss of H3K27me3[107]. Hence bivalent genes present a method for stem cells to make gene expression decisions at a later date. Genes are "poised" for activation or repression depending on what cell fate decisions and developmental path the stem cell takes. For example Bernstein colleagues in 2006 showed that *Nkx2.2*, *Pax5*, *Dixdc1* are all bivalently marked in serum/Lif mESCs. Upon differentiation to neural progenitor cells, *Nkx2.2* become active through loss of H3K27me3, *Pax5* become repressed through loss of H3K4me3 and *Dixdc1* remained bivalent. In mouse and human, *Nkx2.2/NKX2-2* is involved in central nervous system development, *Pax5/PAX5* is involved in B-cell development, and *Dixdc1/DIXDC1* is a regulator of the Wnt

signalling pathway expressed in many different issues including adult brain. When mESCs cell fate is driven to the neural lineage, genes important to the tha lineage are turned on, genes necessary for other cellular lineages are turned off and genes that progenitor cells might need later in development remain bivalent.

While both mouse naïve 2iL ESCs and metastable serum/Lif ESCs are able to contribute to the ICM of a developing embryo, they still harbor some epigenetic differences which aid in defining 2iL as the ground state of pluripotency. One of the ways these differences can be observed in their repressive chromatin structure as observed by Marks *et al.* (2012). Naïve mESCs have overall similar level of H3K27me3 genome wide as detected by immunoblotting and similar expression levels of PRC1 components, PRC2 components, and H3K27 demethylases as serum/Lif mESCs but more of the H3K27me3 is deposited at satellite regions in naïve cells[45]. H3K27me3, Ezh2 and Suz12 occupancy at gene promoters was higher in serum/Lif mESCs than naïve[45]. Bivalency is also reduced in naïve mESCs with just under 1000 genes with bivalent promoters in naïve mESCs, the vast majority of which are also marked bivalently in serum/Lif mESCs[45]. Like in primed ESCs, bivalent genes fall into GO categories involving developmental processes[45]. H3K27me3 is observed flanking, with slight overlap with H3K4me3 modification[45], indicating that these modifications are not found on the same nucleosome, much like in primed ESCs.

Sperber *et al.* (2015) found that naïve ESCs have less H3K27me3 chromatin than primed ESCs in both mouse and human[2]. Naïve hESCs whether derived (Elf1 or naïve mESCs) or reset (3iL-APKMi or NHSM hESCs) have lower levels of H3K27me3 than primed hESCs lines particularly near the TSS of a subset of developmental genes[2]. This could not be explained by EED protein levels or expression levels of other H3K27 and H3K9 methyltransferases or demethylases[2] but is linked to the differential metabolism observed in naïve and primed cells. Naïve hESCs have been shown to have higher NNMT (nicotinamide N-methyltransferase) enzymatic activity which reduces the presence of SAM (S-adenosyl methionine), a metabolite that functions as a major methyl group donor to other macromolecules including histones[102]. The link between metabolism and H3K27me3 was proved as overexpression of *NNMT* in primed hESCs reduces H3K27me3 and inhibition STAT3, an NNMT positive regulator, or direct inhibition of NNMT

increases H3K27me3 and the heterochromatic modification H3K9me3 in Elf1 naïve hESCs[2]. Additionally KO *NNMT* Elf1 lines show a shift towards primed gene expression profile, downregulation of *DNMT3L*, downregulation of the Wnt pathway and upregulation of HIF pathway[2].

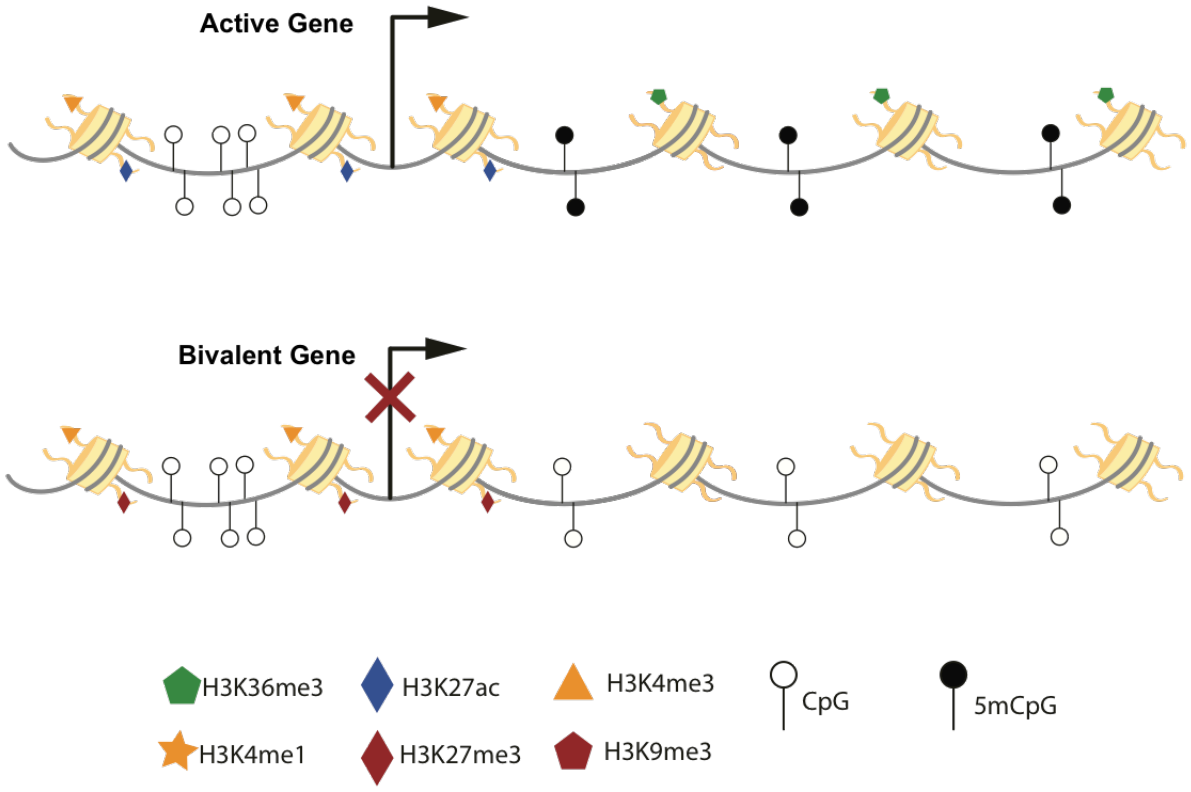


Figure 1.4 Bivalency Poises Genes
Summarizes difference in histone modifications and DNA methylation at actively transcribed genes and bivalent genes. Active genes have H3K4me3 and H3K7ac at their promoters and H3K36me3 in the gene body. DNA methylation in the gene body of transcribed genes is a common feature. However bivalent genes contain both active H3K4me3 and repressive H3K27me3 at the gene promoter. Genes are poised for either activation or repression in the next stage of development.

1.3.7 Heterochromatin and H3K9me3

The H3K9me3 methyltransferase Setdb1 is required for embryogenesis as null mice embryos die shortly after implantation[110]. The pluripotency factor Oct4 complexes with the SUMOylated form of Setdb1/Eset to help maintain pluripotency in serum/Lif mESCs[111]. Setdb1 binds and represses lineage specific genes

and most imprinted genes through H3K9me3[112]. When *Setdb1* is KD in mESC they lose pluripotent phenotype, downregulated *OSN*, *Klf4*, and *Esrrb* and begin expressing differentiation factors like *Cdx2* which is normally repressed by H3K9me3 at its promoter[111,112]. Over half of Setdb1 target genes lose H3K9me3 during *Setdb1* KD in mESCs and KD of *Pou5f1* causes loss of Setdb1 binding at trophoblasts lineage genes *Cdx2* and *Tcfap2a* (*Tfap2a*) but not imprinted gene *H19*[112]. KD of *Setdb1* in mESCs differentiation to trophectoderm cells easier than WT and when injected into mouse embryos incorporate into the trophectoderm of blastocysts, due to upregulation of trophoblasts lineage genes[111,112].

As mentioned previously, naïve ESCs have less repressive chromatin structure, with fewer regions of their genome being marked by repressive histone modifications like H3K9me3. When comparing naïve 2iL mESCs to metastable serum/Lif mESCs, there is little difference observed in where H3K9me3 is deposited in the genome[45]. Many of the same regions are marked in naïve and metastable mESCs, mostly including satellite regions and imprinted genes[45]. Mouse ESCs grown in 2iL have H3K9me3 at methylated regions, even in its hypomethylated genome[43]. Methylated regions in 2iL also cover many IAP elements[43,44]. This presents a dual mechanism of repression of these retroviral elements via DNA methylation and histone modification.

Unlike H3K27me3, H3K9me3 regions identified in primed hESCs do not show the roughly the same DNA methylation levels in human preimplantation ICM as in primed hESCs. H3K9me3 regions are hypomethylated in human ICM (~25%) but have higher methylation levels in postimplantation embryos and primed hESCS (>50%)[54]. Although, what has not been investigated is how much of this similarity was driven by constitutive heterochromatin[54].

1.3.8 Interpretations and Opinions

H3K4me3, H3K4me1 and H3K27ac prove to be very important modifications in establishing active chromatin regions, resetting the early epigenome, and defining naïve and primed ESCs. Due to its biophysical properties, histone acetylation opens up the genome, creating the open chromatin structure of ESC. H3K4me3 broad domains during preimplantation help the embryo activate important genes during

ZGA. Influx of H3K4me1 in the the early zygote help to create the DNA hypomethylated environment of preimplantation due to the exclusion of H3K4 methylation and DNMTs. As ESCs undergo priming, they lose much of the H3K4me1 in their genome and gain DNA methylation. Primed ESCs further lose H3K4me1 as they differentiate but only after losing H3K27ac, which is inhibitory to the H3K3 demethylase.

Repressive modifications are depleted in naïve hESCs but develop in primed and are essential for proper development. The loss of active chromatin marks and gain of repressive chromatin helps to restrict cell fate and suppress pluripotency. The research presented in this thesis will add to the mechanism to how the loss of active chromatin occurs in the naïve hESC genome as cells transition to primed and provide a thorough view of the amount of repressive histone modifications that exist in the naïve hESC epigenome.

1.4 3D Architecture

1.4.1 Chromatin Organization and 3D Architecture

The genome is highly structured in three-dimensional space. In mammals, regulatory elements often are very far in linear space from the genes they regulate. In order to bring two distal regions together, chromatin will loop in on itself creating the basic unit chromatin organization. The inactive X chromosome is excellent example of chromatin organization. It forms an inactive chromatin compartment called the Barr body, named after Murray Barr who discovered it in 1949[113,114]. In addition to associating with HNRNPU (SAF-A), a nuclear scaffold protein, the inactive X also loses active histone modifications, gains H3K27me3 and DNA methylation, and becomes coated in the lncRNA XIST[114]. X inactivation is a specific example of how epigenetic features along with 3D architecture contribute to regulate gene expression.

1.4.2 Imaging the genome architecture

Original methods employed to look at chromatin organization took advantage of optical and electron microscopy. Electron spectroscopic imaging of developing mouse embryos provided great insight into the genome structure over the course of pre- and postimplantation. Initially, male and female pronuclei are structurally similar with only the 10 nm chromatin fiber detected in the embryo[115]. By the 2-cell stage,

chromatin domains and some compaction around the nuclear envelop have started to form[115]. The 4-cell nucleus shows increasing amounts of compaction, some domains present in the nuclear space and at the periphery[115]. The 8-cell nucleus is more distinguishable from the earlier stages. The chromatin is more "dispersed" throughout the nucleoplasm with fibrous regions and an increasing number of compacted domains intermingled throughout the nuclear space[115]. This "dispersed", or open, chromatin phenotype was also observed in epiblast cells of the E3.5 blastocyst[115]. In contrast trophectoderm and primitive endoderm, the first lineage committed cells, display higher levels of compaction and a more dense chromatin structure[115]. Postimplantation in the E5.5 blastocyst, the epiblast cells now look structurally indistinguishable from more differentiated extra-embryonic ectoderm cells, with much larger domains of condensed chromatin and few sparse fibers visible[115].

The genome of mESCs and other early progenitor stem cells *in vitro* mimic the structural nature of their *in vivo* counterparts[115] and the same for hESCs and early human progenitor stem cells who also reflect the structural attributes of their *in vivo* counterparts. Electron spectroscopic imaging of primed hESCs as they differentiate to neural progenitors captured a unique characteristic of stem cell chromatin architecture. Human ESCs contain a loose netting of chromatin spread more uniformly throughout the nucleus with some dense regions of chromatin observed around the nuclear periphery[116]. As cells undergo differentiation, larger compacted clumps of chromatin start forming, again mostly around the nuclear periphery[116]. The dense regions of chromatin observed under the microscope are the same as heterochromatic regions observed based on histone modifications (see Section 1.3). Therefore the imaging data allow visual detection of the gradual increase in repressive chromatin structure as ESC differentiate. The open chromatin feature of ESC can be observed via the enrichment of active histone modifications and less compacted nature of the genome.

This feature of disperse or open chromatin is an essential characteristic of the pluripotent epigenome. *Pou5f1* null mouse preimplantation epiblast cells, which lose their pluripotent potential by the blastocyst stage, show more condensed chromatin structure similar to the lineage restricted trophectoderm cells[115]. As mESCs differentiate into NPCs they also lose their open chromatin structure[117]. Immunostaining for

HP1 and H3K9me3 show few large, undefined regions in serum/Lif mESCs while NPCs have more small foci that covers a greater nuclear area[117]. ESCs also ungo a general loss of acetylated H3 and H4 histones and gain of H3K9me3 as they undergo differentiation[117]. Histones H1, H2B and H3 are more loosely bound in mESC chromatin, diffusing in and out of chromatin at a higher rate than they do in NPCs[117]. The high diffusion rate of histones could also be a result of the looser chromatin structure.

1.4.3 Interacting domains

Proteins involved in 3D genome interactions are important for ESC function and identity. The Mediator and Cohesin complexes bring ESC-specific enhancers and promoters in close proximity through enhancer-promoter looping[118]. Mediator complex, Cohesin complex, and the cohesin loading factor Nipbl all physically interact with each other and bind at the enhancers of ESC master regulators, *Pou5f1*, *Nanog*, and *Sox2*[118]. KD of any of the subunits of Mediator (*Med6*, *Med7*, *Med10*, *Med12*, *Med14*, *Med15*, *Med17*, *Med21*, *Med24*, *Med27*, *Med28* and *Med30*), Cohesin subunits (*Smc1a*, *Smc3* and *Stag2*) or *Nipbl* results in downregulation of *Pou5f1* expression and loss of stem cell characteristics in serum/Lif mESCs[118]. Proper enhancer-promoter interaction is necessary for cell identity. Cohensin-Mediator binding sites are much more cell-type specific than Cohesin-CTCF binding site[118], due to CTCF's more ubiquitous binding dynamics.

Arrangement of the genome within the 3D space of the nucleus is an attribute of cellular identity. Nuclear structural proteins additionally help with cell type identity by sequestering lowly expressed genes to the nuclear lamina in what are called lamina-associated domains (LADs). Roughly 40% of the mouse and human genome, ESCs and differentiated cells, reside in LADs and these LADs are >70% similar (occurring in the same regions of the genome) between ESC and 3 differentiated cell types (mouse embryonic fibroblasts, NPCs, astrocytes)[119]. While LADs are mostly depleted of genes, the genes that are localized to LADs are expressed in a much lower level than genes outside the domains[119]. As serum/Lif mESC differentiate to NPCs and further to astrocytes, genes or regions of genes move in and out of LADs and their transcription decreases or increases respectively[119]. Many of the genes that move between LADs are involved with cell type identity. For example, *Pou5f1*, *Nanog* and *Klf4* relocate to the

nuclear lamina and decrease their expression when mESCs are differentiated to NPC[119].

Of the sequencing methods to look at 3D structure of the embryonic genome, the most widely employed is Hi-C[120]. Additionally, the 3D structure of the genome has been studied more intensely in primed ESCs, progenitor and differentiated cells. The structure of the genome can be broken down into parts where chromatin looping and long-range interactions occur mainly within defined regions or topologically associated domains (TADs). TADs, which are on average ~185kb long[121], have defined boundaries. These boundaries separate interacting regions so that the highest levels of chromatin interaction are observed within a domain[121-123]. The boundaries of TADs are enriched for active chromatin modifications, such as H3K27ac, H3K9ac, H3K4me1, H3K4me3 and H3K36me3[122-125], DHSs, transcribed genes, especially housekeeping genes[121-123], CTCF and cohesin[121-123,126].

Hi-C in preimplantation vs postimplantation mouse embryos has shed light on how the 3D architecture of the developing embryo takes shape and corroborates well with microscopy studies. The male and female gametes are different in how their 3D genome is transformed during their maturation. Mature sperm have clear TAD structure with more extra-long range interactions (>2Mb) than what is observed in other cell types, with more inter-TAD and inter-chromosomal interactions than observed in somatic cells[124,127]. These observations make sense given the highly compacted, protamine-rich composition of the sperm genome. Immature, developing oocytes have chromatin loops and TADs but this structure is not tightly formed when comparing individual cells using single-nuclei Hi-C[128]. As oocytes mature to the MII stage, they lose their TAD structure and remain mostly unstructured after fertilization up to the 4 or 8 cell stage in mouse embryos[124,129]. The preimplantation mouse embryo has distinguishable A and B compartments (discussed further below) with differential CpG methylation, histone modifications and DHSs and the differences in epigenetic features become more telling as preimplantation embryos moves forward in development. Current research also suggests that rather than transcription, DNA replication is essential for the formation of TADs[124,130].

If the 3-dimensional structure of the genome is an epigenomic feature, one would expect it to be flexible to modulate the changes as embryos transition from pre-implantation to post-implantation and beyond. Likewise, you would expect the same in ESCs as they differentiate from naïve to primed and onward. However this is not necessarily the case. Comparison of H1 hESCs, progenitor stem cells and differentiated cells shows that TADs are mostly static between the different cell types with only hundreds changing between any given cell lineage[123]. TADs are also mostly static between serum/Lif mESCs and primed H1 hESCs at syntenic regions of the genome, indicating this is an evolutionarily conserved feature of mammalian genomes[122]. Much like differentiated cells, naïve cells also have a similar TAD structure when compared to primed cells[126]. In 2016, researchers Ji *et al.* came to this conclusion using a slightly different 3D sequencing method called ChIA-PET (chromatin interaction analysis by paired-end tag sequencing)[131]. They performed cohensin ChIA-PET on primed hESCs and compared them to hESCs that were reverted back to the naïve state in 5iLA growth condition. They found that cohensin-bound CTCF-CTCF loop regions could recapitulate TADs from Hi-C data and 80% were shared between naïve and primed cells. Most of the differences, however, were seen in the chromatin looping structure within the TADs. These also corresponded to changes in chromatin modifications, enhancer-promoter looping and changes in gene expression.

TADs are further defined by their active or repressed state and characterized as being in “A” or “B” compartment respectively[120,121]. Compartments can span multiple TADs in linear space for megabases. The active compartment A and repressive compartment B occupy different physical space in the nucleus[120] Like compartments (A with A or B with B) tend to cluster together and have higher interaction frequencies than disalike compartments[120,121]. TADs can switch between compartments in different cell types, the compartment switch is accompanied by a change in chromatin structure and physical rearrangement in the nucleus[121]. Compartments A and B can further be subdivided into subcompartments[121]. However these have not been fully investigated in the context of naïve and primed ESCs.

1.4.4 Interpretations and Opinions

ESCs model embryonic cells post-zygotic genome activation and therefore the changes in 3D structure pre-ZGA may not be observed in naïve and primed lines. In order to have a stable ESC culture, one requirement may be to have a stable 3D genomic structure. If this is the case, studying the establishment of TADs in human cells may not be possible with ESCs. However, so much of mammalian development is conserved, mouse and non-human primate research on 3D genome architecture will provide a lot of insight into our own development.

This manuscript will present the first naïve hESC Hi-C data and compare the naïve TAD structure to primed hESCs. I will show an integrated analysis of TADs and histone modification data, which will explain how boundaries of interacting domains are defined in naïve hESCs.

1.5 Other stem cell systems: ovarian cancer stem cells

1.5.1 Introduction to ovarian cancer stem cells

Ovarian cancer is the fifth leading cause of cancer mortality among women in the United States[132,133]. Despite aggressive surgery and chemotherapy, most ovarian cancer patients experience tumor relapse and develop drug-resistant tumors. This relapse and resistance, along with late stage diagnosis, contribute to the <40% chance of 5 year survival in ovarian cancer patients[132]. Evidence suggests that a small population of ovarian cancer stem cells (OvCSCs) is responsible for tumor relapse and drug resistance[132,134]. OvCSCs have special properties such as the ability to self-renew or divide asymmetrically, enhanced tumorigenicity and inherent chemoresistance. Because of their unique attributes, OvCSCs are thought to be the multipotent drivers behind tumor growth (Figure 1.5).

OvCSCs have been identified identified using different cell surface markers, CD44, CD24, CD117, and CD133[135]. Different studies have used these markers individually or in combination to identify the OvCSC population in a tumor. OvCSCs have also been identified from their aldehyde dehydrogenase activity, rapid efflux of Hoechst dye due to the expression of ABC transporters in the cell surface, or their quiescence (slow division rates)[135]. Moreover, OvCSCs identified using these markers or methods must all pass the gold standard test of multipotency. CSCs must be able to generate the entire heterogeneity of

the parent tumor. This can be tested in the lab by transplanting a pure population of CSCs into a xenograft system and dissecting the resulting tumor.

The different markers for OvCSCs could also be due to the heterogenous nature of the disease itself. There are many different types of ovarian cancer and of epithelial ovarian cancers, the cell type of origin has been proposed to be either the ovarian surface epithelium or fallopian tube epithelium. The latter of the two has shown to be the source for high-grade serous carcinoma ovarian cancer[135]. The multifarious nature of the disease and markers used to identified CSC populations should be taken into account when interpreting genetic data on ovarian cancer. However the unique property of multipotent cells, the ability to differentiate, is best defined as an epigenetic feature.

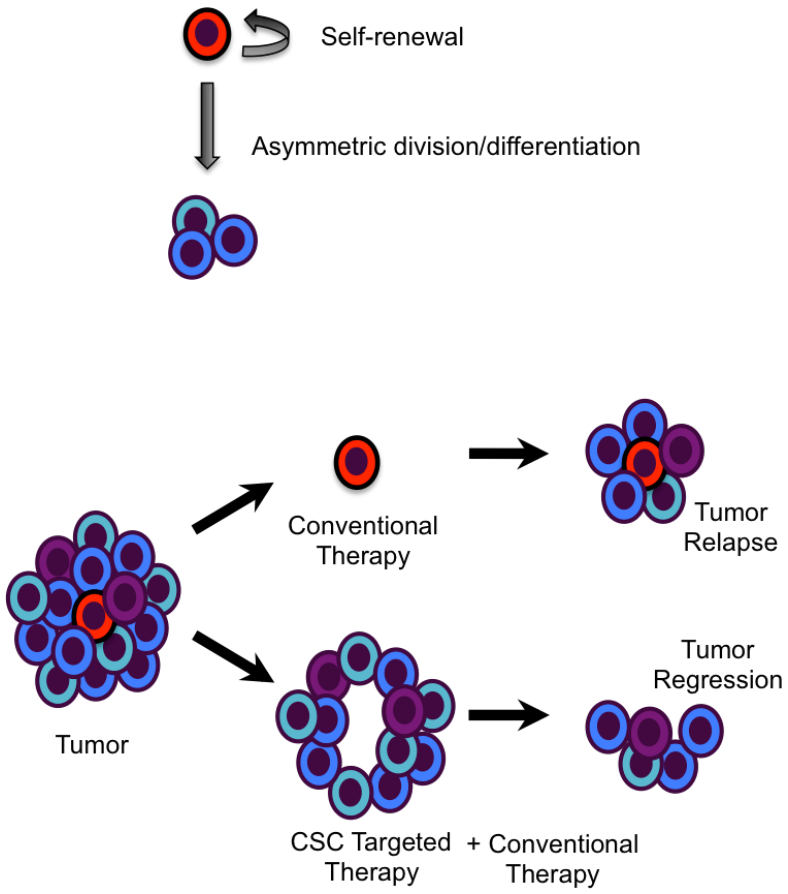


Figure 1.5 Cancer Stem Cells and Tumor Growth
Cancer stem cells have stem cell-like properties such as the ability to self-renew and differentiate. The cancer stem cell hypothesis prostualtes that cancer stem cells are the source of tumor growth and have the ability to generate all the heterogeneous cells of the tumor. Conventional chemotherapy target the rapidly dividing cells that make up the bulk of the tumor but resistant CSCs regenerate tumor mass. However combining chemotherapy with CSC targeted therapy may lead to tumor remission.

1.5.2 Epigenetics of ovarian cancer stem cells

Gene regulation through epigenetic mechanisms helps form cellular identity and cell fate. This is achieved via DNA methylation, non-coding RNAs, and chromatin modifications that uniquely mark regulatory elements. Epigenomes in pluripotent stem cells and cancer have been studied in depth to reveal key regulatory features of these cell states. Much like embryonic and adult stem cells have epigenetic features that distinguish them from differentiated cells, CSCs also have regulatory differences that distinguish them from their daughters. Studying the epigenome of OvCSCs may reveal marks of multipotency and provide insight into OvCSC regulation and differentiation.

In addition to their role in differentiation, epigenetic alterations are known to play an important role in cancer development and progression[134]. Aberrant epigenetic changes in ovarian cancer, such as abnormal DNA methylation (DNAme) and loss/gain of histone modifications, have been studied previously. For example, loss of repressive histone marks is one mechanism for activating genes responsible for ovarian tumor progression, such as claudin-3 (*CLDN3*), claudin-4 (*CLDN4*) and *RASSF1*[136,137]. Other studies have found promoter DNA hypermethylation at *BRCA1* and *PTEN* to be markers for ovarian cancer progression in some patients[138,139]. Hypermethylation of genes involved in the TGF-beta and Wnt pathways may also be potential biomarkers for drug response or clinical outcome in a fraction of ovarian cancers[134]. These studies and many others only focus on the correlation between one epigenetic feature and gene expression to define the current state of a cell.

1.5.3 Interpretations and Opinions

By studying the epigenome of OvCSCs, we will learn more about how these cells are functionally different from the non-stem cancer cells that make up the bulk of the tumor. This study proposes to integrate all elements of the epigenome; therefore defining the cell's current state and what the cell has the potential to do.

Chapter 2:

DNA methylation and Hydroxymethylation in Naïve hESCs

2.1 Motivation

The dynamics of DNA methylation during embryonic development have been well documented. The rapid decrease after fertilization, followed by increase post-implantation of 5mC (mentioned previously) makes it the most dynamic epigenomic modification during early development. Naïve and primed hESCs capture these snapshots of 5mC levels to varying degrees. In contrast to the deluge of immunofluorescence, affinity purification, mass spectrometry, and NGS data generated on 5mC in early development, most of what we know about 5hmC in early development comes from immunofluorescence studies in mouse and more recently human embryos and affinity purification-sequencing in naïve mESCs and primed hESCs. Immunofluorescence studies in fertilized mouse, rabbit and bovine embryos have directly linked the accumulation of 5hmC on the male pronucleus with the timing of 5mC removal[62]. In human zygotes, the paternal genome also has much higher levels of hmC while the maternal genome retains higher levels of 5mC[63]. The difference in maternal-paternal 5hmC and 5mC levels decreases over the course of preimplantation development as the zygotic genome generally becomes more hypomethylated[63].

Yu *et al.* published TAB-Seq in 2012, a method to detect 5hmC at single nucleotide resolution. With this method they were able to confirm many of the previous findings about 5hmC genomic localization in naïve mESCs and primed hESCs including, 5hmC enrichment at regulatory elements. To date, 5hmC has not been studied at single nucleotide resolution in naïve hESCs. Here we present the first TAB-seq data on naïve hESCs, using the Elf1 cell line grown in 2iL+IF conditions. When appropriate we use H1 and Elf1 AF as primed cell comparison groups and naïve mESCs. We find that the naïve epigenome is hypomethylated relative to primed hESCs and hyper-hydroxymethylated. The hydroxymethylation is enriched at regulatory elements, specifically enhancers, an indication that active demethylation is occurring at these regions. We detected a slight gain in methylation at imprinted regions as Elf1 cells are pushed forward to a more primed like state. Taken together these data suggests that Elf1 cells have a

unique DNA methylation profile compared to primed hESCs and may be able to mimic the DNA methylation dynamics of mammalian embryos.

2.2 Methods

2.2.1 Cell growth conditions

ESC culture conditions were as previously described[2], with the following modifications. Naïve growth conditions: 2iL+IF - 1uM Mek inhibitor (PD0325901) [catalog #S1036, Selleck Chemicals 1uM GSK3 inhibitor (CHIR-99021) [catalog #S2924, Selleck Chemicals, 10 ng/mL Leukemia inhibitory factor [catalog #YSP1249, Speed Biosystems], 5ng/mL IGF-1 [catalog #100-11, Peprotech], 10 ng/mL FGF [catalog #PHG0263, Thermo Fisher Scientific]. Elf1 AF growth conditions: TeSR1 [STEMCELL Technologies], 10ng/mL bFGF, 10 ng/mL activin A [Humanzyme].

2.2.2 WGBS libraries preparation and analysis

Purified whole genomic DNA was sonicated using Covaris to ~300bp. DNA fragments were end repaired, A-tailed and ligated to methylated Y-adapters (5'ACACTCTTTCCCTACACGACGCTCTTCCGATCxT and 5'-phosphate-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG). DNA library was bisulfite treated using Thermo Fisher MethylCode Bisulfite Conversion (Cat #MECOV50) and the purified library was subjected to PCR to add a index and full length Illumina sequencing adapter (primer sequence CAAGCAGAAGACGGCATACGAGATXXXXXXXXCGGTCTCGGCATTCCTGCTGAACCG). Quality of libraries was checked on a Bioanalyzer and libraries were sequenced PE100 on an Illumina HiSeq 2000. Libraries were trimmed used Trim Galore v 0.4.1 and mapped using Bismark[140]. Methylation at nucleotide resolution was called using MethPipe[141].

2.2.3 TAB-Seq libraries preparation and analysis

The 5hmC TAB-Seq Kit was purchased from Wisegene and used to create 5hmC libraries from purified whole genomic DNA following manufacturer's protocol. Quality of libraries was checked on Bioanalyzer and libraries were sequenced PE100 on an Illumina HiSeq 2000. Libraries were trimmed used Trim

Galore v 0.4.1 and mapped using Bismark[140]. Hydroxymethylation at nucleotide resolution was called using MethPipe[141].

Confident 5hmC calls for naïve cells were calculated using MethPipe's maximum likelihood methylation level (mlml) program[1]. Briefly, this program combines information from both WGBS and TAB-Seq to estimate 5mC and 5hmC levels. This method is more accurate than looking at TAB-Seq data alone for estimating 5hmC level. CpGs with methylation or hydroxymethylation levels of at least 10% were counted as "methylated" cytosines. To increase our confidence of 5hmC, we choose a 10x coverage cutoff. For 5mC we choose a 5x coverage cutoff. Only CpGs meeting the coverage and methylation cutoff were used in downstream analysis. Confident CpGs were not recalculated for the previously published data but used the same coverage cutoffs as our data, 5x for 5mC and 10x for 5hmC.

2.3 Results

2.3.1 Global 5mC and 5hmC levels in naïve and primed hESCs

DNA methylation patterns have been investigated in naïve hESCs grown in 4iLA, 5iLA and t2iL+Gö previously[33,57]. However to date, DNA methylation and hydroxymethylation have not been investigated in the same naïve hESC at nucleotide resolution, genome-wide. In order to assess DNA methylation and hydroxymethylation, genome-wide at single nucleotide resolution, we performed whole genome bisulfite sequencing (WGBS)[40] and Tet-assisted bisulfite sequencing (TAB-seq)[142] in the Elf1 naïve hESCs and compared it to previously published data from H1 primed hESCs[61,72,141]. In total 403,993,816 million reads of WGBS data were generated and 730,564,043 million reads of TAB-seq data were generated in the Elf1 line (Table A.1). After mapping and filtering for quality, we covered a total of 50,831,088 and 52,334,331 million stranded, autosomal CpGs at an average coverage of 6x in the WGBS library and 11x in the TAB-seq library.

We restricted our analysis to just cytosines in a CpG context. Previously, 5hmC has been shown to be asymmetric across the CG dyad, for this reason we did not merge CpG levels across dyads, keeping the positive and negative strand separate for all downstream analysis. We used the program MethPipe[141] to

call methylation levels genome-wide. Since WGBS cannot distinguish between 5mC and 5hmC, it provides an overestimation of the 5mC level. MethPipe is able to take both the WGBS and TAB-seq libraries to determine confident 5hmC sites as well as calculate methylation and hydroxymethylation levels (Figure 2.1A,B, see methods). Most 5hmC occurs at cytosines that have <40% 5mC level as detected by WGBS (Figure 2.1A). After calculating the 5hmC levels and subtracting it from the WGBS 5mC levels, we observed a reduction in the 5mC levels (Figure 2.1B). The overall distribution of 5mC displayed a bimodal distribution while 5hmC was skewed to mostly low hydroxy levels (Figure A.1).

The preimplantation epigenome is hypomethylated compared to the postimplantation state. Since naïve ESCs are supposed to represent the preimplantation epiblast, we hypothesized that our naïve hESCs would be hypomethylated compared to H1 primed[40,61]. Indeed we found naïve Elf1 hESCs to have a lower global methylation level, 62%, compared to primed hESCs H1 82% (Figure 2.1C). However, we noted that our naïve cells were not as hypomethylated as other naïve lines or human preimplantation blastocysts which have average CG methylation at ~30%[57]. Naïve hESCs were hyper-hydroxymethylated compared to primed, at an average of 15% compared to 2.5%, respectively (Figure 2.1C). The naïve hESC values we observed were similar to the levels reported in naïve mESCs (~20% 5hmC and ~60% 5mC)[72]. Thus, even though the genome has higher average 5mC than other naïve hESCs, it is hypomethylated with respects to primed hESCs. Additionally Elf1 naïve hESCs have a higher average 5hmC level, which is evidence of active demethylation throughout the genome.

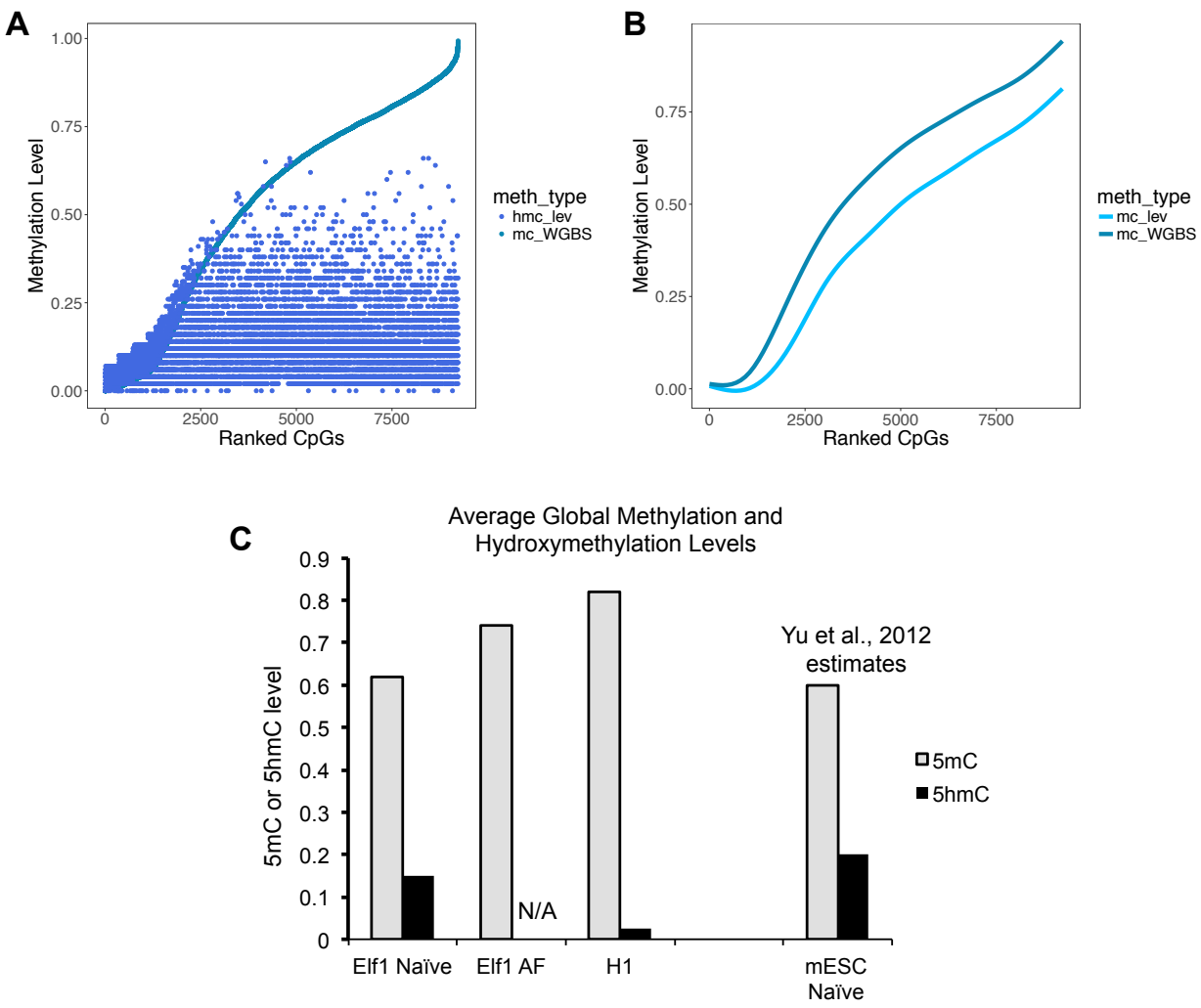


Figure 2.1 5hmC and 5mC level in naïve hESCs
(A) 10,000 CpGs are ranked and plotted based on their WGBS predicted 5mC level. Dots are their calculated 5hmC level from MethPipe's mlml program[1]. (B) Graph of 10,000 CpGs and their 5mC level as predicted based on WGBS data compared to their 5mC calculated level after subtracting 5hmC. (C) Average 5mC and 5hmC levels for each ESC cell type. 5mC cutoff of 5x and 5hmC cutoff of 10x. Mouse ESC averages taken from Yu *et al.* 2012.

2.3.2 Hydroxymethylation at known regulatory elements

We called CpGs with 5mC/5hmC levels above 10% as methylated/hydroxymethylated. We used H1 primed data that were analyzed using MethPipe and provided on UCSC by the Smith Lab. Based on our cutoff there are 16,978,550 naïve and 2,833,400 primed 5hmCs. This is comparable to Yu *et al.* 5hmC calls of 691,414 at 20% hydroxymethylation level as we count 669,292 5hmCs at 20% hydroxy level. We chose a less stringent cutoff of 10% because 5hmC is known to be a transient intermediate in the active

demethylation pathway and exists at very low levels in the genome compared to 5mC. Also by applying a coverage cutoff of 10x, we can more accurately call 5hmC at 10% hydroxymethylation level.

In naïve mESCs and primed hESCs, 5hmC has previously been shown to be enriched at CGIs, promoters, exons, enhancers and flanking TF/protein binding sites like at CTCF, p300, OSN (defined in Table 1.1) and other key pluripotency TF[72,143-148]. We hypothesize that we would observe a similar enrichment pattern in naïve hESCs. We asked what were the average methylation/hydroxymethylation levels at CpG Islands (CGIs), promoters, and enhancers in naïve hESCs. Enhancer regions were identified based on ChIP-seq for H3K4me1 as described in Chapter 3. We further looked at two of the three classes of enhancers, active and poised. Active enhancers have both H3K4me1 and H3K27ac modifications in the same region and poised enhancers contain only H3K4me1 (distinguished from H3K4me1 + H3K27me3 poised enhancers). Our data show a general depletion of 5mC at these regulatory elements (Figure 2.2A). We observed an enrichment of 5hmC at enhancers and CGIs but not at promoters (Figure 2.2A).

Because there is more 5hmC in the naïve genome, we were curious if the 5hmC sites were distributed similarly across genic regions in both cell types. Over 50% of 5hmC sites are found in genes in both naïve and primed hESCs. We identified the number of 5hmC sites in promoter, 5' UTR, first exon, exons, and 3' UTR (Figure 2.2C,D). While the overall number of 5hmC sites is higher in naïve, the fraction of total 5hmC sites in each type of genic region is roughly the same in naïve as it is in primed. This suggests that the expansion of 5hmC in the naïve genome is not specific to a distinct genic category.

We next asked how CpGs were distributed across CPG islands (CGIs), shores and shelves. We didn't observe major differences in the fraction of 5hmCs at CGIs or shelves. Naïve hESCs have 2.8% and 4.4% of their 5hmC in CGIs and shelves respectively while primed hESCs had 4.5% and 5.7% of their 5hmC sites in CGIs and shelves (Figure 2.2E). Interestingly there is less proportionality at shores in naïve and primed. Naïve have 6.6% of 5hmC in shores while primed have 11.4%. The shores of CGIs have

been identified as regions whose dynamic methylation status is more reflective of gene expression levels than methylation status of CGIs themselves and are methylated in a tissue-specific manner[149].

We asked if naïve 5hmC was more frequent at naïve enhancers. We identified 2,180,763 (12.8%) of naïve 5hmC sites in naïve H3K4me1 enhancers and 315,189 (11.12%) of primed 5hmC in primed enhancers (Figure 2.2F). Roughly 95% of naïve enhancers contained at least one 5hmC site. In primed cells 5hmC present in 90% of primed enhancers (Figure 2.2G). Despite there being 2-fold as many poised naïve enhancers than active naïve enhancer (65,334 vs 31,673), there are almost equal number of 5hmC found at both (6.4% and 6.2% respectively; Figure 2.2F). A slightly different trend was observed in primed hESCs. Poised enhancers have 7.7% of 5hmC sites which is >2.5-fold more than at active enhancers (2.8%). Thus while there is roughly the same fraction of 5hmC at enhancers in naïve and primed cells, in naïve cells, there is proportionally more 5hmC at active enhancers.

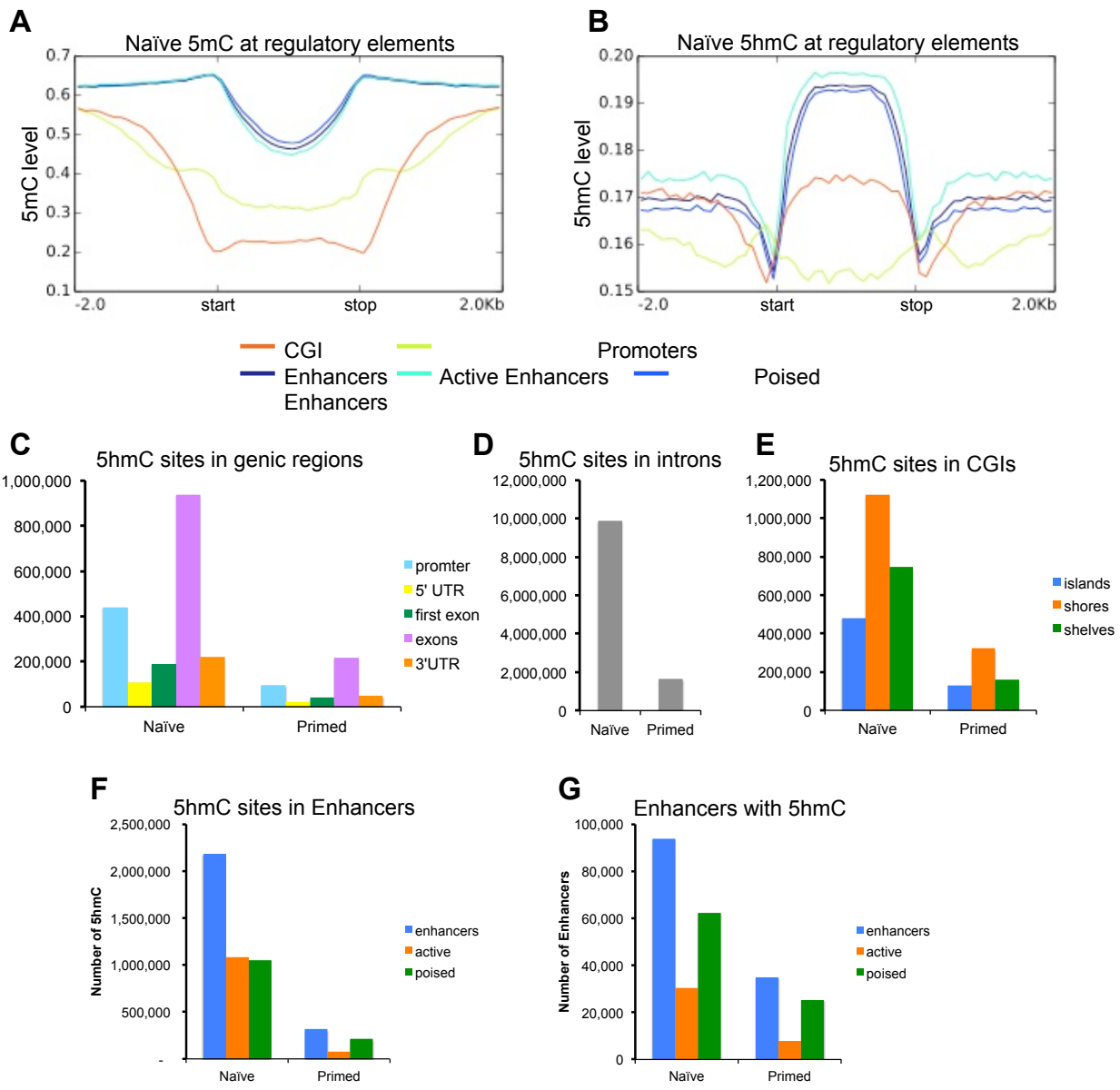
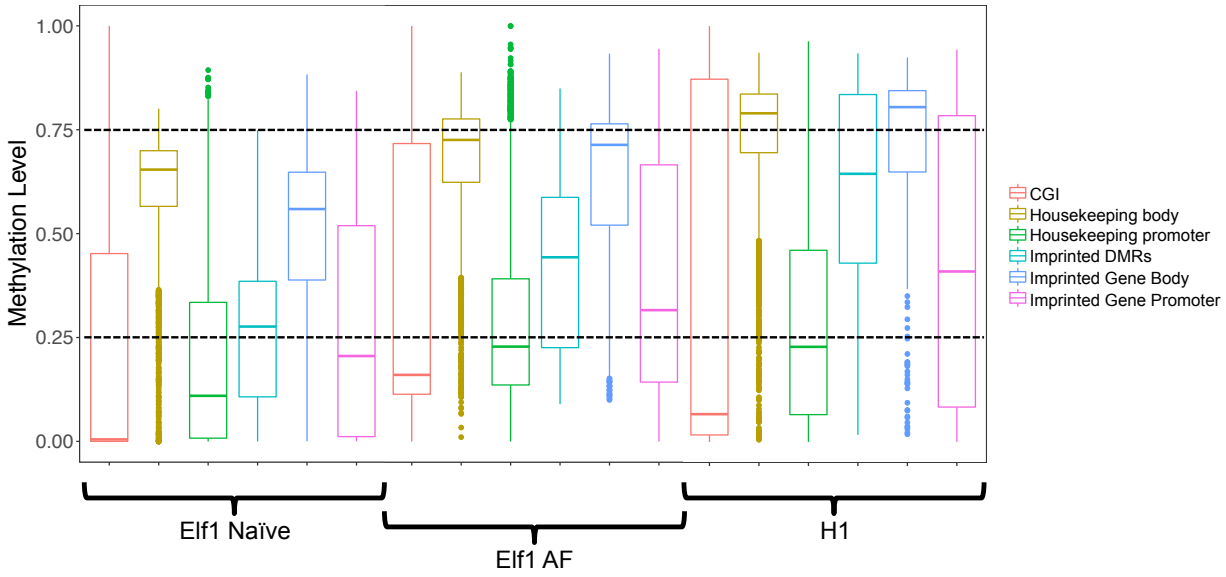


Figure 2.2 5hmC Distribution at Regulatory Elements
(A) Average 5mC level across regulatory elements. Naïve enhances are defined in Chapter 3. (B) Average 5hmC level across regulatory elements. (C) Number of 5hmC sites at genic regions. (D) Number of 5hmC site at introns, shown on a different graph from other genic regions due to scale. (E) Number of 5hmC site at CG islands, shores and shelves. (F) Number of 5hmC sites in enhancers in each cell type (G) Number of naïve or primed enhancers with 5hmC sites.

2.3.3 Imprinting in naïve hESCs

Imprinting in the mammalian epiblast mainly occurs postimplantation although some regions retain their imprinted status from germ cells and throughout early embryonic development[55,150]. Prior to imprinting, there is biallelic expression of the imprinted gene and hypomethylation at its regulatory locus (identified as a differentially methylated region (DMR) between the two alleles) and/or promoter. Upon priming, naïve hESCs should silence one allele and methylated the silenced alleles regulatory elements. Other naïve hESC lines have been shown to not properly undergo imprinting when transitioned to the primed state, that is, they fail to methylated one of the alleles and retain biallelic expression[33,57]. This has not yet been investigated in Elf1 naïve hESCs. We hypothesize that Elf1 cells will show a gain of methylation at the promoters and/or regulatory elements of imprinted genes when pushed forward to a primed state. By examining the 5mC data at imprinted regions of the genome, we can easily assess if there is a gain of DNA methylation in the Elf1 AF condition, which is a semi-primed state. There was a noticeable increase in methylation levels at imprinted differentially methylated regions[33] and imprinted gene promoters (Figure 2.3A). We looked at the expression of imprinted genes using RNA-seq data generated in Chapter 3. For all imprinted gene investigated, there was no significant change in expression between the cell types even though methylation at these gene regulatory elements increases in Elf1 AF and H1 cells (Figure 2.3B). This is likely an indication of the cell regulating the level of expression even though one allele is shut off. The biggest exception to this observation is the gene H19, which is the highest expressed gene in naïve hESCs and has a unique regulatory landscape (discussed more in Chapter 3).

A



B

RPKM values of Imprinted Genes

	Expressed	Elf1 Naïve	Elf1 AF	H1 Primed
*H19***	Maternal	240	9	5
LIN28B	Paternal	38	50	29
PEG10	Paternal	23	14	23
MEST	Paternal	96	89	129
NTM	Maternal	3	3	8
UBE3A	Maternal	44	48	35

Figure 2.3 5mC at Imprinted Regions
(A) Distribution of 5mC level at regulatory elements and imprinted regions for Elf1 naïve, Elf1 AF and H1 primed hESCs. Regulatory elements associated with imprinted regions are hypermethylated in primed cells compared to naïve. Imprinted DMRs are as defined in Theunissen, 2016. (B) Table of RPKM values of select imprinted genes. H19 is one of the highest expressed genes in naïve hESCs and discussed further in Chapter 3.

2.4 Summary and Conclusions

Here, we present the first single nucleotide resolution, genome-wide data on 5hmC in naïve hESCs. We observed 8x as many 5hmC sites in the naïve genome compared to primed. While these sites were proportionately distributed at genic regions, we noticed a larger portion of them at active enhancers in the naïve genome. Co-marking of 5mC and 5hmC has been noted in primed hESCs and naïve mESCs

previously[72] and we observe that many of the lowly methylated 5mC have 5hmC. It is plausible that these sites are dynamically regulated at the level of DNA methylation, constantly teetering between methylated-hydroxymethylated-unmethylated states. They are under active demethylation in naïve hESCs, which likely creates their relatively hypomethylated genome. Additionally we observed that Elf1 cells are the only derived naïve hESCs to date that appear to undergo imprinting. While more work is needed to determine the extent of imprinting as Elf1 cells undergo differentiation, this preliminary work is promising. In order to be able to study early development, researchers need naïve ESCs that model embryonic development. Other naïve hESC lines are resistant to imprinting, a normal developmental process, and were shown to be difficult to maintain long-term in culture[57]. Therefore while other naïve hESCs may more accurately reflect the epigenome of preimplantation blastocysts, at least at the level of DNA hypomethylation, Elf1 hESCs may be better suited as models of development and differentiation.

Chapter 3:

Chromatin and 3D Architecture of Naïve hESCs

Note: Parts of this chapter are available online on BioRxiv as Battle S, Jayavelu N, Azad R, Hesson J, Ahmed F, Zoller J, Mathieu J, Ruohola-Baker H, Ware C, Hawkins R. Epigenomic and 3D genome architecture in naïve and primed human embryonic stem cell states. Aug 26, 2017. doi: https://doi.org/10.1101/181123

3.1 Motivation

Dynamic changes in the epigenome are concerted with morphological and gene expression changes during early embryogenesis. Soon after fertilization DNA methylation is actively removed from the paternal genome, passively lost from the maternal genome and regained in the post-implantation epiblast[54]. In addition to resetting the DNA methylome, the early embryonic epigenome maintains an open chromatin structure as repressive heterochromatin is gained later over the course of development, lineage commitment and differentiation[77,115,151]. These changes in histone modifications correlate with the hypothesis that a more open chromatin structure is a key aspect of pluripotency and allows embryonic cells to respond to a broad array of developmental signaling cues[152,153].

Pre- and postimplantation pluripotent ESCs provide a system to model epigenomic reprogramming during early embryogenesis and to study changes in pluripotency. Mouse ESCs (mESCs) are currently the primary model for studying mammalian pre-implantation embryos and deemed naïve[12], while mouse epiblast stem cells (EpiSCs) model the post-implantation embryo and exist in the primed state of pluripotency[10,11]. Due to a number of similarities between mouse EpiSCs and human ESCs (hESCs), it is now accepted that hESCs exist in the primed state[13]. However, several groups described the first set of naïve hESCs, where primed hESCs or human iPSCs were induced, or reset, to the naïve state[20-24,26,31]. Additionally, new hESC lines were derived, each under a different naïve growth condition[22,23,31,34] (for review see[154]). Similar to mouse, naïve hESCs exhibit DNA hypomethylation and two active X chromosomes[22,31,33], hallmarks of the pre-implantation state.

Given the differences between early human and mouse embryogenesis[29,155], naïve-derived hESC lines provide an opportunity to study changes that are reflective of early human development and pluripotency. The purpose of this project was to better our understanding of epigenomic reprogramming as hESCs transition from the pre-implantation to post-implantation state. We present data from whole transcriptome RNA-seq, ChIP-seq for five histone modifications, and topological associated domains (TADs) from in situ DNaseI Hi-C for the naïve-derived Elf1 line[31] grown in 2i + Lif + IGF1 + FGF (2iL+IF). We include data from cells transitioning from the naïve state (Activin + FGF noted as AF) and compared our results to data from primed H1 hESC[122,153]. Extensive chromatin remodeling occurs at promoters and enhancer elements as cells transition from naïve to primed. Our analysis reveals that naïve hESCs have a more open chromatin structure due to large expansions of H3K4me1 and H3K27ac in the genome. Seventy-seven percent of naïve enhancers are decommissioned in the primed state. TADs are largely stable between pluripotent states, but our data reveal limited naïve specific shifts in TAD boundaries. Overall, these data provide an extensive view of the epigenome and 3D genome for hESC states and a model of epigenomic reprogramming during early human embryogenesis.

3.2 Methods

3.2.1 Human Embryonic Stem Cell Culture

All human ESC culture conditions were as previously described[2], with the following modifications. Growth conditions: 2iL+IF - 1uM Mek inhibitor (PD0325901) [catalog #S1036, Selleck Chemicals, Houston, TX, USA], 1uM GSK3 inhibitor (CHIR-99021) [catalog #S2924, Selleck Chemicals, Houston, TX, USA], 10 ng/mL Leukemia inhibitory factor [catalog #YSP1249, Speed Biosystems, Gaithersburg, MD, USA], 5ng/mL IGF-1 [catalog #100-11 Peprotech, Rocky Hill, NJ], 10ng/mL FGF [catalog #PHG0263, Thermo Fisher Scientific, Waltham, MA, USA]; 4iL+IF - 1uM Mek inhibitor (PD0325901), 1uM GSK3 inhibitor (CHIR-99021), 5uM JNK inhibitor (SP600125) [catalog #S1460, Selleck Chemicals, Houston, TX, USA], 2uMp38 inhibitor (BIRB796) [catalog #S1574, Selleck Chemicals, Houston, TX, USA], 10 ng/mL Leukemia inhibitory factor, 5ng/mL IGF-1, 10ng/mL FGF.

3.2.2 Chromatin Immunoprecipitation and Sequencing (ChIP-seq) hoo

ChIP-seq was performed as previously described[156]. Raw sequence reads from Roadmap Epigenome Project[157]. All sequenced reads were analyzed with the same pipeline and settings. Sequence reads were aligned to genome (version hg19) using Bowtie2[158]. Replicates of aligned files were merged prior to peak calling. For the UCSC genome browser tracks, ChIP-seq signals were normalized by RPKM followed by subtraction of input from ChIP using deepTools suite[159]. Heatmaps and histograms are of normalized ChIP-seq signal: samples are normalized by read count and log2(chip reads/input reads) per 10kb bin is plotted using deepTools suite[159].

3.2.3 Peak Calling

ChIP-seq peaks were called on merged replicates and normalized to input using MACS v1.4[160]. Peak calls with a FDR of 5% or less were used for downstream analysis. Percent of genome covered was defined as total number of bases under the peak divided by 2.7e9, the effective genome size. This was found it to be a better representation of global chromatin structure (e.g. a 10kb region can be covered by one or many ChIP-seq peaks due to peak size; the number of peaks may vary more than the total number of bases under the peaks). Peak comparisons and overlaps were done using the BedTools suite [161].

In order to compare the histone marks (H3K4me1 and H3K27ac) across cell types, we divided the genome into 10 kb bins and counted the reads across these 10 kb genomic regions using featurecounts in Rsubread package[162]. Then, PCA was performed on regularized log transformed read count data obtained using DESeq2[163].

3.2.4 RNA-seq and Gene Expression

Embryonic stem cells were counted and 200,000 cells were pelleted for RNA extraction using the Qiagen All Prep Kit (cat #). RNA-seq libraries were constructed using the Scriptseq RNA-seq Library Preparation Kit on ¾ of total RNA. Libraries were sequenced single-end 75 on Illumina NextSeq. The quality of the reads and contamination of adapter sequences were checked with FastQC tool

(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped to human hg19 genome (UCSC) using TopHat2[164]. Transcript quantification was performed by Cufflinks[165] using GENCODE's comprehensive gene annotation release 19 as reference annotation.

3.2.5 Differential Gene Expression Analysis

The raw read counts were calculated using featurecounts in Rsubread package[162] and GENCODE's release 19 as reference annotation. Differential gene expression analysis was performed with DESeq2[163] using read counts matrix. Two sets of differentially expressed genes (DEGs) are identified with P-value < 0.01, log2FC > |1| and P-value < 0.01, log2FC > |2|. The P-values were adjusted for multiple hypothesis correction. DEGs in all pairwise sample comparisons were identified. PCA was performed on regularized log transformed read count data from autosomes of top 500 highly variant genes obtained using DESeq2[163] and plot was generated using ggplot2 in R[166].

For transposable elements (TE) analysis, transcripts were quantified using hg19 UCSC RepeatMasker TE annotation. We considered unique reads as well as multi mapped reads during quantification of TE transcripts. PCA was performed on regularized log transformed read count data of top 500 highly variant TE transcripts obtained using DESeq2.

3.2.6 Identification of Overrepresented GO Terms and Enriched Pathways

ClueGO[167] was used to identify the overrepresented GO terms and enriched pathways with the data from gene ontology consortium and KEGG pathways database. The input gene lists to the ClueGO were DEGs with P-value < 0.01, log2FC > |1|. We used all genes in the genome as background. The statistically significant GO terms and pathways were filtered with P-value < 0.05 and GO term/pathway should contain at least 5 DEGs. P-values were adjusted with Benjamini Hochberg method for multiple hypothesis correction.

3.2.7 Sankey Plot

We looked at promoter chromatin state transitions from naïve to primed to gain insight into the

establishment of bivalency and other chromatin state changes occurring at gene promoters. In order to accomplish this goal we focused on the over 19,000 autosomal protein-coding gene TSS annotated by GENCODE. We assigned a promoter to a gene if the H3K4me3 peak was within -2kb to +500bp of the TSS. Sankey plot is limited by by the presence of multiple promoters. Sankey plot were created using Google Charts (https://developers.google.com/chart/interactive/docs/gallery/sankey)

3.2.8 In situ DNase Hi-C

Samples were prepared in a manner similar to Deng *et al.*, 2015[168]. Briefly, nuclei from ~5 x 106 cross-linked Elf1 cells were isolated and permeabilized, and chromatin was digested with 4 U DNase I at room temperature for 4 min. Following end-repair and dA-tailing reactions, chromatin ends were ligated to biotinylated bridge adapters, and nuclei were purified with two volumes of AMPure XP beads (Beckman Coulter). Chromatin ends were phosphorylated and ligated in situ, and protein-DNA cross-links were reversed by proteinase K digestion and incubation at 60oC overnight. Following purification, DNA was sonicated to an average size of 400 bp, and chimeric species were enriched via pull-down with streptavidin-coated magnetic beads (Active Motif). Preparation of Hi-C libraries was accomplished by ligating sequencing adapters to the ends of bead-bound DNA fragments and PCR-amplifying the products in the presence of forward and barcoded reverse primers. Libraries were purified with AMPure XP beads, DNA concentrations were determined using a Qubit 2.0 (Thermo Fisher), and size distributions were quantified using a Bioanalyzer with a high sensitivity kit (Agilent). A 10 ng aliquot from each library was digested with BamHI, run on the Bioanalyzer, and compared to an undigested control in order to confirm the presence of a reconstituted BamHI site at the junctions of ligated bridge adapters.

3.2.9 Hi-C Sequencing and Data Processing

Raw Hi-C sequencing reads from H1 hESCs were downloaded from GEO (GSE35156). Reads were aligned using Bowtie2[158] to the hg19 reference genome and filtered for MAPQ ≥ 10, uninformative ligation products, and PCR duplicates using HiC-Pro.

Valid Hi-C read pairs from biological replicates of Elf1 and H1 hESCs were combined, respectively, and used to generate raw chromosome-wide interaction matrices binned at a resolution of 40kb. Raw matrices were ICE-normalized using the HiTC Bioconductor package[169] for R, and TADs and boundaries were identified using TopDom[170] with a window size of 5. X and Y chromosomes were removed for the datasets for all Hi-C analyses.

Insulation scores were calculated for the whole of chr7 from the ICE-normalized matrices of both the Elf1 and H1 hESCs, separately. Insulation vectors were detected via cworld[171] using the script matrix2insulation.pl, and using the following options: (--is 240000 --nt 0.1 --ids 160000 --im median --bmoe 0). Differential insulation scores computing Elf1 score minus H1 score for the whole of chr7 via cworld using the script compareInsulation.pl, with inputs being the two insulation scores above.

High-Confidence SMC1 ChIA-PET interactions for naïve and primed hESCs were downloaded as a supplemental table[126]. A ChIA-PET was considered to span a TAD if both PET termini were located within 40 kb of a TAD boundary.

Spatial compartments and activity status were identified via principal component analysis (PCA) using Homer Tools[172]. Processed Hi-C reads were imported into Homer. For each chromosome, a contact matrix was constructed at 40 kb resolution and normalized using a sliding window of 400 kb as background. Next, the correlation between intra-chromosomal contact profiles was computed and the first principal component (PC1) vector was extracted and saved as a bedGraph file. H3K27ac ChIP-seq peaks served as a seed for determining which regions are active (PC1 > 0). A genomic region was considered cell type-specific if it met the following three criteria: 1) the average PC1 value was positive in one cell type and negative in the other, 2) the difference in the average PC1 value was > 50 and 3) the correlation

between contact profiles was < 0.4. Randomization was achieved by selecting coordinates from a pool of 40 kb regions that had associated PC1 values and were not located within any cell type-specific sub-compartments.

3.3 Results

3.3.1 Gene Expression in Naïve hESCs

Naïve and primed hESCs are expected to have distinct expression profiles, and naïve cells should reflect aspects of human blastocyst gene expression. We performed strand-specific, whole transcriptome RNA-seq in replicate on Elf1 naïve (2iL+IF), Elf1 transitioning (AF) and H1 primed (mTeSR) cells of equal cell numbers (Figure B.1A-C; see methods for growth conditions). We identified differentially expressed genes (DEGs) in a pairwise manner (Figure 3.1A,B). The largest number of DEGs was observed between naïve and primed hESCs (Figure 3.1B and Table B.1), signifying just how distinct these cellular states are. Highlighted are several genes known to be upregulated in the human pre-implantation epiblast[27,29] and other genes of interest, indicating that the characteristics we observe for 2iL+IF naïve cells are reflective of pre-implantation development.

We determined gene ontology (GO) categories and KEGG pathways for naïve DEGs, which were significantly enriched for embryo development and pluripotency signaling pathways along with other pathways important during pre-implantation development (Figure 3.1C,D). In particular, genes in the TGF-beta pathway were found to be upregulated in naïve cells, including *LEFTY1*, *SMAD3* and *NODAL* (Figure B.1D). The TGF-beta pathway was shown to be important for maintenance of NANOG in the human epiblast, whereas inhibition of this pathway has insignificant effects on mouse embryos[29]. PI3K-AKT signaling pathway was also enriched, and is known to promote ESC self-renewal through inhibition of ERK signaling pathway(Figure B.1E)[173]. The WNT signaling pathway was enriched for naïve upregulated genes including *WNT8A*, *WNT5B* and *TCF7* (Figure B.1F)[2]. A number of terms associated with embryonic development and morphogenesis were enriched for naïve upregulated genes. This may foreshadow what happens to the cells of the blastocyst as they prepare to become the embryonic disk of the epiblast.

We identified cell type-specific genes in the different hESC stages by applying a cutoff of a RPKM value greater than or equal to two in one cell type and less than one in the other two cell types (Figure B.2A). Using this cutoff we determined 429 naïve-specific genes, 229 transition-specific genes and 333 primed-specific genes. Compared to the primed states, naïve-specific genes were enriched for GO terms associated with morphogenesis and pattern specification (Figure B.2A). This is due, in part, to the many HOX genes that are uniquely expressed in naïve hESCs and not in transitioning or primed cells. Primed cells were enriched for terms associated with extracellular communication and protein/histone demethylation. Notably, when we constructed a protein-protein interaction (PPI) of the naïve-specific genes 118 proteins had 206 statistically significant known interactions (permutation test, P-value < 0.001) with other proteins from naïve-specific coding genes (Figure B.2B).

Cell-specific gene expression comprises coding and non-coding genes (Figure B.3A,B and Table B.1). We further examined long intergenic non-coding RNAs that were specific to each cell type and used LncRNA2Function [174] to determine if there were enriched pathways represented. Of the 24 lincRNAs specific to the naïve hESC stage, we found they were enriched for GO terms involving the nervous system and cell signaling (Figure B.3C), which is interesting given that this has been shown to be the default lineage for differentiation of mESCs [175]. In contrast, of the limited primed-specific lincRNAs, we could only identify enriched GO terms reflecting sex-specific differences between the lines (Figure B.3E). The pre-primed-specific lincRNAs were enriched for GO terms involving chromatin structure, which is likely indicative of the dramatic chromatin remodeling that must occur in transitioning from naïve to primed (Figure B.3D and see below).

A recent report showed that the transposable element (TE) transcriptome can be used as a state-specific signature in hESCs[33]. Naïve and primed hESCs segregate when clustered on the top 1,000 highly expressed TEs (Figure 3.1E). Lastly, we compared upregulated genes to human embryo RNA-seq data from Yan *et al.*[27]. We find that a similar percentage of upregulated genes from naïve and primed are expressed in pre-zygotic genome activation stages, while naïve hESCs share more upregulated genes

with the post-ZGA embryo than primed (Figure 3.1F). This strengthens reports that naïve cells are a good representative model of the pre-implantation stage of human development[2,31].

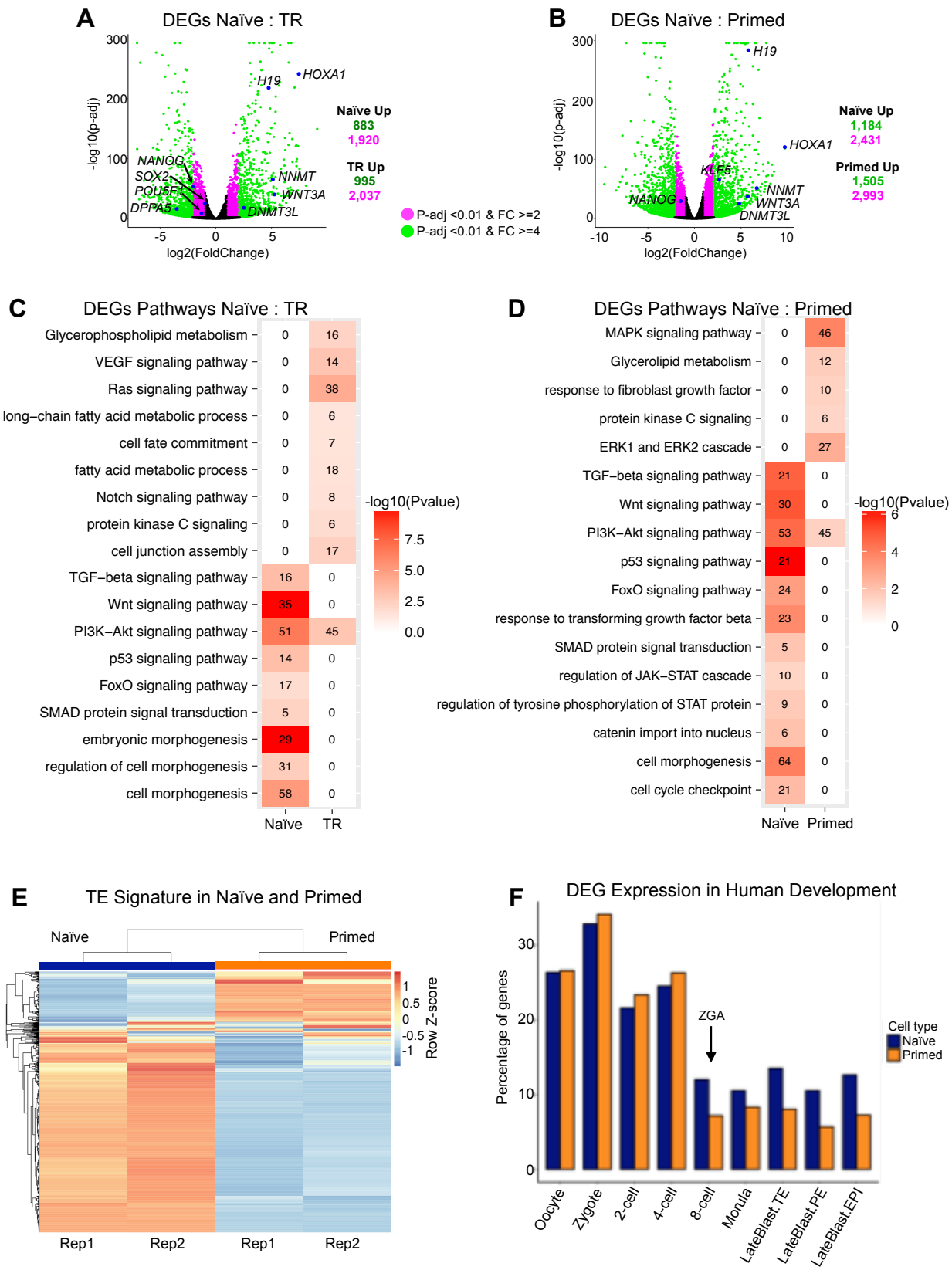

A
DEGs Naïve : TR
-log10(p-adj)
log2(FoldChange)
Naïve Up
883
1,920
TR Up
995
2,037
P-adj <0.01 & FC >=2
P-adj <0.01 & FC >=4
H19
HOXA1
NANOG
SOX2
POU5F1
DPPA5
NNMT
WNT3A
DNMT3L
B
DEGs Naïve : Primed
Naïve Up
1,184
2,431
Primed Up
1,505
2,993
KLF5
C
DEGs Pathways Naïve : TR
Glycerophospholipid metabolism
VEGF signaling pathway
Ras signaling pathway
long−chain fatty acid metabolic process
cell fate commitment
fatty acid metabolic process
Notch signaling pathway
protein kinase C signaling
cell junction assembly
TGF−beta signaling pathway
Wnt signaling pathway
PI3K−Akt signaling pathway
p53 signaling pathway
FoxO signaling pathway
SMAD protein signal transduction
embryonic morphogenesis
regulation of cell morphogenesis
cell morphogenesis
Naïve
TR
-log10(Pvalue)
D
DEGs Pathways Naïve : Primed
MAPK signaling pathway
Glycerolipid metabolism
response to fibroblast growth factor
protein kinase C signaling
ERK1 and ERK2 cascade
TGF−beta signaling pathway
Wnt signaling pathway
PI3K−Akt signaling pathway
p53 signaling pathway
FoxO signaling pathway
response to transforming growth factor beta
SMAD protein signal transduction
regulation of JAK−STAT cascade
regulation of tyrosine phosphorylation of STAT protein
catenin import into nucleus
cell morphogenesis
cell cycle checkpoint
Naïve
Primed
-log10(Pvalue)
E
TE Signature in Naïve and Primed
Naïve
Primed
Row Z-score
Rep1
Rep2
Rep1
Rep2
F
DEG Expression in Human Development
Percentage of genes
ZGA
Cell type
Naïve
Primed
Oocyte
Zygote
2-cell
4-cell
8-cell
Morula
LateBlast.TE
LateBlast.PE
LateBlast.EPI

Figure 3.1. Differential Gene Expression
(A-B) Volcano plot of differentially expressed genes (DEGs) in naïve versus transitioning (A) and naïve versus primed (B) pairwise comparison. Genes in magenta have P-adj < 0.01 and fold change > 2 while genes in green have P-adj < 0.01 and fold change > 4. (C-D) Heatmap showing significantly overrepresented GO terms and KEGG pathways based on DEGs in naïve versus transitioning (C) and naïve versus primed (D) pairwise comparison. (E) Hierarchical clustering of transposable elements gene expression separates naïve from primed hESCs. (F) Percentage of genes upregulated in pairwise comparison of naïve or primed hESCs that are also found to be upregulated in human embryo developmental stages.

3.3.2 Global Chromatin Features of Naïve hESCs

To assess global chromatin dynamics between the cellular states, we performed ChIP-seq on five histone modifications from naïve and transitioning cells (Tables B.2, B.3), and used data previously generated in H1 hESCs for the primed state[153]. These modifications include: H3K4me3 for Pol II-bound promoters[84,176,177], H3K4me1 for enhancers[84,89], H3K27ac for active regions[89,157], H3K27me3 for Polycomb repressed regions[106,178], and H3K9me3 for heterochromatin[176,179]. All five modifications along with ChIP inputs were sequenced in duplicates for both Elf1 naïve and Elf1 transitioning cells for a total of > 270 million and >213 million sequencing reads respectively (Tables B.2, B.3).

We inspected genes with known expression differences during early embryogenesis through the blastocyst/epiblast stage to ensure our chromatin maps reflect changes during differentiation from naïve to primed. *TBX3* was shown to be expressed in naïve ESCs and human epiblasts[29]. The *TBX3* locus exhibits high levels of H3K4me1 and H3K27ac in naïve hESCs, a reduction of H3K27ac in the transitioning state, followed by a reduction of H3K4me1 and a gain of H3K27me3 in primed hESCs (Figure 3.2A). *KLF2*, which was shown not to be expressed in human naïve cells[29], lacks the H3K27ac modification in all three hESC stages (Figure B.4A). *CDX2* has active histone modifications in naïve hESCs but transitions to lost acetylation and gained H3K27me3 in primed hESCs (Figure 3.2A). *CDX2* has been shown to be expressed after blastocyst formation in human embryos and overlaps *OCT4* expression in preimplantation embryos[180]. Expansion of H3K27me3 domains are also shown at the HOXA locus as hESC move from naïve to primed (Figure B.4B). Next, we asked whether these trends observed at specific loci held true genome-wide.

Previous studies, including our own work, in naïve hESCs observed a reduction of H3K27me3 in naïve derived and reset hESCs[2,22,23,26,31], consistent with what was shown in naïve mESCs[45]. Comparisons across cell types reveal a genome-wide depletion of repressive histone modifications in naïve cells (Figure 3.2B,C). H3K27me3 repressed regions are more abundant and broader in primed than in naïve cells, covering ~1.4% of the genome in primed cells compared to 0.5% in naïve (Figure 3.2C, Figure B.4C), which we previously showed is linked to metabolic differences between the cell states[2]. H3K9me3 heterochromatin regions, which are sparse in primed cells[153], are further depleted in transitioning and naïve cells (Figure 3.2B,C, Figure B.4D and Table 3.4, Table 3.5). There is a notable abundance of H3K4me1 regions in naïve hESCs (Figure 3.2B and Table 3.4). Over 9% of the naïve genome is marked by H3K4me1, three times more than primed cells and 1.7 times more than transitioning cells (Figure 3.2C and Table B.5). Monomethylation is present in larger domains, reaching sizes of over 30kb in transitioning cells and over 50kb in naïve cells (Figure B.4E). Acetylation is also more enriched in naïve cells with 3x more peaks than primed, and broad H3K27ac domains reaching over 50kb (Figure 3.2B,C, Figure B.4F and Table B.4, Table B.5). The trends for H3K27 modifications also hold true on the X chromosome (Figure B.4H-J), where both are active in naïve cells[31]. We found H3K4me3 to be the most stable mark though cell-specific peaks exist (Figure 3.2B,C and Figure B.4G).

3.3.3 Promoter Transitions from Naïve to Primed State

We investigated how DEGs were reflected through promoter chromatin states using >19,000 GENCODE defined autosomal protein coding genes. Over 12,000 promoters are marked with H3K4me3 (Figure B.5A). We subdivided promoters into six categories: (1) active - H3K4me3 and H3K27ac; (2) poised - H3K4me3 only; (3) bivalent - H3K4me3 and H3K27me3; (4) H3K27ac - H3K27ac only ; (5) H3K27me3 - H3K27me3 only; and (6) unmarked - lacking all three modifications (Figure 3.2D and Figure B.5B). Although the largest percentages of gene promoters remain static as either active or unmarked across all three stages, many promoters change chromatin state (Figure B.5C), which exemplifies the dynamic nature of the epigenome. To illustrate that chromatin patterns coincide with general trends of expression, we plotted the RPKM values of genes with active, poised and bivalent promoters. As expected, genes with active promoters had overall higher expression levels than genes with promoters in the other two

categories (Figure 3.2E).

Mouse ESCs grown in serum have a greater than three-fold increase in bivalent promoters relative to cells of the mouse ICM [87]. Observing a similar increase in bivalent gene promoters from naïve to primed cells (1,097 vs 2,674), we determined from which epigenetic states the primed bivalent promoters arose. Roughly 60% of primed bivalent promoters are bivalent in transitioning cells, and of those, their promoter states are split between active (42%), bivalent (32%) and poised (20%) in naïve hESCs (Figure 3.2F). Of the ~7% of naïve active gene promoters that become bivalent in transitioning cells, these genes were enriched for GO terms such as morphogenesis and WNT signaling, and includes genes such as *HOXA1*, *HOXA4*, *HOXD8* and *ZEB1*. Naïve bivalent genes fall into categories involving GO terms for synaptic transmission, ion transport and neuron differentiation (Figure 3.2G). Thus, it appears that the neural lineage is the first lineage to be bivalently marked in naïve cells and suggests that naïve hESCs may be an excellent model for further investigation of the establishment of Polycomb repressive regions in the early epigenome.

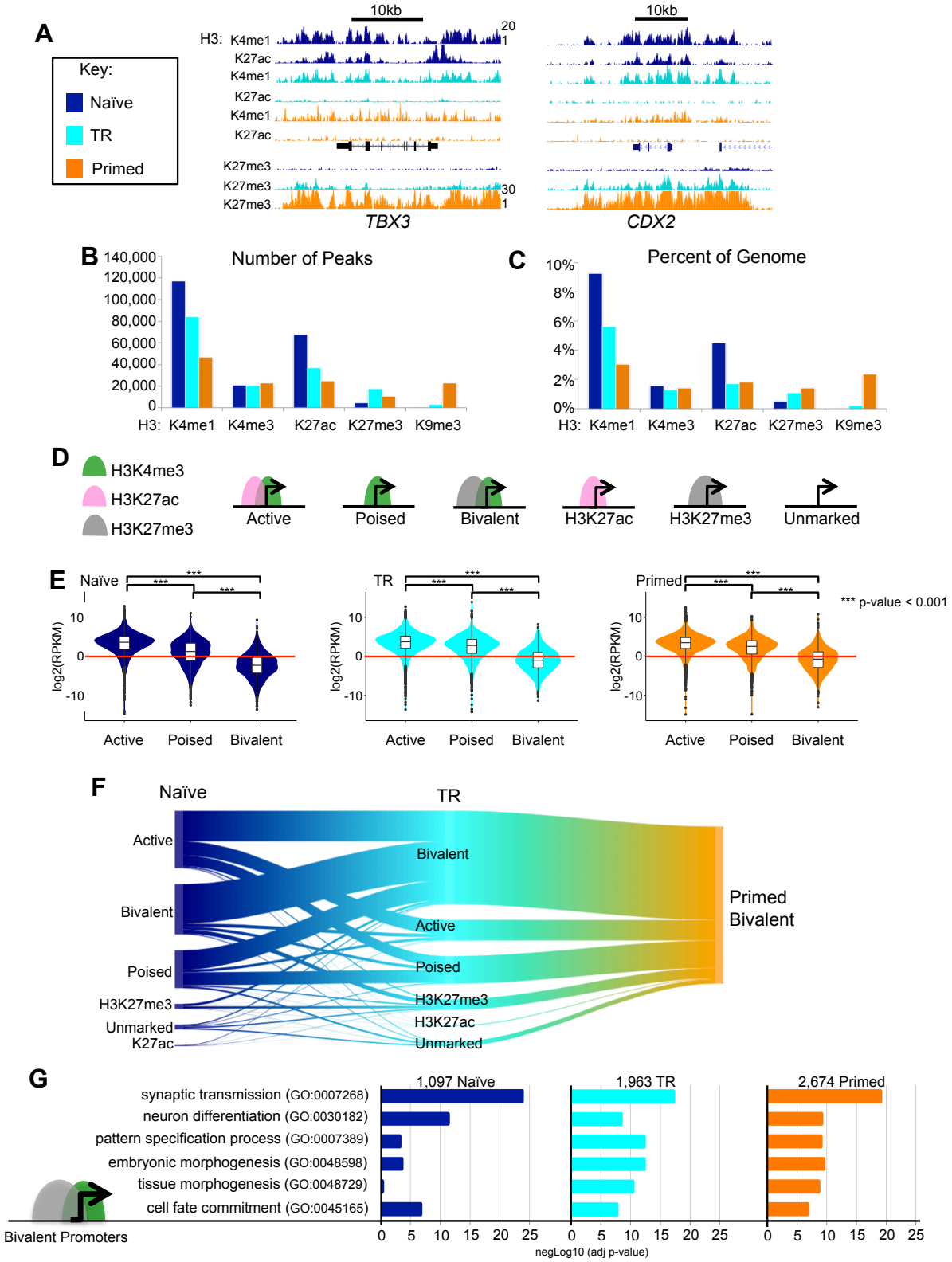

A
Key:
Naïve
TR
Primed
10kb
H3: K4me1
K27ac
K4me1
K27ac
K4me1
K27ac
K27me3
K27me3
K27me3
20
1
30
1
TBX3
CDX2
B
Number of Peaks
140,000
120,000
100,000
80,000
60,000
40,000
20,000
0
H3: K4me1 K4me3 K27ac K27me3 K9me3
C
Percent of Genome
10%
8%
6%
4%
2%
0%
H3: K4me1 K4me3 K27ac K27me3 K9me3
D
H3K4me3
H3K27ac
H3K27me3
Active
Poised
Bivalent
H3K27ac
H3K27me3
Unmarked
E
Naïve
TR
Primed
***
*** p-value < 0.001
log2(RPKM)
10
0
-10
Active Poised Bivalent
F
Naïve
TR
Active
Bivalent
Poised
H3K27me3
Unmarked
K27ac
Bivalent
Active
Poised
H3K27me3
H3K27ac
Unmarked
Primed
Bivalent
G
synaptic transmission (GO:0007268)
neuron differentiation (GO:0030182)
pattern specification process (GO:0007389)
embryonic morphogenesis (GO:0048598)
tissue morphogenesis (GO:0048729)
cell fate commitment (GO:0045165)
1,097 Naïve
1,963 TR
2,674 Primed
Bivalent Promoters
0 5 10 15 20 25
negLog10 (adj p-value)

Figure 3.2 Overview of Chromatin States
Global view of chromatin structure for naïve (navy), transitioning (cyan) and primed (orange) hESCs. These colors are used throughout all figures. (A) UCSC Genome Browser images of TBX3 and CDX2 gene loci showing enrichment of H3K4me1 (RPKM range 1-20), H3K27ac (RPKM range 1-20), and H3K27me3 (RPKM range 1-30) in naïve, transitioning and primed cells. (B) The number of ChIP-Seq peaks called by MACS with FDR cutoff ≤ 0.05. (C) The percent of genome covered by each histone modification (number of bases divided by effective genome size: 2.7e+09). (D) Promoters were classified based on the criteria pictured - also see main text. (E) Violin plots showing the distribution of RPKM values of nearest neighboring genes of active, poised and bivalent promoter peaks in each cell type. P-values for pairwise comparisons are computed using two tailed t-tests with pooled SD. P-values are adjusted with Benjamini-Hochberg method. *** P-value < 0.001;. (F) Sankey plot of primed bivalent gene promoters and their origins from the naïve state. (G) Significance of GO Terms from bivalently marked gene promoters.

3.3.4 Enhancers in the Naïve Embryonic State

Enhancer elements are cis-acting regulatory sequences that control gene expression via interaction with transcription factors and promoters. Enhancer chromatin modifications are highly dynamic and cell type-specific[153]. Here, we defined enhancers as H3K4me1 peaks lacking overlap with H3K4me3 (Table S6). Investigation of the enhancer landscape across hESC states revealed that naïve cells harbor the most cell type-specific enhancers (47,456; Figure 3.3A,B), while transitioning and primed cells had roughly the same number of unique enhancers at 17,308 and 14,376 respectively (Figure B.6A-D). Sixty-four percent of transitioning enhancers and 55% of primed enhancers are marked in the naïve state (Figure B.6A,B). We asked if the expansion of naïve H3K4me1 was random or occurred at known regulatory elements.Using DHS data from 177 ENCODE cells[181], including H1, we found 25-30% of the H3K4me1-marked genome (enhancer-verse) to be hypersensitive in each cell type (Figure 3.3C). Of the 177 cell and tissue types, fetal tissues had the largest collection of DHS overlapping naïve enhancers (Figure 3.3D and Table B.7). Additionally, over 92% of the enhancer base pairs covered by naïve H3K4me1 peaks are utilized as enhancers in 127 Roadmap Epigenome Project cell types, as indicated by H3K4me1 (Figure B.6E-F). Single cell RNA-seq data from early human embryogenesis[27] indicates that 92% of annotated transcription factors[182] are expressed by the late blastocyst stage (Figure B.6G). Their expression provides a plausible means for aiding the localization of H3K4me1 to known enhancers.

Enhancer elements can exist in distinct chromatin states that indicate whether they are active or poised

(Figure 3.3E)[90,94,157]. We characterized differences in the classes of enhancers in each hESC state. We defined active enhancers as regions having H3K4me1 and H3K27ac and poised enhancers as regions with either H3K4me1-only or H3K4me1 and H3K27me3. In all three stages of pluripotency, the majority of enhancers are in the H3K4me1-only poised state (67%, 84%, and 73% in naïve, transitioning and primed cells respectively; Figure 3.3E). There is an increase of H3K27me3 containing poised enhancers moving from naïve to primed (1% to 4%; Figure 3.3E), which correlates with the increase of H3K27me3.

To gain insight into the regulation of the naïve state, we determined which transcription factor binding sites were enriched in naïve active enhancers, and therefore, likely important for regulating the naïve network. Figure B.7 shows representative motif p-values, percentage of target sequences and corresponding RPKM values of the corresponding TFs in each hESC state. Motifs for the ESC embryonic master regulators OCT4, SOX2 and NANOG (Figure B.7) were most significant in primed cells and were present in 20-60% of active enhancers across all three hESCs. STAT3 is largely represented in naïve active enhancers and expressed highest in naïve hESCs. STAT3 is a component of the Lif pathway and a positive regulator of NNMT, a key regulator the naïve hESC metabolic state responsible for reduced H3K27me3[2,9]. *PRDM14* is expressed as early as the 2-cell stage in the mouse embryo and defines cells that become pluripotent cells of the inner cell mass[183], and is important in maintenance of both mESCs and hESCs[26,184]. While PRDM14 has similar expression levels in naïve and primed hESCs, its motif is more highly enriched in the naïve state, suggesting an increased regulatory role. Overall, naïve and primed cells share a number of motifs enriched at active enhancers.

Our comparative analysis of enhancers indicates that both active and H3K4me1-only poised enhancers are largely decommissioned as naïve hESCs transition to the primed state (Figure 3.4A). When assessing overlapping H3K4me1 peaks across hESCs, we see that the chromatin-marked genomic space of naïve enhancers is greatly reduced in primed cells (Figure 3.4A,B). This process happens in a stepwise manner, as is evidenced by the loss of acetylation as cells exit the naïve state followed by the gradual loss H3K4me1. This introduces a different view of development compared to previous studies that showed poised enhancers gain acetylation following differentiation and were often enriched near

genes that became activated later in development[90,94,157]. By using naïve hESCs as a model, we can infer that not only is H3K4me1 likely maintaining open chromatin to aid in the pluripotency phenotype, but that a substantial fraction of enhancers in the human genome are pre-marked early during embryogenesis and subsequently decommissioned during priming.

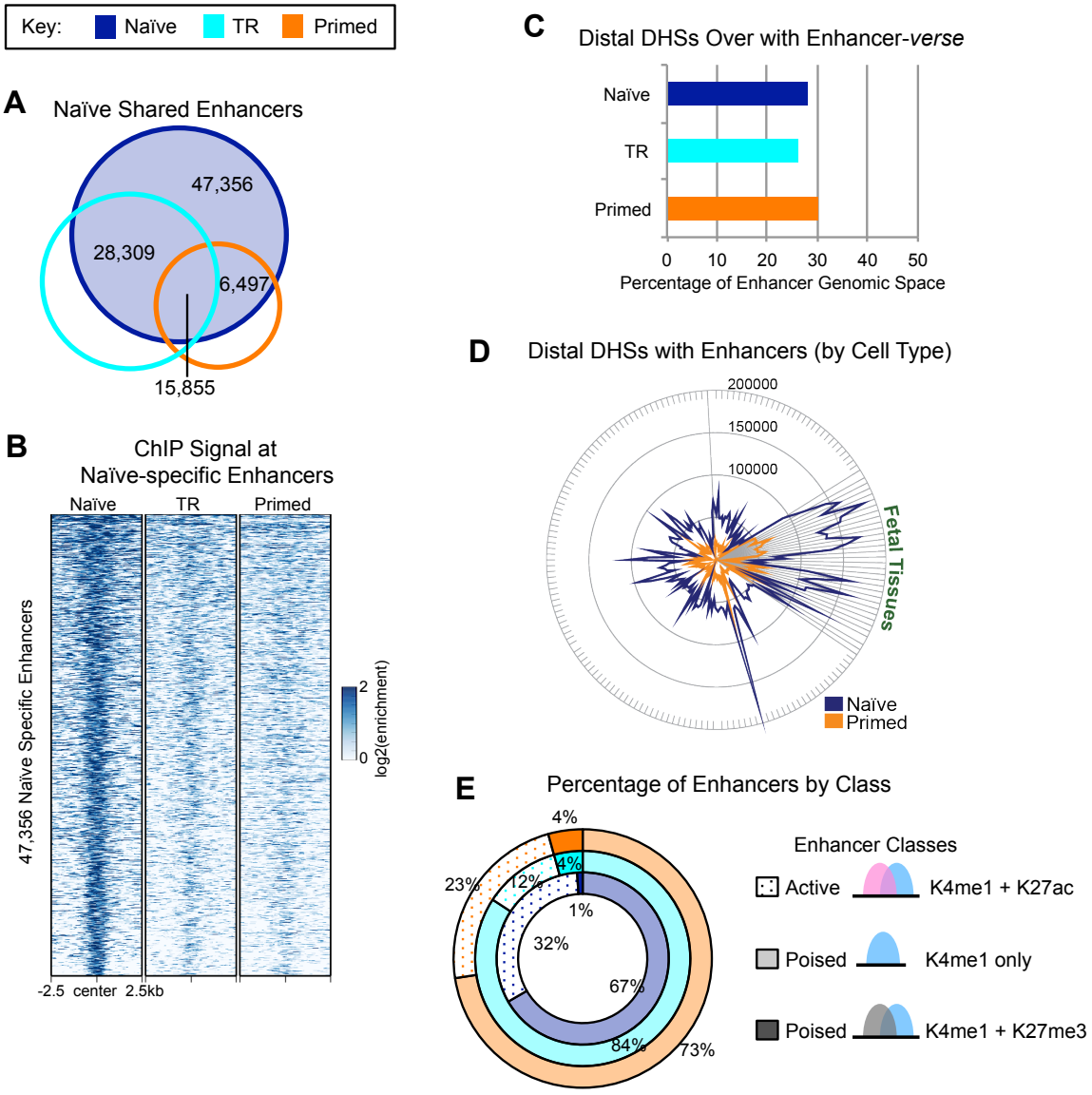


Figure 3.3 Naïve Enhancer Repertoire
(A) Venn diagram of naïve (navy) enhancers overlapped with transitioning (cyan) and primed (orange) enhancers. (B) Heatmap of H3K4me1 normalized ChIP-seq signal centered at naïve-specific enhancers in a 5kb window. (C) Percent of hESC H3K4me1 genomic space (% bases or enhancer-verse) occupied by ENCODE DHSs from 177 cell types (D) Number of ENCODE DHS from 177 cell types overlapping with naïve and primed H3K4me1 enhancers (E) Distribution of active (H3K4me1 + H3K27ac) and poised (H3K4me1 only or H3K4me1 + H3K27me3) enhancer states in each cell type

3.3.5 Broad Enhancer Domains in the Naïve Epigenome

Super[97] and stretch[98] enhancers, which are largely based on H3K27ac, were originally identified in primed ESCs. These regions were shown to upregulate nearby genes and were stronger than conventional enhancers. We asked to what degree these regions were present in our naïve hESCs. To identify both broad H3K4me1 and H3K27ac domains, we identified regions >5kb in all cell types (Figure 3.4C). The H3K4me1 broad enhancers are almost 20 times more abundant in the naïve epigenome compared to the primed hESC stage (7,412 in naïve hESCs compared to 371 in primed) with an average size of 8.1kb compared to 6.1kb in primed (Figure B.6H,I). The number of broad enhancers steadily declines as hESCs transition from naïve to primed. We observed the same trend with H3K27ac broad domains (2,330 in naïve compared to 803 in primed), although the number of broad H3K27ac domains in naïve cells is three times less than the number of H3K4me1 broad enhancers (Figure B.6H). As a control, we looked for broad H3K4me3 peaks, which were limited across the different hESC stages (Figure B.6H).

Next, we determined if H3K4me1 broad enhancers and H3K27ac broad domains occupy the same genomic space. The average number of bases contained within the overlap of broad H3K4me1 and H3K27ac domains is over 70% of the average length of each domain (Figure B.6I). Over 78% of broad H3K27ac domains in naïve cells are found within H3K4me1 broad enhancers (Figure B.6J). In the naïve and primed states 87% and 71% of H3K4me1 broad enhancers, respectively, contained some overlap with H3K27ac, indicating that they are active enhancers (Figure 3.4D and Figure B.6J). The average ChIP-seq signal for H3K4me1 is high at H3K27ac broad domains in all cells except primed hESCs (Figure 3.4D). The active state of broad enhancers is supported by the distribution of expression values of nearest neighboring genes (NNGs; Figure 3.4E). Only in the primed state are there more broad H3K27ac domains than H3K4me1 domains and the difference in the expression distribution of NNGs at broad enhancers versus active broad enhancers in primed cells was the only comparison not found to be significant (Figure 3.4E). This may explain why H3K27ac was originally associated with “super/stretch” enhancers. The frequent occurrence of H3K4me1 and H3K27ac broad domains, where broad H3K27ac domains lie within broad enhancers, provides an additional means of giving the genome its “open

structure" in naïve pluripotency.

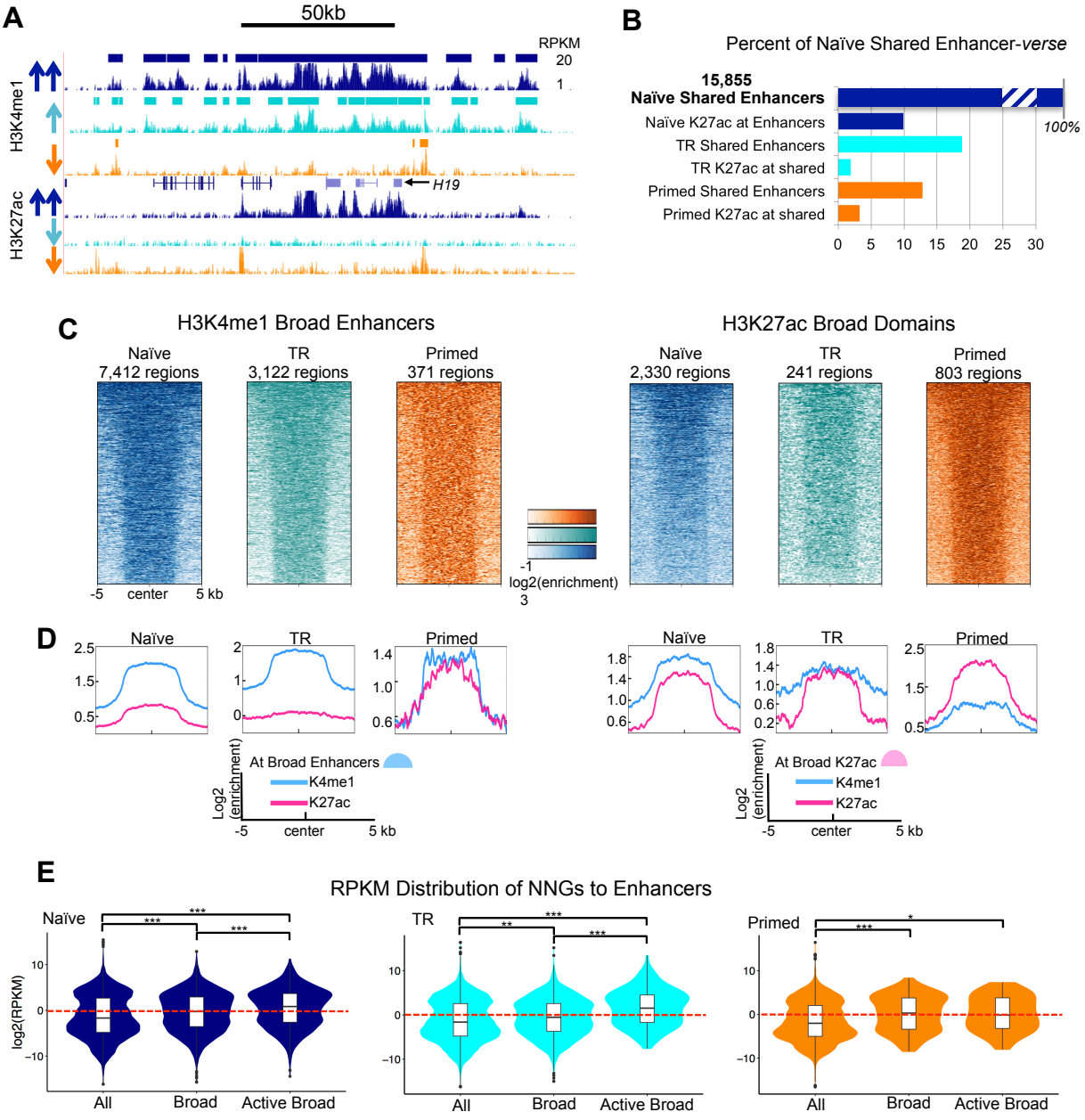


Figure 3.4 Naïve Enhancers are Decommissioned but Active in Other Cell Types
(A) UCSC Genome Browser image illustrating loss of H3K27ac, followed by loss of H3K4me1 at the H19 locus as cells move from naïve (navy), to transitioning (cyan), to primed (orange); RPKM range 1-20 for each track. This region also contains a broad enhancer domain in naïve hESCs. Enhancer peak calls are represented as bars above the H3K4me1 track. (B) Percent of shared naïve enhancers genomic space that is preserved in the follow hESC states. (C) Heatmaps of H3K4me1 and H3K27ac normalized ChIP-seq signal at broad enhancer regions (≥5kb). (D) Histograms of average H3K4me1 and H3K27ac normalized ChIP-seq signal at all broad enhancers or broad H3K27ac domains. (E) Violin plots showing the distribution of RPKM values of nearest neighboring genes of all broad enhancers (H3K4me1 ≥5kb) and active broad enhancers (H3K4me1 ≥5kb overlapping H3K27ac ≥5kb) in each cell type. P-values for pairwise comparisons are computed using two tailed t-tests with pooled SD. P-values are adjusted with Benjamini- Hochberg method. * P-value < 0.05; ** P-value < 0.01; *** P-value < 0.001.

3.3.6 Naïve hESCs Enhancers in Different Growth Conditions

To determine if the expansion of H3K4me1 in the naïve epigenome was indicative of the naïve state and independent of a single growth condition or cell line, we grew three lines in 4i (2i + p38 kinase inhibitor + JNK inhibitor) + Lif + IGF1 + FGF (referred to as 4iL+IF): Elf1, H1 reset to naïve and the naïve derived LIS1 line[22], which grew slightly better in 4iL+IF compared to the original growth conditions (Figure B.8A). In order to determine the effect of growth conditions and genetic background on the enhancer landscape, we compared the enhancer profiles from H3K4me1 ChIP-seq data across cell types and conditions. Overall, all naïve cells have a similar enhancer profile (Figure 3.5A). Cells grown in 4iL+IF exhibit a stronger enhancer signal at Elf1 2iL+IF naïve-specific enhancers (Figure 3.5B,C), and less enrichment at primed- and transitioning-specific enhancers (Figure 3.5B and Figure B.8B). PCA of gene expression data shows a clear separation between naïve and primed cells(Figure B.8C). PCA of H3K4me1 signal reveals that all lines grown in 4iL+IF are largely indistinguishable, and most similar to 2iL+IF (Figure 3.5D). Transitioning cells (Elf1 AF) have naïve-like enhancer profiles as mentioned above, transitioning cells have lost naïve H3K27ac but have not yet lost H3K4me1 to primed levels. Our analysis suggests that naïve 2iL+IF enhancers do not vary greatly in 4iL+IF naïve conditions, although 2iL+IF naïve hESCs have some distinct H3K4me1 features. The expansion of H3K4me1 regardless of cell line or growth condition confirms this as a new signature of the naïve hESC state. The acquired expansion upon resetting primed H1 cells to naïve may suggest that this epigenetic feature is necessary for maintenance in the naïve state. Further experiments will be needed to confirm this hypothesis.

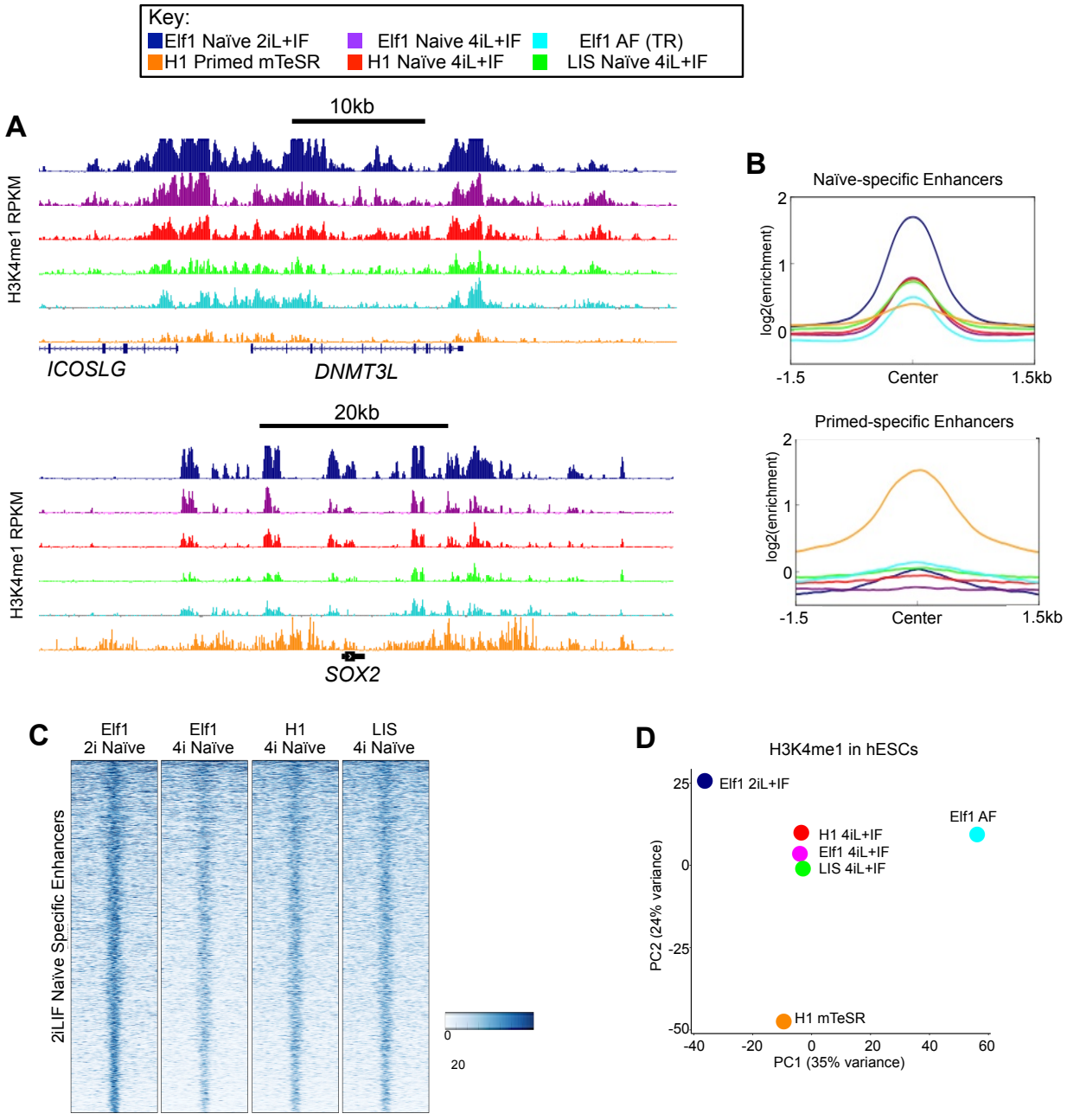


Figure 3.5 Naïve Enhancers from Various Naïve Culture Conditions
ChIP-Seq of naïve cells grown in different culture conditions including naïve Elf1 naïve (navy), Elf1 4iL+IF (purple), Elf1 AF (transitioning - cyan), primed H1 mTeSR (orange), naïve H1 4iL+IF (red), naïve LIS1 4iL+IF (green). (A) H3K4me1 enrichment in different growth conditions at DNMT3L and SOX2 loci (first panel) (B) Average ChIP-Seq signal at naïve-specific enhancers (top panel) and primed-specific enhancers (bottom panel). (C) Heatmap of H3K4me1 ChIP-Seq signal at naïve specific enhancers. (D) PCA of top 500 10kb bins of H3K4me1 with largest variance.

3.3.7 3D Genome Architecture in Naïve hESCs

Genome architecture is an important component of gene regulation. Topological associated domains

(TADs) identified in primed hESCs proved to be surprisingly stable upon differentiation to distinct cell types in spite of diverse changes to chromatin structure[123]. Similarly, recent ChIA-PET data for cohesin in primed and naïve reset cells showed a similar recovery of primed TADs[126]. However, domain-scale 3D genome architecture is still missing for the naïve state. To characterize TADs in naïve Elf1 2iL+IF hESCs, we generated deeply sequenced in situ DNase Hi-C maps[168] (Figure B.9A), which exhibited characteristic reductions in contact frequency as a function of linear distance between two loci (Figure 3.6A). We processed raw Hi-C read pairs produced from H1 primed hESCs[122,123] and compared the architectural features identified in each cell type at 40kb resolution. A total of 6,119 TADs were identified in naïve hESCs compared to 5,822 TADs in primed hESCs (Figure B.9B), consistent with previous observations in primed hESCs[170]. We defined boundaries as regions between two adjacent TADs and found that 7.3% and 6.2% of boundaries were greater than 40kb in naïve and primed cells, respectively (Figure B.9C). To give confidence in our TAD calls, we calculated insulation scores [171,185]. Insulation scores are calculated at each Hi-C bin by aggregating the contact measurements in a fixed window around each Hi-C bin. The insulation score represents how insulated each bin is from TAD boundaries. It is expected that TAD boundaries occur at the valleys/minima of insulation scores, and TAD centers occur at the peaks/maxima. We found that boundary insulation scores were significantly different from TAD center scores (Figure 3.6B). However, we could detect differences in location of TAD boundaries as illustrated by naïve-specific boundaries exhibiting an enrichment of primed Hi-C signal (Figure 3.6C).

Overall, TAD size distributions are similar (Figure 3.6D, first panel), with means of 420kb in naïve and 444kb in primed. We observed 2,024 TADs whose genomic coordinates are identical at 40kb resolution while the remaining overlapping TADs differ by at least 40kb (Figure 3.6D, second panel). We asked if the higher number of naïve Elf1 TADs may be due to better resolution of our in situ data, as the two datasets were generated using different Hi-C protocols, and indeed we found that some H1 TADs were split into two or more Elf1 TADs, which accounts for an “extra” 427 naïve TADs (Figure B.9D). The average overlap between naïve and primed TADs is 319kb, suggesting that the overall TAD structure remains intact between the naïve and primed states (Figure B.9E),

We investigated if there was a relationship between higher-order chromatin structure at differential TAD boundaries and changes in chromatin modifications. We observe a significant enrichment for H3K4me1 and H3K27ac across differential TAD boundaries in the naïve state relative to random (naïve H3K4me1 and H3K27ac P-value < 5x10-5), and a similar enrichment for primed H3K27me3 (P-value < 1x10-4) (Figure 3.7A). A clear example illustrating these differences in TAD and chromatin structure is the HOXA cluster, where a broad boundary spans the HOXA cluster in primed hESCs and is enriched for H3K27me3 (Figure 3.7B). In naïve hESCs, where HOXA genes are expressed, the TAD to the left of the boundary in primed cells is extended across the cluster and marked by H3K4me1 and H3K27ac. We asked if there was a significant difference in the TAD structure around the HOXA locus between naïve and primed cells. In order to do this, we calculated the differential insulation scores by comparing the naïve minus primed insulation scores. The differential insulation score represents the differential TAD structure between two samples. We examined the differential insulation score around the HOXA locus (Figure 3.7B), and observed that there was a noticeable drop in the signal at the HOXA locus, confirming that the TAD structure at the HOXA locus is different between naïve and primed cells.

Cohesin ChIA-PET and CTCF ChIP-seq data from primed and reset naïve hESCs revealed that looping structures can change in a stable TAD background[126]. Most TADs had a CTCF binding site near their boundaries. We asked if the reset naïve CTCF ChIP-seq signal was also enriched at our naïve derived hESCs. Indeed we found the CTCF signal to be enriched near our boundaries and this enrichment was present in both our naïve derived and primed hESCs (Figure 3.6E). This makes sense as Ji *et al.* found that 80% of CTCF binding sites were just between their reset naïve and primed hESC lines[126]. This helps confirm our in situ DNaseI Hi-C data as accurately capturing the 3D structure of the naïve genome.

Cohesin ChIA-PET data from primed and reset naïve hESCs could recapitulate Hi-C TADs in H1 primed hESCs[126]. The authors note that their ChIA-PET data was undersaturated, even still we asked if the cohesin PETs could help to additionally validate our TAD calls. An overlap analysis with cohesin ChIA-PET data yielded 1,363 naïve and 1,818 primed TADs with at least one PET, from the respective cell type, whose termini are located within 40 kb of each boundary of a given TAD (Figure 3.6F). This

corresponds to 22% of our naïve TADs having a naïve PET and 31% of primed TADs having a primed PET within 40kb of the TAD boundary. We looked to see if any of the PETs were near (within 40kb) differential TAD boundaries, those having different boundaries in naïve and primed of 80kb or greater. Of 1,363 PETs near a naïve boundary, 529 (39%) are near a naïve differential TAD (Figure 3.6F). This helps confirm some of the structural differences observed in the naïve 3D genome.

Finally, to compare the spatial organization of chromatin within the nuclei of naïve and primed hESCs, we partitioned the genome into active and inactive (A/B) compartments by performing a PCA of each intra-chromosomal contact matrix[120,123]. Compartments identified using the first principal component (PC1) ranged in size from 40 kb to over 49 Mb in both cell types, with means of 3.6 Mb in naïve cells and 3.4 Mb in primed. An overwhelming majority of compartments are static, with only 23 switching from being active in naïve cells to inactive in primed (A to B), and 124 switching from being active in primed cells to inactive in naïve (B to A; Figure 3.7C,D). While there is enrichment of primed-specific active compartments, a previous study showed that inactive B sub-compartments are largely devoid of histone modifications, including H3K27me3 and H3K9me3[121]. It is therefore likely that the primed-specific active compartments are driven by the lack of repressive modifications in naïve hESCs (alternatively, these are naïve-specific inactive B compartments). Additionally, cell-specific active compartments are enriched for TE expression relative to stable compartments (Figure 3.7E), and gene expression to a lesser extent (Figure B.9F).

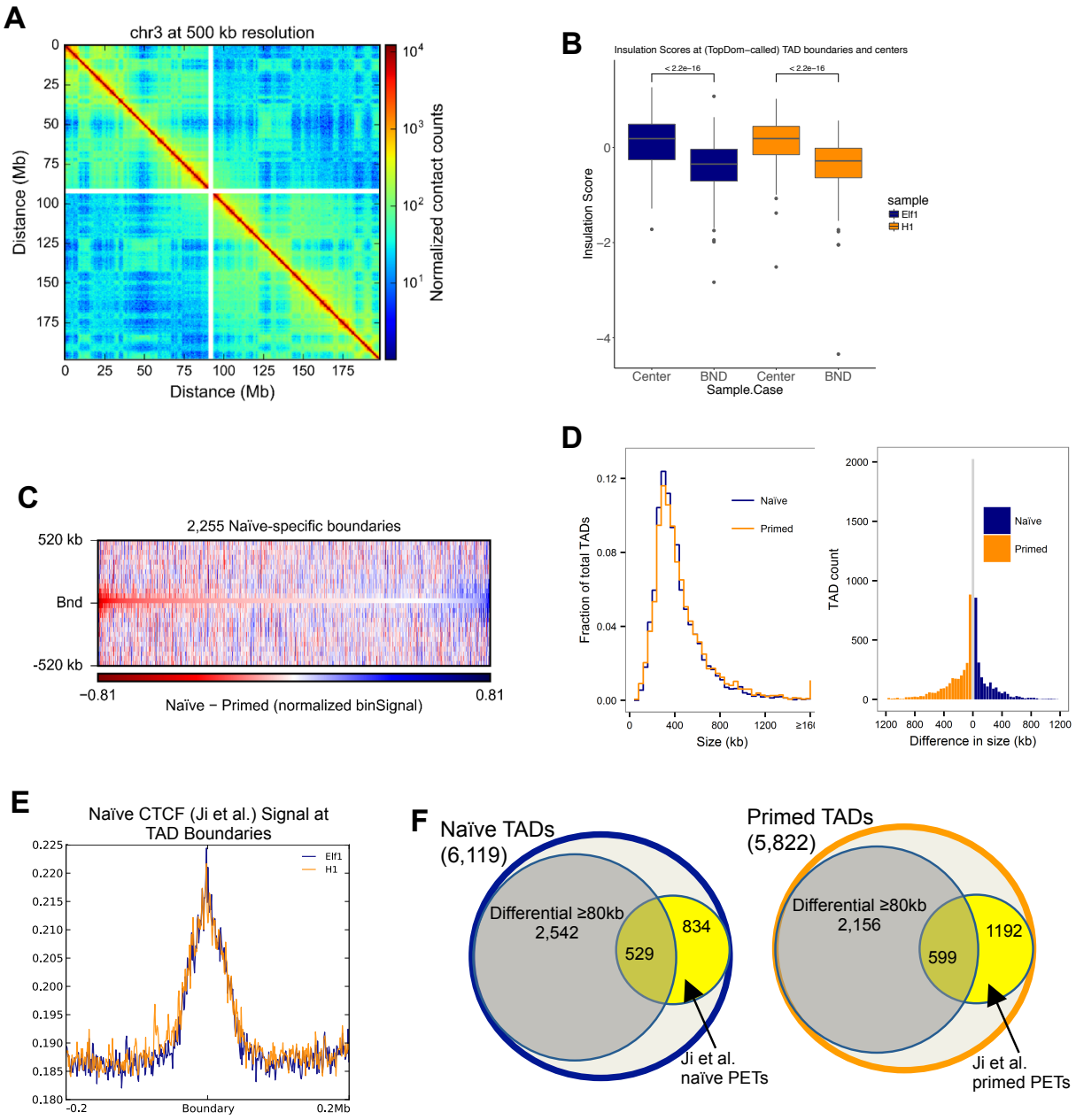
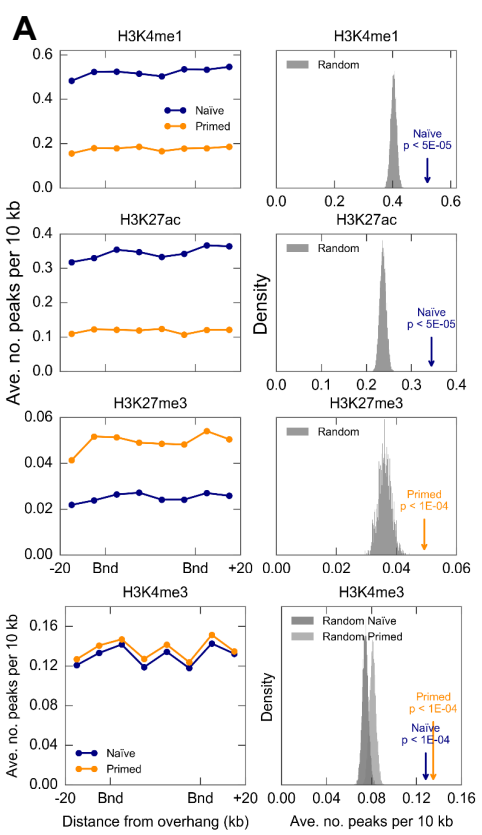
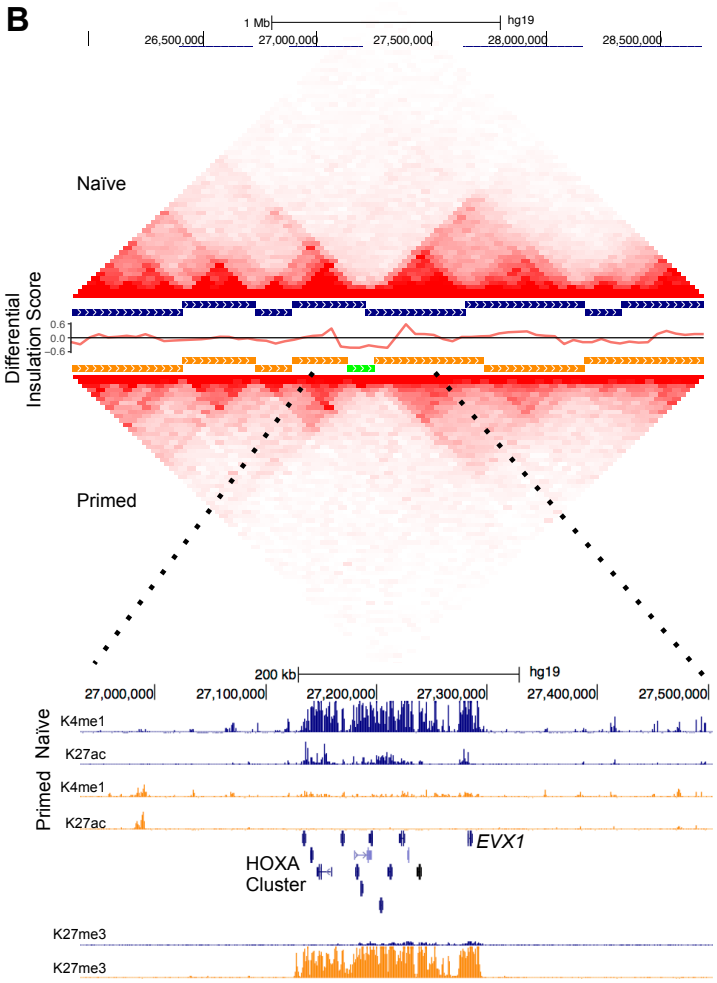
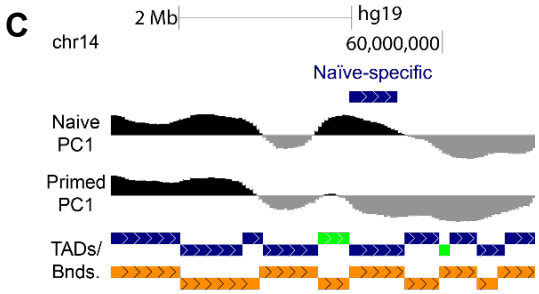
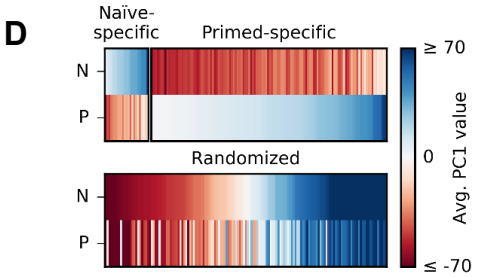


Figure 3.6 3D Genome Architecture in Naïve hESCs
(A) Hi-C contact heatmap of chromosome 3 in naïve cells at 500kb resolution. (B) Boxplots of the insulation scores along chr7 at both TAD centers and boundaries for both naïve and primed cells. P-values are computed using individual Wilcoxon signed-rank tests. (C) Differential heatmap of naïve minus primed Hi-C bin signal centered at naïve-specific boundary regions. Negative (red) values indicate a stronger bin signal in primed cells relative to naïve cells. (D) Global size distributions of TADs within naïve and primed cells (left panel) and size differences of overlapping TADs (40kb bin resolution) in naïve and primed cells (right panel). (E) Naïve CTCF ChIP-seq signal from Ji et al. 2016, centered at TAD boundaries. (F) Number of TADs or differential TADs with cohesin ChIA-PETs within 40kb of boundary


A
H3K4me1
H3K4me1
Random
Naïve
Primed
Naïve
p < 5E-05
H3K27ac
H3K27ac
Random
Naïve
p < 5E-05
Density
Ave. no. peaks per 10 kb
H3K27me3
H3K27me3
Random
Primed
p < 1E-04
-20
Bnd
Bnd
+20
H3K4me3
H3K4me3
Random Naïve
Random Primed
Primed
p < 1E-04
Naïve
p < 1E-04
Distance from overhang (kb)
Ave. no. peaks per 10 kb


B
1 Mb
hg19
26,500,000
27,000,000
27,500,000
28,000,000
28,500,000
Naïve
Primed
Differential Insulation Score
200 kb
hg19
27,000,000
27,100,000
27,200,000
27,300,000
27,400,000
27,500,000
Naïve
K4me1
K27ac
Primed
K4me1
K27ac
EVX1
HOXA Cluster
K27me3
K27me3


C
2 Mb
hg19
chr14
60,000,000
Naïve-specific
Naive PC1
Primed PC1
TADs/ Bnds.


D
Naïve-specific
Primed-specific
N
P
Randomized
N
P
≥ 70
0
≤ -70
Avg. PC1 value

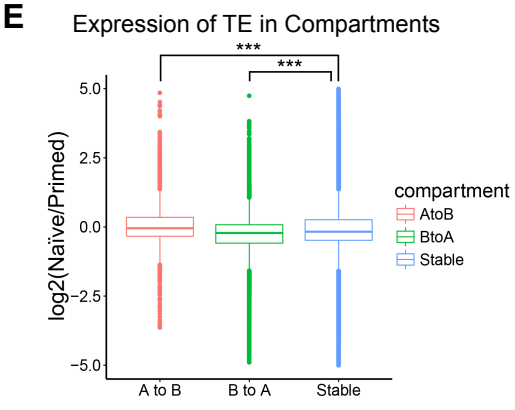

E
Expression of TE in Compartments
***
***
log2(Naïve/Primed)
5.0
2.5
0.0
−2.5
−5.0
A to B
B to A
Stable
compartment
AtoB
BtoA
Stable

Figure 3.7 Active Histone Modifications at TAD Boundaries
(A) Enrichment of ChIP peaks for histone marks H3K4me1, H3K27ac, H3K27me3 and H3K4me3 at overhanging TAD regions. (B) Interaction matrices of region of chr7 containing HOXA locus. Between matrices, horizontal bars with a vertical offset represents an individual TAD, naive in navy and primed in orange. The green bar indicates a boundary region > 40kb. Track of differential insulation score of naïve vs. primed cells around the HOXA locus nested in between TAD calls. ChIP-seq signal (RPKM) scaled from 0 to 20 for H3K4me1 and H3K27ac, scaled 0 to 30 for H3K27me3, naive in navy and primed in orange. (C) Example of a naïve-specific A compartment relative to primed. PC1 scale from -60 to 60. (D) Heatmap of PC1 values at naïve- and primed-specific A compartments. PC1 values at randomized compartments are displayed underneath. "N" and "P" denote naïve and primed, respectively. (E) Boxplot of transposable elements expression (RPKM) overlapping A to B compartment switches. "A to B" and "B to A" are naïve to primed directions. Stable are compartments that do not switch. P-values are computed using two-sample t-test with one sided alternative. *** P-value < 2.2 x 10-16.

3.4 Summary and Conclusions

Based on equivalent cells in mouse, naïve and primed hESCs are thought to be reflective of early human embryogenesis, capturing the pre-implantation and post-implantation states respectively[13]. Our transcriptome analysis is consistent with these states. We observe novel features of the early epigenome. We hypothesize that widespread deposition of H3K4me1 and histone acetylation are part of the mechanism to reset the zygotic genome along with known global DNA demethylation, and that this aids to open chromatin structure for or in response to zygotic genome activation In support of this, previous studies in mice showed that histone modifications and variants change dramatically during the earliest stages of embryogenesis[77,78,80,186]. H3K4me1 and hyperacetylation increase on the genome, prior to zygotic genome activation[78,186-188]. More recently, ChIP-seq results for H3K27ac in the mouse embryo and serum-maintained mESCs showed enrichment of H3K27ac genome-wide post-ZGA followed by a decline in mESCs[85]. Our comparison of single-cell RNA-seq from the developing human blastocyst [27] to naïve and primed hESCs demonstrated an enrichment of key developmental genes and pathways, including WNT and TGF-beta, in 2iL+IF naïve hESCs. In addition, we described the enrichment of hundreds of naïve-specific genes that form known protein-protein interactions, and are expressed during early embryogenesis. Coupled with our and other reports that the genome of naïve hESCs are also hypomethylated, naïve versus primed hESCs can be used to model epigenomic reprogramming that occurs as cells shift between these developmental states.

Enhancer decommissioning is required for proper ESC differentiation through LSD1 activity[99], which is inhibited by acetylation[100,189]. De-acetylation must, therefore, precede the removal H3K4me1 by LSD1. This stepwise decommissioning was observed as cells exited the naïve state, lending support to our hypothesis that enhancer pre-marking is a likely component of epigenetic reprogramming during embryogenesis. Lastly, changes in chromatin modifications are reflected in changes in 3D genome architecture. Although primed TADs are largely unchanged upon hESC differentiation[123], most naïve TAD boundaries are unique to this pluripotent state, subsets of which are validated by recent ChIA-PET data. Our findings suggest that TAD structures are still formalizing prior to implantation. Collectively, these characteristics are likely to shape naïve pluripotency and provide new insights on epigenetic reprogramming through this model of development.

Chapter 4:

Ovarian Cancer Stem Cells as a Model for Studying Multipotent Stem Cell Regulatory Networks

4.1 Motivation

In ovarian cancer, ovarian cancer stem cells (OvCSC) may make up about 3-5% of cells in the tumor[132]. This percentage may vary, however, from tumor to tumor[190]. OvCSCs are characterized by their ability to self-renew and divide asymmetrically to produce progeny with high proliferative potential (non-stem cancer cells)[132,191]. In some studies, OvCSCs have been shown to express transporter proteins that can pump molecules (like chemotherapy drugs) out of the cell, while other studies have shown that OvCSCs can exist in a dormant, quiescent state[132,190-192]. These characteristics may contribute to their ability to survive chemotherapeutic attack[132,191]. CSC have been identified in many cancers, first in AML, but also in breast, lung, colon, skin, liver, prostate, pancreatic, and some brain cancers[193,194].

The goal of this project is to understand how the epigenome contributes to the unique cellular properties of OvCSC. Defining the regulatory network of OvCSCs may help explain tumor growth, drug resistance and metastasis. Strauss *et al.* has isolated patient ovarian tumor cells (denoted ovc316) and identified a CD133+ CSC population in these cells. These cells expressed epithelial and mesenchymal markers, indicative a cells undergoing epithelial-mesenchymal transition (EMT) and shown to be highly metastatic[195,196]. Importantly Strauss *et al.* functionally validated the CD133+ CSC population by showing that they are more efficient than CD133- cells at generating tumors in immunocompromised mice[195]. The ovc316 xenograft system is a way to expand the ovc316 cells *in vivo* and allows the isolation of CD133+ and CD133- tumor cells. Here we use RNA-Seq, ChIP-Seq and WGBS to identify gene expression, enhancer and DNA methylation difference between OvCSCs and their daughter cells that make up the bulk of the tumor. We identify enhancers and differentially methylated regions that potentially regulated genes associated with the CSC phenotype.

4.2 Methods

4.2.1 Growing xenograft ovarian tumors in mice

Secondary tumors were grown from the cell strain ovc316 described in Strauss, 2011 and from the cell line OVCAR5. Cancer cells were injected into the mammary fat pad of 10 CD17-SCID- beige mice and allow the tumors to grow for 8-10 weeks. Tumor dissociation protocol was performed as described in Strauss, 2011 with some exceptions. Briefly, tumors were harvested and dissociated using the MACS Preparation of Single-Cell Suspensions from Implanted Mouse Tumors protocol developed by Miltenyi Biotec. Dissociated tumor cells were allowed ~4 hours *in vitro* culture in Mammary Epithelial Basal Medium supplemented with EGF, insulin, hydrocortisone, bovine pituitary extract (MEGM) and 1-2% FBS to re-express cell surface markers. The MACS magnetic bead cell separation kit was used to select for CD133+ cells using 1 purification step. For the CD133- cells, the MACS Magnetic Bead Blood Cell Lineage Depletion Kit was used to get rid of any contaminating mouse blood cells. OVCAR5 tumors grew much more vigorously, therefore we split the cell collection into 2 set, 5 tumors comprised each replicate 1 and replicate 2. Cells that were to be used for chromatin isolated were cross-linked with 1% formaldehyde at room temperature for 10 min with constant rotation. Cells were washed 3x with cold PBS and stored at -80°C. Cells pellets for DNA and RNA were collected and stored at -80°C.

4.2.2 Growth of normal human ovarian surface epithelial cells

Human ovarian surface epithelial cells (HOSE) were purchased from ScienCell [cat# 7310] and grown until ~80% confluent in manufacturer's media. Cells were harvested using trypsin. Cell pellets were frozen at -80℃ until DNA and RNA was extracted using Qiagen All Prep Kit. Formaldehyde cross-linked cells were stored at -80°C.

4.2.3 Isolation of DNA, RNA, and Chromatin

DNA and RNA was isolated using Qiagen's DNeasy and RNeasy kits and stored at 4°C, -20°C or -80°C (RNA) until ready to use. To isolate chromatin, cross-linked cells were thawed on ice and sonicated using

Covaris to break down chromatin to mono/dinucleotides. Chromatin was used directly for IP or stored with 10% glycerol at -80°C.

4.2.4 Whole genome bisulfite sequencing and data analysis

Whole genome bisulfite sequencing (WGBS) libraries were built using two comparable methods. Ovc316 purified DNA was sonicated to ~300bp using Covaris, followed by end repair, A-tailing and ligation to custom methylated Illumina compatible adapters (sequences in Section 2.2.2). The adapter-ligated DNA is treated to sodium bisulfite treatment using the Thermo Fisher MethylCode Bisulfite Conversion (Cat #MECOV50). Bisulfite DNA is purified and amplified to create final full length sequencing adapters and add indices. Ovc316 WGBS libraries were sequenced PE100 on an Illumina HiSeq 2000. All OVCAR5 WGBS libraries were built using Swift Accel-NGS Methyl Seq Kit following manufacturer's protocol. Libraries were sequenced SE150 on HiSeq 2000.

Data were processed as described in Section 2.2.2. Briefly, libraries were trimmed used Trim Galore v 0.4.1 and mapped using Bismark[140]. Methylation at nucleotide resolution was called using MethPipe[141].

4.2.5 Poly-A selected RNA-seq and data analysis

Dynal Oligo-dT beads were used to enrich the poly-A fraction from total RNA (manufacturer's protocol). Poly-A enriched RNA was treated with DNase to get rid of any contaminating DNA prior to building RNA-seq libraries. Ovc316 replicate 1 samples and HOSE replicates 1 and 2 were built using the TotalScript kit from Epicentre and sequenced SE50 on an Illumina HiSeq 2000. All other libraries were built using ScriptSeq Kit from Epicentre and sequenced SE75 on an Illumina HiSeq 2000. Data were mapped to hg19 using Kalisto[197] and differential gene expression called using DEseq2[163].

Differential gene expression lists were put into Enrichr[198,199] to obtain significant GO Biological Processes categories and significance values. Protein-protein interaction maps were generated using Cytoscape[200].

4.2.6 ChIP-Seq

ChIP-Seq analysis was performed as mentioned in Section 3.2.2. Briefly, chromatin is immunoprecipitated using antibodies specific to histone modification of interest. Antibodies are tested for specificity via dot blot. DNA is purified from IP and sequenced on an Illumina sequencing platform. ChIP-seq is mapped using Bowtie2 and MACS was used to identify ChIP-Seq peaks.

4.3 Results

4.3.1 Gene expression in OvCSCs and tumor cells

I collected xenograft tumors grown from two ovarian cancer cell populations; one from the OVCAR5 cell line previously shown to have a CD133+ population[201], and the other from a patient also shown to have a CD133+ OvCSC population, Ovc316[195]. Tumors were dissociated using following the MACS tumor dissociation protocol. All RNA-seq libraries were sequenced SE50 or SE75 on the Hiseq 2000 platform. The number of transcripts detected by Kallisto with TPM greater than one varies from 19,215 to 48,904 across all samples (Table C.1). Technical replicates were combined and DESeq2 was used to identify differentially expressed genes in a pairwise comparison (Figure 4.1). Differentially expressed genes with a p-value of <0.01 were used for all downstream analysis. For ease, CD133+ cells are referred to as OvCSCs throughout the rest of this text.

First, we asked how different the ovarian cancer cells' gene expression profiles were from normal human ovarian surface epithelial cells (HOSE cells), one of two possible/accepted origins of ovarian cancer. We compared OvCSC expression from both OVCAR5 and ovc316 cells to HOSE cells and did the same for the CD133- population. In the OvCSC-to-HOSE pairwise comparison, 1,703 genes were upregulated in OvCSCs and 1,111 downregulated (Figure 4.1A). CD133- cells upregulated 1,390 genes compared to HOSE cells and downregulated 712 genes (Figure 4.1B). The ovarian cancer cells have roughly the same difference from normal ovarian cells, regardless if they are OvCSCs or CD133-. Next, we compared OvCSCs to CD133- tumor cells, and fewer genes were differentially expressed. Of differentially expressed genes, 404 were upregulated in OvCSCs and 262 were downregulated (Figure 4.1C). GO categories such as transmembrane transport and leukocyte cell-cell adhesion were included in the list of upregulated genes, while genes in pathways associated with proliferation and cell death were

downregulated (Figure 4.1D,E). Transmembrane transport pathway was enriched due to the expression of the ABC family of transporter genes in our OvCSCs, while *VCAM1* expression likely drove pathways like leukocyte cell-cell adhesion (and other pathways, Table C.2) to significance. The ABC transporter protein have been shown to be expressed on OvCSCs and are predicted to aid in transporting chemotherapy drugs out of the cell, thus giving OvCSCs their drug resistance phenotype[135]. *VCAM1* was previously shown to be expressed in CD133+ ovarian cancer cells that undergo EMT[195]. The downregulated gene pathways are likely caused by the quiescent nature of OvCSC[135]. Overall, our gene expression data support existing knowledge of the biology of OvCSCs.

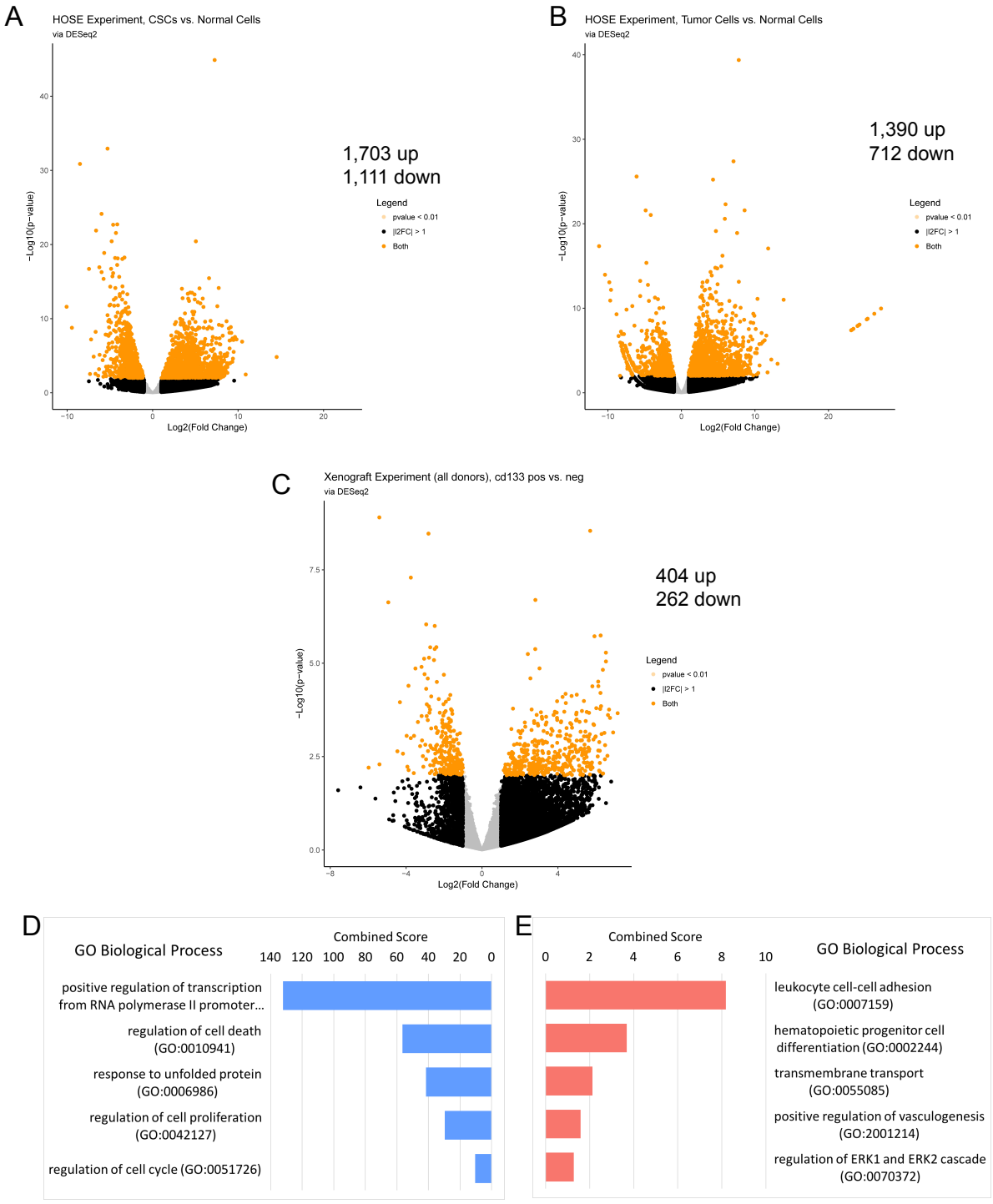


Figure 4.1 Differential Expression and Pairwise Comparisons between OvCSCs, Bulk Tumor and HOSE cells

Volcanoes plot of pairwise gene expression comparisons between (A) CSCs and HOSE, (B) CD133- cells and HOSE and (C) CSC and CD133- cells. (D) GO Biological Categories for upregulated CD133- bulk tumor cell genes compared to CSC. (E) GO Biological Categories for upregulated CSC genes compared to CD133- cells.

4.3.2 Protein-protein interactions maps predict key regulators

Another way to predict which genes are important in regulating cell identity is to look at protein-protein interactomes. We generated protein-protein interaction maps to identify which proteins were both highly expressed and highly connected with other expressed proteins (Figure 4.2). Within OvCSC upregulated genes, several proteins stood out as biological relevant in the interactome (Figure 4.2A; gene descriptions from www.genecards.org). The most connected protein is PTPRC, a signaling molecule that regulates cellular processes including growth, differentiation and oncogenic transformation. VCAM1 also came up in our protein-protein interactome. It is important cell-cell recognition, adhesion and signal transduction. CFTR is an ABC transporter protein, previously shown to be involved in drug-resistance, is highly connected in the OvCSC network, where many of ABC transporters can also be found. ATM a gene known to be mutated in cancer, particularly in breast cancer[202] is highly expressed in the OvCSC protein network. Lastly, OCT4 (*POU5F1 gene)* also shows up in our protein-protein network. The identification of OCT4/*POU5F1*, a stem cell transcription factor, in OvCSCs is strong evidence of a stem cell-like regulatory network in OvCSCs.

The CD133- upregulated gene list provided a less informative protein-protein interaction network. There was one large node around UBC, the ubiquitin c protein, and all of the edges forming a large network (Figure 4.2B). We postulate that the CD133- population of cancer cells was probably much more heterogeneous, making distinction of any relevant protein-protein networks more challenging.

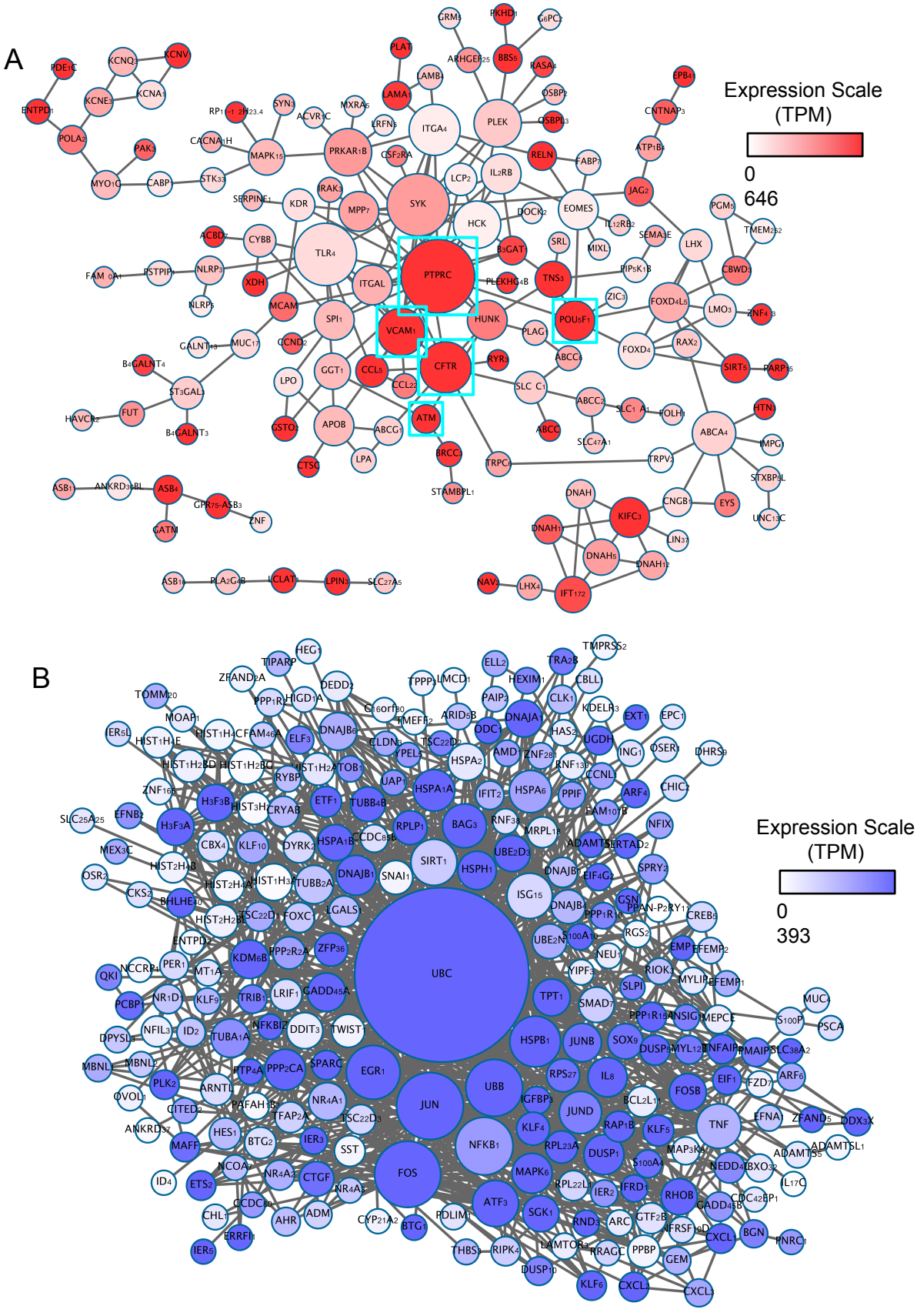

A
Expression Scale
(TPM)
0
646
B
Expression Scale
(TPM)
0
393

Figure 4.2. PPI Networks in CSCs and CD133- Bulk Tumor Cells
(A) Protein-protein interaction network taken from upregulated CD133+ OvCSC gene list. Color represents expression level and size of circle represents number of connections to other proteins on the map. Highlighted in cyan boxes are key genes mentioned in the text. (B) Protein-protein interaction network take from upregulated CD133- tumor cells.

4.3.3 Differentially Methylated Regions and Cell-type specific enhancers

One general feature of cancer cells is a hypomethylated genome[203]. We asked what differences existed between CD133+ OvCSCs and CD133- bulk tumor cells at the level of DNA methylation and could these differences in DNA methylation result in differential gene regulation. We performed whole genome bisulfite sequencing on one replicate of ovc316 and two replicates of OVCAR5 CD133+ and CD133- cells from dissociated xenograft tumors. Cytosine methylation was calculated using the program Methpipe[141] on each sample individually and methylation across all OvCSC samples and all CD133- samples was merged. Differentially methylated regions (DMRs) were identified between OvCSCs and CD133- samples in order to identify regions in OvCSCs that have higher and lower DNA methylation levels compared to CD133- bulk tumor cells, called hyperDMRs and hypoDMRs respectively. We identified 447 hyperDMRs and 1,244 hypoDMRs in OvCSCs (Figure 4.3A).

We wanted to know if any of the DMRs had potential regulatory function. Regulatory elements function differently when they are in a hyper- or hypomethylated states. As explained in Chapter 1, promoters that are hypomethylated are associated with gene expression and gene silencing when hypermethylated. To address this, we asked how many DMRs overlapped with other regulatory elements in the genome (Figure 4.3B). In addition to having more hypoDMRs, a higher fraction of OvCSC hypoDMRs are at genomic regions with regulatory potential than hyperDMRs. For example, 13.6% of hypoDMRs fall at promoters and 12.7% at first exons compared to 8.9% and 5.4% of hyperDMRs (Figure 4.3B). A similar trend is observed at CGI elements with 25.7% and 26.6% of hypoDMRs overlapping CGIs and shores but only 4.3% and 16.1% of hyperDMRs in OvCSCs. A striking contrast to this trend is observed at shelves where 14.8% of hyperDMRs and only 8.1% of hypoDMRs are located.

In order to identify enhancer regions in OvCSCs and the bulk tumor, we generated H3K4me1 ChIP-seq data in Ovc316 xenograft tumor CD133+ and CD133- cells. There are 74,830 H3K4me1 enhancer peaks identified in OvCSC and only 17,602 enhancers identified in CD133- cells (Figure 4.3C). We searched the OvCSC enhancers for TF motifs and found ESC TF motifs to be significantly enriched(Figure 4.3D). OCT4 was identified in our PPI and its motif also was enriched in the OvCSC enhancers, along with its ESC binding partner NANOG. PRDM14 and ESRRB motifs are also enriched suggesting a role of the ESC pluripotency network in regulating OvCSCs.

We were especially interested in the cell-type specific (CTS) enhancers as these would potentially regulate genes involved in CTS processes. In OvCSCs 82% of enhancers are CTS while only 30% of CD133- enhancers were cell-type specific. Predicted gene targets of the CTS enhancers are significantly upregulated in the cell types of interest (Figure 4.3E). A small fraction, 0.1% - 0.42%, of CTS enhancers overlap DMRs. Conversely, when we asked how many DMRs overlapped OvCSC CTS enhancers, a roughly equal percentage of hyperDMRs (21%) were found at OvCSC-specific enhancers compared to hypoDMRs (18%; Figure 4.3B). These data, along with the data presented above, suggest many of the DMRs could have regulatory activity. We further investigated what genes the DMRs could be regulating.

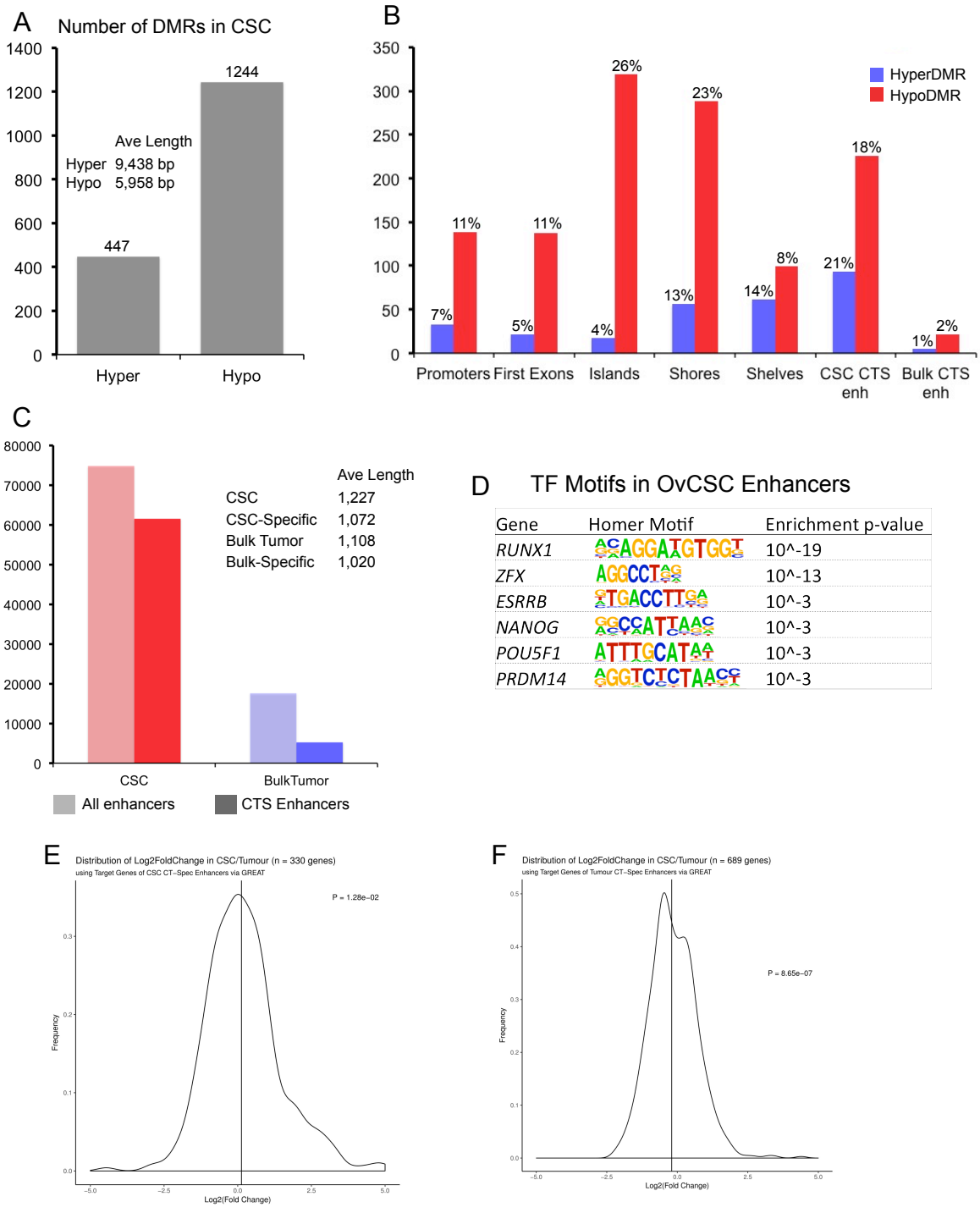


Figure 4.3. DMRs in CSCs
A) Number of hypoDMRs and hyperDMRs identified in CSC – CD133- comparison (B) Number of DMRs at known regulatory elements (C) Number of enhancers in CSCs and CD133- cell. (D) Enriched TF motifs in OvCSC H3K4me1 enhancers. (E) and (F) Density plots of log2 fold change of CSC CTS enhancers putative gene targets (E) and Tumor CTS enhancer putative gene targets (F) as identified by GREAT.

4.3.4 DMRs and their putative gene targets

Since many of OvCSC DMRs occurred at genomic locations that also had known regulatory function, we asked what genes could DMRs be regulating. We used GREAT[204] to identify the genes that are potentially regulated by OvCSC hyperDMRs and hypoDMRs (Figure 4.4 and Table 4.2). Hypermethylation at regulatory regions is generally thought to prevent TFs and activating factors from binding to their sequence motifs. For OvCSC hyperDMRs, these are regions that gained methylation and likely downregulate expression in OvCSCs, or are hypomethylated in tumor cells and should be positively regulating gene expression. We used GREAT to find which genes the hyperDMRs potentially regulated and then filtered this list based on genes that were downregulated in OvCSCs (Figure 4.4). This left us with 73 genes that were downregulated targets of hypermethylated regions in OvCSCs. These genes were included in GO Terms involving regulation of transcription and proliferation and negative regulation of apoptosis. These processes make sense in the context of tumor biology as a cancer needs to needs to upregulated cellular growth signals in order to expand. OvCSC hypermethylate the regulatory regions associated with these gene and downregulated their expression. This potentially provides an explanation to how OvCSCs develop resistance to therapeutics that target rapidly dividing cells.

We next looked at OvCSC hypoDMR regulated genes, using GREAT to identify genes potentially regulated by these regions and filtering for the genes that are upregulated in OvCSCs. This resulted in 201 genes that are upregulated in OvCSCs and potentially regulated by a hypoDMR. Some of the same GO categories came up as in the hyperDMR tumor upregulated genes list, including negative regulation of apoptosis and regulation proliferation pathways (Figure 4.4). It is interesting that genes involved in signal transduction and angiogenesis pathways arose as our protein-protein interactome identified VCAM1 as a major node in the network (Figure 4.4).

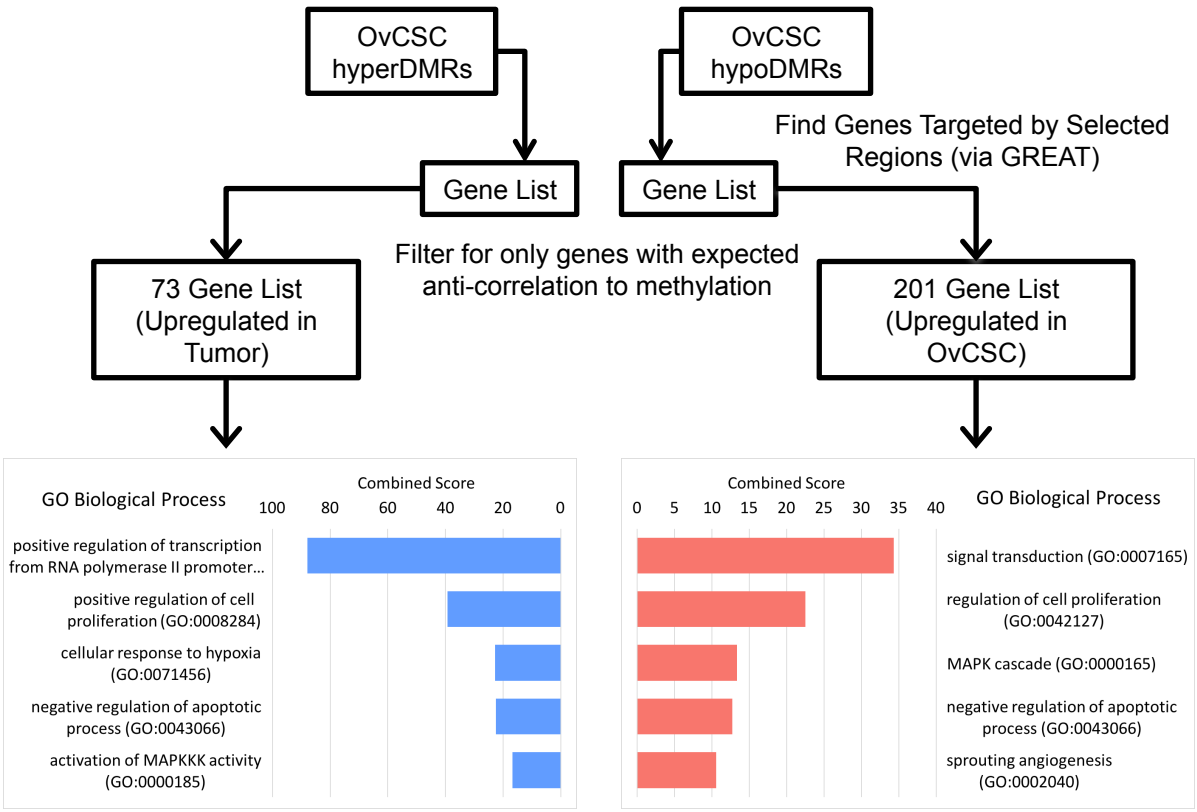


Figure 4.4. Workflow to identify potential DMR regulated genes.
OvCSC hyperDMRs were put into GREAT to identify potential gene targets. Putative gene targets were further filtered based on gene expression in relevant cell-type. A selection of GO categories for gene lists is shown along with significance scores.

4.4 Summary and Conclusions

Much like ESCs, CSCs use epigenetic mechanisms to make cell fate decisions. We attempted to look at the gene expression and epigenome of OvCSCs to gather evidence that could explain some of their biological properties. We identified several proteins that are potential key regulators in OvCSCs including the stem cell factor OCT4 (*POU5F1*). Identifying OCT4 as a key protein in the protein-protein interactome, and not just as an upregulated gene, implies that OvCSCs repurpose normal stem cell regulatory networks for tumor growth and development. Cancer is previously known to have genome-wide DNA methylation abnormalities. By narrowing our focus on the differences between CD133+ OvCSCs and the CD133- bulk tumor cells, we were able to identify pathways that are specifically differentially regulated between the two cell types.

There are a few limitations of this study. First, we have a very limited sample size. Ideally, we would have deeper sequencing of the libraries and more patient tumors and cell lines would be included to identify more general feature of OvCSCs. By including OvCSCs from different patients we need to be aware that OvCSCs from different individuals may use different epigenetic mechanisms to affect the same pathway. Therefore secondly, I propose taking a bottom up approach, where we identify the proteins/genes that are misregulated and then ask which epigenetic mechanism did each patient's OvCSC use to alter it. Third, functional validation of the identified pathways and proteins is necessary in order to gain clinical relevance. This must be accomplished through both *in vitro* and *in vivo* studies.

The work presented in this chapter should be viewed as a stepping stone for further experimentation. This type of hypothesis generating research is necessary in order to derive new questions about CSCs. Continued work on the epigenomic regulatory network in OvCSCs is ongoing.

Chapter 5:

Reflection

This body of work investigates epigenetic regulation in pluripotent and multipotent stem cells. Pluripotent stem cells, such as ESCs, are able to give rise to all the tissue of the embryo and developing organism. I explored the regulatory network of naïve Elf1 hESCs, a newly derived hESCs that represents the preimplantation stage of development. Multipotent stem cells, like CSCs, are able to give rise to cells of a tissue. I probed the regulatory network of OvCSCs, and tried to link known attributes of CSC biology to feature of the regulatory network. Despite their apparent differences, stem cells are characterized by their potential to give rise to differentiate cells. Exploring epigenetic features of stem cells can tell us how stem cell potential is defined and how stem cells make cell fate decisions. This work comprehensively combines two major aspect of the epigenome, histone modifications and DNA methylation to understand gene regulation in naïve hESCs and OvCSCs. Unexpected parallels in the TF that are active in these networks were found that hint at a universal stem cell network. Further research on naïve hESCs and OvCSCs will need to be done to refine the role of the regulatory network in defining these unique cellular states.

References

1 Qu, J., Zhou, M., Song, Q., Hong, E. E. & Smith, A. D. MLML: consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics* **29**, 2645-2646, doi:10.1093/bioinformatics/btt459 (2013).

2 Sperber, H. *et al.* The metabolome regulates the epigenetic landscape during naive-to-primed human embryonic stem cell transition. *Nat Cell Biol* **17**, 1523-1535, doi:10.1038/ncb3264 (2015).

3 Bongso, A., Fong, C. Y., Ng, S. C. & Ratnam, S. Isolation and culture of inner cell mass cells from human blastocysts. *Hum Reprod* **9**, 2110-2117 (1994).

4 Thomson, J. A. *et al.* Isolation of a primate embryonic stem cell line. *Proc Natl Acad Sci U S A* **92**, 7844-7848 (1995).

5 Thomson, J. A. *et al.* Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145-1147 (1998).

6 Reubinoff, B. E., Pera, M. F., Fong, C. Y., Trounson, A. & Bongso, A. Embryonic stem cell lines from human blastocysts: somatic differentiation in vitro. *Nat Biotechnol* **18**, 399-404, doi:10.1038/74447 (2000).

7 James, D., Levine, A. J., Besser, D. & Hemmati-Brivanlou, A. TGFbeta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development* **132**, 1273-1282, doi:10.1242/dev.01706 (2005).

8 Daheron, L. *et al.* LIF/STAT3 signaling fails to maintain self-renewal of human embryonic stem cells. *Stem Cells* **22**, 770-778, doi:10.1634/stemcells.22-5-770 (2004).

9 Sato, N., Meijer, L., Skaltsounis, L., Greengard, P. & Brivanlou, A. H. 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**, 55-63, doi:10.1038/nm979 (2004).

10 Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196-199, doi:10.1038/nature05972 (2007).

11 Brons, I. G. *et al.* Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191-195, doi:10.1038/nature05950 (2007).

12 Silva, J. & Smith, A. Capturing pluripotency. *Cell* **132**, 532-536, doi:10.1016/j.cell.2008.02.006 (2008).

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

14 Reijo Pera, R. A. *et al.* Gene expression profiles of human inner cell mass cells and embryonic stem cells. *Differentiation* **78**, 18-23, doi:10.1016/j.diff.2009.03.004 (2009).

15 Abeyta, M. J. *et al.* Unique gene expression signatures of independently-derived human embryonic stem cell lines. *Hum Mol Genet* **13**, 601-608, doi:10.1093/hmg/ddh068 (2004).

16 Huang, Y., Osorno, R., Tsakiridis, A. & Wilson, V. In Vivo differentiation potential of epiblast stem cells revealed by chimeric embryo formation. *Cell Rep* **2**, 1571-1578, doi:10.1016/j.celrep.2012.10.022 (2012).

17 Mascetti, V. L. & Pedersen, R. A. Human-Mouse Chimerism Validates Human Stem Cell Pluripotency. *Cell Stem Cell* **18**, 67-72, doi:10.1016/j.stem.2015.11.017 (2016).

18 Li, W. *et al.* in *Cell Stem Cell* Vol. 4 16-19 (2009).

19 Buecker, C. *et al.* A murine ESC-like state facilitates transgenesis and homologous recombination in human pluripotent stem cells. *Cell Stem Cell* **6**, 535-546, doi:10.1016/j.stem.2010.05.003 (2010).

20 Hanna, J. *et al.* Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proc Natl Acad Sci U S A* **107**, 9222-9227, doi:10.1073/pnas.1004584107 (2010).

21 Takashima, Y. *et al.* Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* **158**, 1254-1269, doi:10.1016/j.cell.2014.08.029 (2014).
22 Gafni, O. *et al.* Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 282-286, doi:10.1038/nature12745 (2013).
23 Theunissen, T. W. *et al.* Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15**, 471-487, doi:10.1016/j.stem.2014.07.002 (2014).
24 Valamehr, B. *et al.* Platform for induction and maintenance of transgene-free hiPSCs resembling ground state pluripotent stem cells. *Stem Cell Reports* **2**, 366-381, doi:10.1016/j.stemcr.2014.01.014 (2014).
25 Irie, N., Tang, W. W. C. & Azim Surani, M. in *Reprod Med Biol* Vol. 13 203-215 (2014).
26 Chan, Y. S. *et al.* Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell Stem Cell* **13**, 663-675, doi:10.1016/j.stem.2013.11.015 (2013).
27 Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* **20**, 1131-1139, doi:10.1038/nsmb.2660 (2013).
28 Vassena, R. *et al.* Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development* **138**, 3699-3709, doi:10.1242/dev.064741 (2011).
29 Blakeley, P. *et al.* Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3151-3165, doi:10.1242/dev.123547 (2015).
30 Ware, C. B. *et al.* Histone deacetylase inhibition elicits an evolutionarily conserved self-renewal program in embryonic stem cells. *Cell Stem Cell* **4**, 359-369, doi:10.1016/j.stem.2009.03.001 (2009).
31 Ware, C. B. *et al.* Derivation of naive human embryonic stem cells. *Proc Natl Acad Sci U S A* **111**, 4484-4489, doi:10.1073/pnas.1319738111 (2014).
32 Zhou, W. *et al.* HIF1alpha induced switch from bivalent to exclusively glycolytic metabolism during ESC-to-EpiSC/hESC transition. *Embo j* **31**, 2103-2116, doi:10.1038/emboj.2012.71 (2012).
33 Theunissen, T. W. *et al.* Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell* **19**, 502-515, doi:10.1016/j.stem.2016.06.011 (2016).
34 Guo, G. *et al.* Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass. *Stem Cell Reports* **6**, 437-446, doi:10.1016/j.stemcr.2016.02.005 (2016).
35 Lengner, C. J. *et al.* Derivation of pre-X inactivation human embryonic stem cells under physiological oxygen concentrations. *Cell* **141**, 872-883, doi:10.1016/j.cell.2010.04.010 (2010).
36 Huang, K., Maruyama, T. & Fan, G. The naive state of human pluripotent stem cells: a synthesis of stem cell and preimplantation embryo transcriptome analyses. *Cell Stem Cell* **15**, 410-415, doi:10.1016/j.stem.2014.09.014 (2014).
37 Bestor, T., Laudano, A., Mattaliano, R. & Ingram, V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol* **203**, 971-983 (1988).
38 Lei, H. *et al.* De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* **122**, 3195-3205 (1996).
39 Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257 (1999).

40 Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322, doi:10.1038/nature08514 (2009).
41 Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-1022, doi:10.1101/gad.2037511 (2011).
42 Smith, Z. D. *et al.* DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**, 611-615, doi:10.1038/nature13581 (2014).
43 Habibi, E. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13**, 360-369, doi:10.1016/j.stem.2013.06.002 (2013).
44 Leitch, H. G. *et al.* Naive pluripotency is associated with global DNA hypomethylation. *Nat Struct Mol Biol* **20**, 311-316, doi:10.1038/nsmb.2510 (2013).
45 Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590-604, doi:10.1016/j.cell.2012.03.026 (2012).
46 Jackson, M. *et al.* Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. *Mol Cell Biol* **24**, 8862-8871, doi:10.1128/mcb.24.20.8862-8871.2004 (2004).
47 Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol Cell Biol* **23**, 5594-5605 (2003).
48 Tsumura, A. *et al.* Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes Cells* **11**, 805-814, doi:10.1111/j.1365-2443.2006.00984.x (2006).
49 Dodge, J. E., Ramsahoye, B. H., Wo, Z. G., Okano, M. & Li, E. De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. *Gene* **289**, 41-48 (2002).
50 Ng, R. K. *et al.* Epigenetic restriction of embryonic cell lineage fate by methylation of Elf5. *Nat Cell Biol* **10**, 1280-1290, doi:10.1038/ncb1786 (2008).
51 Ooi, S. K. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714-717, doi:10.1038/nature05987 (2007).
52 Arand, J. *et al.* In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet* **8**, e1002750, doi:10.1371/journal.pgen.1002750 (2012).
53 Hu, Y. G. *et al.* Regulation of DNA methylation activity through Dnmt3L promoter methylation by Dnmt3 enzymes in embryonic development. *Hum Mol Genet* **17**, 2654-2664, doi:10.1093/hmg/ddn165 (2008).
54 Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606-610, doi:10.1038/nature13544 (2014).
55 Okae, H. *et al.* Genome-wide analysis of DNA methylation dynamics during early human development. *PLoS Genet* **10**, e1004868, doi:10.1371/journal.pgen.1004868 (2014).
56 Molaro, A. *et al.* Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146**, 1029-1041, doi:10.1016/j.cell.2011.08.016 (2011).
57 Pastor, W. A. *et al.* Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory. *Cell Stem Cell* **18**, 323-329, doi:10.1016/j.stem.2016.01.019 (2016).
58 Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci U S A* **97**, 5237-5242 (2000).
59 Ziller, M. J. *et al.* Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet* **7**, e1002389, doi:10.1371/journal.pgen.1002389 (2011).

60 Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868-5877, doi:10.1093/nar/gki901 (2005).
61 Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134-1148, doi:10.1016/j.cell.2013.04.022 (2013).
62 Wossidlo, M. *et al.* 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* **2**, 241, doi:10.1038/ncomms1240 (2011).
63 Efimova, O. A. *et al.* Chromosome hydroxymethylation patterns in human zygotes and cleavage-stage embryos. *Reproduction* **149**, 223-233, doi:10.1530/rep-14-0343 (2015).
64 Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935, doi:10.1126/science.1170116 (2009).
65 Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-1303, doi:10.1126/science.1210597 (2011).
66 He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-1307, doi:10.1126/science.1210944 (2011).
67 Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J Biol Chem* **286**, 35334-35338, doi:10.1074/jbc.C111.284620 (2011).
68 Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129-1133, doi:10.1038/nature09303 (2010).
69 Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One* **5**, e8888, doi:10.1371/journal.pone.0008888 (2010).
70 Jin, S. G., Kadam, S. & Pfeifer, G. P. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res* **38**, e125, doi:10.1093/nar/gkq223 (2010).
71 Song, C. X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nat Biotechnol* **30**, 1107-1116, doi:10.1038/nbt.2398 (2012).
72 Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368-1380, doi:10.1016/j.cell.2012.04.027 (2012).
73 Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934-937, doi:10.1126/science.1220671 (2012).
74 Sim, Y. J. *et al.* 2i Maintains a Naive Ground State in ESCs through Two Distinct Epigenetic Mechanisms. *Stem Cell Reports* **8**, 1312-1328, doi:10.1016/j.stemcr.2017.04.001 (2017).
75 Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705, doi:10.1016/j.cell.2007.02.005 (2007).
76 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260, doi:10.1038/38444 (1997).
77 Sarmento, O. F. *et al.* Dynamic alterations of specific histone modifications during early murine development. *J Cell Sci* **117**, 4449-4459, doi:10.1242/jcs.01328 (2004).
78 Lepikhov, K. & Walter, J. Differential dynamics of histone H3 methylation at positions K4 and K9 in the mouse zygote. *BMC Dev Biol* **4**, 12, doi:10.1186/1471-213x-4-12 (2004).
79 Nonchev, S. & Tsanev, R. Protamine-histone replacement and DNA replication in the male mouse pronucleus. *Mol Reprod Dev* **25**, 72-76, doi:10.1002/mrd.1080250113 (1990).
80 Santos, F., Peters, A. H., Otte, A. P., Reik, W. & Dean, W. Dynamic chromatin modifications characterise the first cell cycle in mouse embryos. *Dev Biol* **280**, 225-236, doi:10.1016/j.ydbio.2005.01.025 (2005).

81 Erhardt, S. *et al.* Consequences of the depletion of zygotic and embryonic enhancer of zeste 2 during preimplantation mouse development. *Development* **130**, 4235-4248 (2003).
82 Cowell, I. G. *et al.* Heterochromatin, HP1 and methylation at lysine 9 of histone H3 in animals. *Chromosoma* **111**, 22-36 (2002).
83 Jackson, V., Shires, A., Tanphaichitr, N. & Chalkley, R. Modifications to histones immediately after synthesis. *J Mol Biol* **104**, 471-483 (1976).
84 Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318, doi:10.1038/ng1966 (2007).
85 Dahl, J. A. *et al.* Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature* **537**, 548-552, doi:10.1038/nature19360 (2016).
86 Zhang, B. *et al.* Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature* **537**, 553-557, doi:10.1038/nature19361 (2016).
87 Liu, X. *et al.* Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature* **537**, 558-562, doi:10.1038/nature19362 (2016).
88 Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**, 825-837, doi:10.1016/j.molcel.2013.01.038 (2013).
89 Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112, doi:10.1038/nature07829 (2009).
90 Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283, doi:10.1038/nature09692 (2011).
91 Zentner, G. E., Tesar, P. J. & Scacheri, P. C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res* **21**, 1273-1283, doi:10.1101/gr.122382.111 (2011).
92 Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642-654, doi:10.1016/j.cell.2012.12.033 (2013).
93 Yeom, Y. I. *et al.* Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. *Development* **122**, 881-894 (1996).
94 Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936, doi:10.1073/pnas.1016071107 (2010).
95 Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87**, 953-959 (1996).
96 Yao, T. P. *et al.* Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. *Cell* **93**, 361-372 (1998).
97 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).
98 Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A* **110**, 17921-17926, doi:10.1073/pnas.1317023110 (2013).
99 Whyte, W. A. *et al.* Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* **482**, 221-225, doi:10.1038/nature10805 (2012).
100 Forneris, F., Binda, C., Vanoni, M. A., Battaglioli, E. & Mattevi, A. Human Histone Demethylase LSD1 Reads the Histone Code. *The Journal of Biological Chemistry*, doi:10.1074/jbc.M509549200 (2005).
101 Montgomery, N. D. *et al.* The murine polycomb group protein Eed is required for global histone H3 lysine-27 methylation. *Curr Biol* **15**, 942-947, doi:10.1016/j.cub.2005.04.051 (2005).

102 Mozzetta, C., Boyarchuk, E., Pontis, J. & Ait-Si-Ali, S. Sound of silence: the properties and functions of repressive Lys methyltransferases. *Nat Rev Mol Cell Biol* **16**, 499-513, doi:10.1038/nrm4029 (2015).
103 Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128**, 669-681, doi:10.1016/j.cell.2007.01.033 (2007).
104 Faust, C., Lawson, K. A., Schork, N. J., Thiel, B. & Magnuson, T. The Polycomb-group gene eed is required for normal morphogenetic movements during gastrulation in the mouse embryo. *Development* **125**, 4495-4506 (1998).
105 Pasini, D., Bracken, A. P., Jensen, M. R., Lazzerini Denchi, E. & Helin, K. Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *Embo j* **23**, 4061-4071, doi:10.1038/sj.emboj.7600402 (2004).
106 Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315-326, doi:10.1016/j.cell.2006.02.041 (2006).
107 Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* **8**, 532-538, doi:10.1038/ncb1403 (2006).
108 Zhao, X. D. *et al.* Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* **1**, 286-298, doi:10.1016/j.stem.2007.08.004 (2007).
109 Voigt, P. *et al.* Asymmetrically modified nucleosomes. *Cell* **151**, 181-193, doi:10.1016/j.cell.2012.09.002 (2012).
110 Dodge, J. E., Kang, Y. K., Beppu, H., Lei, H. & Li, E. Histone H3-K9 methyltransferase ESET is essential for early development. *Mol Cell Biol* **24**, 2478-2486 (2004).
111 Yeap, L. S., Hayashi, K. & Surani, M. A. ERG-associated protein with SET domain (ESET)-Oct4 interaction regulates pluripotency and represses the trophectoderm lineage. *Epigenetics Chromatin* **2**, 12, doi:10.1186/1756-8935-2-12 (2009).
112 Yuan, P. *et al.* Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. *Genes Dev* **23**, 2507-2520, doi:10.1101/gad.1831909 (2009).
113 Barr, M. L. & Bertram, E. G. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature* **163**, 676 (1949).
114 Heard, E. & Disteche, C. M. Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev* **20**, 1848-1867, doi:10.1101/gad.1422906 (2006).
115 Ahmed, K. *et al.* Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PLoS One* **5**, e10531, doi:10.1371/journal.pone.0010531 (2010).
116 Hiratani, I. *et al.* Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res* **20**, 155-169, doi:10.1101/gr.099796.109 (2010).
117 Meshorer, E. *et al.* Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev Cell* **10**, 105-116, doi:10.1016/j.devcel.2005.10.017 (2006).
118 Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435, doi:10.1038/nature09380 (2010).
119 Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol Cell* **38**, 603-613, doi:10.1016/j.molcel.2010.03.016 (2010).
120 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

121 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
122 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).
123 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).
124 Ke, Y. *et al.* 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell* **170**, 367-381.e320, doi:10.1016/j.cell.2017.06.029 (2017).
125 Jung, Y. H. *et al.* Chromatin States in Mouse Sperm Correlate with Embryonic and Adult Regulatory Landscapes. *Cell Rep* **18**, 1366-1382, doi:10.1016/j.celrep.2017.01.034 (2017).
126 Ji, X. *et al.* 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell* **18**, 262-275, doi:10.1016/j.stem.2015.11.007 (2016).
127 Battulin, N. *et al.* Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol* **16**, 77, doi:10.1186/s13059-015-0642-0 (2015).
128 Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110-114, doi:10.1038/nature21711 (2017).
129 Du, Z. *et al.* Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232-235, doi:10.1038/nature23263 (2017).
130 Hug, C. B., Grimaldi, A. G., Kruse, K. & Vaquerizas, J. M. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. *Cell* **169**, 216-228.e219, doi:10.1016/j.cell.2017.03.024 (2017).
131 Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64, doi:10.1038/nature08497 (2009).
132 Guddati, A. K. Ovarian cancer stem cells: elusive targets for chemotherapy. *Med Oncol* **29**, 3400-3408, doi:10.1007/s12032-012-0252-6 (2012).
133 Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2012. *CA Cancer J Clin* **62**, 10-29, doi:10.3322/caac.20138 (2012).
134 Gloss, B. S. & Samimi, G. Epigenetic biomarkers in epithelial ovarian cancer. *Cancer Lett* **342**, 257-263, doi:10.1016/j.canlet.2011.12.036 (2014).
135 Lupia, M. & Cavallaro, U. Ovarian cancer stem cells: still an elusive entity? *Mol Cancer* **16**, 64, doi:10.1186/s12943-017-0638-3 (2017).
136 Kwon, M. J. in *Int J Mol Sci* Vol. 14 18148-18180 (2013).
137 Abbosh, P. H. *et al.* Dominant-negative histone H3 lysine 27 mutant derepresses silenced tumor suppressor genes and reverses the drug-resistant phenotype in cancer cells. *Cancer Res* **66**, 5582-5591, doi:10.1158/0008-5472.can-05-3575 (2006).
138 Schondorf, T. *et al.* Hypermethylation of the PTEN gene in ovarian cancer cell lines. *Cancer Lett* **207**, 215-220, doi:10.1016/j.canlet.2003.10.028 (2004).
139 Baldwin, R. L. *et al.* BRCA1 promoter region hypermethylation in ovarian carcinoma: a population-based study. *Cancer Res* **60**, 5329-5333 (2000).
140 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572, doi:10.1093/bioinformatics/btr167 (2011).
141 Song, Q. *et al.* A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* **8**, e81148, doi:10.1371/journal.pone.0081148 (2013).
142 Yu, M. *et al.* Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat Protoc* **7**, 2159-2170, doi:10.1038/nprot.2012.137 (2012).

143 Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402, doi:10.1038/nature10008 (2011).
144 Pastor, W. A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-397, doi:10.1038/nature10102 (2011).
145 Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol* **12**, R54, doi:10.1186/gb-2011-12-6-r54 (2011).
146 Szulwach, K. E. *et al.* Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet* **7**, e1002154, doi:10.1371/journal.pgen.1002154 (2011).
147 Wu, H. *et al.* Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes Dev* **25**, 679-684, doi:10.1101/gad.2036011 (2011).
148 Xu, Y. *et al.* Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Mol Cell* **42**, 451-464, doi:10.1016/j.molcel.2011.04.005 (2011).
149 Ndlovu, M. N., Denis, H. & Fuks, F. Exposing the DNA methylome iceberg. *Trends Biochem Sci* **36**, 381-387, doi:10.1016/j.tibs.2011.03.002 (2011).
150 Hudson, Q. J., Kulinski, T. M., Huetter, S. P. & Barlow, D. P. Genomic imprinting mechanisms in embryonic and extraembryonic mouse tissues. *Heredity (Edinb)* **105**, 45-56, doi:10.1038/hdy.2010.23 (2010).
151 Liu, H., Kim, J. M. & Aoki, F. Regulation of histone H3 lysine 9 methylation in oocytes and early pre-implantation embryos. *Development* **131**, 2269-2280, doi:10.1242/dev.01116 (2004).
152 Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat Rev Mol Cell Biol* **7**, 540-546, doi:10.1038/nrm1938 (2006).
153 Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479-491, doi:10.1016/j.stem.2010.03.018 (2010).
154 Ware, C. B. Concise Review: Lessons from Naïve Human Pluripotent Cells. *STEM CELLS*, doi:10.1002/stem.2507 (2016).
155 Rossant, J. Mouse and human blastocyst-derived stem cells: vive les differences. *Development* **142**, 9-12, doi:10.1242/dev.115451 (2015).
156 Hawkins, R. D. *et al.* Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity* **38**, 1271-1284, doi:10.1016/j.immuni.2013.05.011 (2013).
157 Hawkins, R. D. *et al.* Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. *Cell Res* **21**, 1393-1409, doi:10.1038/cr.2011.146 (2011).
158 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
159 Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-191, doi:10.1093/nar/gku365 (2014).
160 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
161 Quinlan, A. R. & Hall, I. M. in *Bioinformatics* Vol. 26 841-842 (2010).
162 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

163 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).
164 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).
165 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515, doi:10.1038/nbt.1621 (2010).
166 Wickham, H. (Springer-Verlag New York, 2009).
167 Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-1093, doi:10.1093/bioinformatics/btp101 (2009).
168 Deng, X. *et al.* Bipartite structure of the inactive mouse X chromosome. *Genome Biol* **16**, 152, doi:10.1186/s13059-015-0728-8 (2015).
169 Servant, N. *et al.* HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics* **28**, 2843-2844, doi:10.1093/bioinformatics/bts521 (2012).
170 Shin, H. *et al.* TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res* **44**, e70, doi:10.1093/nar/gkv1505 (2016).
171 Giorgetti, L. *et al.* Structural organization of the inactive X chromosome in the mouse. *Nature* **535**, 575-579, doi:10.1038/nature18589 (2016).
172 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).
173 Chen, Y.-G., Li, Z. & Wang, X.-F. Where PI3K/Akt Meets Smads: The Crosstalk Determines Human Embryonic Stem Cell Fate. *Cell Stem Cell* **10**, 231-232, doi:10.1016/j.stem.2012.02.008 (2012).
174 Jiang, Q. *et al.* LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. *BMC Genomics* **16 Suppl 3**, S2, doi:10.1186/1471-2164-16-s3-s2 (2015).
175 Tropepe, V. *et al.* Direct neural fate specification from embryonic stem cells: a primitive mammalian neural stem cell stage acquired through a default mechanism. *Neuron* **30**, 65-78 (2001).
176 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837, doi:10.1016/j.cell.2007.05.009 (2007).
177 Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77-88, doi:10.1016/j.cell.2007.05.042 (2007).
178 Boyer, L. A. *et al.* Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-353, doi:10.1038/nature04733 (2006).
179 Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-124, doi:10.1038/35065138 (2001).
180 Niakan, K. K. & Eggan, K. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Dev Biol* **375**, 54-64, doi:10.1016/j.ydbio.2012.12.008 (2013).
181 Consortium, E. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
182 Zhang, H. M. *et al.* AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res* **43**, D76-81, doi:10.1093/nar/gku887 (2015).

183 Burton, A. *et al.* Single-cell profiling of epigenetic modifiers identifies PRDM14 as an inducer of cell fate in the mammalian embryo. *Cell Rep* **5**, 687-701, doi:10.1016/j.celrep.2013.09.044 (2013).
184 Ma, C., Li, W., Xu, Y. & Rizo, J. Munc13 mediates the transition from the closed syntaxin-Munc18 complex to the SNARE complex. *Nat Struct Mol Biol* **18**, 542-549, doi:10.1038/nsmb.2047 (2011).
185 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240-244, doi:10.1038/nature14450 (2015).
186 van der Heijden, G. W. *et al.* Asymmetry in histone H3 variants and lysine methylation between paternal and maternal chromatin of the early mouse zygote. *Mech Dev* **122**, 1008-1022, doi:10.1016/j.mod.2005.04.009 (2005).
187 Wiekowski, M., Miranda, M., Nothias, J. Y. & DePamphilis, M. L. Changes in histone synthesis and modification at the beginning of mouse development correlate with the establishment of chromatin mediated repression of transcription. *J Cell Sci* **110 (Pt 10)**, 1147-1158 (1997).
188 Adenot, P. G., Mercier, Y., Renard, J. P. & Thompson, E. M. Differential H4 acetylation of paternal and maternal chromatin precedes DNA replication and differential transcriptional activity in pronuclei of 1-cell mouse embryos. *Development* **124**, 4615-4625 (1997).
189 Lee, M. G. *et al.* in *Mol Cell Biol* Vol. 26 6395-6402 (2006).
190 Chaffer, C. L. & Weinberg, R. A. A perspective on cancer cell metastasis. *Science* **331**, 1559-1564, doi:10.1126/science.1203543 (2011).
191 Chen, J. *et al.* A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature* **488**, 522-526, doi:10.1038/nature11287 (2012).
192 Balch, C., Fang, F., Matei, D. E., Huang, T. H. & Nephew, K. P. Minireview: epigenetic changes in ovarian cancer. *Endocrinology* **150**, 4003-4011, doi:10.1210/en.2009-0404 (2009).
193 Mak, A. B. *et al.* Regulation of CD133 by HDAC6 promotes beta-catenin signaling to suppress cancer cell differentiation. *Cell Rep* **2**, 951-963, doi:10.1016/j.celrep.2012.09.016 (2012).
194 Cho, R. W. & Clarke, M. F. Recent advances in cancer stem cells. *Curr Opin Genet Dev* **18**, 48-53, doi:10.1016/j.gde.2008.01.017 (2008).
195 Strauss, R. *et al.* Analysis of epithelial and mesenchymal markers in ovarian cancer reveals phenotypic heterogeneity and plasticity. *PLoS One* **6**, e16186, doi:10.1371/journal.pone.0016186 (2011).
196 Toh, T. B., Lim, J. J. & Chow, E. K. Epigenetics in cancer stem cells. *Mol Cancer* **16**, 29, doi:10.1186/s12943-017-0596-9 (2017).
197 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-527, doi:10.1038/nbt.3519 (2016).
198 Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128, doi:10.1186/1471-2105-14-128 (2013).
199 Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97, doi:10.1093/nar/gkw377 (2016).
200 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).
201 Cioffi, M. *et al.* Identification of a distinct population of CD133(+)CXCR4(+) cancer stem cells in ovarian cancer. *Sci Rep* **5**, 10357, doi:10.1038/srep10357 (2015).
202 Ahmed, M. & Rahman, N. ATM and breast cancer susceptibility. *Oncogene* **25**, 5906-5911, doi:10.1038/sj.onc.1209873 (2006).

203 Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43**, 768-775, doi:10.1038/ng.865 (2011).
204 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501, doi:10.1038/nbt.1630 (2010).

Appendix A

Supplemental material to Chapter 2

	Raw Reads	Mapped reads
Elf1 naïve rep1 TAB-Seq	365,008,884	295,470,450
Elf1 naïve rep2 TAB-Seq	365,555,159	292,379,050
Elf1 naïve rep1 WGBS	204,346,075	141,407,484
Elf1 naïve rep2 WGBS	199,647,741	151,732,283
Elf1 AF rep1 WGBS	106,262,562	83,544,369
Elf1 AF rep2 WGBS	117,429,313	92,663,532

Table A.1 TAB-Seq and WGBS Sequencing Statistics

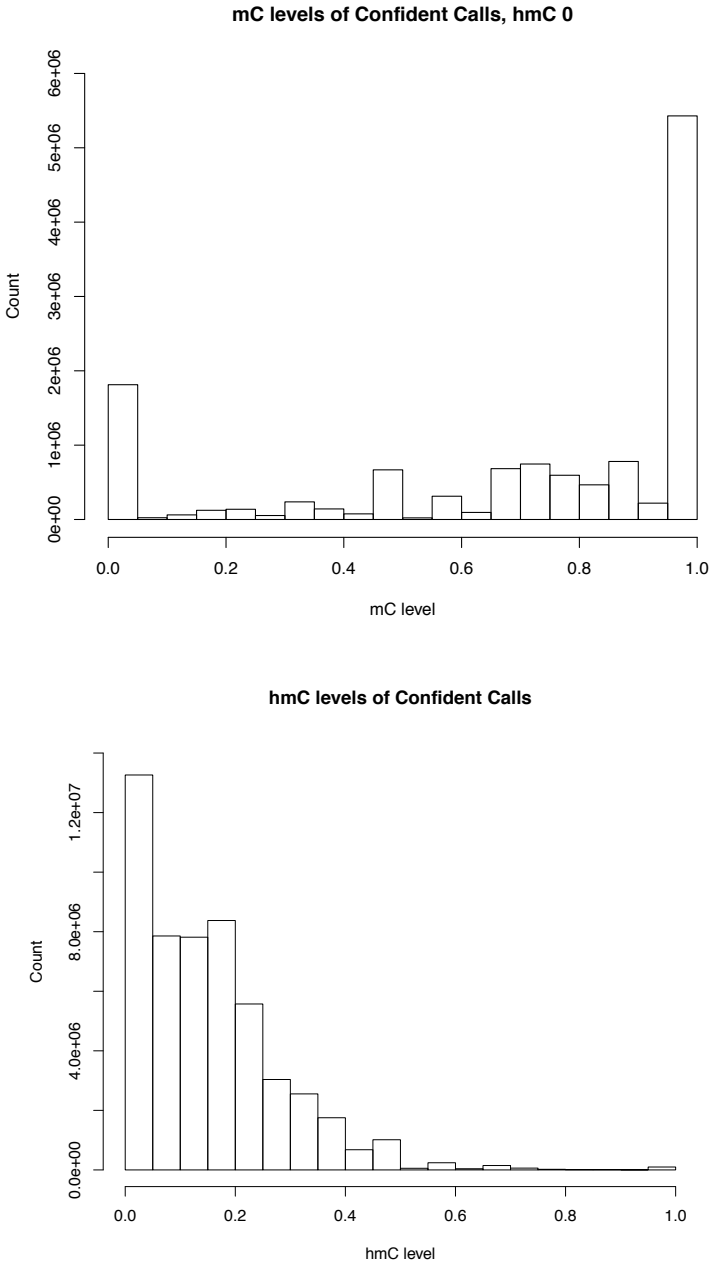


Figure A.1 Distribution of 5mC and 5hmC methylation/hydroxymethylation levels
Distribution of 5mC and 5hmC levels based on confident calls calculated from WGBS and TAB-Seq libraries

Appendix B

Supplemental material to Chapter 3

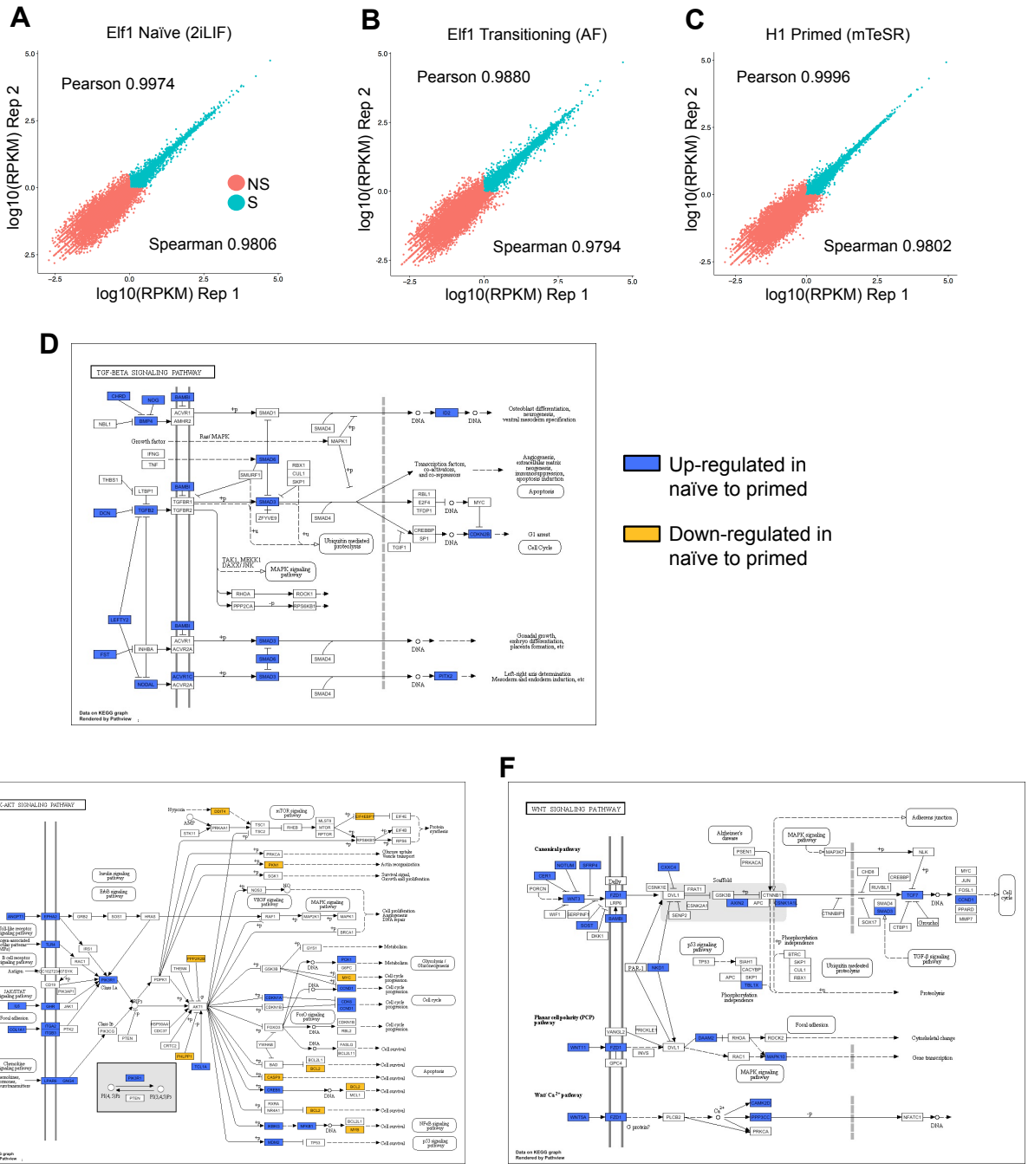


Figure B.1 RNA-seq replicates and DEGs in Key Pathways
(A-C) Correlation between RNA-seq replicates, log10(rpkm is plotted). On figure NS means “not significant” and S means “significant” (D-F) DEGs involved in the TGF-beta pathway (D), PI3K-AKT

signaling pathway (E) and WNT signaling pathway (F). Genes are colored based on up-regulation (blue) or down-regulation (light orange) in naïve compared to primed hESCs.

Available as an Excel file

Table B.1 DEGs and DEG Pathways
List of differentially expressed genes and Gene Ontology pathways

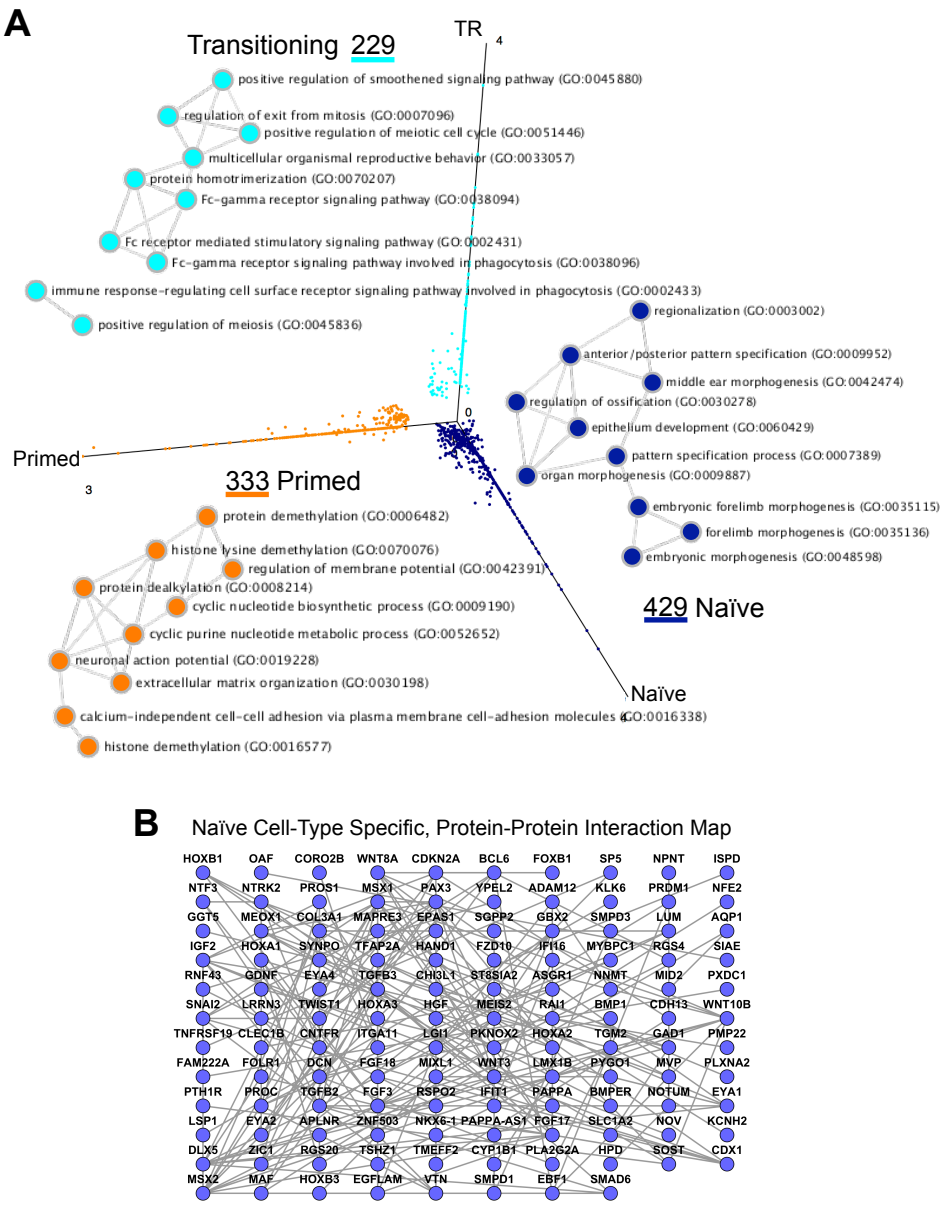


Figure B.2 Cell-type Specific Genes and GO Categories
Cell type-specific genes in the different hESC stages by applying a cutoff of a RPKM value greater than or equal to two in one cell type and less than one in the other two cell types. GO term shown in network, connected by shared genes between terms.

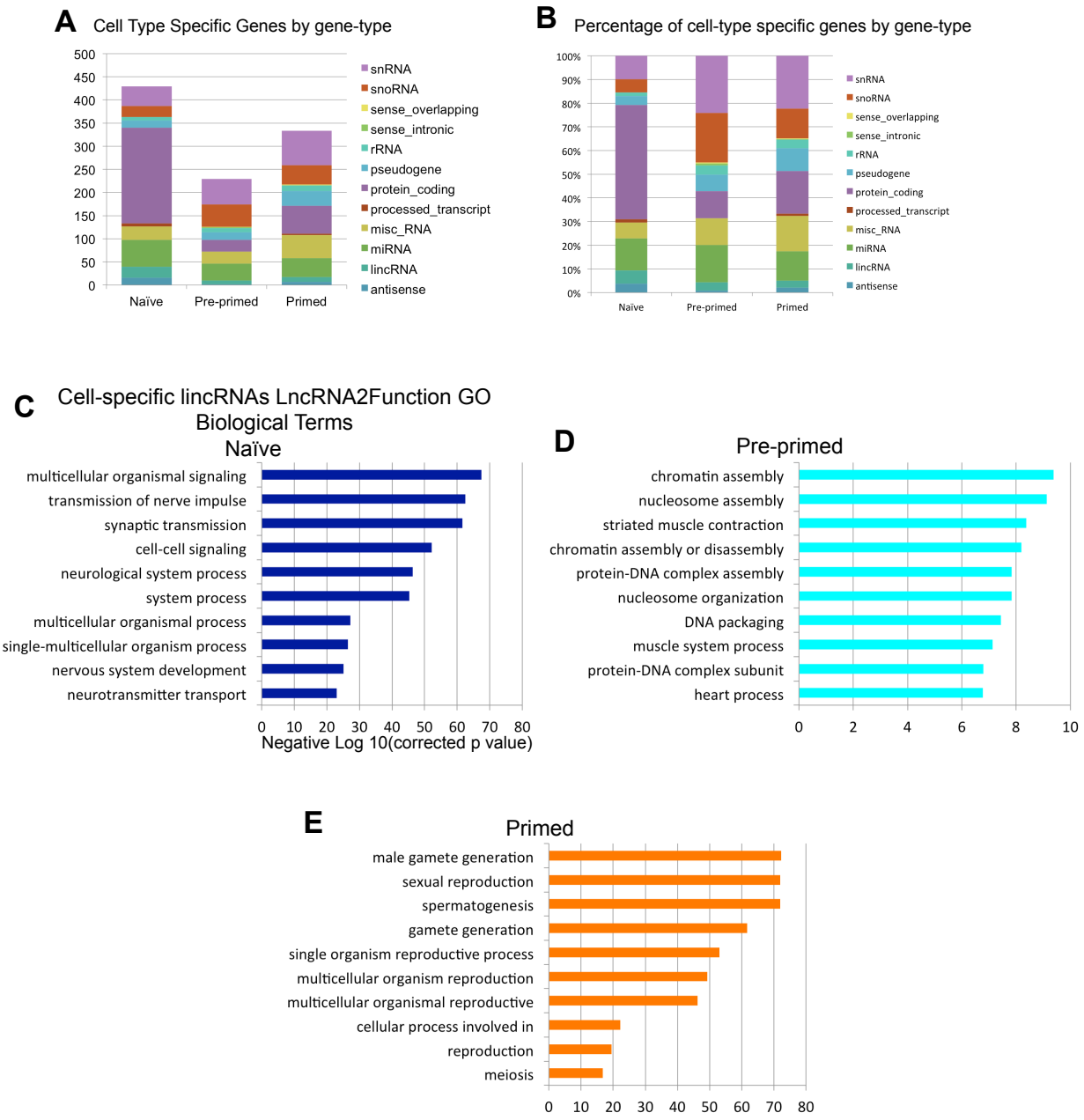


Figure B.3 Cell-Type Specific Genes by Gene Type
(A) Cell-type specific genes, breakdown by gene-type as defined by Gencode. (B) Cell-type specific genes, breakdown as percentages by gene-type as defined by Gencode. (C-E) Gene Ontology Biological Processes Terms of cell-type specific lincRNAs in each hESC cell type.

Elf1 Naive Sample	Number of Reads Sequenced	Mapped reads	Mapping Efficiency
ENK27acRep1	10,289,987	9,885,369	96.1%
ENK27acRep2	34,868,010	31,613,836	90.7%
ENK27me3Rep1	11,131,576	10,256,807	92.1%
ENK27me3Rep2	28,435,049	25,806,462	90.8%
ENK4me1Rep1	29,070,786	26,338,406	90.6%
ENK4me1Rep2	23,139,610	22,376,410	96.7%
ENK4me3Rep1	15,960,227	14,359,079	90.0%
ENK4me3Rep2	25,423,585	23,909,154	94.0%
ENK9me3Rep1	28,486,276	25,936,935	91.1%
ENK9me3Rep2	42,803,514	38,115,104	89.0%
ENInputRep1	31,149,021	23,377,765	75.1%
ENInputRep2	20,116,865	18,585,106	92.4%

	Reads	Bases
Total EN genome	270,560,433	1.35E+10

Table B.2 Summary of Mapping Statistics for Elf1 naïve Cells
ChIP-seq DNA libraries were sequenced on Illumina platform, SE75 and mapped using Bowtie2.

Elf1 Transitioning Sample ID	Number of Reads Sequenced	Mapped reads	Mapping Efficiency
EPK27acRep1	16,542,791	15,466,596	93.5%
EPK27acRep2	19,045,610	14,495,788	76.1%
EPK27me3Rep1	14,661,174	12,484,843	85.2%
EPK27me3Rep2	17,360,465	10,303,499	59.4%
EPK4me1Rep1	13,466,280	12,643,726	93.9%
EPK4me1Rep2	14,628,581	13,524,908	92.5%
EPK4me3Rep1	14,142,256	12,502,355	88.4%
EPK4me3Rep2	12,112,327	9,124,628	75.3%
EPK9me3Rep1	37,350,962	32,818,042	87.9%
EPK9me3Rep2	42,244,794	36,811,767	87.1%
EPInputRep1	8,581,717	7,920,395	92.3%
EPInputRep2	37,772,483	35,291,325	93.4%

	Reads	Bases
Total EP genome	213,387,872	1.07E+10

Table B.3 Summary of Mapping Statistics for Elf1 transitioning cells
ChIP-seq DNA libraries were sequenced on Illumina platform, SE75 and mapped using Bowtie2

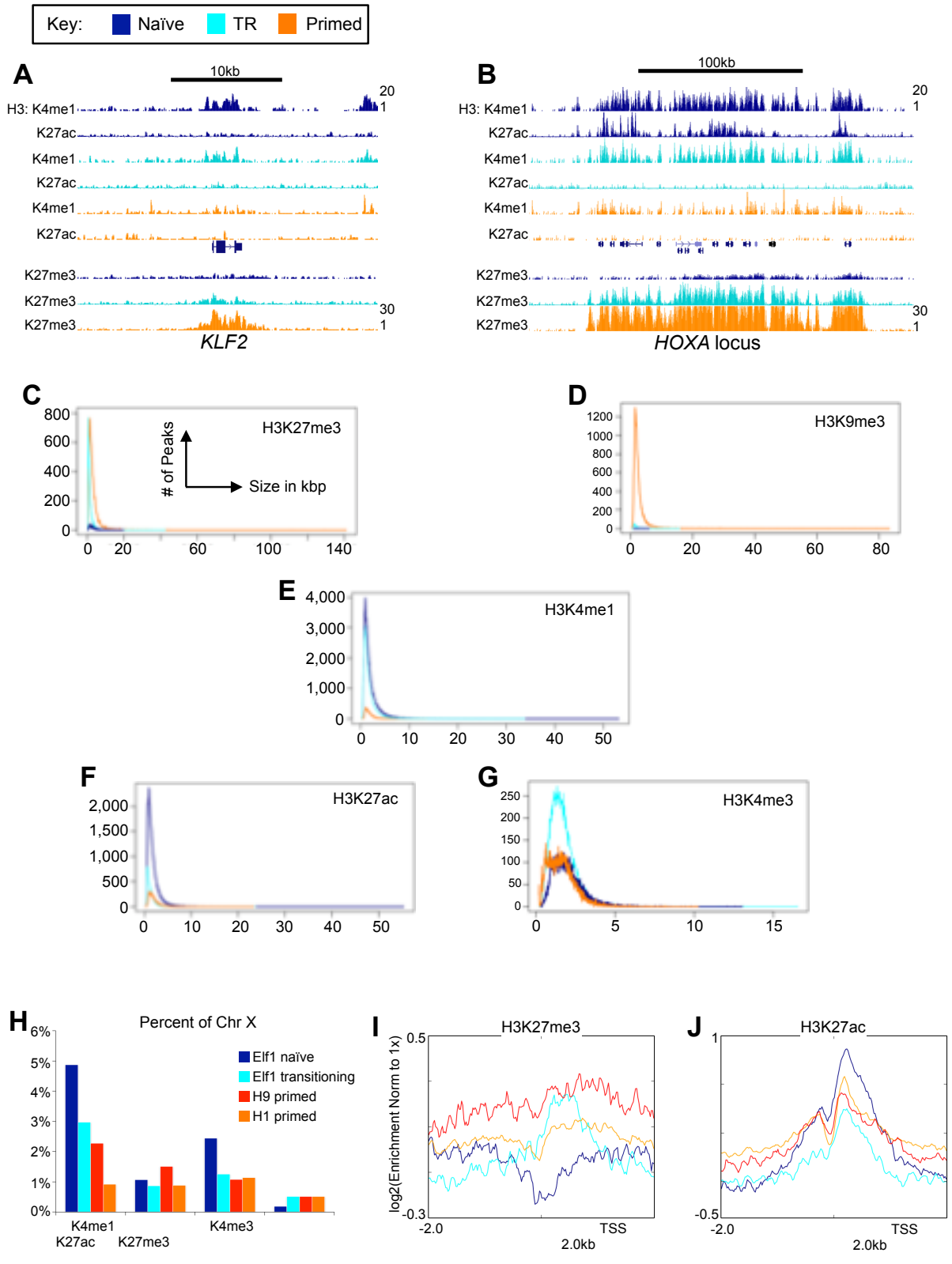


Figure B.4 Histone Modifications globally and Chromosome X
(A-B) UCSC Browser image of the KLF2 (A) and HOXA (B) loci. H3K4me1 and H3K27ac (top 4 tracks, RPKM scale 1-20) and H3K27me3 (bottom 3 tracks, RPKM scale 1-30) are shown for naïve, transitioning

and primed ChIP-seq data. (C-D) Histograms showing distribution of peak lengths for (C) H3K27me3, (D) H3K9me3, (E) H3K4me1, (F) H327ac and (G) H3K4me3. (H) Percent of chromosome X bases covered by histone modifications. (I-J) Average enrichment of ChIP-seq signal for H3K27me3 (I) and H3K27ac (J) at all gene TSS on chromosome X.

Number of Peaks

H3	Elf1 Naïve	Elf1 TR	H1 Primed
K4me1	116999	83940	46843
K4me3	21071	20541	22866
K27ac	67500	36701	24668
K27me3	4512	17379	10454
K9me3	12	2831	22911

Table B.4 Number of ChIP-Seq Peaks by Histone Modification
Numbers in bar chart for Figure 3.2B. Number of ChIP-Seq peaks called by MACS that pass FDR 5% for each cell type.

Percent of Genome

H3	Elf1 Naïve	Elf1 TR	H1 Primed
K4me1	9.25%	5.61%	3.04%
K4me3	1.58%	1.29%	1.41%
K27ac	4.49%	1.71%	1.81%
K27me3	0.51%	1.09%	1.39%
K9me3	0.00%	0.22%	2.35%

Table B.5 Percent of Genome Covered by Histone Modifications
Numbers in bar chart for Figure 3.2C. Percent of genome covered by each histone modification

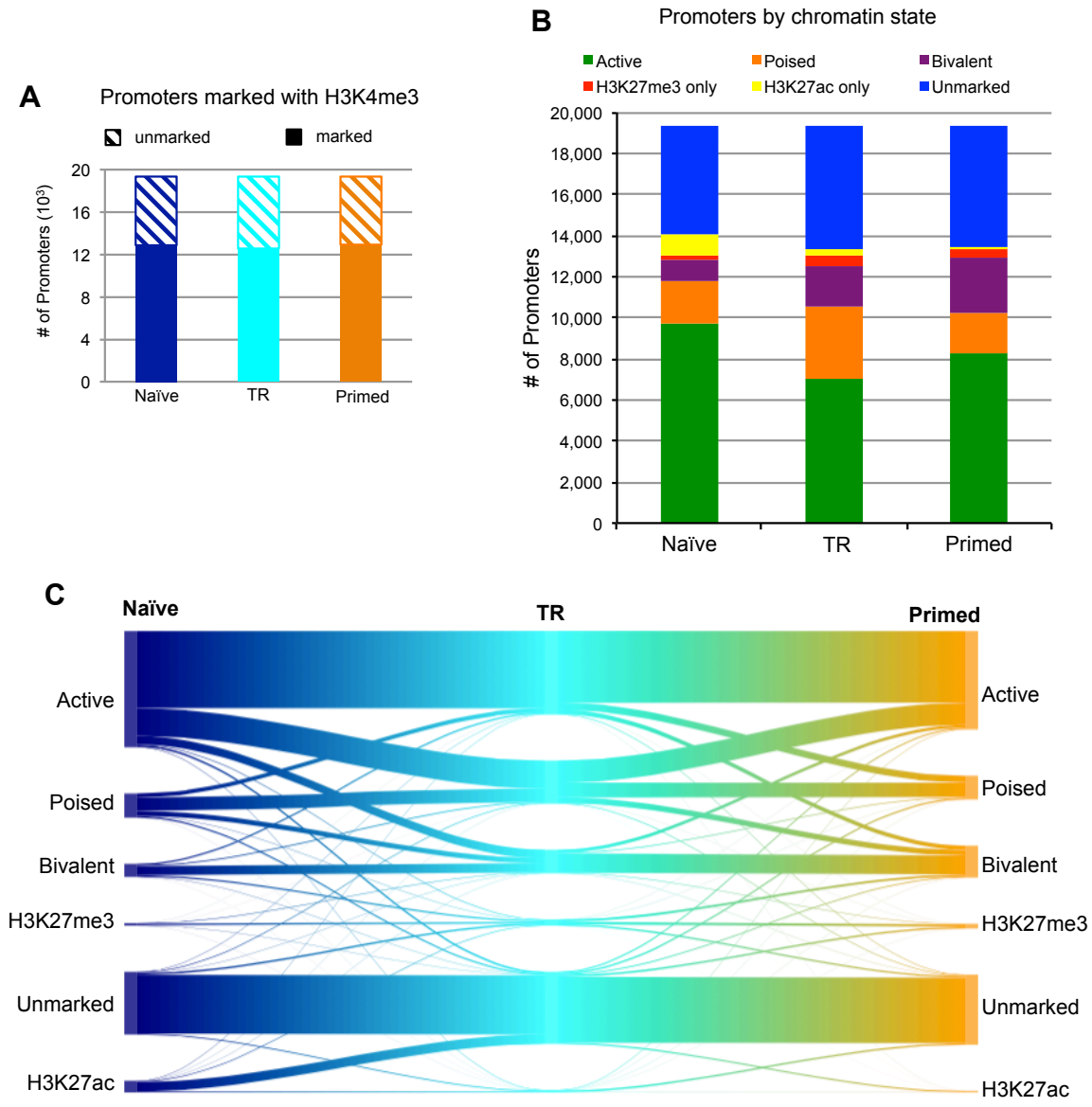
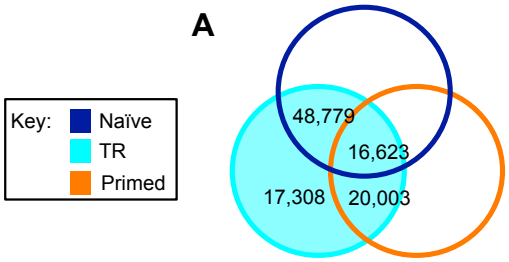
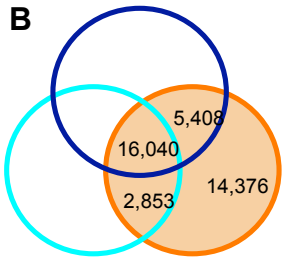


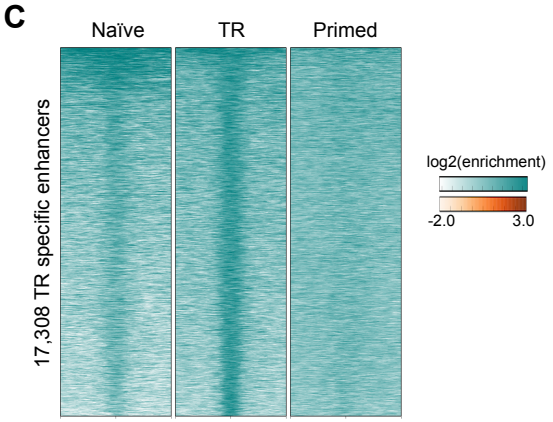
Figure B.5 Histone Modifications at Promoters and Bivalent Gene Ontology
(A) Number of GENCODE coding gene promoters marked with H3K4me3 in each hESC stage. (B) Breakdown of promoter chromatin state categories. Chart shows how many gene promoters are found in each category. (C) Sankey plot of all promoter state transitions between three hES cell types.

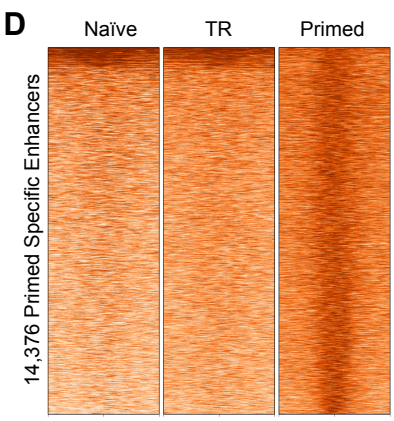
	Called Peaks	After Removing overlapping K4me3
Naïve Enhancers	116999	98020
TR Enhancers	83940	69467
Primed Enhancers	46843	38677

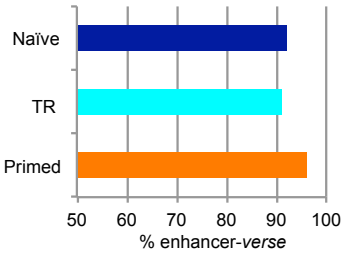
Table B.6 Number of Enhancers Genome-wide
Number of K4me1 peaks after excluding peaks that overlap K4me3.

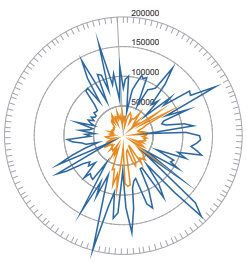


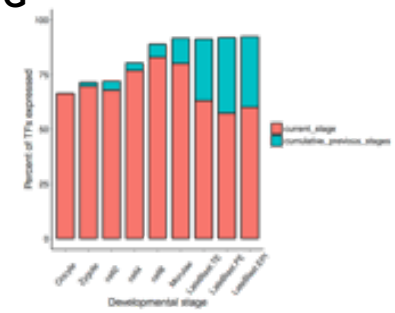








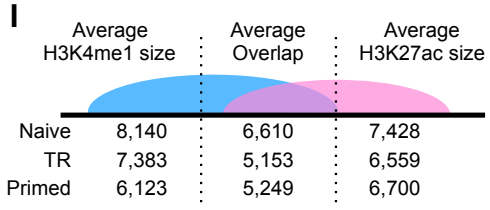




H

	H3K4me1				H3K27ac				H3K4me3
	< 5kb	≥ 5kb	≥10kb	≥20kb	< 5kb	≥ 5kb	≥10kb	≥20kb	≥ 5kb
	Short	**Broad**	**Grande**	**Venti**	**Short**	**Broad**	**Grande**	**Venti**	**Broad**
Naïve	109,587	5,999	1,256	157	65,170	2,036	274	20	283
TR	80,818	2,731	373	18	36,460	231	9	1	74
Primed	46,472	361	10	0	23,865	750	50	3	144

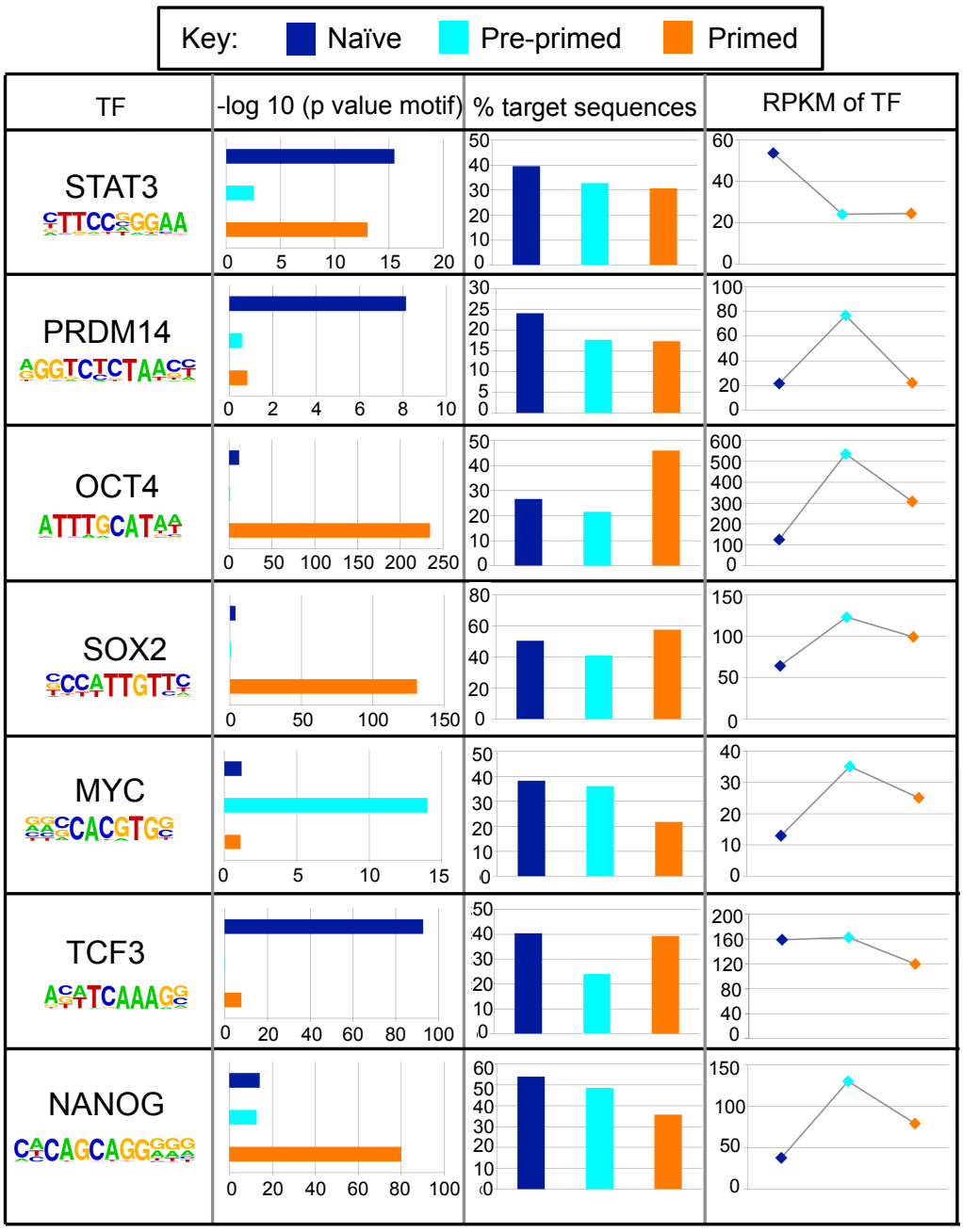
I



J

	Naïve	TR	Primed
Broad K4me1 w/ K27ac	87.72%	49.01%	71.16%
Broad K4me1 w/ Broad K27ac	23.46%	4.13%	22.64%
Broad K27ac w/ Broad K4me1	78.58%	52.70%	10.46%

Figure B.6 Characteristics of Enhancers in hESCs
(A-B) Venn diagrams of enhancer overlaps for transitioning (A) and primed (B) cells with other cell types. (C-D) Heatmaps of H3K4me1 ChIP-Seq signal at transitioning- (C) and primed- (D) specific enhancers. (E) Percent of hESC H3K4me1 genomic space (% bases or enhancer-verse) occupied by Roadmap H3K4me1 from 127 cell types. (F) Number of Roadmap H3K4me1 peaks from 127 cell types overlapping with naïve and primed H3K4me1 enhancers. (G) Percent of transcription factors, annotated from Animal TFDB, expressed at each stage of embryogenesis and cumulatively across stages using single cell RNA-Seq data from Yan *et al*.[27]. (H) Number of H3K4me1, H3K27ac and H3K4me3 peaks, broken down by size. (I) Average length of broad (≥ 5kb) H3K4me1 and H3K27ac domains and the average number of bases overlapped in shared regions. (J) Percent of broad peaks that overlap in these comparisons: broad H3K4me1 with any H3K27ac overlap, broad H3K4me1 with only broad H3K27ac overlap, or broad H3K27ac.

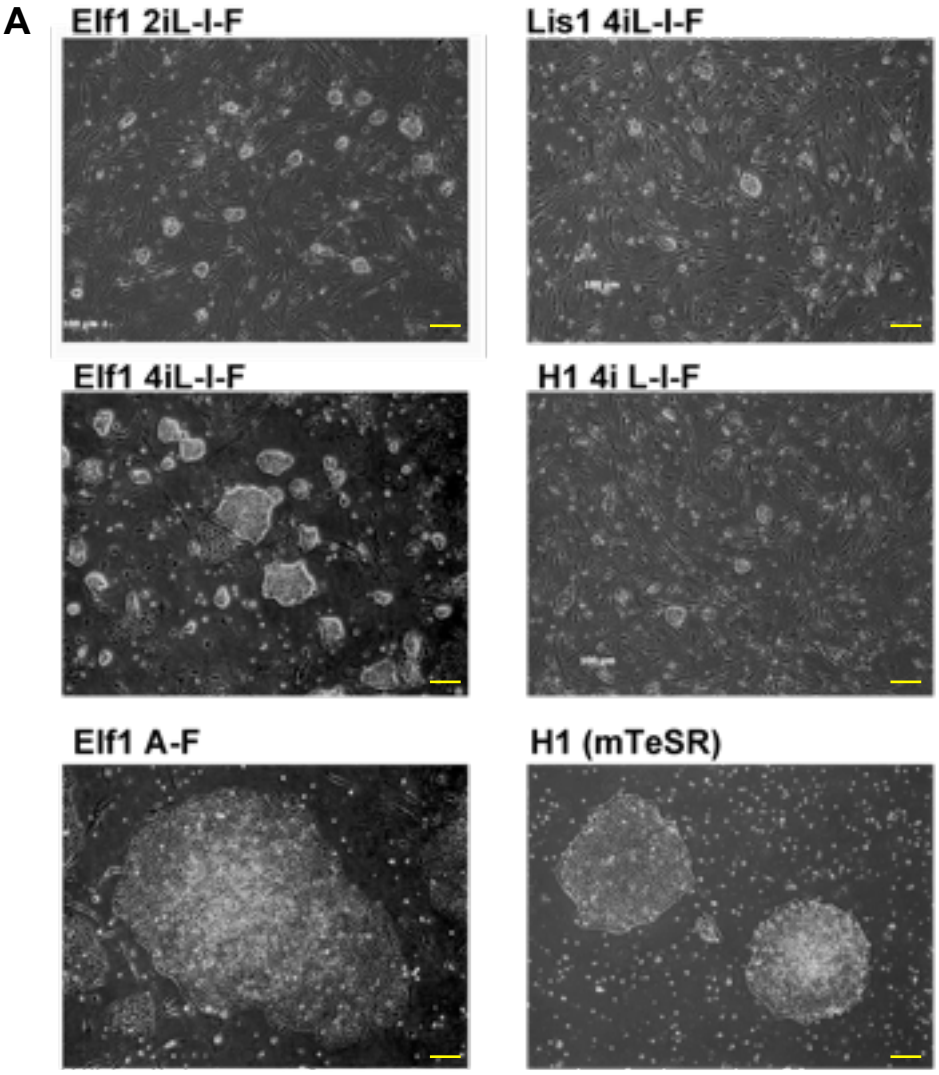


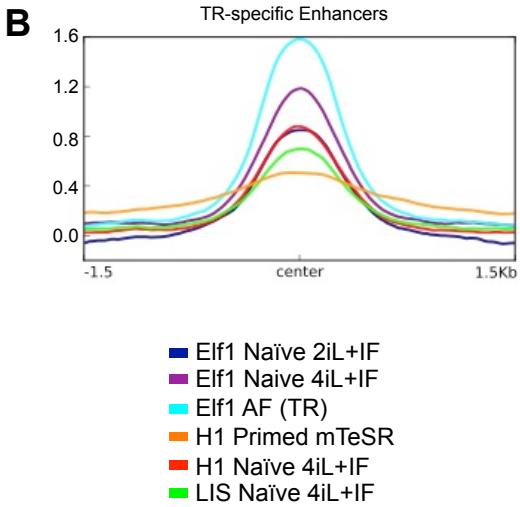
B.7 TF at Naive Active Enhancers
Examples of TF motifs from active enhancers showing significance of motif enrichment, percent of target sequences, and RPKM of respective gene.

Available as an Excel File

Table B.7 Enhancer overlaps with ENCODE DHS and Roadmap H3K4me1

Charts of the number or fraction of a cell type's peaks were found in naïve, transitioning or primed hESCs from 177 ENCODE DHS cell types or 127 Roadmap Epigenome Project cell types.


A
Elf1 2iL-I-F
Lis1 4iL-I-F
Elf1 4iL-I-F
H1 4i L-I-F
Elf1 A-F
H1 (mTeSR)


B
TR-specific Enhancers
1.6
1.2
0.8
0.4
0.0
-1.5
center
1.5Kb
Elf1 Naïve 2iL+IF
Elf1 Naive 4iL+IF
Elf1 AF (TR)
H1 Primed mTeSR
H1 Naïve 4iL+IF
LIS Naïve 4iL+IF

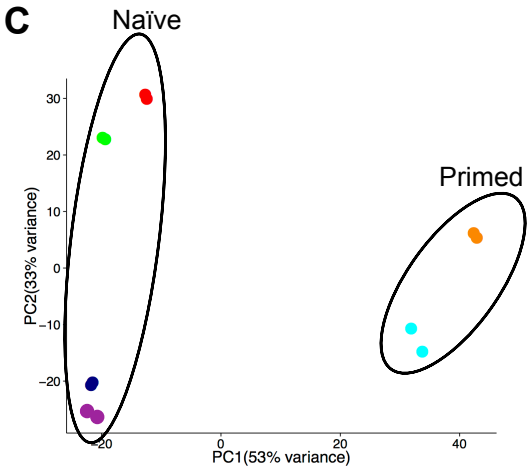

C
Naïve
Primed
PC2(33% variance)
PC1(53% variance)
30
20
10
0
−10
−20
−20
0
20
40

Figure B.8 RNA-Seq and ChIP-Seq of hESCs in Different Growth Conditions
(A) Representative images of cell lines in each growth condition used for this study. Yellow bar is 200um. (B) Average normalized ChIP-Seq signal in all naïve and primed hESCs at transitioning-specific enhancers. (C) PCA of autosomal gene expression, naïve cells cluster separately from primed.

A

Hi-C library	Raw read pairs	Processed read pairs	Cis interactions	Cis interactions within 20 kb
Elf1 Replicate 1	798,771,830	318,226,786	224,980,721	65,753,953
Elf1 Replicate 2	770,256,147	324,524,006	221,137,452	62,491,765
H1 Replicate 1	237,662,270	35,822,152	19,713,347	3,369,816
H1 Replicate 2	496,522,946	160,344,628	124,240,997	20,935,372

B

Feature	Naïve	Primed
TAD	6,119	5,822
Boundary	5,487	5,279
Boundary > 40 kb	433	351

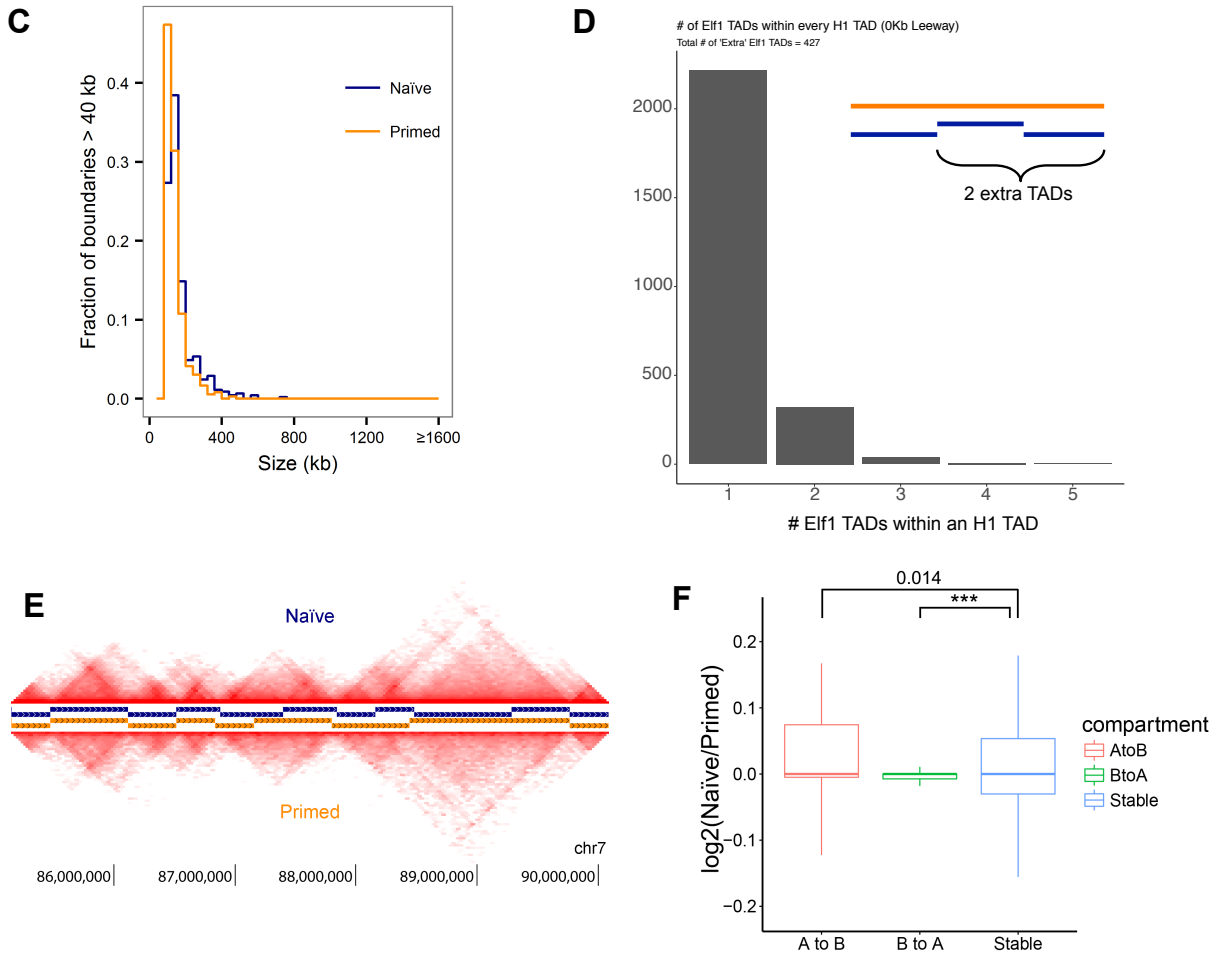


Figure B.9 HiC libraries and TAD structure
(A) Summary of Hi-C sequencing read pairs and interactions in naïve and primed hESCs. (B) Total counts of TADs and boundaries identified in naïve and primed hESCs after discarding X and Y chromosome interactions. (C) Global size distributions of boundaries > 40 kb. (D) Bar chart of the number of Elf1 TADs within an H1 TAD and depiction of how “extra” TADs are defined. (E) Contact heatmaps of a region in chromosome 7. Navy and dark orange tracks denote TADs in naïve and primed hESCs, respectively. (F) Boxplot of gene expression (RPKM) overlapping A to B compartment switches. “A to B” and “B to A” are naïve to primed directions. Stable are compartments that do not switch. P-values are computed using two-sample t-test with one sided alternative. *** P-value < 1.34 x 10-10.

Appendix C

Supplemental material to Chapter 4

RNA-Seq	***Raw reads***	***Number of transcripts TPM>0 (TPM>1)***
*Ovc316 CD133+ rep 1	32,303,391	56,348 (48,904)
Ovc316 CD133+ rep 2	29,264,858	44,505 (34,752)
*Ovc316 CD133- rep 1	49,751,671	35,630 (24,848)
Ovc316 CD133- rep 2	33,171,574	23,878 (19,215)
Ovcar CD133+ rep 1	26,593,321	33,631 (27,091)
Ovcar CD133+ rep 2	24,034,854	37,501 (28,861)
Ovcar CD133- rep 1	36,055,165	33,003 (26,298)
Ovcar CD133- rep2	30,198,642	33,765 (26,687)
*HOSE rep 1	53,423,293	42,972 (24,811)
*HOSE rep 2	34,457,038	39,356 (24,182)
Methylomes		***Mapped Reads***
§Ovc316 CD133+	153,824,798	118,807,258
§Ovc316 CD133-	156,050,202	123,934,205
Ovcar CD133+ rep 1	47,174,414	27,219,637
Ovcar CD133+ rep 2	164,576,081	102,201,746
Ovcar CD133- rep 1	112,471,907	92,002,020
Ovcar CD133- rep 2	162,574,588	137,212,952

ChIP-Seq		***Mapped Reads***
Ovc316 CD133+ H3K4me1	11,155,646	10,143,242
Ovc316 CD133+ Input	15,573,052	14,076,294
Ovc316 CD133- H3K4me1	10,572,967	9,553,500
Ovc316 CD133- Input	8,966,258	8,078,046

Table C.1 OvCSC Sequencing Statistics
Sample with * were sequenced SE50. Samples with § were sequenced PE100.

Term	Overlap	P-value	Adjusted P-value	Z-score	Combined Score	Genes
OvCSC Upregulated Gene Categories						
leukocyte cell-cell adhesion (GO:0007159)	5/23	8.1766 1E-05	0.059280442	-2.8944	8.1781	VCAM1;SYK;ITGA4;CCL5;ITGAL
low-density lipoprotein particle remodeling (GO:0034374)	3/11	0.0011 96476	0.244263282	-3.5406	4.9905	LPA;APOB;ABCG1
cellular response to calcium ion (GO:0071277)	5/39	0.0010 73259	0.244263282	-2.7380	3.8592	ACER1;RASA4;CPNE7;GPLD1;RYR3
hematopoietic progenitor cell differentiation (GO:0002244)	3/12	0.0015 71499	0.244263282	-2.6082	3.6763	PTPRC;PLEK;MIXL1
potassium ion transport (GO:0006813)	6/63	0.0016 84574	0.244263282	-2.4334	3.4299	KCNV1;KCNA1;KCNJ15;KCNQ3;KCNN1;ABCC9
negative regulation of interleukin-6 production (GO:0032715)	3/16	0.0037 67312	0.314915918	-2.4849	2.8712	NCKAP1L;IRAK3;TLR4
lipopolysaccharide-mediated signaling pathway (GO:0031663)	3/16	0.0037 67312	0.314915918	-2.4651	2.8483	HCK;CCL5;TLR4
inflammatory response (GO:0006954)	9/209	0.0266 17512	0.446191568	-3.4205	2.7604	CCL22;SYK;PLA2G4B;CCL5;NLRP3;CYBB;TNFRSF25;ITGAL;TLR4
xenobiotic metabolic process (GO:0006805)	6/85	0.0074 95378	0.386291392	-2.5056	2.3832	SULT1B1;GSTO2;AKR1C1;CYP46A1;GGT1;ACSM2B
transmembrane transport (GO:0055085)	7/140	0.0237 6937	0.446191568	-2.6250	2.1184	ABCC2;SLC47A1;ABCA4;ABCC6;ABCC9;ABCG1;CFTR
positive regulation of T cell proliferation (GO:0042102)	4/36	0.0057 98748	0.382190191	-2.1058	2.0254	VCAM1;PTPRC;CCL5;NCKAP1L
neutrophil chemotaxis (GO:0030593)	4/51	0.0194 39486	0.446191568	-2.1344	1.7224	CCL22;SYK;CCL5;NCKAP1L
defense response to bacterium (GO:0042742)	6/119	0.0338 51091	0.446191568	-2.1224	1.7128	WFDC12;HTN3;SYK;LPO;EPPIN;TLR4
negative regulation of tumor necrosis factor production	3/27	0.0167 07103	0.432594626	-1.9142	1.6040	IRAK3;TLR4;HAVCR2

(GO:0032720)						
positive regulation of vasculogenesis (GO:2001214)	2/8	0.010514698	0.432320158	-1.8924	1.5869	KDR;ASB4
regulation of membrane potential (GO:0042391)	4/59	0.031260543	0.446191568	-1.8946	1.5290	SLC26A1;KCNA1;CACNA1H;CNGB1
bile acid biosynthetic process (GO:0006699)	3/26	0.015072131	0.432594626	-1.8108	1.5174	OSBPL3;CYP46A1;SLC27A5
regulation of cell shape (GO:0008360)	5/96	0.045139828	0.446191568	-1.8604	1.5013	FGD3;HCK;EPB41;KDR;SEMA3E
glycerol metabolic process (GO:0006071)	2/7	0.007992236	0.386291392	-1.5103	1.4366	GK5;GK
protein autophosphorylation (GO:0046777)	6/145	0.074175732	0.446191568	-1.7643	1.4238	HCK;STK33;GRK7;KDR;ATM;MAPK15
retinoid metabolic process (GO:0001523)	4/59	0.031260543	0.446191568	-1.6461	1.3284	AKR1C1;ABCA4;BCO2;APOB
positive regulation of actin filament polymerization (GO:0030838)	3/34	0.030830672	0.446191568	-1.6408	1.3242	HCK;TENM1;NCKAP1L
embryonic hemopoiesis (GO:0035162)	2/10	0.016453591	0.432594626	-1.5748	1.3196	KDR;FLT3LG
response to virus (GO:0009615)	5/100	0.052163932	0.446191568	-1.6120	1.3009	CCL22;CCL5;IRAK3;APOB;TRIM22
regulation of ERK1 and ERK2 cascade (GO:0070372)	3/23	0.010733466	0.432320158	-1.5150	1.2705	PKHD1;SYK;ROS1
CD133- Bulk Tumor Upregulated Gene Categories						
negative regulation of transcription from RNA polymerase II promoter (GO:0000122)	36/473	1.4157 2E-17	1.31237E-14	-5.4095	172.9107	FOXC1;NFIX;CITED2;UBE2D3;NEDD4L;TWIST1;TNF;ETS2;ZFP36;NFIL3;UBB;UBC;EPC1;HES1;JUNB;HEXIM1;KLF10;TFAP2A;OSR2;JUND;CBX4;ZNF281;ARID5B;NR1D1;KLF4;SIRT1;NFKB1;SMAD7;EFNA1;PER1;DDIT3;BHLHE40;SNAI1;FOSB;MXD1;ATF3
positive regulation of transcription from RNA polymerase II	36/712	3.7397 4E-12	1.15558E-09	-6.4287	132.2936	ARF4;CSRNP1;FOXC1;NFIX;DDX3X;CITED2;TWIST1;AHR;T

promoter (GO:0045944)						NF;ARNTL;UBB;UBC;EPC1;HES1;SOX9;CCNL1;JUNB;IER5;TFAP2A;KDM6B;EGR1;OSR2;JUN;JUND;FOS;KLF4;SIRT1;NFKB1;SMAD7;PER1;NR4A1;KLF6;KLF5;DDIT3;NCOA7;ATF3
negative regulation of transcription, DNA-templated (GO:0045892)	26/408	3.2182E-11	7.45819E-09	-5.0475	94.4586	CITED2;TWIST1;AHR;TNF;DEDD2;ARNTL;CCDC85B;ZNF703;EPC1;HES1;SOX9;HEXIM1;KLF10;TFAP2A;JUN;CBX4;ZNF281;ARID5B;NR1D1;KLF4;SIRT1;PER1;ELF3;DDIT3;BHLHE40;ID4
negative regulation of apoptotic process (GO:0043066)	22/323	3.22799E-10	4.98724E-08	-4.7217	79.3890	TFAP2A;ARF4;DDX3X;CBX4;CITED2;PLK2;HSPB1;HIGD1A;SIRT1;NFKB1;TNFRSF10D;DNAJA1;UBB;BAG3;UBC;PPIF;SPRY2;SOX9;CRYAB;HSPA1B;TPT1;HSPA1A
positive regulation of transcription, DNA-templated (GO:0045893)	25/464	2.56903E-09	3.40213E-07	-5.0969	75.9111	FOXC1;CITED2;AHR;TNF;ETS2;ARNTL;EPC1;SOX9;TFAP2A;EGR1;OSR2;JUN;ZNF281;FZD7;NR1D1;FOS;KLF4;NFKB1;KLF6;ELF3;ID2;DDIT3;SNAI1;CKS2;CREB5
regulation of cell death (GO:0010941)	7/9	2.15183E-12	9.97375E-10	-2.7233	56.4433	JUN;JUND;CRYAB;JUNB;HSPA1B;HSPA1A;IER3
response to unfolded protein (GO:0006986)	9/29	7.89064E-11	1.46292E-08	-2.2994	41.4825	DNAJA1;DNAJB1;HSPH1;DDIT3;DNAJB4;HSPA6;HSPB1;HSPA2;HSPA1A
positive regulation of apoptotic process (GO:0043065)	15/231	4.23835E-07	3.92895E-05	-3.8555	39.1128	MOAP1;JUN;DDX3X;GADD45B;GADD45A;IGFBP3;SIRT1;TNF;IFIT2;RHOB;DNAJA1;BCL2L11;UBB;UBC;PMAIP1
response to lipopolysaccharide (GO:0032496)	10/78	7.21437E-08	7.4308E-06	-2.6139	30.8704	JUN;JUND;SLPI;CXCL1;PPBP;CXCL3;TRIB1;CXCL2;JUNB;TNFRSF10D

regulation of cell proliferation (GO:0042127)	11/124	7.2458 3E-07	6.10626E-05	-3.0562	29.6560	JUN;JUND;PPBP;SOX9;SGK1;CXCL3;JUNB;IER5;SIRT1;CXCL2;TNFRSF10D
negative regulation of cell proliferation (GO:0008285)	14/276	1.7938 2E-05	0.001004286	-3.9822	27.4908	KLF10;TFAP2A;BTG2;BTG1;IGFBP3;CXCL1;KLF4;TOB1;RNF139;SST;ADAMTS1;HSPA1B;HIST1H2AC;HSPA1A
neutrophil degranulation (GO:0043312)	18/479	6.6111 8E-05	0.002451426	-4.5175	27.1552	GSN;DDX3X;HSPA6;CXCL1;PPBP;TUBB4B;NFKB1;RAP1B;SLPI;TCN1;CEACAM6;NEU1;S100P;PTX3;YPEL5;LAMTOR3;HSPA1B;HSPA1A
positive regulation of gene expression (GO:0010628)	12/214	2.6938 4E-05	0.001248597	-3.4493	23.0610	TFAP2A;OSR2;GSN;DDX3X;CITED2;ID2;TWIST1;SPRY2;KLF4;TNF;HSPA1B;HSPA1A
negative regulation of inclusion body assembly (GO:0090084)	5/9	4.4830 4E-08	5.19472E-06	-1.8411	22.4021	DNAJB1;DNAJB6;HSPA2;HSPA1B;HSPA1A
protein ubiquitination (GO:0016567)	12/224	4.2124 1E-05	0.001740051	-3.2525	20.6661	MYLIP;RNF139;UBB;RNF38;UBE2D3;UBC;UBE2N;NEDD4L;FBXO32;SIRT1;CBLL1;HIST1H2BD
negative regulation of cell growth (GO:0030308)	9/102	7.8487 2E-06	0.000485051	-2.5986	19.8302	PPP2CA;DDX3X;BTG1;CCDC85B;SERTAD2;SIRT1;ING1;HSPA1B;HSPA1A
circadian regulation of gene expression (GO:0032922)	7/51	4.3257 5E-06	0.000334164	-2.3605	18.8931	PER1;ID2;BHLHE40;AHR;NR1D1;SIRT1;ARNTL
regulation of cellular response to heat (GO:1900034)	8/76	6.8097 1E-06	0.0004509	-2.4503	18.8778	DNAJB1;HSPH1;DNAJB6;BAG3;SIRT1;CRYAB;HSPA1B;HSPA1A
negative regulation of protein ubiquitination (GO:0031397)	6/34	4.7180 4E-06	0.000336433	-2.3206	18.5579	DNAJA1;TNFAIP3;ISG15;HSPA1B;SMAD7;HSPA1A
negative regulation of gene expression (GO:0010629)	8/97	4.1202 E-05	0.001740051	-2.6103	16.5852	TIPARP;CITED2;ZNF281;ID2;KLF4;TNF;SIRT1;NFKB1
extrinsic apoptotic signaling pathway via death domain receptors (GO:0008625)	5/26	1.9500 7E-05	0.001004286	-2.2421	15.4780	MOAP1;DDX3X;BAG3;TNF;DEDD2
apoptotic process	10/195	0.0002	0.006531765	-	14.5880	PPP2CA;PPP1R15A

(GO:0006915)		60707		2.8996		;CSRNP1;BCL2L11;RRAGC;GADD45A;PMAIP1;AHR;IER3;RHOB
inflammatory response (GO:0006954)	10/209	0.000451923	0.008549638	-2.9947	14.2602	ELF3;CXCL1;PTX3;PPBP;FOS;CXCL3;TNF;CXCL2;NFKB1;TNFRSF10D
cellular response to heat (GO:0034605)	5/26	1.95007E-05	0.001004286	-2.0548	14.1853	BAG3;HSPA6;IER5;HSPA1B;HSPA1A
positive regulation of NF-kappaB transcription factor activity (GO:0051092)	8/120	0.000185033	0.005717507	-2.5934	13.3928	UBB;UBC;RIPK4;UBE2N;TNF;HSPA1B;NFKB1;HSPA1A

Table C.2 Top 25 GO Terms from CSC and CD133- gene expression Comparison
Significance scores from Enrichr.

Term	Overlap	P-value	Adjusted P-value	Z-score	Combined Score	Genes
Putative upregulated OvCSC gene targets of OvCSC HypoDMRs						
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	23/712	8.56336E-07	0.000366756	-6.462	51.1229	DLX1;TFAP2C;GSX1;ONECUT2;BCL11B;FOXF1;WNT3A;PAX6;ASCL1;PKD2;ISL1;HOXC11;GREM1;PLSCR1;SHH;SALL1;TAL1;THRAP3;PLAGL1;HNF4A;IRF8;NEUROG1;MIXL1
positive regulation of cell proliferation (GO:0008284)	17/326	3.5612E-08	3.05551E-05	-4.845	50.3735	EGR4;EPO;WNT3A;IRS1;NTRK3;FLT4;IGF2;POU3F3;FGF3;TBX3;INS;FGF5;GREM1;GHRHR;SHH;PTK2B;FGFR4
negative regulation of transcription from RNA polymerase II promoter	17/473	6.26387E-06	0.000895734	-5.264	36.9399	DLX1;URI1;TFAP2C;HDAC10;EPO;PTCH1;ANKRD33;ASCL1;IS

(GO:0000122)						L1;TBX3;NR2F6;NCOR2;SHH;SALL1;NSD1;ZNF536;IRF8
signal transduction (GO:0007165)	22/861	5.8098 4E-05	0.004493377	-6.350	34.3243	LINGO1;LRSAM1;BCL11B;EPO;IRS1;KCNIP2;SPHK1;RASGRF1;LHB;CDC42BPB;SMAD5;FGF3;NR2F6;BCR;GREM1;SALL3;SALL4;PTK2B;ZNF536;TNFRSF25;INHA;BTRC
positive regulation of transcription, DNA-templated (GO:0045893)	15/464	7.4640 2E-05	0.004923383	-4.926	26.1750	EPO;FOXF1;WNT3A;PAX6;SMAD5;POU3F3;RAI1;SHH;SALL1;TAL1;THRAP3;NSD1;HNF4A;SPP1;BTRC
regulation of cell proliferation (GO:0042127)	9/124	4.5807 5E-06	0.000895734	-3.212	22.5384	ZAP70;TFAP2C;SHH;GRAP2;PTK6;TNFRSF25;TNFRSF1B;INHA;PKD2
heart development (GO:0007507)	8/79	1.2823 6E-06	0.000366756	-2.739	21.6670	SHH;SALL1;FOXF1;NTRK3;NF1;PKD2;ZFPM1;MIXL1
regulation of heart contraction (GO:0008016)	5/26	5.4039 6E-06	0.000895734	-2.565	17.9976	KCNIP2;KCNQ1;TNNT2;FXYD1;TPM1
neurogenesis (GO:0022008)	5/29	9.5198 5E-06	0.001056766	-2.289	15.6866	SALL1;BCL11B;SALL3;SALL4;ASCL1
renal system development (GO:0072001)	3/10	0.0001 13922	0.005144469	-2.651	13.9729	SHH;FOXF1;PTCH1
MAPK cascade (GO:0000165)	9/229	0.0005 32569	0.016923854	-3.270	13.3400	FGF5;SPTBN4;IRS1;RASGRF1;NF1;GFRA1;FGFR4;FGF3;INS
negative regulation of apoptotic process (GO:0043066)	10/323	0.0016 45943	0.029836656	-3.628	12.7399	GREM1;SHH;SPHK1;FLT4;BNIP3;PTK2B;ASCL1;MPO;POU3F3;TBX3
muscle filament sliding (GO:0030049)	5/38	3.7389 6E-05	0.003564477	-2.197	12.3842	ACTA1;MYBPC2;TNNT2;TNNT3;TPM1
transmembrane receptor protein tyrosine kinase	6/72	8.6073 1E-05	0.004923383	-2.226	11.8302	ZAP70;GRAP2;NTRK3;FLT4;PTK6;MTSS1

signaling pathway (GO:0007169)						
embryonic hemopoiesis (GO:0035162)	3/10	0.000113922	0.005144469	-2.143	11.2924	TPO;TAL1;ZFPM1
sprouting angiogenesis (GO:0002040)	4/22	6.28444E-05	0.004493377	-1.953	10.5556	FLT4;RSPO3;PTK2B;SEMA3E
sympathetic nervous system development (GO:0048485)	3/10	0.000113922	0.005144469	-1.957	10.3148	SEMA3A;NF1;ASCL1
positive regulation of gene expression (GO:0010628)	8/214	0.001485813	0.028329504	-2.866	10.2132	ACTA1;PLSCR1;WNT3A;NTRK3;PAX6;FGFR4;POU3F3;INS
glucose homeostasis (GO:0042593)	5/58	0.000290091	0.010370739	-2.211	10.1031	IRS1;HNF4A;PAX6;FGFR4;INS
positive regulation of apoptotic process (GO:0043065)	8/231	0.002395232	0.036054543	-2.723	9.0490	ITGB1;OBSCN;PLEKHG5;BNIP3;KCNMA1;NF1;SCRIB;BOK
sarcomere organization (GO:0045214)	4/28	0.000167797	0.006855686	-1.760	8.7671	MYBPC2;OBSCN;TPM1;OBSL1
erythrocyte differentiation (GO:0030218)	4/27	0.00014496	0.006218797	-1.627	8.2632	TAL1;EPO;INHA;ZFPM1
positive regulation of glycogen biosynthetic process (GO:0045725)	3/12	0.000205777	0.007676371	-1.689	8.2261	IRS1;IGF2;INS
response to hypoxia (GO:0001666)	5/67	0.000568402	0.017417459	-1.948	7.8886	RYR1;BNIP3;KCNMA1;NF1;CBFA2T3
wound healing (GO:0042060)	4/36	0.000453362	0.014960955	-1.874	7.8741	SCARB1;TPM1;NF1;INS
Putative upregulated CD133-cell gene targets of OvCSC HyperDMRs						
positive regulation of transcription from RNA polymerase II promoter (GO:0045944)	16/712	4.47549E-09	1.24866E-06	-6.462	87.8464	TFAP2A;RARG;FOXC1;JAG1;TNF;BARX2;MEOX1;MED6;FOSL2;GLI2;RUNX1;VEGFA;IL6;RGCC;SOX17;MYC
positive regulation of transcription, DNA-templated (GO:0045893)	14/464	1.14113E-09	6.36752E-07	-5.268	75.1556	TFAP2A;FOXC1;USP22;GATA6;NFATC2;CTCF;TNF;GLI2;RUNX1;IL6;SOX17;MYC;HIVEP3;MAPRE3
negative regulation of	12/473	1.3251	2.46476E-05	-5.351	56.7804	TFAP2A;RARG;

transcription from RNA polymerase II promoter (GO:0000122)		4E-07				FOXC1;TGIF2;MYC;GATA6;NEDD4L;CTCF;BCOR;TNF;GLI2;VEGFA
positive regulation of cell proliferation (GO:0008284)	9/326	2.8253 4E-06	0.00022522	-4.681	39.3168	IL6;EDN2;MYC;CUL3;ZNF703;PDGFA;S1PR3;IL6ST;VEGFA
positive regulation of gene expression (GO:0010628)	7/214	1.3036 1E-05	0.000808241	-3.706	26.3894	TFAP2A;IL6;RGCC;MYC;CTCF;TNF;VEGFA
kidney development (GO:0001822)	5/42	4.3163 4E-07	6.02129E-05	-2.648	25.7315	TFAP2A;FOXC1;SULF2;VEGFA;GLI2
cellular response to hypoxia (GO:0071456)	5/55	1.7013 1E-06	0.000158222	-2.599	22.7475	RGCC;MYC;GATA6;NDRG1;VEGFA
negative regulation of apoptotic process (GO:0043066)	7/323	0.0001 75619	0.006959258	-4.530	22.5045	TFAP2A;IL6;MYC;GATA6;ASNS;IL6ST;VEGFA
positive regulation of p38MAPK cascade (GO:1900745)	3/15	2.0568 9E-05	0.001147742	-2.837	19.2092	GADD45A;GADD45G;VEGFA
activation of MAPKKK activity (GO:0000185)	3/7	1.6158 1E-06	0.000158222	-1.915	16.7599	GADD45A;TNF;GADD45G
positive regulation of mesenchymal cell proliferation (GO:0002053)	3/10	5.4964 4E-06	0.000383377	-1.942	15.2771	MYC;PDGFA;VEGFA
angiogenesis (GO:0001525)	4/93	0.0003 73143	0.008675585	-2.726	12.9409	JAG1;SOX17;PDGFA;VEGFA
membrane organization (GO:0061024)	5/223	0.0013 37269	0.016230229	-3.076	12.6751	SGCE;SNX9;ACTB;RAB11A;VAMP3
vascular endothelial growth factor receptor signaling pathway (GO:0048010)	4/68	0.0001 11481	0.005655112	-2.378	12.3053	PIK3CA;ROCK1;ACTB;VEGFA
heart development (GO:0007507)	4/79	0.0001 99549	0.006959258	-2.454	12.1913	FOXC1;TAB2;BCOR;GLI2
positive regulation of apoptotic process (GO:0043065)	5/231	0.0015 61985	0.01743175	-2.892	11.7110	RARG;IL6;GADD45A;TNF;GADD45G
negative regulation of transcription, DNA-templated (GO:0045892)	6/408	0.0037 42598	0.034806157	-3.477	11.6772	TFAP2A;ZNF703;GATA6;CTCF;BCOR;TNF
MAPK cascade (GO:0000165)	5/229	0.0015 03377	0.017372115	-2.803	11.3622	SHC3;MYC;CUL3;PDGFA;TNF
positive regulation of epithelial to mesenchymal transition (GO:0010718)	3/34	0.0002 57372	0.007121204	-2.149	10.6266	FOXC1;RGCC;ZNF703

cell cycle arrest (GO:0007050)	4/111	0.000728521	0.011292073	-2.242	10.0503	GADD45A;RHEB;MYC;CUL3
positive regulation of MAP kinase activity (GO:0043406)	3/41	0.000450165	0.00927251	-2.100	9.8274	PDGFA;TNF;VEGFA
hemopoiesis (GO:0030097)	3/35	0.000280764	0.007121204	-1.979	9.7834	CD164;JAG1;RUNX1
axon guidance (GO:0007411)	4/125	0.001134529	0.015028609	-2.201	9.2411	SHC3;PIK3CA;GRB10;GLI2
regulation of transcription from RNA polymerase II promoter (GO:0006357)	5/301	0.004871318	0.040570084	-2.831	9.0732	TFAP2A;MYC;MED6;FOSL2;VEGFA
positive regulation of tyrosine phosphorylation of STAT protein (GO:0042531)	3/45	0.000593001	0.010674015	-1.862	8.4524	IL6;IL6ST;VEGFA

Table C.3 Top 25 GO Terms of Putative Gene Targets of CSC Hyper- and HypoDMRs
Significance scores from Enrichr.

Vita

Stephanie Lauren Battle began her educational journey at the University of Maryland – Baltimore County (UMBC), where she studied Biochemistry and Molecular Biology, was she was a MARC U*STAR Scholar and Meyerhoff Scholar. She graduated in 2010 with *cum laude* honors. During her undergraduate years, she studied defense related genes in the model plant *Arabidopsis thaliana* in the lab of Dr. Hua Lu. Her research excellence earned her numerous awards, including the Excellence in Biochemistry Award and 1st place in poster presentation at the UMBC Undergraduate Research Symposium. Between 2010 and 2011, Stephanie worked at the National Institutes of Health (NIH) as part of the NIH Academy, which emphasized the study of health disparities in the United States. While at the NIH, she worked in the lab of Dr. David Bodine and studied the affects of the chemotherapy drug 5-Azacytidine on global DNA demethylation in CMML patients.

In September 2011 she began her Ph.D. work in the Department of Genome Sciences at the University of Washington. During this time she worked on epigenomic characterization of stem cells, with particular emphasis placed on DNA methylation, DNA hydroxymethylation and histone modifications. This work advanced how we understood pluripotency, stem cell behavior, and cellular differentiation.

While in graduate school, Stephanie presented her research at numerous conferences around the world including China, Canada, and Spain, as well as many conferences across the United States. Her research accomplishments and exceptional work ethic earned her the Ford Fellowship and support from multiple traning grants. Stephanie has ben a co-author on many publications, is publishing her own first authorship work, and has written a book chapter on embryonic stem cell epigenetics. Stephanie defended her Ph.D. thesis on November 17, 2017.

Dr. Stephanie Lauren Battle, is a highly motivated scientist who does not shy away from a challenge. As she moves forward in her career, she will continue to push the boundaries of scientific knowledge.