The dynamics of bivalent chromatin during development in mammals

Anna Mantsoki

Thesis presented for the degree of Doctor of Philosophy

University of Edinburgh

November 2016

Declaration

I declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated below. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

The work presented in Chapter 2 was previously published in the book with title IWBBIO 2015: Bioinformatics and Biomedical Engineering as 'Comparative Analysis of Bivalent Domains in Mammalian Embryonic Stem Cells' by Anna Mantsoki (student, thesis author) and Anagha Joshi (supervisor) under DOI:10.1007/978-3-319-16483-0_39. This study was conceived by all of the authors. I carried out the analysis, wrote the manuscript and participated in the design of this study. Chapter 2 is copyright (c) 2016 of Springer (reproduced here under license number: 3954271432719) and may not be reproduced without written agreement from the copyright holder.

The work presented in Chapter 3 was previously published in Scientific Reports as 'CpG island erosion, Polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells' by Anna Mantsoki, Guillaume Devailly and Anagha Joshi under DOI:10.1038/srep16791. This study was conceived by Anagha Joshi and Anna Mantsoki. I carried out the analysis, wrote the manuscript and participated in the design of this study. Chapter 3 is copyright (c) 2016 of Macmillan Publishers Limited reproduced here under the terms of the associated CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

The work presented in Chapter 4 was previously published in Computational Biology and Chemistry as 'Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data' by Anna Mantsoki, Guillaume Devailly and Anagha Joshi under DOI: 10.1016/j.compbiolchem.2016.02.004. This study was conceived by Anagha Joshi. I carried out the analysis, wrote the manuscript and participated in the design of this study. Chapter 4 is copyright (c) 2016 of Elsevier reproduced here under the terms of the associated CC-BY 4.0 license.

Anna Mantsoki

November 2016

Acknowledgements

If it wasn’t for the many people that have helped me along the way this thesis would not have been a reality and I have to thank them for that.

First and foremost, my supervisor Anagha Joshi that has trusted me with this project and gave me the opportunity to fulfil one of my long standing ambitions! Her support and guidance provided me with all the skills and knowledge to complete this task. Our post-doc Guillaume Devailly, for his immense patience and understanding, his valuable feedback and quick response on my countless questions! My second supervisor Tom Freeman for helping me settle in Roslin the first few months of my PhD and his invaluable feedback during our meetings. My PhD committee, Professor Dave Burt and Professor Wendy Bickmore for their advice and meaningful conversations during my yearly progress meetings. Doug Vernimmen for his encouragement and allowing me to participate in the journal club of his group. The scientists and staff (especially the IT department) in Roslin Institute that have always made sure we are on the right track and kept us busy with many interesting seminars.

My friends and fellow PhD students in the Roslin Institute for lifting up my spirit and making me part of their ‘Phriends Phor Life’ family! My friends in Greece and abroad for keeping me sane and cheering me up with just a few texts that meant the world!

My father, mother and brother for their unconditional love and support! Thank you for being by my side since the first day at school and believing in me so strongly! You are the reason that I keep on asking questions that still do not have answers!

Finally, Jorge because without him I would not feel confident enough not only to complete this PhD but get up and fight for each day! Your love is empowering me to be a better person, scientist, friend, sister and daughter! I dedicate to you this thesis, but frankly you deserve the world!

Abstract

Mammalian cell types and tissues have diverse functional roles within an organism but can be derived by the differentiation of the embryonic stem cells (ESCs). ESCs are pluripotent cells with self-renewal properties. During development subsets of genes in ESCs are activated or silenced for manifestation of the cell type specific function. Gene expression changes occur transiently in early developmental stages, through signals received and executed by a variety of transcription factors (TFs), regulatory elements (promoters, enhancers) and epigenetic modifications of chromatin.

Post-translational modifications of the histone tails are regulated by chromatin modifiers and transform the chromatin architecture. Polycomb (PcG) and Trithorax (TrxG) group proteins are the most commonly studied histone modifiers. They were first discovered as repressors (H3K27me3) and activators (H3K4me3) respectively of Homeobox (Hox) genes in Drosophila and they are conserved in mammals. Bivalent chromatin is defined as the simultaneous presence of silencing (H3K27me3) and activating (H3K4me3) histone marks and was first discovered as a feature of many developmental gene promoters of ESCs. Bivalent promoters are thought to be in a 'poised' state for later activation or repression during differentiation due to the presence of the two counter-acting histone modifications and a pausing variant of RNA polymerase II (RNAPII) accompanied with intermediate-low levels of expression.

By integrative analysis of publicly available ChIP sequencing (ChIP-seq) datasets in murine and human ESCs, we predicted 3,659 and 4,979 high–confidence (HC) bivalent promoters in mouse and human ESCs respectively. Using a peak-based method, we acquire a set of bivalent promoters with high enrichment for developmental regulators. Over 85% of Polycomb targets were bivalent and their expression was particularly sensitive to TF perturbation. Moreover, murine HC bivalent promoters were occupied by both Polycomb repressive component classes (PRC1 and PRC2) and grouped into four distinct clusters with different biological functions. HC bivalent and active promoters were CpG rich while H3K27me3-only promoters lacked CpG islands. Binding enrichment of distinct sets of regulators distinguished bivalent from active promoters and a 'TCCCC' sequence motif was specifically enriched in bivalent promoters.

Using the recent technology of single cell RNA sequencing (scRNA-seq) we focused on gene expression heterogeneity and how it may affect the output of differentiation. We collected single cell gene expression profiles for 32 human and 39 murine ESCs and studied the correlation between diverse characteristics such as network connectivity and coefficient of variation (CV) across single cells. We further characterized properties unique to genes with high CV. Highly expressed genes tended to have a low CV and were enriched for cell cycle genes. In contrast, High CV genes were co-expressed with other High CV genes, were enriched for bivalent promoters and showed enrichment for response to DNA damage and DNA repair.

Bivalent promoters in ESCs grouped in four distinct classes of variable biological functions according to Polycomb occupancy and three RNAPII variants. To study the dynamics of epigenetic and transcription control at promoters during development, we collected ChIP-seq data for two chromatin modifications (H3K4me3 and H3K27me3) and RNAPII (8WG16 antibody) as well as expression data (RNA-seq) across 8 cell types (ESCs and seven committed cell types) in mouse. Hierarchical clustering of 22,179 unique gene promoters across cell types, showed that H3K4me3 peaks are in agreement with the expression data while H3K27me3 and RNAPII peaks were not highly consistent with the hierarchical tree of gene expression. Unsupervised clustering of ChIP-seq and RNA-seq profiles has resulted in 31 distinct profiles, which were subsequently narrowed down to nine major profile groups across cell types. TF enrichment at individual clusters using ChIP sequencing data did not fully agree with the classification of 8 major profile groups.

Considering all the above results, three major epigenetic profiles (active, bivalent and latent) seem to be conserved across the species and cell types in our study. These states could recapitulate only a fraction of the transcriptional information - adding other chromatin marks could enrich it - since they are seemingly unaffected by their respective expression profiles. H3K27me3 only state has low CpG density and shows stronger signatures at differentiated cell types. Transcriptional control is tighter in active than bivalent promoters and the different occupancy levels of PcG subunits and RNAPII can be reflected at the expression variance of bivalent genes, where a fraction of them are involved in developmental functions while others are more tissue-specific. Last, there is a striking similarity in the pausing patterns of RNAPII in the progenitor cell types, which suggests that RNAPII pausing is correlated with the developmental potential of the cell type.

Finally, this analysis will serve as a resource for future studies to further understand transcriptional regulation during development.

Lay of summary

All the different types of organs and tissues in a mammalian organism, derive from one single cell type named Embryonic Stem Cells (ESCs). ESCs have some unique properties that distinguish them from other cell types, such as their ability to give rise to all the cells of the organism and self-renew indefinitely. The diverse characteristics of the different cell types in the body, even though they all contain the same genetic code (DNA), are acquired by expression of different subsets of genes. Expression of genes in each cell type is guided by a specific set of instructions known as epigenetic control of development. Special developmental proteins called transcription factors (TFs) bind to specific areas of the genome (promoters) and are involved in the regulation of transcription of their nearby genes.

DNA is wrapped around proteins called histones, forming the so called chromatin, getting compacted and more easily stored at the limited space of the cell nucleus. Protruding tails of histone proteins, are susceptible to being modified by proteins called chromatin modifiers. Modifications of histone tails lead to subsequent recruitment of other transcription factors that either create a more compacted chromatin structure (silent chromatin) or a more permissive structure (active chromatin) where the transcriptional machinery can initiate transcription of the adjacent gene. Two histone modifications associated with silencing and activating of chromatin respectively are H3K27me3 and H3K4me3. Surprisingly, those two histone marks were found co-existing at the promoter regions of multiple developmental genes in ESCs, raising questions for their functionality and significance in differentiation. Gene promoters associated simultaneously with both of these marks were called bivalent. Bivalent genes are thought to have this mixed chromatin state, known as poised, so that they are easily recognised by the transcriptional machinery and become easily activated or repressed depending on the differentiation signals.

A high-throughput sequencing technique called Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is used for the accurate mapping of the genomic location where histone marks and TFs are found. Integrating multiple ChIP-seq datasets from previously published studies in human and mouse ESCs, we managed to detect bivalent promoters that were found in the majority of the studies, thus increasing their confidence levels. 3,659 high confidence (HC) bivalent promoters in mouse and 4,979 HC bivalent promoters in human ESCs, were particularly enriched for developmental protein functions and were shown to be easily perturbed when the expression levels of many TFs was altered. Bivalency seems to be the rule rather than the exception for H3K27me3 silenced genes in ESCs. A ‘TCCCC’ sequence motif, specific to bivalent promoters, was also detected, possibly allowing regulators to bind those regions with high specificity.

To assess gene expression heterogeneity in ESCs, we also gathered expression data derived from single cells (scRNA-seq) in both human and mouse ESCs. This allowed us to classify genes dependent on their variation of gene expression and co-expression patterns with other genes. A subset of bivalent genes was represented in the group of genes with high variation, which were found being co-expressed with other highly variant genes and were associated with DNA damage and DNA repair functions.

In an effort to follow the fate of bivalent promoters in other cell types we gathered chromatin modification (ChIP-seq) and gene expression (RNA-seq) data in ESCs and seven other committed cell types like progenitor motor neurons, Macrophages and B cells. Using machine learning techniques, we clustered the promoters across cell types, according to their chromatin marks and expression profiles. H3K4me3 mark was fully agreeing with the gene expression across cell types, whereas H3K27me3 was not. Nine major profile groups emerged, ranging from fully active to fully silenced in terms of expression and chromatin status.

Contents

List of Figures

List of tables

List of abbreviations

2i	Two inhibitors
BAM	Binary alignment map
BMDM	Bone marrow derived macrophages
BP	Base pair
CBSs	CTCF binding sites
CBX	Chromobox
ChIP	Chromatin immunoprecipitation
COMPASS	Complex of proteins associated with Set1
CTD	C-terminal domain
CV	Coefficient of variation
DCs	Dendritic cells
DDBJ	DNA Databank of Japan
DNMT3A	DNA (Cytosine-5)-Methyltransferase 3 Alpha
DNMT3B	DNA (Cytosine-5)-Methyltransferase 3 Beta
DPPA3	Developmental Pluripotency Associated 3
DSIF	DRB sensitivity-inducing factor
EBI	European Bioinformatics Institute
EED	Embryonic Ectoderm Development
ESCs	Embryonic stem cells
EST	Expressed Sequenced Tags
EZH1	Enhancer of Zeste 1
EZH2	Enhancer of Zeste 2
FDR	False discovery rate
FISH	Fluorescent in Situ Hybridization
FPKM	Fragments per Kilobase of transcript per Million mapped reads
GEO	Gene Expression Omnibus
GFF	General Feature Format
GO	Gene Ontology
GRO-seq	Global run-on sequencing
GSK3	Glycogen Synthase Kinase-3

HC	High confidence
HCNEs	Highly conserved non-coding elements
HKTMs	Histone Lysine (K) methyltransferases
HMT	Histone methyltransferase
HOX	Homeobox
IP	Immunoprecipitation
JARID	Jumongi/ARID domain
KDM2A	Lysine (K)-Specific Demethylase 2A
KDM2B	Lysine (K)-Specific Demethylase 2B
L3MBTL2	Lethal(3) Malignant Brain Tumor-like Protein 2
LEFTY	Left-Right Determination Factor
MAE	Monoallelic expression
MBs	Myoblasts
MEFs	Mouse embryonic fibroblasts
MEK1	Mitogen-Activated Protein Kinase 1
miRNAs	microRNAs
MLL	Mixed lineage leukemia
MNase	Micrococcal nuclease
MS	Mass spectrometry
MTs	Myotubes
NANOG	Nanog Homeobox
NCBI	National Center for Biotechnology Information
NELF	Negative Elongation Factor
NGS	Next Generation Sequencing
NPCs	Neural progenitor cells
P-TEFb	Positive Transcription Elongation Factor b
PCA	Principal component analysis
PcG	Polycomb Group
PCL	Polycomb-like
PIC	Pre-initiation complex
piRNAs	piwi-interacting RNAs
PMNs	Progenitor motor neurons
PPI	Protein-protein interactions

PRC1	Polycomb Repressor Complex 1
PRC2	Polycomb Repressor Complex 2
RA	Retinoic acid
REX1	RNA Exonuclease 1 Homolog
RING1A	Ring Finger Protein 1
RING1B	Ring Finger Protein 2
RNAPII	RNA polymerase II
RYBP	RING1 and YY1 Binding Protein
SAM	Sequence alignment map
SBL	Sequencing by ligation
SBS	Sequencing by synthesis
scRNA-seq	Single cell RNA-seq
SET	Su(var)3-9, Enhancer of Zeste, Trithorax
SET1	SET Domain Containing 1
SNA	Single nucleotide addition
SNV	Single nucleotide variant
SRA	Sequence Read Archive
SUZ12	Suppressor of Zeste 12
TAD	Topologically associated domain
TES	Transcription Ending Site
TET	Ten-eleven translocation methylcytosine dioxygenase
TF	Transcription factor
TFBS	Transcription Factor Binding Sites
TrxG	Trithorax Group
TSS	Transcription Start Site
UMIs	Unique molecular identifiers
ZFX	Zinc Finger Protein, X-Linked
ZIFA	Zero-inflated factor analysis

Chapter 1 Introduction

1.1.1 Gene regulation during development

The diverse range of mammalian organs and tissues is a product of underlying differences in the gene expression programme of different cell types with the same DNA sequence. Subsets of genes are activated or silenced during development according to a set of instructions which includes epigenetic control mechanisms (Reik, 2007). Throughout development and differentiation, the fate of each cell type is primarily controlled by gene regulation (Pearson et al., 2005) where genomic regulatory elements receive and execute transcriptional signals, dependent on their epigenetic state and chromatin accessibility, controlling the expression of key developmental factors (Wilson et al., 2010).

Gene expression changes occur transiently during the early stages of development, influenced by transcription factors (TFs) and epigenetic modifications (Bird, 2002; Li, 2002; Morgan et al., 2005; Turner, 2007), such as DNA methylation at CpG dinucleotides (Bird, 2002; Li, 2002) and histone modifications at the tails of nucleosomal histones (Turner, 2007). To gain a better view of the changes taking place during development, we need a deeper understanding of the basic elements of chromatin. DNA is wrapped around the core histone proteins, creating a structure of 8 histone proteins (2 copies of H2A, H2B, H3, H4) and 147 base pairs (bp) of DNA around them, named the nucleosome (Kornberg, 1974). Higher-order chromatin is formed through compaction of the nucleosomes, with the assistance of various assembly and packaging related proteins (Li, 2002). Two main distinct chromatin states are ‘Euchromatin’ and ‘Heterochromatin’. A more open chromatin environment in which the nucleosomes are spaced far apart and the DNA becomes accessible to transcriptional machinery characterizes euchromatin where the majority of active genes localize. However, euchromatin is not marked uniformly with epigenetic and transcriptional signals i.e. more histone modifications are present in regions with high density of transcription factor binding sites (TFBS), typically regions where either regulation or transcription takes place (Barski et al., 2007). On the other hand, heterochromatin is characterised by a more compact environment where inactive genes, non-coding DNA and repetitive elements reside (Bannister and Kouzarides, 2011). Like euchromatin, heterochromatin has non-uniform epigenetic and transcription status distribution that can be distinguished into two groups, facultative and constitutive. Genes with high differential expression throughout development often reside within facultative heterochromatin. These regions are switched off when the cell acquires its new identity. In contrast, constitutive heterochromatin is gene poor, rich in repetitive elements, mainly found in centromeres and telomeres and silenced indefinitely.

Chromatin structure is tightly organized with the assistance of the numerous histone modifications and a potential cross-talk between them provides an extra level of complexity to the chromatin architecture (Kouzarides, 2007). Histone amino (N)-terminal tails extend beyond the main nucleosome body, interacting with neighbouring nucleosomes and are subject to modifications that can influence the inter-nucleosomal relationship and chromatin structure. Of various histone modifications, the most well studied types are methylation, acetylation, phosphorylation and ubiquitination (Bannister and Kouzarides, 2011).

Histone modifications influence chromatin mainly in two ways. Firstly and primarily, the modifications affect directly the structure of the chromatin over a long or short distance. Histone modifications can recruit DNA binding proteins and chromatin re-modellers, consequently leading to the relocation of nucleosomes (Margueron et al., 2005). Hence, nucleosome removal could open the chromatin and a possible binding motif could be revealed, or instead, newly recruited nucleosomes could conceal a binding motif, affecting transcriptional machinery recruitment at the locus. Histone modifications also work jointly with DNA methylation, to allow or inhibit specific protein binding. For example KDM2A (Lysine (K)-Specific Demethylase 2A) binds only to nucleosomes that present histone H3 Lysine 9 tri-methylation (H3K9me3) where DNA is un-methylated (Bartke et al., 2010).

A number of studies have connected specific histone modifications to a variety of processes with discrete functionalities (Margueron et al., 2005; Nightingale et al., 2006). The emergent term ‘histone code’ or ‘epigenetic code’ tries to assign an associated function to multiple combinations of histone modifications and DNA methylation, linking them to the presence or absence of transcriptional activity and genomic functional elements (Jenuwein, 2001; Turner, 2007). For example, H3K4me1 is present at regulatory elements called enhancers (defined later in this section, page 5) and is widely used to predict their location (Hon et al., 2009). H3K4me3 is highly enriched at the Transcription Start Site (TSS) of actively transcribed genes (Barski et al., 2007; Schneider et al., 2004) and H3K36me3 is found on the gene body of genes under transcription (Bannister et al., 2005). Also, high levels of H3K9me3 are associated with constitutive heterochromatin (Trojer and Reinberg, 2007). Figure 1.1 is a graphical representation of the 'histone code'.

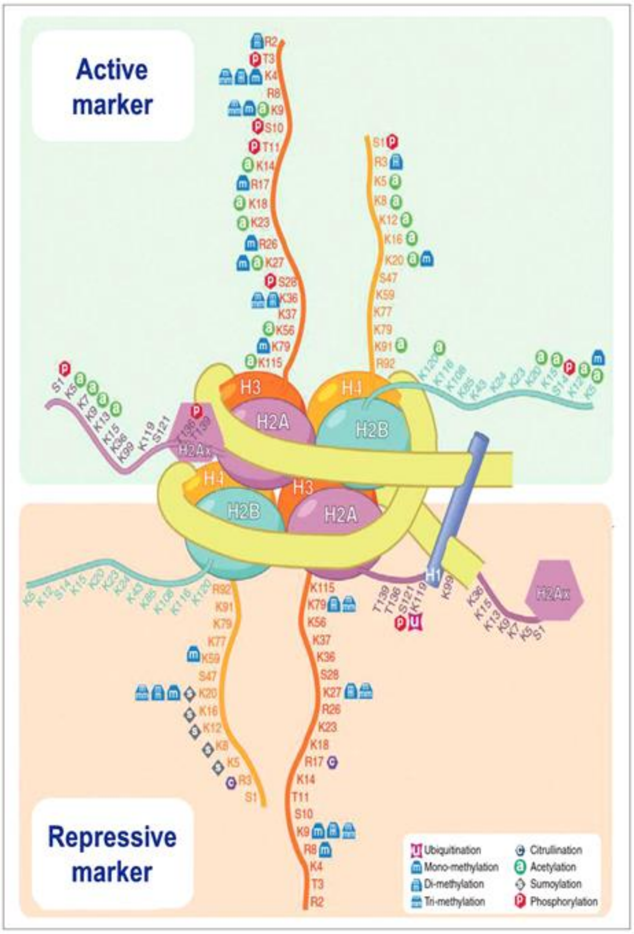


Figure 1.1 Representation of the post-translational modifications of the histone tails. The modifications are divided according to their association with activation (Active marker – green panel) and repression (Repressive marker – orange panel) of transcription. Image taken from (Kim, 2014).

Histone modifications occupy various regulatory sequence elements across the genome in a dynamic fashion across development (Zhou et al., 2011). One of these elements is the promoter element, which overlaps with the TSS of a gene and is typically composed of two main regulatory regions: the core promoter and the region immediately upstream of the core promoter, the proximal promoter. The core promoter area (including the TSS) is necessary for the initiation of transcription. The RNA polymerase II (RNAPII) is recruited at the core promoter of many protein-coding genes. The proximal promoter is the region where many TFs bind; it acts in collaboration with the core promoter (Juven-Gershon et al., 2006).

The architecture of the promoters is dynamic during differentiation and evolution, acquiring distinct functional and regulation patterns according to the type of genes. The promoters associated with RNAPII are divided in 3 different classes (Lenhard et al., 2012). Genes expressed uniquely in mature cell types mostly belong to the Type I promoters, with characteristics such as TATA-box enrichment, sharp TSS, great distance from CpG islands and ‘fuzzy’ nucleosomes. Housekeeping genes belong to Type II promoters, which have a wide TSS region, well-positioned nucleosomes and are close to CpG islands. Type III promoters are mainly allocated to developmentally regulated genes. Type III promoters demonstrate a sharper TSS (more so than Type I promoters), they usually have more than one CpG island in their direct proximity and often in the gene body, and they are associated with silencing by Polycomb Group proteins (PcG). Moreover, some developmental TFs, cell adhesion genes and mediator genes feature some very specific characteristics in their loci that allow them to create their own unique promoters category (Akalin et al., 2009).

One cannot fully explain the innumerable gene expression patterns observed throughout development and differentiation by only focussing on the promoter types, since most of regulation of the metazoan genome happens with the assistance of enhancer regions, which are another well studied group of regulatory elements (Heintzman et al., 2009; Thurman et al., 2012; Yip et al., 2012). Enhancer elements can be up to hundreds of base pairs long and they can bind numerous TFs and chromatin regulators, regulating the level of expression of their target genes in a unique spatiotemporal pattern. Distance or direction of enhancers with respect to their target gene appear unrelated with their efficiency (Maston et al., 2006; Vavouri et al., 2006), although recent evidence suggests CTCF binding sites (CBSs) may influence enhancer/promoter interactions within the topologically associated domain (TAD) in which they reside, regarding their orientation (Guo et al., 2015; Narendra et al., 2015). Enhancers can be found in intergenic regions (upstream or downstream) (Sanyal et al., 2012), in the introns of the same gene that is regulated or the neighbouring genes (Kikuta et al., 2007a) or in coding exons of their own or of neighbouring genes (Birnbaum et al., 2012; Lampe et al., 2008). In most of the cases enhancers remain close to their target genes, even after genome duplication (Kikuta et al., 2007b). A simulation study analysing the promoter-enhancer interactions in 12 human cell types, used real enhancer-promoter pairs with median distance around 15-17 kb as training sets for their prediction (He et al., 2014). Also, there are some trans-acting enhancer elements that are found on a different chromosome from their target gene (Bateman et al., 2012). It has been documented that genes associated with cell lineage commitment are regulated by super-enhancers (Whyte et al., 2013). Super-enhancers are fairly larger than normal enhancers, are bound by the mediator and important master TFs and they mainly regulate cell identity genes. The underlying biology of the super-enhancers remains unclear and some view them as clusters of classical enhancers with no additional properties (Hay et al., 2016; Whyte et al., 2013).

In Embryonic Stem Cells (ESCs), the majority of promoters with high CpG content have un-methylated DNA. During differentiation however, some of them become methylated, acquiring their final transcriptionally silenced identity (Mohn et al. 2008). The promoters that remain methylated through differentiation are mainly un-transcribed and are related with Type I promoters/tissue specific genes (Isagawa et al., 2011). Developmental genes show high inconsistency at the level of their methylation and they seem to be regulated from multiple enhancers, found both near and far (Mikkelsen et al., 2010). Active genes are located at euchromatin regions and they interact with a restricted number of enhancers (Soler et al., 2011). The causal relationship of the interaction between promoters and enhancers across the various differentiation pathways and the implication of histone modifications in it has not yet been deciphered.

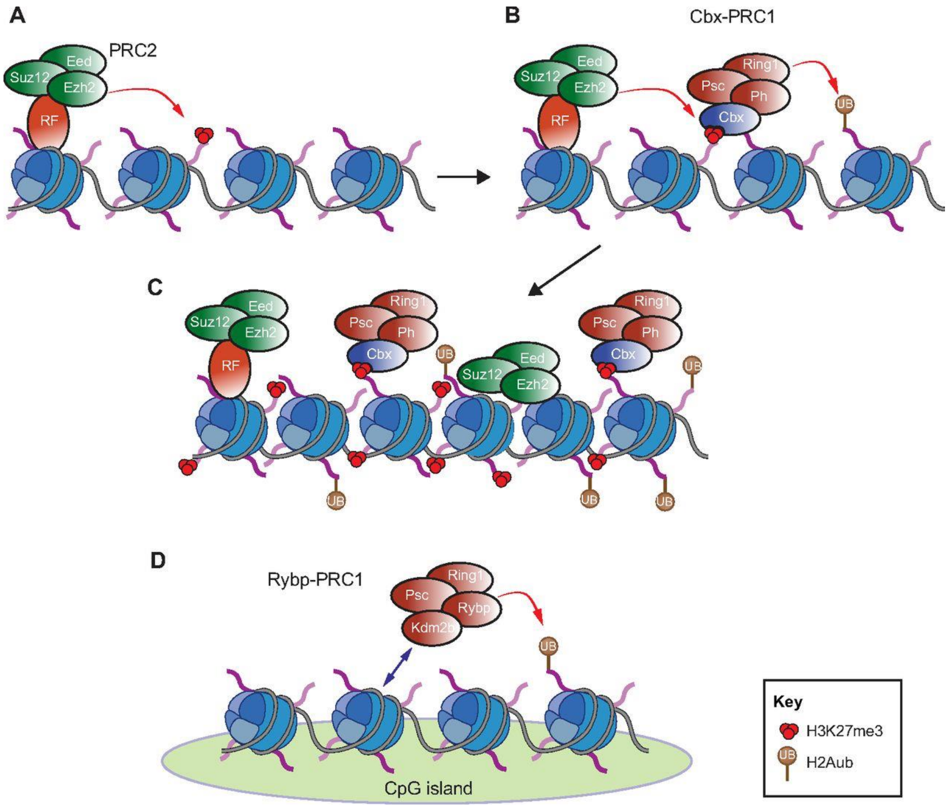
1.1.2 Polycomb group and Trithorax group families in ESCs

To unravel key developmental transitions that lead to different cell identities, ESCs offer a valuable model for examination (Thomson et al., 1998). ESCs can replicate themselves indefinitely and give rise to progenies with the same developmental pluripotency i.e. the capability of differentiating into all the cell types and tissues of an adult organism during development and adult life (Thomson et al., 1998; Voigt et al., 2013). Azuara et al.(2006) proposed that particular histone modifications and the chromatin structure features (Thomson et al., 1998; Voigt et al., 2013) assist in the formation of the special ESC properties. Histone methylation takes place predominantly on the side chains of Lysine and Arginine amino acid residues of the histone tails. It does not generate any difference in the charge of the histone protein but acts as a binding site for effector proteins (Chromo, Tudor and WD40-repeat domains interact with histone lysine methylation) that subsequently bear changes in chromatin or transcriptional output (Ng et al., 2009). Lysines are subject to mono-, di- or tri-methylation while Arginines are subject to mono- and symmetrical or asymmetrical di-methylation (Ng et al., 2009). The lysine methyltransferases (KMTs) methylate the Lysine with the assistance of a SET (Su(var)3-9, Enhancer of Zeste, Trithorax) domain, which is responsible for the enzymatic activity (Rea et al., 2000), using an S-adenosyl-L-methionine (SAM) as a methyl donor (Lanouette et al., 2014). The SET-domain proteins present sequence and domain similarities and are roughly classified in seven groups (Dillon et al., 2005): SUV3/9, SET1, SET2, SMYD, EZ, SUV4-20 and RIZ (Lanouette et al., 2014) . SUV3/9, SET1, SET2, SMYD and EZ family proteins are capable of catalysing the methylation of both histone and non-histone proteins (He et al., 2012; Huang et al., 2006; Lu et al., 2010; Rathert et al., 2008; Zhang et al., 2005) . The role of histone KMTs (HKTMs) appears to be very specific, since they catalyse the methylation of a particular lysine of the histone tail and to a certain degree (i.e. mono-, di- or tri-methylation) using a catalytic domain involved in determining the degree of the methylation (Zhang et al., 2003).

Histone modifications are regulated by many TFs that act as chromatin modifiers (Niwa, 2007). Two of the most commonly studied histone modifications are H3K4me3 and H3K27me3. They are associated with activation (H3K4me3) and repression (H3K27me3) of chromatin (Bannister and Kouzarides, 2011). Polycomb (PcG) and Trithorax (TrxG) group proteins catalyse H3K27me3 and H3K4me3 respectively, regulating development and differentiation (Ringrose and Paro, 2004).

PcG proteins were first described as suppressors of Homeobox (Hox) genes in *Drosophila* (Kennison, 1995; Lewis, 1978; Schuettengruber et al., 2007) and there is a strong conservation of their function in mammals. The PcG proteins can form various complexes. Polycomb Repressor Complex 1 (PRC1) and Polycomb Repressor Complex 2 (PRC2) are the most well-studied (Margueron and Reinberg, 2011). PRC2 di- and tri-methylates the lysine 27 of histone H3, establishing transcriptional repression at those sites (Czermin et al., 2002) (Figure 1.2). On the other hand, PRC1 catalyses the mono-ubiquitination of lysine 119 of histone H2A, which also represses gene transcription (Endoh et al., 2012) and more specifically stops transcriptional elongation (Stock et al., 2007a). PRC1 also compacts chromatin (Endoh et al., 2012) to impair transcription but its function is not only limited to transcriptional repression. The H2AK119ub modification that PRC1 facilitates is also responsible for the proper removal of PRC1 from chromatin, so that under the appropriate developmental cues gene transcription can be activated (Richly et al., 2010) (Figure 1.2).

Molecular functions of PRC1 and PRC2.



Luigi Aloia et al. Development 2013;140:2525-2534

Figure 1.2 A) PRC2 catalyses H3K27me3 and then B, C) recruits canonical PRC1 that mono-ubiquitinates H2A, D) KDM2B facilitates the recruitment of non-canonical PRC1 to un-methylated CpG regions, without the PRC2 involved. Taken from (Aloia et al., 2013)**.**

The core proteins that provide PRC2 its enzymatic activity are the histone methyltransferases Enhancer of Zeste 1 (EZH1) and Enhancer of Zeste 2 (EZH2) (Margueron et al., 2008). Furthermore, Suppressor of Zeste 12 (SUZ12) and Embryonic Ectoderm Development (EED) facilitate the assembly of PRC2 complex and when bound with EZH1 or EZH2, the methyltransferase property of the complex is activated (Blackledge et al., 2015; Cao and Zhang, 2004). PRC2 recruitment to chromatin has been attributed to Jumongi/ARID domain containing (JARID) protein and the Polycomb-like family members (PCL) proteins (Peng et al., 2009). These two protein families target a different set of genes with different mechanisms (Walker et al., 2010) and they do not co-exist in the same complex (Ballaré et al., 2012), suggesting that in mammals there are several mechanisms and complexes for different sets of genes. JARID2 binds to un-methylated GC and GA dinucleotides through its ARID domain, while PCL proteins, through their TUDOR domain, bind at sites that are enriched with H3K36 methylation, which is related to elongation of transcription (Ballaré et al., 2012). Thus, PCL proteins target genes that are already actively transcribed whereas JARID2 protein is recruited at sites where DNA methylation is removed, through the action of Ten-eleven translocation methylcytosine dioxygenase (TET) proteins (Tan and Shi, 2012).

PRC1 can be found in different compositions corresponding to the cell environment (Gao et al., 2012). The proteins Ring Finger Protein 1 (RING1A) and Ring Finger Protein 2 (RING1B) are members of all the PRC1 complexes. They are E3 ubiquitin ligases that are responsible for the addition of a ubiquitin group at the lysine 119 of histone H2A (de Napoles et al., 2004; Wang et al., 2004). There are two categories of PRC1 complexes based on the presence of Chromobox (CBX) proteins in the complex. The canonical PRC1 complex interacts with PRC2 complex after recognizing and binding to the H3K27me3 mark through the CBX proteins, leading to gene repression. It is also hypothesized that the interaction between PRC1 and PRC2, stabilizes the repressive effects in key genes during development and differentiation (Bracken et al., 2006; Schuettengruber et al., 2007). On the contrary, non-canonical PRC1 complexes usually have RING1 and YY1 Binding Protein (RYBP), Lethal(3) Malignant Brain Tumor-like Protein 2 (L3MBTL2) or Lysine-specific Demethylase 2B (KDM2B) (García et al., 1999; He et al., 2013; Qin et al., 2012) and they are known to target different genes from the canonical PRC1 (Morey et al., 2013). In stem cells though, there exists a shared set of genes regulated by both canonical and non-canonical PRC1 complexes, indicating possible overlap at a molecular level (Morey et al., 2013).

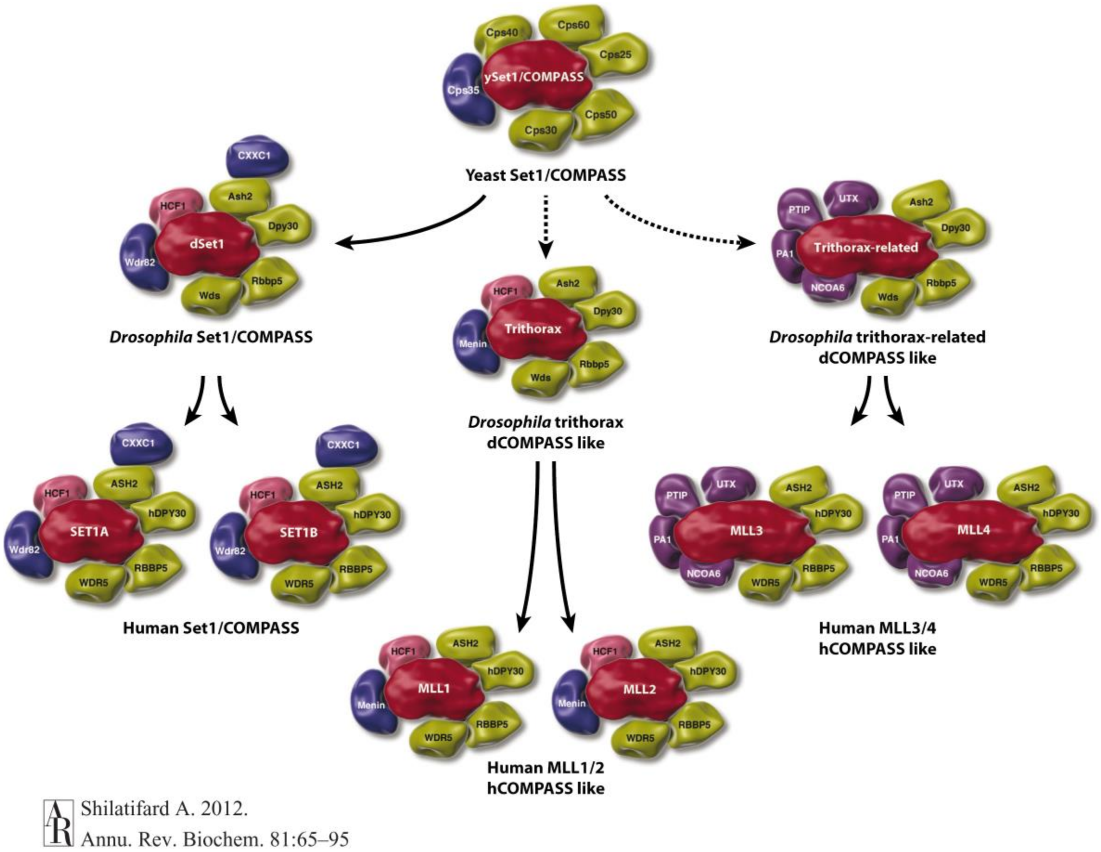


Figure 1.3 COMPASS family in yeast, *Drosophila* and human is divided into three sub-groups: Set1/COMPASS, trithorax-containing and trithorax-like. Red shows the SET-domain containing enzymes, green highlights common in eukaryotes, blue and purple show the specific subunits to each complex and magenta highlights Host Cell Factor 1 (HCF1) that is found in some members of the Complex Proteins Associated with SET1 (COMPASS) family (figure taken from Shilatifard 2012a).

TrxG proteins, first discovered activating Hox genes in *Drosophila*, are a highly conserved group of proteins from yeast to mammals, consisting of H3K4 methyltransferases (Schuettengruber et al., 2007; Shilatifard, 2012a). The mixed lineage leukemia (MLL) gene was the first recognised homologue of the *Drosophila trx* gene in mammals. Initial studies of MLL's *Saccharomyces cerevisiae* (yeast) homologue, SET Domain Containing 1 (Set1), confirmed its methyltransferase activity. To isolate Set1 in yeast, a complex of associated proteins was extracted and was given the name complex of proteins associated with Set1 (COMPASS) (Miller et al., 2001). SET1/COMPASS methyltransferases can mono-, di- and tri- methylate H3K4 residues, with SET1 gaining its methyltransferase activity only when it is a member of the COMPASS complex (Krogan et al., 2002; Miller et al., 2001).

Some of the enzymes that TrxG complex consists of in mammals are SET1A-B and MLL1-4 (Figure 1.3). MML1 and MLL2 are considered crucial for the H3K4me3 modification at the promoters of mouse ESCs (Hu et al., 2013). MLL2 specifically is necessary for the deposition of H3K4me3 at all Homeobox gene clusters, the most well-studied developmental genes (Hu et al., 2013). MLL1 drives the program of haematopoiesis through the mid to late stages of development (Ernst et al., 2004) whereas MLL2 is detected at the very early stages of development regulating genes that are very important for cell commitment and differentiation (Hu et al., 2013).

1.1.3 Bivalent promoters

Several specific genomic loci in ESCs exhibited simultaneous counteracting histone modifications at promoter regions (Zhou et al., 2011). The observed combinatorial signals of activating (H3K4me3) and repressing (H3K27me3) chromatin were named “bivalent” and they appeared to mark developmental gene promoters in ESCs (Bernstein et al., 2006a). Initially, Bernstein et al. (2006) were intrigued by the highly conserved non-coding elements (HCNEs) that were found in the proximity of genes encoding for developmental transcription factors. They used chromatin immunoprecipitation (ChIP) and tilling array (ChIP-chip) techniques, studying the patterns of the co-existing H3K4me3/H3K27me3 marks in HCNEs in mouse ESCs (mESCs). They performed sequential ChIP, which verified that promoters of certain genes carried H3K4me3 and H3K27me3 marks at the same time, which until then were considered mutually exclusive.

Initially, bivalent domains were assumed as a distinctive feature of ESCs, since during differentiation developmental gene promoters that were occupied by both marks in ESCs, were typically occupied by a single mark (monovalent) which typically expanded in size. In 2007, Mikkelsen et al. combined ChIP and next generation sequencing (ChIP-seq)[1] to examine the bivalent marks and construct genome-wide chromatin state maps for various cell types such as mESCs, neural progenitor cells

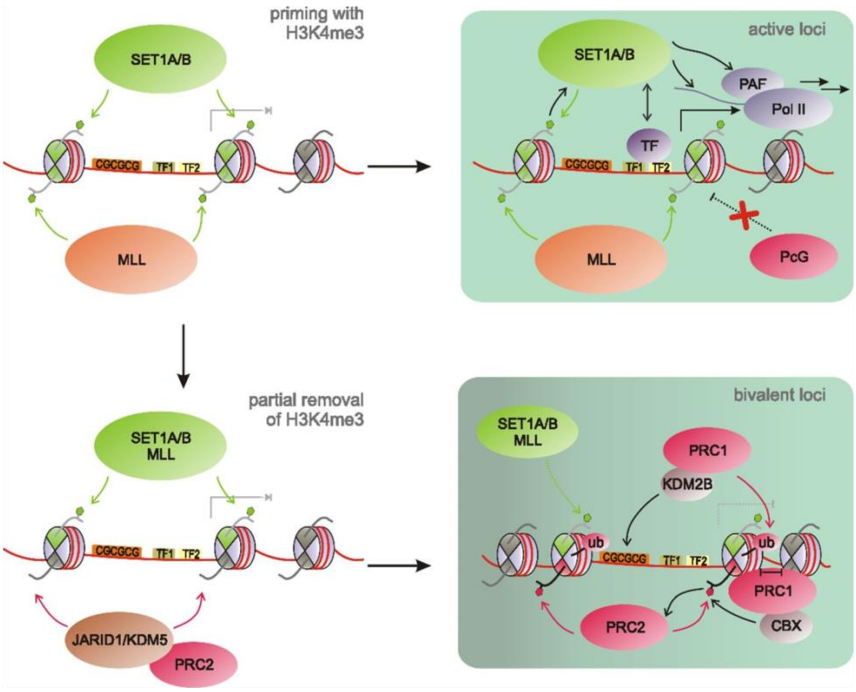
[1] ChIP-seq method and challenges using this technology are described in detail in section **1.2.1**

(NPCs) and mouse embryonic fibroblasts (MEFs). They distinguished three categories of gene promoters, determined by their chromatin marks (H3K4me3/H3K27me3), namely: expressed (only H3K4me3), poised for expression (both marks, i.e. bivalent) and repressed (only H3K27me3) (Mikkelsen et al., 2007). Their study also showed for the first time, that bivalent domains exist also in cells of restricted potency. During the differentiation of mESCs to NPCs and mESCs to MEFs, 8% (~202) and 43% (~1085) of bivalent domains, respectively, retained their bivalent mark (Mikkelsen et al., 2007). Moreover, Mohn et al. (2008) indicated that in mESCs that were terminally differentiated (ESCs to NPCs to neurons), ~1000 bivalent domains were lost, whereas ~340 new bivalent domains emerged, suggesting that reduced potency cells may have bivalent genes that are not present in the pluripotent cells. Hence, bivalent domains are not specific to ESCs, but they appear in unipotent cells as well, as demonstrated by a number of studies (Adli et al., 2010; Barski et al., 2007; Cui et al., 2009; Roh et al., 2006). Bivalent genes were also detected in human ESCs (hESCs) (Pan et al., 2007; Zhao et al., 2007) and the majority of them were shared with the bivalent genes in mESCs. Specifically, there were ~ 2,000 bivalent genes overlapping between mouse and human (consensus number made from more than 60% of the studies) (Mikkelsen et al., 2007; Pan et al., 2007; Sharov and Ko, 2007; Zhao et al., 2007). Consistent with the studies in mice, hESCs bivalent genes are functionally enriched with developmental transcription factors and genes, with most of them gradually losing the repressive H3K27me3 mark during differentiation (Pan et al., 2007; Zhao et al., 2007).

Bivalent chromatin was also found in epiblast stem cells of mouse embryos (Rugg-Gunn et al., 2010) confirming their presence in developing organisms where pluripotency is transient and not artificially pluripotent as in cultured ESCs. However, the H3K9me3 mark seemed to replace H3K27me3 in bivalent domains of other pluripotent cell lines derived from the blastocyst (trophoblast and extraembryonic endoderm stem cells), possibly due to lower efficiency of PRC2 silencing mechanisms (Rugg-Gunn et al., 2010).

Ku et al. (2008) found that in mouse ESCs there are two distinct categories of bivalent domains according to the occupation of PcG complex proteins. The first class consists of domains where only PRC2 exists (“PRC2 only”) and the second one, called “PRC1-positive”, where PRC2 domains are also occupied by PRC1. “PRC2 only” bivalent domains include non-developmental groups of genes and they are not highly conserved. There is an association of PRC1 with clearly broader bivalent regions, which are highly conserved, which have high maintenance levels of H3K27me3 and which are linked to numerous developmental promoters (Ku et al., 2008a). These two distinct types of bivalent domains may suggest that different classes of bivalent promoters do exist, requiring the recruitment of PcG proteins in a different way.

A step-wise model for the generation of bivalent domains.



Voigt P et al. Genes Dev. 2013;27:1318-1338

Figure 1.4 Members of the COMPASS family are priming all the un-methylated CpG loci with H3K4me3. If there are enough activators and transcription factors in the regions, gene is activated, while PcG proteins are depositing repressing marks in the absence of activators, leading to the formation of bivalent domains (figure taken from Voigt et al., 2013).

1.1.4 Functional relevance of bivalent chromatin

Various regulatory mechanisms prevent ESCs from losing their pluripotency, e.g. DNA methylation that would silence important genes indefinitely is prevented. Bivalent genes include developmental factors that are thought to be poised for activation or repression at the right moment during the differentiation process (Voigt et al., 2013). H3K4me3 impedes the activity of DNA (Cytosine-5)-Methyltransferase 3 Alpha (DNMT3A) and DNA (Cytosine-5)-Methyltransferase 3 Beta (DNMT3B), both de-novo methyltransferases capable of catalysing the transfer of methyl-groups at cytosine residues of the DNA (Ooi et al., 2007; Zhang et al., 2010). Moreover, TET enzymes safeguard CpG islands from DNA methylation, ensuring the plasticity of bivalent genes is retained (Williams et al., 2011). Nevertheless, an ultra-permissive chromatin would allow RNAPII and associated TFs to be recruited at the loci and initiate transcription. Bivalent genes were found to produce abortive transcripts (Brookes et al., 2012a; Kanhere et al., 2010; Min et al., 2011; Walker et al., 2010), yet PcG proteins were actively regulating the binding of RNAPII in both its initiation and elongation forms (Chopra et al., 2011; Min et al., 2011; Stock et al., 2007a). H3K27me3 is deposited by PRC2 to counterbalance the effects of H3K4me3 and TrxG machinery, adjusting the levels of expression. The repressive mark may inhibit the deposition of H3K36me3 at the same nucleosome, since these two histone modifications have an opposing effect and cannot be present simultaneously at the same histone tail (Schmitges et al., 2011; Voigt et al., 2012). Furthermore, H3K27me3 assists at the recruitment of PRC1 complex, which in turn catalyses H2Aub1, creating a barrier for RNAPII and the pre-initiation complex, preventing their recruitment at the highly compacted chromatin (Francis et al., 2004; Grau et al., 2011; Lehmann et al., 2012; Min et al., 2011) .

Voigt et al. (2013) proposed a model for the generation of bivalent domains in ESCs, where CpG rich promoters are marked with different levels of H3K4me3 with the assistance of the SET1A/B and MLL complexes. In their hypothetical model, activation occurs only at those loci where there is an abundance of TFs recruited by regulatory elements (Voigt et al. 2013). Bivalency arises at loci with insufficient transcriptional machinery, where the PRC2 complex is able to deposit its repressing mark at the opposite tail of the deposited H3K4me3. In order to strengthen the repression, PRC1 is recruited at some bivalent domains by PRC2. It is observed that the higher the CpG density, the more efficient the PRC2 recruitment is at loci already occupied by activation marks (Voigt et al. 2013). This seems to help the PRC2 complex to compete against SET1/MLL complexes only at the bivalent genes, but not at the active genes where the transcription is well established and protected with many TFs. One way of controlling bivalency is by controlling the load of transcription activation that the gene is subject to, until the appropriate environmental cues initiate the differentiation process. PRC1 and PRC2 recruit each other and establish stable bivalent domains (Figure 1.4). The bivalent state may protect the plasticity of developmental genes between the anticipated activation or repression. When genes need to be activated, activating stimuli recruit all the activating TFs, the H3K27 demethylases and the H2A de-ubiquitinases, and transform bivalent regions to active regions. Correspondingly, H3K4 demethylases are reinforced and robust silencing machinery is gathered at the bivalent loci in case of repression necessity (Voigt et al., 2013).

Voigt et al. (2013) further proposed that as soon as the differentiation process begins, genes should switch on only after a specific threshold of developmental cues is reached. Repressed genes demand a particularly high developmental signal in order to become activated in an efficient manner. Genes lacking both repressing and activating marks could start being transcribed before the required threshold. Bivalent genes, however, are not fully repressed or constitutively active, since they contain repressive marks at levels that could easily be removed in order for transcription to be induced after the desirable developmental threshold (Voigt et al., 2013).

Many studies have argued against the ambiguous function of “poised” bivalent genes, finding the original hypothesis too simple to be accurate. Interestingly, even though PRC2 ablation in ESCs has caused developmental factors to be abnormally expressed (Boyer et al., 2006), it did not affect their pluripotent properties (Chamberlain et al., 2008). Likewise, various trxG components (Dpy-30, RbBP5 and WDR5) knocked down in ESCs caused variable effects with conflicting phenotypes. There was hindering of differentiation in one case (Jiang et al., 2011a) and failure of cells to self-renew in another (Ang et al., 2011). Additionally, the observation that bivalency is present not only in ESCs but also committed cell types (Adli et al., 2010; Barski et al., 2007; Cui et al., 2009; T. Mikkelsen et al., 2007; Mohn et al., 2008; Roh et al., 2006), arguably poses a question to the functional relevance of bivalent chromatin in association to ESC differentiation.

Using ChIP-qPCR and expression analysis, in both bulk populations and single cells, in a well-defined bivalent locus in ESCs like α globin, De Gobbi et al. (2011) showed a clear positive correlation between gene expression and H3K4me3/H3K27me3 ratio. Basal levels of transcription at the bivalent α globin promoter, backed up by variable levels of H3K4me3 at bivalent promoters of multiple genes (Barski et al., 2007; Roh et al., 2006) suggest an alternative scenario for bivalency. Bivalent genes may be mainly regulated by PcG proteins and their corresponding repressive marks, whereas H3K4me3 might appear in variable intensities, reflecting the low transcriptional signal. Hence, the apparent bivalency of some genes can be attributed to low levels of stochastic expression of those genes due to multi-lineage priming (De Gobbi et al., 2011). For example, in many highly-potent and progenitor cells it is observed that stochastic gene expression of tissue-specific genes occurs (Hu et al., 1997). In hematopoietic multipotent cells, for example, enhancer elements are subject to multi-lineage priming before the cells commit to the lymphoid or myeloid fate (Mercer et al., 2011). This comes in contrast to the proposed scenario of competing TrxG and PcG proteins where both TrxG and PcG proteins are present at the specific locus so that bivalency is retained in every cell division until the cell activity changes. Alternatively, De Gobbi et al. (2011) propose that PcG proteins could lose their efficiency in the role of suppression and elimination of transcriptional noise along the various differentiation pathways. Hence, PcG proteins may constitute the main controllers of bivalent genes, subsequently clearing the way for TFs to act upon a transcriptional plan based on the cell lineage commitment (Raser and O’Shea, 2004).

Interestingly, MLL2 knock-down in mouse ESCs has resulted in reduction of H3K4me3 at bivalent loci, but there was no notable difference in their induction kinetics when treated with retinoic acid (RA) and forced to differentiate (Hu et al., 2013). Moreover, H3K27me3 levels at bivalent promoters of Hox genes remained unchanged, suggesting a more peripheral role for bivalency and the proposed competition between TrxG and PcG proteins. Undoubtedly, genome-editing tools such as CRISPR-Cas9 will assist in experiments where the absence of both PRC2 and MLL2 at bivalent targets will be simultaneously assessed in vivo, and the functional relevance of bivalency will be contested in the physiological context of development.

1.1.5 Controversies around bivalency

Co-existence of both H3K4me3 and H3K27me3, at the same allele or nucleosome, cannot be validated through ChIP assays conducted independently for each mark. Bivalency might be due to cellular heterogeneity of the bulk cell population used in ChIP experiments. Nevertheless, the heterogeneous cell population argument does not seem to fully explain the occurrence of bivalent domains at committed cell lineages (sorted populations of T cells and MEFs) that are more homogeneous (Pan et al., 2007; Roh et al., 2006).

Arguments supporting cellular heterogeneity have blamed this diverse epigenetic landscape of ESCs on the serum culture where various components allow for heterogeneous expression of several pluripotency factors like Nanog Homeobox (NANOG) (Chambers et al., 2007; Singh et al., 2007), RNA Exonuclease 1 Homolog (REX1) (Toyooka et al., 2008) and Developmental Pluripotency Associated 3 (DPPA3, STELLA) (Hayashi et al., 2008). To establish a better understanding of the heterogeneous cell population that acts as the inducer of bivalency, Marks et al. (2012) have examined the landscape of epigenetic factors in naïve pluripotent ESCs. The serum obstacle can be avoided using two inhibitors (2i), signalling proteins Mitogen-Activated Protein Kinase 1 (MEK1) and Glycogen Synthase Kinase-3 (GSK3) (Ying et al., 2008). These 2i conditions result in a more homogeneous ESC population and keep the expression of developmental genes consistently low. These naïve ESCs show remarkably lower levels of H3K27me3 at the promoters, leading to the detection of fewer bivalent genes (Marks et al., 2012). Even though the computationally imposed, arbitrary signal cut-off is definitely affecting the number of identified bivalent genes, it is clear that confident bivalent domains can still be detected in highly homogeneous populations of ESCs.

Brookes et al. (2012) have combined genome-wide ChIP-seq data of histone marks (H3K4me3, H3K27me3) and RNAPII in multiple conformations according to its phosphorylated C-terminal domain (CTD). They identified ~3600 bivalent genes in mouse ESCs grown under normal serum conditions, with the majority of them (~2400) being bound by RNAPII phosphorylated at Ser5 (S5P), a conformation found at promoters during initiation of transcription. Sequential ChIP on multiple bivalent loci confirmed the co-occurrence of Polycomb subunits PRC1 and PRC2 with RNAPII S5P. The remaining 1/3 of bivalent genes, were also bound by RNAPII S2P, which is related with elongation of transcription, and their transcription was significantly elevated compared with the rest of bivalent genes. Since H3K27me3 and transcriptional elongation were considered incompatible (Schmitges et al., 2011), Fluorescent in Situ Hybridization (FISH) analysis was performed on Left-Right Determination Factor (LEFTY) locus and the results indicated that its promoter in some cells was marked by the silencing PRC2 subunit (repressed) and in others by the elongating RNAPII S2P form (active transcription) (Brookes et al., 2012a). Further integration of expression data and functional enrichment analysis, has uncovered multiple groups of Polycomb regulated genes initially classified bivalent, only due to cell population differences. In particular, the PRC target group accompanied by the RNAPII S2P form showed alternate active and PRC-silenced states within the cell population and was significantly enriched for genes involved in metabolic processes.

Additionally, cell-intrinsic heterogeneity of allele mark variation might be a contributing factor in the numbers of genes detected as bivalent. By predicting monoallelic expression (MAE) using chromatin signatures of H3K36me3 and H3K27me3, Nag et al. (2013) have found ~20% of house-keeping genes and >30% of tissue-specific genes could confer MAE signature at their locus across multiple human cell lines. Intriguingly, more than 80% of bivalent genes in hESCs were also predicted as MAE in at least one of the used cell lines. This high overlap suggests that there is a need for single-cell allele specific approaches in order to unravel accurately the bivalent landscape.

Despite the evidence provided by sequential ChIP in T cells (Roh et al., 2006), human ESCs (De Gobbi et al., 2011; Pan et al., 2007), mouse ESCs (Voigt et al., 2012) and other cell types or organisms (Alder et al., 2010; Seenundun et al., 2010; Vastenhouw et al., 2010; Xie et al., 2012) using sonication of chromatin or micrococcal nuclease (MNase) digested mononucleosomes, their results cannot extend beyond the scope of a few single genes, making the assay inappropriate for the confirmation of the large number of bivalent genes detected through ChIP-seq. In an effort to address quantitatively the issue, Voigt et al. (2012) used mass spectrometry (MS) of ChIP-ed mononucleosomes and discovered a quite significant number of H3 histones in ESCs that carry both histone modifications. Further analysis of isolated H3 histones showed that the majority of bivalent nucleosomes in ESCs were having the tails of the opposite H3 copies modified concomitantly by H3K4me3 and H3K27me3, suggesting an asymmetrical deposition of the competing marks (Voigt et al., 2012). However, MS cannot answer the question of exact genomic location. In a new method, isolated mononucleosomes with ligated biotinylated adaptors at their DNA ends, were combined with antibody-based histone mark detection that was followed by an in-situ single molecule sequencing-by-synthesis reaction (Shema et al., 2016). This allowed the detection of combinatorial modification state of a mononucleosome along with its respective DNA sequence (Shema et al., 2016). About 0.5% of total nucleosomes in ESCs were marked by both histone modifications, being enriched relative to random expectation based on H3K4me3 and H3K27me3 abundance (Shema et al., 2016).

More strikingly, Weiner et al. (2016) have developed combinatorial ChIP (co-ChIP), a method for the genome-wide identification of co-incident histone modifications at the same nucleosome. Each of the two rounds of immunoprecipitation (IP) (each antibody separately, second round uses a pooled set of nucleosomes) are followed by DNA barcoding which allows for the identification of the specific histone marks after the tags are mapped back to the genome. Mutually exclusive histone modifications (H3K27ac and H3K27me3) showed a random distribution of reads across the genome and the order of the antibodies used in the IPs did not alter significantly the detected regions of co-existing histone marks (Weiner et al. 2016). They used co-ChIP to assess bivalent marks in a comparative manner across naïve (2i) and primed (serum) ESCs as well as in various differentiated tissues. They observed that bivalency is more widespread at the primed pluripotent state and it disappears or re-forms in a highly tissue specific manner (Weiner et al. 2016).

Despite the general credit that bivalency has received as an important regulatory characteristic of development, there is still a certain degree of controversy around it. Emergence of single nucleosome ChIP techniques accompanied by singe cell expression measurement will undoubtedly shed more light in the field.

1.1.6 RNAPII pausing and poising

In order to understand the regulatory mechanisms that produce and maintain the pluripotent state of ESCs or the precise differentiation pathways, it is pivotal to use methods that can capture the complete transcriptional activity at different steps of transcription, as the regulation of transcription happens at multiple stages in eukaryotes (Min et al., 2011). Formation of the pre-initiation complex (PIC) at the promoter, with recruitment of general TFs and the hypo-phosphorylated RNAPII, is usually followed by initiation of transcription and release of RNAPII from the promoter region. Many genes are regulated at the stage of RNAPII recruitment (Nevado et al., 1999), but genome-wide studies of RNAPII chromatin immuno-precipitation have shown that ~40% of genes maintain high levels of RNAPII localized at their 5' end (Guenther et al., 2007; Kim et al., 2005; Muse et al., 2007; Rahl et al., 2010; Zeitlinger et al., 2007). This is due to either the formation of a paused RNAPII complex (with the assistance of DRB sensitivity-inducing factor (DSIF) and Negative Elongation Factor (NELF) protein complexes) or a transcriptionally arrested complex immediately after the PIC formation. Even though these are distinct transcriptional regulation steps, they are effectively indistinguishable through the ChIP assays (Adelman et al., 2005; Rougvie and Lis, 1988).

When the early transcription elongation complex is controlled by NELF and DSIF, RNAPII stops elongating after producing a short nascent transcript of about 25-50 nucleotides and is held firmly at the promoter proximal region, a phenomenon known as RNAPII pausing (Cheng and Price, 2007; Williams et al., 2015). RNAPII pausing has been proposed as a mechanism which facilitates the poised state of several bivalent promoters of developmental genes. Preliminary RNAPII loading of those promoters would offer them an advantage in the anticipation of activation according to the appropriate developmental signals. (Adelman and Lis, 2012; Bernstein et al., 2006a; Brookes et al., 2012a; Ku et al., 2008a).

The CTD of the biggest RNAPII subunit comprises of 52 repeats of a heptapeptide sequence (Y1-S2-P3-T4-S5-P6-S7) that is subject to modifications by the Positive Transcription Elongation Factor b (P-TEFb). The distinct CTD modifications can attract chromatin modifying enzymes and RNA processing factors that could lead to gene activation (Brookes and Pombo, 2009). Phosphorylation of Serine 5 residue (S5P) is recognized by histone methyltransferase (HMT) SET1, which deposits the H3K4me3 mark, and by the RNA capping machinery (Komarnitsky et al., 2000; Ng et al., 2003). Serine 2 phosphorylation (S2P) is associated with elongation, H3K36me3 HMTs recruitment, splicing and polyadenylation (Krogan et al., 2003; Proudfoot et al., 2002).

ChIP-seq has been widely performed on cell populations using various antibodies to capture the conformations of these distinct RNAPII complexes thus providing the genome wide distribution profiles of RNAPII in many cell types. Brookes and Pombo (2009) have classified genes in ESCs into three categories according to their transcription levels and RNAPII variant profiles at their promoters (Figure 1.5). Low levels of RNAPII S5P and 8WG16 (recognizes hypo-phosphorylated CTD of RNAPII) (Komarnitsky, Cho, & Buratowski, 2000a) are observed in a very small confined region at the promoters of paused genes. Active genes, on the other hand, show increased levels of RNAPII S5P and S2P that extend into the gene body. 8WG16 is also present, but only at the promoter region. Lastly, poised genes show high levels of RNAPII S5P solely. All three transcriptional states show also characteristic histone mark profiles. H3K4me3 is ubiquitous in all three states, changing its position slightly depending on the promoter of the gene. Paused genes have a distribution of H3K4me3 that matches exactly the confined promoter profiles of S5P and 8WG16. In active genes, H3K4me3 is wider at the promoter and is accompanied by H3K36me3 at the gene body. The poised promoters show a characteristic bivalent histone mark combination of H3K4me3 and H3K27me3. Thus, there is a clear association between histone modifications and RNAPII pausing characteristics with a likelihood that some of the marks (i.e. H3K36me3) are deposited due to the effect of RNAPII elongation and not the opposite (Brookes and Pombo, 2009). The terms ‘paused’ and ‘poised’ are used in a variety of contexts in scientific literature. Here, we define poised promoters as occupied by H3K4me3, H3K27me3 and having preloaded RNAPII, being in a ready state for transcription while paused promoters show variable levels of RNAPII S5P, high pausing index and are mostly H3K4me3 marked, but sometimes bivalent as well.

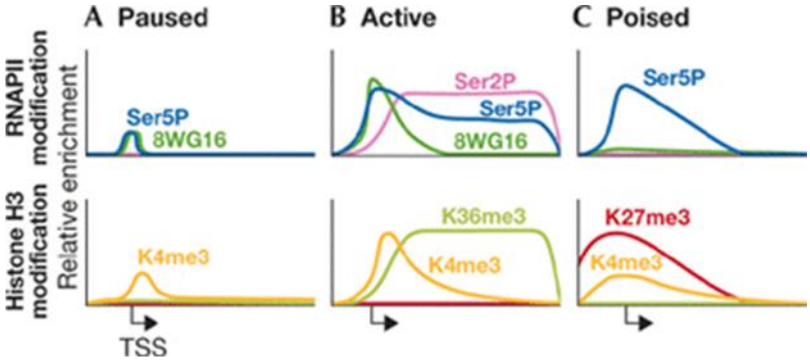


Figure 1.5 RNAPII variants and histone modification profiles at promoters classified according to their expression potential (figure taken from Brookes & Pombo, 2009)

The typical measurement widely used to infer the relationship between pausing and elongation is the pausing index. The pausing index or travelling ratio is given by the ratio of RNAPII density at the promoter to the density in the gene body, defined in the following formula (Muse et al., 2007):

$$S = log_2\left(d\left(RNAPII_{promoter}\right)\right) - log_2(d(RNAPII_{genebody}))$$

where d stands for the number of reads per nucleotide (nt) in the given region.

The difference between the densities in log base 2 units equals to the ratio of fold enrichment in these regions, meaning a value of 1 would represent a 2-fold greater enrichment of RNAPII signal at the promoter region rather than in the gene body (Muse et al., 2007). The genes with S value greater than 2 standard deviations from the mean (for the distribution of S across all genes), are the ones that present promoter proximal enrichment of RNAPII or promoter proximal pausing (Muse et al., 2007).

Due to inherent difficulty of imposing a threshold on continuous data and the differences in the underlying methods used in RNAPII experiments, there have been studies reporting that ~30% to ~90% of mouse ESCs genes present promoter proximal pausing (Adelman and Lis, 2012). This variance might not reflect actual biological differences, but the various statistical thresholds imposed to define RNAPII pausing. Using the global run-on sequencing (GRO-seq)[2] method (Core et al., 2008), there has

[2] GRO-seq is a method used to measure the RNAPII elongation rate genome-wide. Short nascent RNAs associated with engaged RNAPII are tagged with bromo (BrU) domains, isolated and subsequently sequenced with Next Generation sequencing techniques (Core et al., 2008; Jonkers and Lis, 2015).

been a consistent ~30% of genes displaying promoter proximal pausing of RNAPII across species and developmental stages (Adelman and Lis, 2012).

In contrast with the perception that RNAPII pausing represents a gene silencing mechanism, RNAPII pausing has been found to occur at genes with wide gene expression range (Core et al., 2008; Min et al., 2011). RNAPII pausing reduction at the promoter does not always lead to increased gene expression, but it can often induce the opposite results (Min et al., 2011). Bivalent genes occupied by PRC components show variable levels of RNAPII variants, being mostly interconnected with the RNAPII-S5P species (Brookes et al., 2012a). Interestingly, there are also genes that not only feature PRC occupancy but also relatively high levels of gene expression, the elongating form of RNAPII (S2P) and H3K36me3 (related with transcription elongation as mentioned before). This intriguing conformation at the promoters of several metabolic and developmental genes could be attributed to differences in the alleles of the genes or even cellular heterogeneity in ESCs (Brookes et al., 2012a). Furthermore, using ESCs grown in 2i media, where ESCs show high levels of homogeneity since the expression of lineage markers is low (no priming) (Marks et al., 2012), RNAPII pausing was mainly found occurring at cell cycle and signal transduction genes (Williams et al., 2015). There was no silencing of developmental genes observed due to RNAPII poising, but there was attenuation of differentiation pathways due to lack of it (Williams et al., 2015).

Overall, these data suggest that RNAPII pausing and poising are associated with the fine-tuning of expression of genes that participate in signalling networks, regulating in turn developmental genes marked as bivalent and affecting the cell’s differentiation potential.

1.2 Next Generation Sequencing and applications

Over the past decade, Next Generation Sequencing (NGS) technologies have increased tremendously the range of genomic analyses that are available to laboratories around the world. One of the important factors being the cost, NGS technologies have dramatically reduced sequencing costs in comparison with the automated Sanger method, almost 100-fold, from $10.00 to $0.10 per finished base pair (Metzker, 2010;

Morozova and Marra, 2008; Wetterstrand K., 2013). In the recent years, costs dropped significantly more and the record low of $1000 for sequencing an entire human genome was achieved, making sequencing one of the most promising clinical tools (Goodwin et al., 2016).

NGS technologies can be sub-grouped in the short-read and long-read approaches. Short read sequencing approaches are further divided in two categories: sequencing by ligation (SBL) and sequencing by synthesis (SBS). SBL platforms include SOLiD (Valouev et al., 2008) and Complete Genomics (Drmanac et al., 2010), whereas SBS platforms include Illumina and Qiagen (cyclic reversible termination - CRT)(Guo et al., 2008; Ju et al., 2006), 454 and Ion Torrent (single nucleotide addition - SNA) (Margulies et al., 2005; Rothberg et al., 2011). In long read sequencing platforms, single-molecule long-read sequencing platforms (PacBio and Oxford Nanopore Technologies) (Clarke et al., 2009; Eid et al., 2009) compete with synthetic long-read technologies (Illumina and 10x Genomics) (Voskoboynik et al., 2013).

Illumina currently holds the largest share in the short read sequencing industry, due to its mature technology, flexibility between platforms and wide application range. Diversified sequencing instruments such as MiniSeq (lower throughput) to HiSeq X (the latest high throughput sequencer), offer many options to laboratories, adapted to their needs for runtimes, read lengths and budgets (Goodwin et al., 2016; Metzker, 2010). Illumina short-read technology and its applications relevant to the thesis are described below.

The Illumina CRT system belongs in the category of clonal template generation approaches. A DNA template (after DNA fragmentation) binds covalently through the adapter sequence to oligos found on a glass slide containing a number of lanes (flow cell). Amplification of templates (solid-phase bridge amplification) leads to the formation of clusters of templates, placed in great proximity with each other, but they do not overlap. After the completion of several amplification rounds, tens to hundreds of millions of clusters are formed on the flow cell, depending on the sequencing platform (Goodwin et al., 2016). The sequencing process then starts with an addition of a sequence complementary to the adapter region, to facilitate polymerase binding to the double stranded DNA. Each cycle is comprised of a terminally blocked, fluorophore-labelled nucleotide addition, followed by imaging in four or two laser channels of the colour emitted, depending on the added base. Lastly, fluorophore cleavage and washing from the flow cells is followed by 3'-OH group regeneration (Goodwin et al., 2016). The flow cell clusters are then sequenced in a massively parallel manner.

NGS raw data are stored as image data, sequence reads (FASTQ format[3]) or as aligned reads to the genome (SAM/BAM format[4]) (Park, 2009). Storing the image data could be useful, to keep up to date with the development of new base calling technologies, albeit expensive. It is a common policy to discard the image data and keep only the sequence reads in FASTQ format. Labs producing data from various NGS experiments, are obliged to deposit them to public databases such as Gene Expression Omnibus (GEO) accompanying their publications (Barrett et al., 2013). Because of large file sizes, FASTQ data uploading or downloading from GEO repository can result in failure. Thus, the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the DNA Databank of Japan (DDBJ), generated databases assuring that the research communities worldwide could retrieve sequencing data along with other useful details for each experiment (meta-data) in a secure and robust way. They have created the Sequence Read Archive (SRA) which allows retrieval, movement and storage of large scale NGS datasets (Cochrane et al., 2009; Sayers et al., 2009).

Out of the many applications of NGS, here we introduce epigenetic applications such as Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) (Park, 2009) and transcriptomics applications such as RNA-seq (Wang et al., 2009). Both applications were used in our project.

[3] FASTQ format was initially developed by the Wellcome Trust Sanger Institute as a means of storing a FASTA sequence along with its quality score per base. It has now been recognised as the standard file format for storing the output of NGS instruments (Cock et al., 2010).

[4] SAM format stands for Sequence Alignment Map and it is a text-based representation of biological sequences mapping to a reference genome. BAM stands for Binary Alignment Map and has been developed as a
means of compressing data from SAM format (Li et al., 2009).

1.2.1 ChIP-seq

The ChIP sequencing technology successfully achieves mapping of protein-DNA interaction at the exact chromosomal locations that they occur. ChIP-seq (Barski et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007) has been used for the profiling of histone modifications and protein-binding sites, allowing a better mapping of TF target regions (promoters, enhancers) and subsequent identification of specific sequence motifs in case of TFs.

In a typical ChIP-seq experiment, the cross-linking of a DNA-binding protein to DNA in vivo is followed by chromatin shearing using sonication. DNA fragments range between 200-600 bp in length. IP of the DNA-protein complexes is carried out using an antibody that specifically recognises the protein of interest or one of its isoforms. Experiments targeting specific histone modifications can be achieved without cross-linking, and sonication is frequently substituted by MNase digestion for the chromatin fragmentation. Thus, a more accurate representation of each nucleosome is achieved. In the event of cross-linking, after a reverse cross-linking step, DNA fragments are purified and the sequencing library is constructed according to the steps of the sequencing platform (Park, 2009). Illumina Genome Analyzer (I and II) and HiSeq 2000 are the most prevalently used platforms for most ChIP-seq experiments (Park, 2009). Issues concerning experimental designs (lack of replicates, lack of positive and negative controls, lack of validation, etc) plague the ChIP-seq assay, like many other assays, often affecting the integrity of results. Some of the factors are elaborated below.

Antibody specificity: Variable sensitivity of antibodies that fluctuate considerably between batches can easily affect the quality of a ChIP-seq experiment. Cross-reactive antibodies recognizing more than one histone modification (i.e. di- and tri- methylation of lysine residues) or TFs is a common phenomenon that needs thorough testing. Validating different antibodies (for example through western blotting) before the choice of the most suitable one is highly recommended (Park, 2009). There have been cases where comparison between experiments that have used different antibodies for the same protein has not yielded a high degree of overlap (Devailly et al., 2015).

Sample sizes: The average amount of DNA necessary for a ChIP-seq experiment using the Illumina platform ranges around 50 ng. However, the initial DNA material for an experiment is highly dependent on the quality of antibody used and quantity of the target protein or histone modification (Park, 2009).

Control experiments: Various steps along the ChIP-seq pipeline are prone to incorporate some artefacts. DNA fragmentation, for example, is highly biased towards open chromatin regions, resulting in a non-uniform distribution of DNA fragments. In order to tackle this limitation, labs have been using control samples. These samples can be: 1. Input DNA, where part of the initial DNA material is used before IP, 2. Mock IP DNA, which is taken after IP but without the use of antibodies and 3. DNA from non-specific IP, where an antibody such as immunoglobulin G is used, targeting a protein not involved in DNA binding or identification of a histone modification (Park, 2009). Unfortunately, inherent difficulties in purification of adequate material from the input experiments, in order to acquire high numbers of reads (trying to avoid sampling bias), lead to higher costs. Consequently, some labs might avoid it and try to deal with the problem at a later stage during the computational analysis.

Sequencing depth: The number of sequenced reads for a ChIP-seq experiment is highly reliant on the financial resources of each lab. It is expected that if a protein binds more across the genome or a histone modification covers broader regions, more reads should be necessary to cover completely the corresponding regions, resulting in higher enrichment levels compared with a control sample. One would expect that after a certain number of reads, there should not be more significantly enriched binding sites (or histone marks) to be detected. Hence, there should be a sequencing depth value where a saturation point could be met. Although there have been arguments against this saturation point using simulations after read sampling, peak enrichment threshold between experiment and control can confirm that saturation exists only at peaks with significantly higher enrichment than the control (Kharchenko et al., 2008). Currently, most ChIP-seq experiments in public domain have a depth between 10 and 100 million reads.

Spike-in normalisation: The nature of ChIP-seq method is not quantitative enough to allow for a direct comparison between samples coming from non-identical cell types or cells that have been chemically treated (Orlando et al., 2014). Hence, differential TF binding or global histone modification landscape changes cannot be adequately assessed only with a sample size normalisation. This would merely quantify the distinct sample signals as a percentage of the total number of mapped reads, potentially equating them incorrectly and missing valuable information (Chen et al., 2015).

Adding ‘spike-in’ epigenome molecules as a reference can internally normalise the read counts of each sample. The choice of the reference molecules should be from a distant species from the species of interest, where a minimal mapping of one genome to another should take place. A known abundance of the reference epigenome can accurately determine when the global level of a histone modification would decrease in the sample of our interest due to the increasing number of total reads mapping back to the reference epigenome or ‘spike-in’ (Orlando et al., 2014). However, this method is still dependent on the antibodies used for IP, which they need to recognise the protein isoforms from both species (Orlando et al., 2014).

Data analysis and management: The most important step in the downstream computational analysis of ChIP-seq data is the alignment to the reference genome of the organism. All the subsequent results are directly dependent upon the alignment outcome. There are several aligning algorithms suitable for this kind of data (Trapnell and Salzberg, 2009) such as Eland (default Illumina pipeline aligner), MAQ (Li et al., 2008) (more suitable for single nucleotide variant (SNV) calling) and Bowtie (Langmead et al., 2009) which is known for its speed and efficiency and was the chosen aligner in our study. Interestingly, many aligners have options that do not allow the alignment of reads that map to more than one site in the genome. Considering that many areas of the mammalian genomes can be quite repetitive, for example TFBSs and promoters of lineage specific genes (Huda et al., 2009; Polavarapu et al., 2008), one needs to be cautious with the parameter choices of the alignment step.

The enriched regions are subsequently called with the assistance of peak calling algorithms (Kharchenko et al., 2008; Xu et al., 2014; Zhang et al., 2008). Genome regions selected for identification of greater numbers of reads than in the control are called peaks. A simple fold enrichment ratio of ChIPed tags in a sample in comparison to the control does not contain enough information or account for bias of the region such as the underlying chromatin structure. There are peak calling algorithms using distribution models or the inherent bi-directionality of the reads to tackle the problem.

The Poisson model used by MACS, a peak caller widely used for protein binding peak detection, effectively combines the fold enrichment ratio and the absolute number of reads at the region (Zhang et al., 2008). Using the inherent characteristics of ChIP-seq experiments, such as read bi-directionality and size, distributions of reads mapping to both strands are formed (Schmid and Bucher, 2007). In a putative peak site, the reads mapping at each strand are added together after they have been extended towards the centre intersection of the two distributions, and the combined signal is smoothed. From the comparison of the underlying signal probability distributions (using dynamic λ for the Poisson distribution, which allows for capturing of the local biases) of the actual experiment and the input control, false discovery rates (FDR) are calculated and account for the significance of the enriched sites.

Protein binding sites usually have a sequence length of 4 to 30 bp (Borneman et al., 2007), which is not clearly portrayed at the combined fragment of ChIP-seq reads that can reach up to a few hundred base pairs. MACS empirically models the distance between the aligned reads at the forward and reverse strands, shifting all the reads by “distance/2” towards the 3’ prime, bringing them closer to the putative binding site (Zhang et al., 2008). The peak resolution is highly improved and a more accurate distance from the peak summit is computed.

However, the majority of peak callers do not account for the variant width of histone modifications, which can be sharp, broad or mixed. Histone modification peaks for H3K27me3 are mostly broad, extending in many kilobases, in contrast with H3K4me3 which is located sharply at promoter proximal regions (Park, 2009). SICER (Xu et al., 2014) uses a spatial clustering method specialized on the identification of histone modification patterns. By combining multiple signals from nucleosomes that are recognised with the same histone modification, it increases the signal to noise ratio, especially for regions where the signal is not sufficiently higher than the input but covers a wide region. Enrichment scores are subsequently calculated for each putative peak, comparing the signals of the ChIP experiment with the signal in the respective input control.

Scores accounting for statistical significance of the peaks in the experiment are highly dependent on the quality of the input control experiment. When there is no control experiment available, randomised reads from the ChIP experiment can be used as the null distribution instead. In a ChIP experiment the reads that map to adjacent regions on the genome are not entirely independent, affecting directly the generated random distribution used as a control and subsequently the number of significantly enriched regions (Park, 2009).

In this thesis, we have chosen to use SICER for the histone marks ChIP-seq experiments and MACS for the protein-binding sites.

1.2.2 Transcriptome sequencing

Modern transcriptomics studies aim to discover all transcript species including mRNAs, non-coding RNAs or small RNAs, and their accurate quantification between distinct developmental stages or conditions (Wang et al., 2009). The advent of NGS methods has led to the development of RNA-seq, a method for transcriptome mapping and quantification that is superior to the previously used hybridization-based approaches, like genomic tilling microarrays (Bertone et al., 2004; Clark et al., 2002) or Expressed Sequenced Tags (ESTs) that require complementary DNA (cDNA) cloning (Adams et al., 1991). In a typical RNA-seq experiment, an mRNA population (depleted of ribosomal RNA (rRNA) or enriched for poly-adenylated RNAs) is fragmented and converted to a cDNA library. Next, adaptors are ligated to one or both sides of the cDNA fragments, which are later amplified and sequenced with the respective sequencing platform. Short read sequences can be obtained either from one or both ends of the fragment, hence the names single and paired end sequencing (Wang et al., 2009). RNA-seq can be utilized for discovery of novel transcripts, differential expression between samples and alternative splicing studies (Sims et al., 2014). Despite the advantages of RNA-seq to older transcript quantification methods, there are some challenges using this technology.

Sequencing depth: Coding and non-coding transcripts show different levels of expression, requiring a minimal number of reads to be precisely quantified. The ability to detect transcripts of lower abundance is determined by the sequencing depth of a sample and is not known a priori. According to Tarazona et al. (2011), more than 200 million paired-end reads would adequately detect all transcripts and alternative isoforms in a human sample.

Spike-in normalisation: In all genome-wide expression analyses, there is an underlying hypothesis that wants the populations of cells in comparison, to generate total RNA at very similar levels (Lovén et al., 2012). Transcription signal normalisation is akin to the fact that some cells can produce two to three times more RNA in total, making them unsuitable to compare with other cells of lower RNA production (Lovén et al., 2012). Spike-in control can safeguard and address this issue with a normalisation to cell number of each population. The use of synthetic spike-ins (Jiang et al., 2011b) or known transcripts of known quantity from a distant species can prevent the introduction of bias that the usual normalisation to the average brings (reads per million of transcripts).

Biological replicates: Given cost limitations, in a differential expression analysis set up, it is a common practice to sacrifice the precision feasible by higher sequencing depth for the sake of biological replicates. Even though this compromise offers a more precise portrayal of the biological variation, thus allowing for a more robust detection of differentially expressed genes, the number of replicates must be carefully considered (Sims et al., 2014). In a recent study, investigative analysis of 48 biological replicate experiments has shown that 3 or more replicates could be enough to avoid the misinterpretations caused by unsuccessful replicates (Gierliński et al., 2015).

Library construction: Large RNA molecules (>200 bp) have to be fragmented before adaptor ligation and subsequent sequencing. On the other hand, many small RNAs, like microRNAs (miRNAs) or piwi-interacting RNAs (piRNAs) directly bind to the adaptor without prior fragmentation. RNA fragmentation or cDNA fragmentation are known for incorporating different degrees of bias in the library construction process by favouring the identification of the transcript body (Mortazavi et al., 2008) or the 3’ prime end of the transcript (Nagalakshmi et al., 2008) respectively.

Bioinformatics challenges: Data generated from RNA-seq technology faces similar limitations as other high throughput sequencing applications, namely storage, fast retrieval and easy processing of the data are the main issues. Fast processing pipelines have been implemented to align the data when a reference genome is available and complete a successful statistical analysis in a reasonable length of time.

Identification and quantification of transcripts is a challenging task, starting from the step of RNA-seq read alignment. There are two routes depending on the availability of a reference genome: mapping to the reference genome or to the annotated transcriptome (Conesa et al., 2016). Genome mapping allows for the identification or discovery of novel transcripts and their quantification, whereas transcriptome mapping is limited to identify and quantify already annotated transcripts. When mapping to a reference genome, gapped aligners such as TopHat (Trapnell et al., 2009) or STAR (Dobin et al., 2013) are used as a first step. Then using an annotation file (GFF[5]), known transcripts are identified and counted. For novel transcripts no annotation file is necessary. Cufflinks (Trapnell et al., 2012) is one of the traditionally used programs for this step of the analysis. The common measure for the transcript abundance is Fragments per Kilobase of transcript per Million mapped reads (FPKM), which is a way to normalize the data by RNA length and total number of reads enabling the comparison between samples.

Mapping to a reference transcriptome (FASTA file of annotated transcripts) is done by first using an un-gapped aligner such as Bowtie (Langmead et al., 2009). Programs such as RSEM (Li et al., 2011) and kallisto (Bray et al., 2016) are used for the next step of transcript identification and counting. In both methods described above, there is a major limitation when reads map to more than one location. This happens extremely often, due to high repetitiveness of the genome and existence of paralogous genes. The transcriptome mapping is more prone to genomic multi-mapping, since a read can map multiple times to transcripts that share one exon (Conesa et al., 2016).

When there is no reference genome available for the organism studied, de novo transcript aligning algorithms are used. Packages such as Trinity (Brown et al., 2012) are used to reconstruct ab-initio the reads into transcriptome contigs. Next, the reads

[5] GFF format stands for General Feature Format and is used for gene annotation, containing tabular information with 9 fields per line. Names of the features in order of their appearance are: sequence, source, feature (i.e. “gene” or “exon”), start, end, score, strand, frame or phase and attributes.

are remapped to the resulting transcriptome from the first step using and un-gapped aligner, followed by quantification and functional annotation (Conesa et al., 2016).

Overall, RNA-seq is a great tool in the hands of modern biology scientists. Until recent years, the transcriptome profiles of multiple types of cells and tissues were generated using bulk cell populations and quantifying an average number of transcripts across the population (Shapiro et al., 2013). Nevertheless, in many cases in biology the underlying cell population is composed of distinct cell types and their stochastic expression needs to be systematically assayed (Raj et al., 2008). Single cell RNA-seq (scRNA-seq) (Tang et al., 2010) is a recently developed advanced technique that opens new ways in answering posed biological questions that were previously impossible to reach.

1.2.3 Single cell transcriptomics (scRNA-seq)

Single cell RNA-seq (scRNA-seq), is a high throughput gene expression profiling technology that is based on material provided by a single cell (Islam et al., 2011; Ramsköld et al., 2012; Tang et al., 2009). Novel biological insights can be obtained from a deeper characterization of a cell population, uncovering hidden heterogeneity. Small numbers of rare cells can now be clustered according to their expression levels and comprise potential categories of previously undocumented cells (Stegle et al., 2015). Using single cell RNA-seq, cells are classified according to their developmental pathways and put in order across the differentiation cascade. Moreover, differential expression analysis and estimation of alternative transcript usage among the cell types can identify new marker genes, coming up with novel regulatory networks impossible to detect without the scRNA-seq technology.

Another interesting aspect of transcription that is explored using scRNA-seq technology is kinetics of gene expression. Even though the method is not appropriately measuring expression changes in one gene over time (Raj et al., 2006), overall rate of transcription between individual cells can be acquired and approximately represent the stochasticity of expression of a vast number of genes. Allelic biases in gene expression can also be investigated with studies already examining stochastic allelic expression in early embryogenesis (Deng et al., 2014).

Protocols used for this high-throughput technique require single cell isolation either sorting them manually or using microfluidics (Kalisky and Quake, 2011). The cell is then lysed and the captured polyadenylated-mRNAs are reversed transcribed to cDNA. PCR amplification, or in vitro transcription of the cDNA, increases the material that is going to be used for sequencing with the appropriate NGS technology (Tang et al., 2009). Bulk RNA-seq data analysis has given us multiple tools that can be used for scRNA-seq data as well. However, the new technology carries some unique features that create challenges specific to it.

Quality control of individual cells: The same tools can be used in order to map the reads to the reference genome, as in bulk RNA-seq data. Inspection of the fraction of reads mapping back to the genome can be a valuable metric of the library quality for that cell. Another complementary approach uses principal component analysis (PCA) or a similar approach such as zero-inflated factor analysis (ZIFA- takes into account the high numbers of zero values), performed on the matrix of the obtained gene expression values. High quality cells are expected to form tight clusters whereas low-quality ones tend to be the outliers. There are however cases, where poor-quality cells form their own clusters, blurring the lines. A comprehensive quality control with more than one metric is sensible for discarding true low quality samples (Stegle et al., 2015). Discarding results from multiple cells can have a negative effect on someone’s study, since the number of singe cells used in any experiment is already limited due to high costs. A simulation study tried to uncover the threshold up to which single cells could approach the transcriptome library complexity of bulk cell populations, and has concluded that 30 cells could be enough (Marinov et al., 2014).

Technical variation: scRNA-seq expression outcomes are highly susceptible to technical noise dependent on the single-molecule capture efficiency, i.e. the fraction of mRNA molecules that are captured, amplified and subsequently sequenced from each cell (Marinov et al., 2014; Stegle et al., 2015). To address this limitation spike-in quantification of molecules of known abundance and sequence have been used (Mortazavi et al., 2008). Whole-transcriptome spikes from a distantly related organism or a set of artificial spike-in mix (ERRC) (Jiang et al., 2011b) is added to each cell extract. Another layer of control is the use of unique molecular identifiers (UMIs), which are short DNA sequences added before the step of amplification uncovering related biases and estimating the absolute number of some of the transcribed molecules (Islam et al., 2013). Extending the measurement of technical noise, pool/split experiments are yet another way of control. RNA pooled from multiple cells is subsequently divided in separate reactions consisting of equal material, which is used for the libraries' construction. Variation in the pool/split experiment will be solely attributed to technical noise whereas in the single cell libraries with spike-ins is both technical and biological (Marinov et al., 2014). Genes with low levels of expression tend to suffer from high experimental noise, in contrast with highly expressed genes.

Batch effects: In contrast with the conventional bulk RNA-seq methods, in single cell transcriptomics, cells representing one condition are not prepared for sequencing along with the cells of another condition. There is no parallel preparation of their libraries, which are not dispersed in multiple lanes of a flow cell as it would happen in a bulk RNA-seq experiment in an effort to moderate batch effect as much as possible (Stegle et al., 2015). This inevitably leads to unwanted mixing of batch effects with biological variation and consequently wrong conclusions. An impossible in-parallel capture of cells for multiple conditions could only be resolved with several repetitions of each experiment, isolating cells for each condition multiple times. In this way, modelling of confounding variation could be more feasible since there are multiple replicates of cells for the same condition.

Cell cycle variation: Cell heterogeneity in a sample of differentiating cells reflect biological variation related with differentiation signals obscuring the other biological variability under study. If the cells in the sample are not synchronized for a cell cycle stage, there is a need for modelling of the cell cycle noise. Recently, cell-cycle variation was modelled using Gaussian processes, followed by linear regression, thus allowing the removal of noise caused by cell cycle (Buettner et al., 2015).

1.3 Aims of study

The advent of low-cost next generation sequencing technologies has contributed to the development of a plethora of genomic, epigenomic and transcriptomic datasets in multiple cell types, tissues and species. Bivalent chromatin is one of the most studied chromatin signatures, however the concerns raised around its conformation and importance during development have not been adequately answered. The associated studies published so far, have mainly investigated bivalency in isolation, therefore a comprehensive comparison between species and cell types is pertinent.

In this study we set with a goal to investigate bivalent chromatin in mammalian cells and reveal its characteristics, through a meta-analysis of publicly available relevant datasets. Our primary aim is to develop a method for the identification of a robust list of bivalent promoters that would prevent outliers of each incorporated study permeating in our results. Initially, the use of data derived from mouse and human ESCs is imperative, as we need a point of reference for the detection of high-confident bivalent promoters. Next, we characterize bivalent promoters regarding their conservation across species, biological function, TF occupancy, sequence motif and expression. We also include single cell transcriptomics to address the heterogeneity of ESC populations and inspect the expression variation of bivalent genes. Last, we investigate the dynamics of epigenetic states and RNAPII pausing at promoters across pluripotent and differentiated cell lineages.

We aim to develop a useful resource for further studies trying to unveil the aspects of bivalent chromatin and its association with transcriptional regulation.

Chapter 2 Comparative analysis of bivalent domains in mammalian embryonic stem cells

2.1 Chapter Introduction

This chapter was published as a conference publication (Third International Conference, IWBBIO 2015, Granada, Spain, April 15-17, 2015, Proceedings, Part I) in the book with title Bioinformatics and Biomedical Engineering, under the DOI: 10.1007/978-3-319-16483-0_39. Here we introduce a comparison between two approaches we have used for the detection of high confident bivalent promoters. The conclusions of this analysis were subsequently used in Chapter 3 where we present a more detailed characterization of bivalent promoters.

2.2 Introduction

The key cellular processes determining the fate of each cell type during development and differentiation are thought to be controlled by gene regulation (Pearson et al., 2005). Genomic regulatory elements such as promoters receive and execute transcriptional signals, dependent on their epigenetic state and chromatin accessibility, controlling the expression of key developmental factors (Wilson et al., 2010). Apart from the transcription control at the promoters and enhancers, gene expression is also controlled epigenetically, by post-translational histone modifications, which transform the chromatin structure and thereby control gene expression (Bannister and Kouzarides, 2011).

To unravel key developmental transitions that lead to different types of cell identities, embryonic stem cells (ESCs) offer a valuable model (Thomson et al., 1998) as they have an unlimited potential to self-renew as well as to differentiate in specific lineages when suitable external stimuli are provided. In ESCs, the majority of promoters with high CG content are un-methylated. During differentiation though, some of them become methylated, assisting to the acquisition of their final cell identity (Mohn et al., 2008). Azuara et al., 2006 proposed that particular histone modifications and chromatin structure (Thomson et al., 1998; Voigt et al., 2013) are characteristic of ESCs. Two of the most commonly studied histone modifications related to activation and repression of chromatin respectively are H3K4me3 and H3K27me3 (Bannister and Kouzarides, 2011). Polycomb (PcG) and Trithorax (TrxG) group proteins catalyze H3K27me3 and H3K4me3 respectively, regulating genes involved in development and differentiation (Ringrose and Paro, 2004). Bernstein et al., 2006 observed activating (H3K4me3) and repressing (H3K27me3) chromatin signals in promoters of several developmentally regulated genes in murine ESCs. These activating and repressive marks were previously thought to be mutually exclusive and therefore the promoters marked with both modifications were named ‘bivalent’. Mikkelsen et al., 2007 used the ChIP sequencing technique to examine the bivalent status and construct chromatin state maps across three cell types: mouse ESCs, mouse neural progenitor cells (NPCs) and mouse embryonic fibroblasts (MEFs). Their study showed for the first time, that bivalent domains also exist in cells of restricted potency and 8-43% of them retained their signature during differentiation (Mikkelsen et al., 2007). Moreover, Mohn et al., 2008 indicated that bivalent genes that are not present in the pluripotent cells may arise in reduced potency cells.

Bivalent genes were detected also in human ESCs (Pan et al., 2007; Zhao et al., 2007) and the majority of them were shared with bivalent genes in mouse ESCs. Specifically, in two out of three studies there were 2,157 common bivalent genes (Mikkelsen et al., 2007; Pan et al., 2007; Sharov and Ko, 2007; Zhao et al., 2007). In agreement with the studies in mice, human ESCs bivalent genes were functionally enriched with developmental transcription factors and genes and most of them lose the repressive H3K27me3 mark during differentiation (Pan et al., 2007; Zhao et al., 2007).

ESCs employ various mechanisms to avoid losing their pluripotency. For example, they manage to prevent DNA methylation that would silence important genes indefinitely. Bivalent genes belong to an important category of genes full of developmental factors that need to be poised for activation or repression at the right moment during the differentiation process (Voigt et al. 2013). The bivalent state preserves the plasticity of the developmental genes until certain environmental cues lead to proper differentiation.

Though bivalent genes have been identified across multiple species in ESCs as well as differentiated cells, there is no study so far collecting multiple data sets to build a high-confidence bivalent gene set. We therefore collected genome wide ChIP-seq data of H3K4me3 and H3K27me3 in murine ESCs from eight different studies. We then used two complementary approaches; peak-based and cutoff-based approach to define high confidence bivalent promoters. The high confidence bivalent promoters detected by peak-based method were more enriched for developmental genes than the cutoff-based. Finally, we collected data to identify bivalent promoters in human ESCs and pig induced pluripotent cells (iPSCs) to study the evolutionary conservation of bivalency. By performing the comparative analysis of bivalent domains across three species we highlighted the functional relevance of coexistence of these marks on the developmental promoters.

2.3 Methods

2.3.1 Data collection and processing

Murine ChIP sequencing data for H3K4me3 and H3K27me3 histone marks in ESCs was obtained in fastq format from Gene Expression Omnibus (GEO) database (Barrett et al., 2013). Accession numbers for mouse are: SRX001923, SRX001921 (Mikkelsen et al., 2007), SRX185810, SRX085431 (Yue et al., 2014), SRX122629, SRX122633 (Yu et al., 2013), SRX172574, SRX172569 (Jia et al., 2012), SRX266816, SRX266814, SRX266817, SRX266815 (Cao et al., 2013), SRX305910, SRX305921, SRX305911 and SRX305922 (Wamstad et al., 2012). Details for accession numbers, antibodies used and cell lines shown in Table 2.2.

Human ChIP-Seq data (fastq format) for H3K4me3 and H3K27me3 histone marks in hESCs was obtained from Roadmap Epigenomics (Bernstein et al., 2010) and Gene Expression Omnibus (GEO). Accession numbers for human are: SRX003843, SRX003845 (Ku et al., 2008a), SRX006874, SRX006237, SRX012368, SRX012501, SRX027857, SRX027864, SRX040598, SRX027865, SRX056700, SRX056719 (Bernstein et al., 2010; Hawkins et al., 2010), SRX007379, SRX007385, SRX019898, SRX019896 (Bernstein et al., 2010), SRX027484, SRX027487 (Rada-Iglesias et al., 2011), SRX064487, SRX064486 (Kim et al., 2011), SRX189254, SRX189253 (Akdemir et al., 2014). Details for accession numbers, antibodies used and cell lines shown in Table 2.1.

HUMAN						
Samples	H3K27me3 SRA EXPERIMENT	Antibody	H3K4me3 SRA EXPERIMENT	Antibody	Cell line	Growth medium
GSE13084_1	SRX003843	Upstate 07-449	SRX003845	Abcam 8580	H1	DMEM
GSE16256_1	SRX006874	Upstate 07-449	SRX006237	Millipore 04-745	H1	CDI-protocol
GSE16256_2	SRX012368	Upstate 07-449	SRX012501	Abcam 8580	H1	mTESR
GSE16256_3	SRX027857	Millipore 07-449	SRX027864	Millipore 04-745	H1	mTESR
GSE16256_4	SRX040598	Millipore 07-449	SRX027865	Millipore 04-745	H9	mTESR
GSE16256_5	SRX056700	Millipore 07-449	SRX056719	Millipore 04-745	H9	mTESR
GSE17312_1	SRX007379	Upstate 07-449	SRX007385	Abcam 8580	H1	TESR
GSE17312_2	SRX019898	Millipore 07-449	SRX019896	Millipore 07-473	H1	TESR
GSE24447_1	SRX027484	Active Motif 39536	SRX027487	Active Motif 39159	H9	mTESR
GSE29422_1	SRX064487	Upstate 07-449	SRX064486	Abcam 8580	H9	DMEM
GSE39912_1	SRX189254	Millipore 07-449	SRX189253	Millipore 04-745	H9	DMEM

Table 2.1 Accession numbers, type of cell line, growth media and antibodies for the samples gathered for human ESCs (Millipore 07-449 and Upstate 07-449 have no difference)

MOUSE						
Samples	H3K27me3 SRA EXPERIMENT	Antibody	H3K4me3 SRA EXPERIMENT	Antibody	Cell line	Growth medium
GSE12241_1	SRX001921	Upstate 07-449	SRX001923	Abcam 8580	v6.5	DMEM
GSE31039_1	SRX185810	Millipore 07-449	SRX085431	Millipore 07-473	ES-Bruce4	DMEM
GSE38596_1	SRX122629	Millipore 07-449	SRX122633	Abcam 1012	ES-E14	DMEM
GSE39513_1	SRX172574	Abcam 6002	SRX172569	Cell Signaling 9751	V6.5	DMEM
GSE46134_1	SRX266816	Millipore 07-449	SRX266814	Millipore 07-473	-	DMEM
GSE46134_2	SRX266817	Millipore 07-449	SRX266815	Millipore 07-473	-	DMEM
GSE47949_1	SRX305910	Millipore 07-449	SRX305921	Millipore 07-473	ES-E14	IMDM+HAM'S F12
GSE47949_2	SRX305911	Millipore 07-449	SRX305922	Millipore 07-473	ES-E14	IMDM+HAM'S F12

Table 2.2 Accession numbers, type of cell line, growth media and antibodies for the samples gathered for mouse ESCs (Millipore 07-449 and Upstate 07-449 have no difference)

ChIP sequencing data for H3K4me3 and H3K27me3 in pig (Sus Scrofa) induced pluripotent stem cells (iPSCs) was downloaded from a published study with accession number GSE36114 (Xiao et al., 2012). After downloading all the raw sequence files for all the experiments, each technical and biological replicate was imported into FastQC 0.10.1 (S. Andrews, 2010) for quality control. Alignment of reads was done using Bowtie 0.12.9 (Langmead et al., 2009) using reference genomes mm10 for mouse, hg19 for human and susScr3 for pig. For all the species we used single end alignment, seed length=28. We then performed the bowtie execution using custom bash scripts and the samtools (Li et al., 2009) pipeline to convert the sam format files to bam format for each sample. The bam files that belonged to the same experiment (technical replicates) were merged into a common bam file in order to proceed with the further analysis. The biological replicates of each experiment were not merged. We downloaded the Gencode (Harrow et al., 2012a) genes for human (Gencode 19) and mouse (Gencode M2). We filtered out and kept only the genes from the initial GTF files. Also, we created bed files for the promoter regions, keeping the areas that were (-1000 bp, +2000 bp) from the Transcription Start Site (TSS). For mouse there were 38,922 promoter regions and for human 57,818. Since there was not a Gencode file available for pig, we downloaded the ensembl gene file available from Biomart (Haider et al., 2009). After doing the same procedure as mentioned above in order to keep only the promoter regions, we ended up with 21,116 regions for pig promoters.

Using BEDtools (Quinlan & Hall 2010) and the bedGraphToBigWig (Kuhn et al. 2009) script from UCSC database, we created bigwig format files for each sample and we uploaded them to UCSC genome browser (Kuhn et al. 2009). Representative tracks for mouse and human datasets are shown below in Figure 2.1 and 2.2.

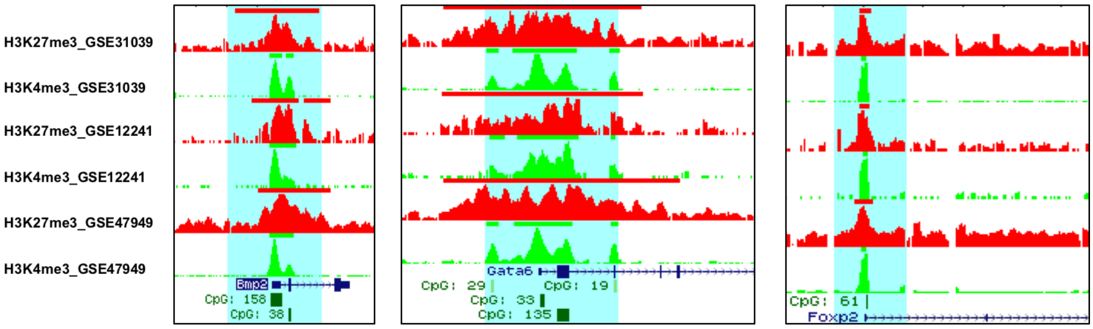


Figure 2.1Representative tracks of normalized number of reads (histogram-tracks) in H3K4me3 (green) and H3K27me3 (red) samples, for the genes Bmp2, Gata6 and Foxp2 in mouse ESCs. The line marks on top of the histograms represent the detected peaks.

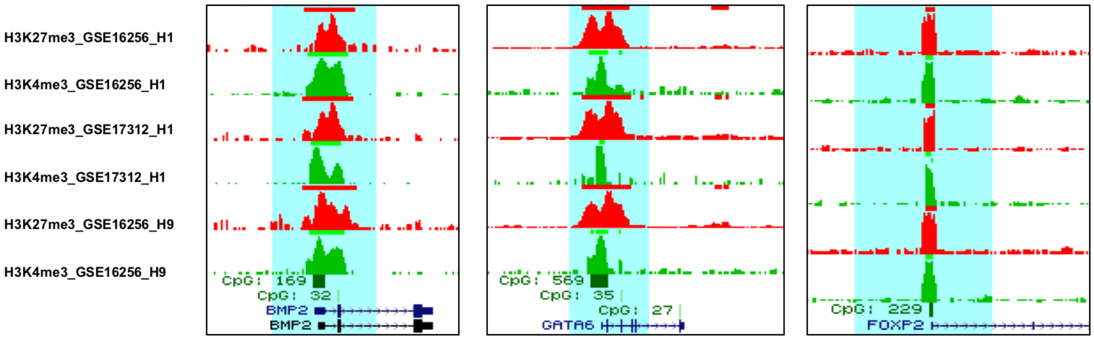


Figure 2.2 Representative tracks of normalized number of reads (histogram-tracks) in H3K4me3 (green) and H3K27me3 (red) samples, for the genes BMP2, GATA6 and FOXP2 in human ESCs. The line marks on top of the histograms represent the detected peaks.

To inspect the trends of the signal over the gene promoters, for each of the samples in both species, we created plots for the average normalized ChIP-seq signal around the TSS region (-2500 bp, +2500 bp). Figure 2.3 and 2.4 show the average signal at the promoter regions for mouse and human ChIP-seq samples respectively.

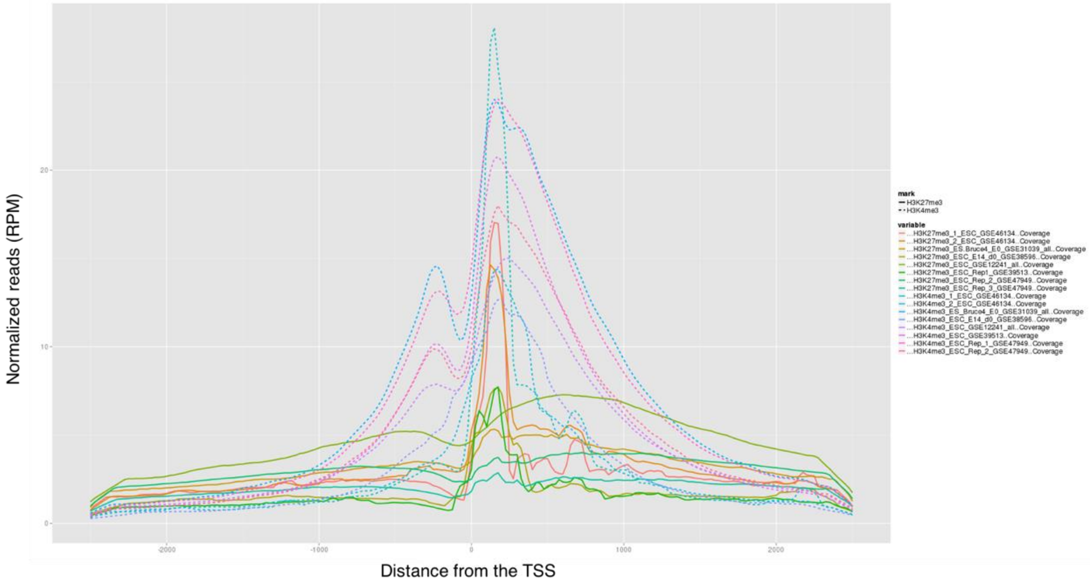


Figure 2.3 Average normalized ChIP-seq signal (reads per million-RPM) across the gene promoters (±2500bp) for each of the samples in mouse ESCs.

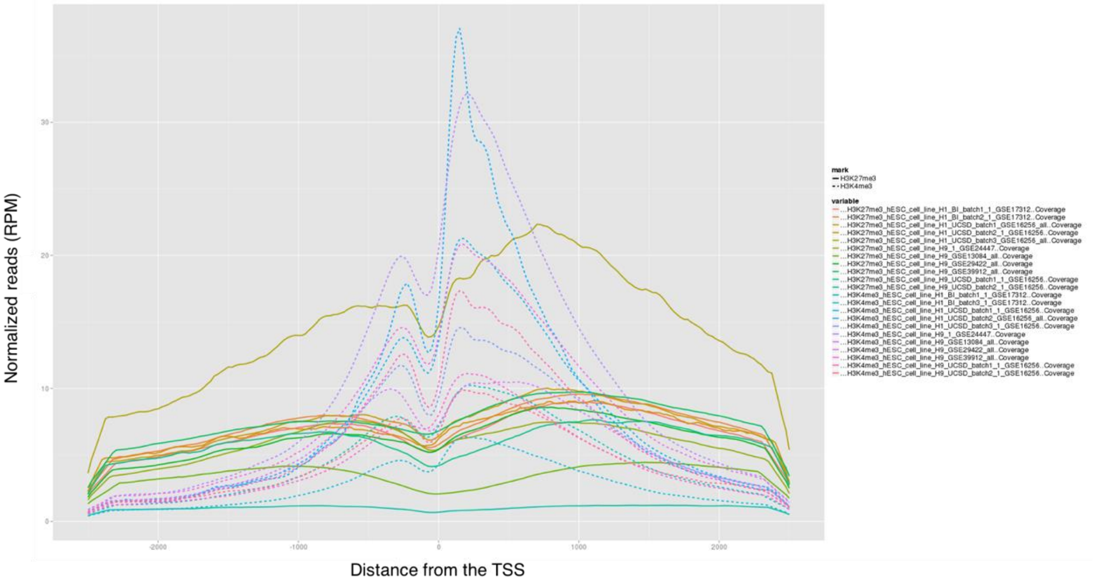


Figure 2.4 Average normalized ChIP-seq signal (reads per million-RPM) across the gene promoters (±2500bp) for each of the samples in human ESCs.

2.3.2 Peak calling method

We used SICER (Zang et al., 2009), a tool that is recommended for enrichment analysis of histone modification data, since it outperforms MACS and FindPeaks in its category for peak calling specific to histone modifications of higher peak width (Zang et al., 2009). The input controls were used when they were provided. When input was available, the SICER parameters were: for H3K4me3, window=200 and gap size=200. For H3K27me3, window=200 and gap size=2x300, since this histone mark is found covering wider chromatin domains. The rest of the parameters (same for both H3K4me3 and H3K27me3) were effective genome fraction =0.7, false discovery rate (FDR) = 0.01, redundancy threshold = 1 and fragment size = 150 (the fragment size was chosen having in mind the nucleosome size where approximately 146 bp of DNA are wrapped around the histone octamer). When a control library was unavailable, the FDR value parameter was replaced by the E-value parameter equal to 100. We intersected the resulting files after peak calling with the promoter files using the intersect command from BEDtools (Quinlan and Hall, 2010).

2.3.3 Cutoff method

We obtained the read density only at the regions we were interested in, the promoters. Using custom scripts and the coverageBed (BEDtools) (Quinlan and Hall, 2010) command, we created bed files for each sample. In the resulting bed files, the column that we kept was the one that contained the number of reads in the promoter regions. We applied logarithmic scale (natural logarithm - ln) to the read densities of all samples, followed by quantile normalization for H3K4me3 and H3K27me3 samples separately to define a threshold that would reveal the real enrichment for H3K4me3 and H3K27me3 and even out the variability across samples. We generated scatterplots of the same histone modification samples against each other to examine what type of normalization to choose. To further increase the accuracy of the cutoff method, we created promoter files with sliding windows. Every promoter region was divided in windows of 200bp, with a sliding step of 50bp. For all the window regions corresponding to the initial promoter region, the maximum coverage value was chosen as the representative for this region. The distribution pattern of H3K4me3 reads is very close to the bimodal distribution. Following that, we used the mixtools package (Benaglia et al., 2009) in order to fit the bimodal distribution to all of our samples, both for H3K4me3 and H3K27me3. Bimodal distribution was fitted successfully for most of H3K4me3 samples. In contrast, most H3K27me3 samples were not following the bimodal distribution. For the successfully fitted H3K4me3 samples we kept the mean and standard deviation of the second curve of each distribution. After subtracting the standard deviation value from its respective mean value, we obtained the initial threshold values for each sample. The final threshold value for all the H3K4me3 samples was the average of all the initial values. In the case of H3K27me3 distributions, since we had no successful fitting bimodal distribution, we chose empirically 3 different thresholds and chose the one that would give results best matching to previous studies. The final threshold values used were 4.57 for H3K4me3 and 3.00 for H3K27me3. We used the study of Mikkelsen et al., 2007 to compare and assess how accurate were the peak-based and the cut-off methods.

2.3.4 Functional enrichment analysis

We conducted gene ontology functional analyses for the bivalent promoters for both approaches, using DAVID (Dennis et al., 2003).

2.3.5 Overlap between species

To obtain a list of common bivalent, expressed and repressed genes between the species, we used only the orthologous genes that mouse and pig share with human (18,255 genes). We got the common list of genes for all three species, but also for each combination by two (human-mouse, human-pig, mouse-pig).

2.3.6 P-value calculation

To calculate if the overlap of two gene lists can happen due to random chance, we used the hypergeometric test. Specifically, to compare two lists we used the phyper function in R. When we were comparing more than two lists we used random permutation of the rows and columns of the results table (species in columns, genes in rows) simulated for 1000 times. We used the permatfull function from the vegan package (Oksanen et al., 2013) in R. Then we compared the mean of all the simulations with our result of common genes in order to assess whether there is significant difference between them.

2.4 Results

2.4.1 Peak-based method to detect high-confidence bivalent promoters

Bivalent promoters are defined by the presence of both active (H3K4me3) and repressive (H3K27me3) chromatin modifications. In ESCs, they are highly enriched for developmental genes and therefore the identification of high confidence bivalent promoters might lead to discovery of novel developmental regulators. With this rationale, we set to look for high confidence bivalent marked promoters in murine ESCs and collected data for eight paired (H3K4me3 and H3K27me3) ChIP sequencing samples from eight studies in GEO (methods for details). The samples varied in read length, ranging from 27 bp to 115 bp and their total number of mapped reads to the mouse genome, ranging from 14 million to 200 million reads per sample. We called peaks using SICER (Zang et. al., 2009), the best suited algorithm for peak detection in histone modification data. For eight samples of H3K4me3, between 16 thousand and 66 thousand peaks were identified while for H3K27me3, between 9 thousand and 26 thousand peaks were identified. To check if this variation in peak number can be attributed to the variability in total number of reads across samples, we calculated Pearson’s correlation coefficient between number of reads and number of peaks detected across eight samples and found a high correlation. The correlation coefficient for H3K4me3 was 0.84 while for H3K27me3 was 0.75. The only way to adjust for the sequencing depth using a peak based method would be to consider the same number of reads across samples (i.e. same as the sample with the fewer number of reads) for peak calling. However, this would not allow us to use most of the available data. Thus, we defined an approach complementary to the peak-based approach – a cutoff-based method (described in detail in the following section). As annotation in mouse we used 38,922 genes from GENCODE (Harrow et al., 2012a) and defined promoters as -1kb and +2kb region around the transcription start site of each transcribed unit. We then intersected these promoters with H3K4me3 and H3K27me3 peaks. Despite the large variance in the number of H3K4me3 peaks identified in individual samples, the number of peaks within promoters was very consistent across samples ranging from 18 thousand to 20 thousand H3K4me3 marked promoters. This suggests that most promoters have a high peak height of H3K4me3 and therefore H3K4me3 is a distinguishing mark for promoters. In contrast, the number of H3K27me3 promoter peaks showed a large variability ranging from 3 thousand to 9 thousand peaks. The Pearson’s correlation coefficient value between the total H3K27me3 peaks and the fraction of these in promoters was 0.5. This suggests that H3K27me3 does not show preference to promoters and therefore is not a distinguishing mark for promoters. The number of bivalent marked promoters varied between 2 thousand and 7 thousand across eight samples. Pearson’s correlation coefficient between the number of H3K4me3 promoters and bivalent promoters was 0.58 while between H3K27me3 promoters and bivalent promoters was 0.98. This shows that the classification of a promoter as a bivalent promoter highly depends upon identification of H3K27me3 modification rather than H3K4me3 modification.

To identify the high confidence bivalent promoters, we calculated cumulatively the number of promoters identified with the H3K4me3 modification in ‘n’ or more samples. Over 20 thousand promoters were H3K4me3 marked in at least one sample, while about 15 thousand promoters were H3K4me3 marked in all eight samples. This demonstrates that H3K4me3 modification on promoters across samples is quite stable (Table 2.3). On the contrary, over 11 thousand H3K27me3 promoters were detected in at least one sample of which only about 2 thousand were H3K27me3 marked in all samples (Table 2.3). The rate of decrease in the number of bivalent promoters (ratio of six or more to one or more) was 0.44, in H3K4me3 promoters was 0.81 and in H3K27me3 promoters was 0.37 in ‘n’ or more samples. This again demonstrates that the number of high confidence bivalent promoters is dependent on the H3K27me3 histone mark. We noticed that over 80% of H3K27me3 promoters were consistently marked bivalent (Table 2.3). This means that most H3K27me3 marked promoters also have H3K4me3 modification present. This demonstrates that the co-existence of these two chromatin modifications on promoters initially thought as a surprise, is rather a rule than exception. ChIP enrichment signals can be missed during peak calling procedure or by experimental error in an individual sample. Peaks detected in all samples are likely to miss true bivalent promoters. As the ratio of bivalent to H3K27me3 marked promoters was highly consistent when 4, 5 or 6 or more samples are taken into account, we used an arbitrary cut off of six or more to define high confidence bivalent promoters. This resulted into identification of 16,885 high confidence H3K4me3 marked, 4,239 high confidence H3K27me3 marked and 3,740 high confidence bivalent promoters (Table 2.3).

We then investigated whether the high confidence detection was biased towards any individual study or was true representative of all eight studies. About 50% of high confidence peaks were present in individual H3K4me3 samples while the fraction of high confidence H3K27me3 peaks in individual sample varied between 40 and 70%. This again demonstrates that H3K4me3 is consistent while H3K27me3 varies on the promoters.

MOUSE (WITH PEAK CALLING METHOD)				
Samples	**H3K4me3**	**H3K27me3**	**Bivalent**	**Bivalent/H3K27me3**
1 or more	20761	11610	8515	0.73
2 or more	19980	8931	7252	0.81
3 or more	19358	7413	6343	0.85
4 or more	18523	6198	5458	0.88
5 or more	17848	5175	4679	0.90
6 or more	**16885**	**4239**	**3740**	**0.88**
7 or more	16062	3287	2764	0.84
8 or more	14720	2236	1555	0.69

Table 2.3 Cumulative count of three categories of promoters in mESCs with the peak based method. The cells with bold font (6 or more) represent the high confidence cut off chosen.

2.4.2 Cutoff-based method to detect high-confidence bivalent promoters

As the peak calling method is highly sensitive to the sequencing depth, we defined another independent method to identify enriched genomic regions for a specific histone modification, henceforth called cutoff-based method. We calculated the number of reads mapping to each promoter in each H3K4me3 and H3K27me3 sample by using custom scripts and the BEDtools suite (Quinlan & hall 2010). To normalize the reads across multiple samples, the logarithmic scaled promoter read counts across all H3K4me3 and H3K27me3 experiments were quantile normalized (separately for each histone mark, see Methods). The H3K4me3 normalized promoter read density followed a clear bimodal distribution separating H3K4me3 unmarked from marked promoters (Figure 2.5a). We further noticed that the H3K4me3 positive and H3K4me3 negative sets were conserved across samples. On the contrary, the normalized promoter read density for H3K27me3 did not show a clear bimodal distribution making it hard to distinguish between the H3K27me3 positive and H3K27me3 negative sets (Figure 2.5b). Moreover, although the H3K27me3 mark was coherent across samples, the distinction of two groups was not clear as in the case of H3K4me3

(Figure 2.5). We fitted a bimodal distribution to the normalised H3K4me3 promoter read density and consistently obtained a cut-off of 4.57 to distinguish between H3K4me3 positive and negative promoters (Figure 2.5a). On the other hand, the bimodal distribution failed to fit to the normalised H3K27me3 read density, thus we defined an arbitrary cut-off of 3.00 to distinguish between H3K27me3 positive and negative promoters (Figure 2.5b). The cutoff based method identified consistently about 7 thousand H3K27me3 marked promoters and about 13 thousand H3K4me3 marked promoters.

To identify high confidence bivalent promoters using the cutoff method, we calculated cumulatively the number of promoters identified with a given modification in 'n' or more samples. Both H3K4me3 and H3K27me3 marks showed a large variability across samples. Over 15 thousand promoters were H3K4me3 marked in at least one sample while only about 11 thousand promoters were H3K4me3 marked in all eight samples (Table 2.4). Similarly, over 16 thousand H3K27me3 promoters were detected in at least one sample from which only about 3 thousand were H3K27me3 marked in all samples (Table 2.4). The ratio levels were not as consistent as in the case of the peak calling method but for most of the cases (except for the extremes) more than 50% of the bivalent promoters were part of the H3K27me3 marked promoters. Like the peak calling procedure, we used a threshold of six or more to define high confidence bivalent promoters. This resulted into the identification of 13,034 high confidence H3K4me3 marked, 4,660 high confidence H3K27me3 marked and 2,396 high confidence bivalent promoters.

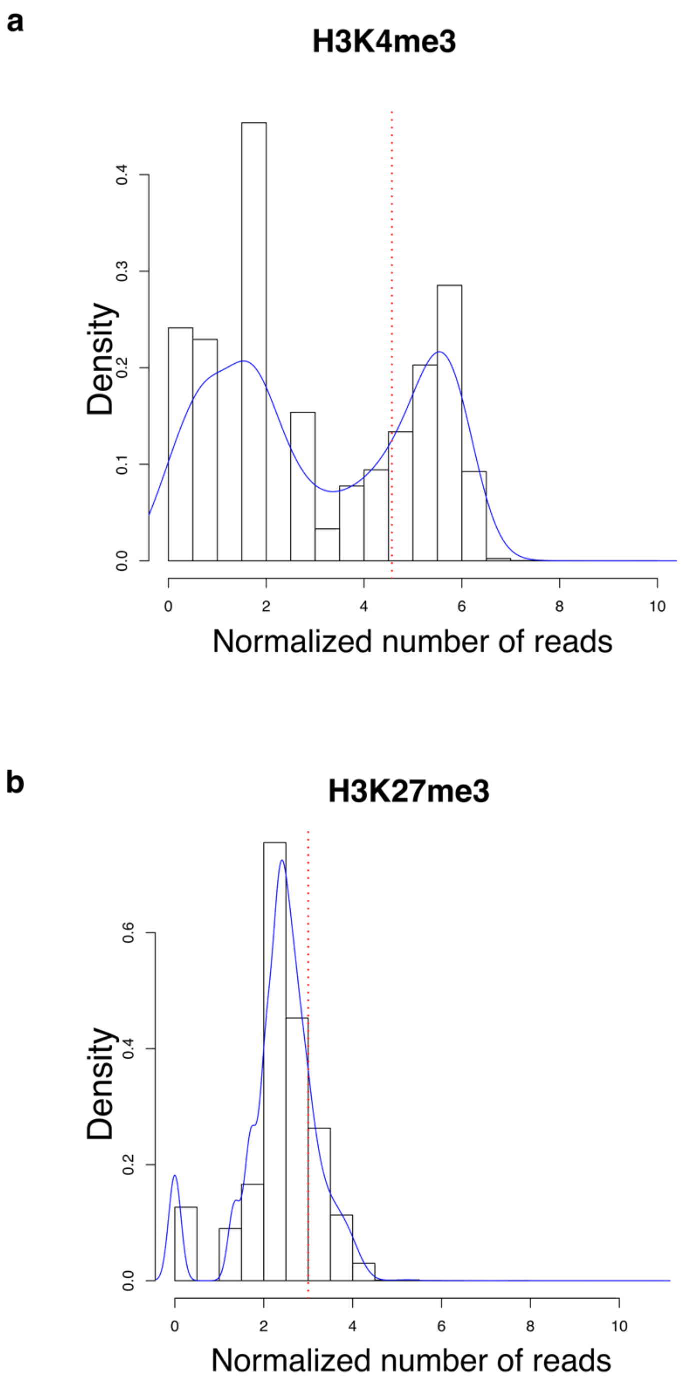


Figure 2.5 Representative histograms and density plots for normalised number of reads in a) H3K4me3 and b) H3K27me3 samples in mouse ESCs. The vertical dotted red line marks the threshold used.

MOUSE (WITH CUTOFF BASED METHOD)				
Samples	**H3K4me3**	**H3K27me3**	**Bivalent**	**Bivalent/H3K27me3**
1 or more	15668	16624	7711	0.46
2 or more	14895	10327	5428	0.52
3 or more	14389	7748	4400	0.56
4 or more	13942	6419	3685	0.57
5 or more	13478	5479	3027	0.55
6 or more	**13034**	**4660**	**2396**	**0.51**
7 or more	12378	3846	1708	0.44
8 samples	11190	2829	945	0.33

Table 2.4. Cumulative count of three categories of promoters in mESCs with the cutoff based method. The cells with bold font (6 or more) represent the high confidence cutoff chosen.

2.4.3 Systematic comparison of peak-based and cutoff-based method

We performed a systematic comparison of the peak-based and cutoff-based method. Across individual samples, the variability in the total number of peaks identified by cutoff-based method was much lower compared to the peak-based method for both H3K4me3 and H3K27me3 data sets. Though the cutoff-based method showed high consistency across samples for both modifications, it showed higher variability when the cumulative analysis was performed (Tables 2.3 & 2.4). We then compared the high confidence bivalent promoters obtained by both methods by defining the same threshold of six or more samples. The cutoff-based method concluded that only about 50% of H3K27me3 marked promoters were bivalent whereas the peak-based method predicted this fraction to be over 80%. The peak-based method results are thus in agreement with the literature (Mikkelsen et al., 2007). This is expected as peak calling approaches are widely used in the literature. Over 80% of bivalent peaks detected by the cutoff method, were also found by the peak calling method. The peak-based method is therefore able to identify high confidence bivalent promoters missed by the cutoff method (Figure 2.6a). Finally, we calculated functional enrichment for bivalent promoters using both approaches. Although both set of high confidence promoters were enriched for developmental categories such as anatomic structure development and developmental process as expected, the enrichment was higher for the peak method than the cutoff one (Figure 2.6b). Taken together, the peak-based method was more reliable in detecting high confidence bivalent promoters.

2.4.4 Comparison of high-confidence bivalent promoters in serum-grown ESCs and 2i ESCs

Having established that the peak detection method reliably predicts high confidence bivalent promoters, we used the bivalent promoters detected by the peak-based method for further analysis. Murine ESCs can be maintained in two distinct culture conditions in vitro, 2i (with inhibitors of two kinases Mek and GSK3) and serum. All eight samples used for high confidence bivalent promoter detection were grown in serum culture condition. Marks et al., 2012 identified 1,014 bivalent genes in murine ESCs grown under 2i media and 2,936 bivalent genes grown in serum and stated that the identification of fewer bivalent genes in ‘2i’ was in agreement with the postulated naïve ground state of ESCs grown in ‘2i’ and not in serum. If this were the case, the high confidence bivalent promoters should show a higher overlap with 2i grown bivalent genes than bivalent genes detected in a serum grown sample. 76% of 2i-grown bivalent genes and 68% of serum-grown bivalent genes overlapped with our high confidence bivalent promoters respectively. This suggests that the high confident bivalent promoters defined in this study show greater similarity with the ones found at the naïve pluripotent state. A fraction of 2i-grown bivalent genes were not identified bivalent in any of the samples grown in serum. This suggests that there are genes specifically bivalent marked in 2i and not in serum culture condition.

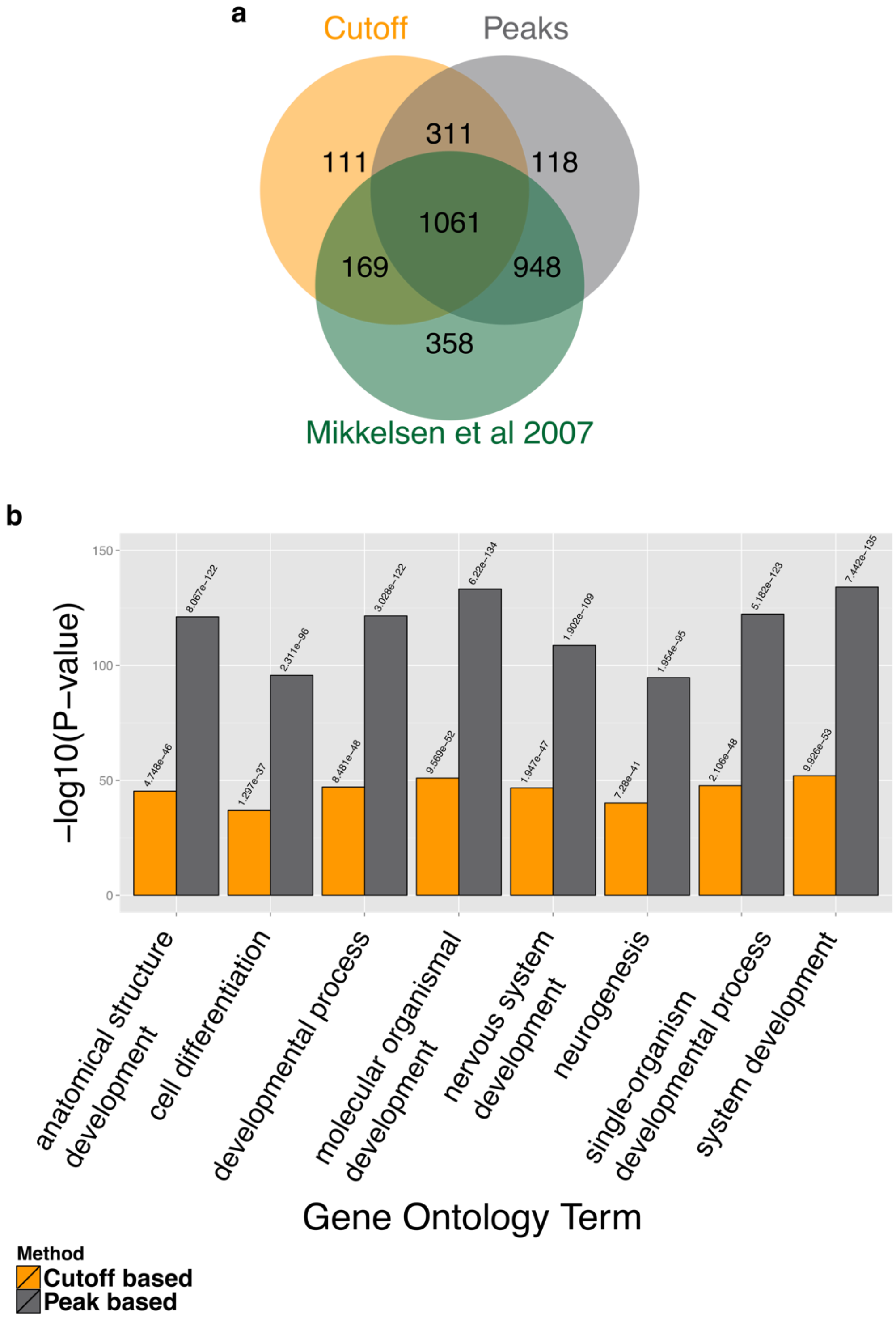


Figure 2.6 a) Common bivalent promoters between the cutoff based method, the peak based method and from Mikkelsen et al., 2007 b) Functional enrichment values (-log10Pvalue) for the most enriched gene ontology terms for the two methods (P-value indicated on top of each bar, Fisher's exact test).

2.4.5 Identification of bivalent regions in other mammalian species

To investigate if the high confidence bivalent regions are conserved across species, we collected genome wide ChIP-seq data for H3K4me3 and H3K27me3 in human ESCs and pig induced pluripotent cells (iPSCs). We gathered 11 paired samples in humans from six studies with reads ranging from 13 million to 60 million in individual samples. We used the peak based method to call peaks in individual samples. These peaks were then mapped to promoters of 57,818 transcribed units defined by GENCODE (Harrow et al., 2012a). Similar to mouse, the number of H3K4me3 promoter peaks were highly consistent across samples (mean 19,219.73, SD 462.88) while the number of H3K27me3 promoter peaks was more variable (mean 8,035.73, SD 2,626.27). To identify high confidence human bivalent promoters, we considered bivalent promoters identified in 'n' or more samples. The rate of decrease for the number of bivalent promoters (ratio of six or more to one or more) was 0.39, for H3K4me3 promoters was 0.89 and for H3K27me3 promoters was 0.31 (Table 2.5). The fraction of bivalent to H3K27me3 promoters was consistently higher than 80% (Table 2.5). We used an arbitrary threshold of eight or more samples to define high confidence bivalent promoters. This resulted into the identification of 18,744 high confidence H3K4me3 marked, 5,841 high confidence H3K27me3 marked and 5,116 high confidence human bivalent promoters (Table 2.5).

In pig (Sus Scrofa), only one study was available in the public domain hindering detection of high confidence bivalent promoters in pig. Using 21,116 promoter regions in pig we detected 8,383 H3K4me3 marked, 2,816 H3K27me3 marked and 1,561 bivalent marked promoters again demonstrating that over half of H3K27me3 marked promoters also contain an H3K4me3 modification.

HUMAN (WITH PEAK CALLING METHOD)				
Samples	**H3K4me3**	**H3K27me3**	**Bivalent**	**Bivalent/H3K27me3**
1 or more	21167	18701	13206	0.70
2 or more	20275	12066	9778	0.81
3 or more	19865	9825	8236	0.83
4 or more	19602	8560	7308	0.85
5 or more	19341	7789	6713	0.86
6 or more	19123	7102	6177	0.86
7 or more	18944	6480	5660	0.87
8 or more	**18744**	**5841**	**5116**	**0.87**
9 or more	18489	5171	4505	0.87
10 or more	18189	4087	3495	0.85
11 samples	17678	2771	2202	0.79

Table 2.5 Cumulative count of three categories of promoters in mESCs with the peak based method. The cells with bold font (8 or more) represent the high confidence cut off chosen.

2.4.6 Comparative analysis of bivalent and promoters across three species

Finally, we computed the overlap of bivalent promoters across three species by considering only one-to-one mapping orthologues. The bivalent promoters were less conserved across three species compared to the active promoters (Figure 2.7a and 2.7b). Specifically, less than 10% of human bivalent promoters were conserved across three species while over 25% of H3K4me3 marked promoters were conserved across three species. The functional enrichment of common bivalent genes resulted in development processes, more specifically embryogenesis, such as pattern specification process, embryonic morphogenesis and embryonic organ development, suggesting that the three species have more commonalities during embryonic development.

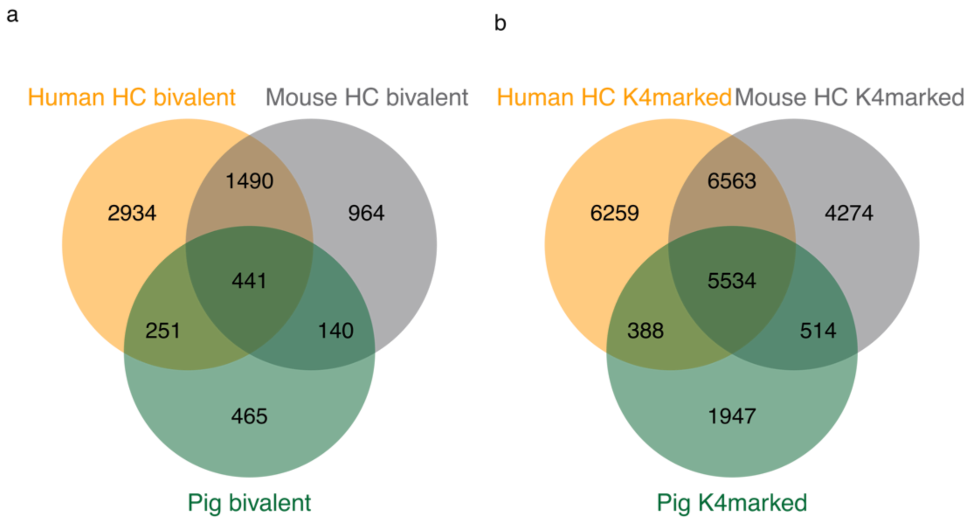


Figure 2.7 Venn diagram of a) bivalent and b) K4marked promoters between human, mouse and pig using the peak calling method.

2.5 Conclusion

In summary, we identified high confidence bivalent domains in murine ESCs by integrating data across eight studies using two methods; peak-based and cutoff-based, and demonstrated that the peak-based method is more reliable. We then identified bivalent promoters in human and pig and performed a multi-species comparative analysis of bivalent promoters to show that the conserved bivalent promoters were highly enriched for embryonic developmental processes.

Chapter 3 CpG island erosion, Polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells

3.1 Chapter Introduction

This chapter was published in 2015 in Scientific Reports under DOI: 10.1038/srep16791. Here we use the peak calling method, which was evaluated to be more suitable for the purpose of our study (Chapter 2), to detect a robust list of bivalent promoters in human and mouse ESCs. We also proceed with characterization of the properties of high confidence bivalent promoters. Variant expression levels among distinct groups of bivalent promoters, have subsequently led us to incorporate single cell transcriptomics data. Detailed transcriptomics analysis, under the scope of its relevance with bivalency, is presented in Chapter 4.

Abstract

In embryonic stem cells (ESCs), developmental regulators have a characteristic bivalent chromatin signature marked by simultaneous presence of both activation (H3K4me3) and repression (H3K27me3) signals and are thought to be in a 'poised' state for subsequent activation or silencing during differentiation. We collected eleven pairs (H3K4me3 and H3K27me3) of ChIP sequencing datasets in human ESCs and eight pairs in murine ESCs, and predicted high-confidence (HC) bivalent promoters. Over 85% of H3K27me3 marked promoters were bivalent in human and mouse ESCs. We found that (i) HC bivalent promoters were enriched for developmental factors and were highly likely to be differentially expressed upon transcription factor perturbation; (ii) murine HC bivalent promoters were occupied by both Polycomb repressive component classes (PRC1 and PRC2) and grouped into four distinct clusters with different biological functions; (iii) HC bivalent and active promoters were CpG rich while H3K27me3-only promoters lacked CpG islands. Binding enrichment of distinct sets of regulators distinguished bivalent from active promoters. Moreover, a 'TCCCC' sequence motif was specifically enriched in bivalent promoters. Finally, this analysis

will serve as a resource for future studies to further understand transcriptional regulation during embryonic development.

3.2 Introduction

Embryonic stem cells (ESCs) have the unique ability to self-renew indefinitely as well as to differentiate in response to internal as well as external stimuli (O'Shea, 2004). These two properties of ESCs pose specific constraints on the genome, as self-renewal requires maintenance of cellular memory that specifies its pluripotent capacity, while differentiation potential requires pluripotent ESCs to be highly plastic to enter any one distinct differentiation pathway. While the pluripotent state of ESCs is controlled through a network of core transcription factors (Takahashi and Yamanaka, 2006), emerging data point to a key role for epigenetic mechanisms such as chromatin dynamics and histone modifications in pluripotency (Meshorer and Misteli, 2006). Histone proteins and their post-translational modifications define the chromatin status of a cell and are correlated with the transcriptional status of genes. Mono-methylation of lysine 4 of histone protein 3 (H3K4me1) and acetylation of lysine 27 of histone protein 3 (H3K27ac) mark active enhancers while H3K4me3 and H3K27me3 mark active and repressed promoters, respectively (Bannister and Kouzarides, 2011). Other epigenetic marks are also associated with promoters and enhancers. For example, H4K16 acetylation marks active genes and enhancers in ESCs (Taylor et al., 2013). Set/MLL histone methyltransferases, the mammalian homologues of the trithorax group proteins (trxG), catalyse the H3K4me3 marks and Polycomb (PcG) group proteins catalyse H3K27me3. Both complexes are thought to regulate expression of important differentiation and developmental genes (Schuettengruber et al., 2007; Shilatifard, 2012b). These two chromatin modifications previously thought to be mutually exclusive were observed co-existing on promoters in murine ESCs and were named 'bivalent' promoters (Bernstein et al., 2006b). Bivalent genes are typically silenced or expressed at a very low level in ESCs, and by the presence of both active and repressive marks, are thought to be poised for activation or repression during the differentiation process (Azuara et al., 2006; Mikkelsen et al., 2007). Bivalent genes in ESCs either lost the H3K27me3 mark and were expressed, or lost H3K4me3 and were silenced when differentiated into the neuronal lineage (Mikkelsen et al., 2007). Upon receiving endoderm differentiation signals, the bivalent BRACHYURY and NODAL promoters in human ESCs were unilaterally resolved to activation of the associated genes by losing H3K27me3 (Loh et al., 2014).

Bivalency of chromatin has therefore become an important property to investigate the functional relevance of a gene through development, and the presence of bivalent genes in human and mouse ESCs has been validated by many studies independently (Jia et al., 2012a; Ku et al., 2008a; Mikkelsen et al., 2007; Pan et al., 2007; Zhao et al., 2007). Here we performed a systematic identification and characterisation of bivalent genes and their functions by integrating all publicly available pairs (H3K4me3 and H3K27me3 measured on the same samples) of ChIP sequencing datasets in human and mouse ESCs, and identified and characterised a set of 4,979 and 3,659 high–confidence (HC) bivalent promoters respectively.

3.3 Methods

3.3.1 Data collection and processing

Covered in section 2.3.1 of Chapter 2. However, pig datasets were not used in this chapter.

3.3.2 Peak Calling Method

Covered in section 2.3.2 of Chapter 2. However, addition of input controls that were not previously integrated in Chapter 2, led in slight differences in the numbers of detected peaks and subsequently numbers of HC promoters.

3.3.3 Detection of High Confidence (HC) bivalent, H3K4me3-only, H3K27me3-only and latent promoters

As mentioned before, we acquired ChIP-seq data (H3K4me3 and H3K27me3) from 8 studies for mouse and 11 studies for human. The resulting files after peak calling were intersected with the promoter files. Our aim was to find whether or not the peaks were overlapping with the regions around the promoters. We used the intersect command from BEDtools (Quinlan and Hall, 2010). The resulting bed files for each sample contained the peaks that were found only in promoter regions. In Tables 3.1 and 3.2, we present the peaks at promoter areas for human and mouse respectively. After the intersection with the promoter areas, we created a matrix that contained the values of all the samples (the rows represent each region and the columns represent the number of peaks that overlap with each region for each sample). For the further analysis, we created R scripts in order to keep the bivalent promoters where both histone mark peaks were identified, the H3K4me3-only promoters where H3K4me3 peak was identified and not H3K27me3, and the H3K27me3-only promoters where only H3K27me3 peak was identified and the latent promoters where peak for neither mark was identified. We obtained all the possible numbers of bivalent regions taking into account combinations for 1 or more data samples until the total numbers of samples (Tables 3.3 and 3.4). The level of stringency was increased as we took into consideration more samples. We defined high-confidence (HC) bivalent promoters as ones identified in 70% or more samples. Therefore, for mouse we would consider a locus as bivalent if it was found in 6 or more samples (6/8 studies) and in human if it was found in 8 or more samples (8/11 studies). We applied the same definition for the H3K4me3-only, H3K27me3-only and latent high-confidence promoters.

3.3.4 Read density at the promoter regions

Using BEDtools (8) (coverageBed command) we calculated the coverage at the promoter regions in all different groups for each histone mark sample.

3.3.5 Peak height and overlap with top peaks

Using BEDtools (Quinlan and Hall, 2010) we intersected the peak files for all the samples in both species with the high-confidence (HC) bivalent regions we had previously detected. We classified the peaks in bivalent and non-bivalent depending on whether they were found or not in HC bivalent promoters. We performed peak height (read density) normalization in each sample then converting it in the logarithmic scale (log10). After checking the significance of the difference of peak height between bivalent and non-bivalent promoters (Student’s t-test), we also checked whether the top peaks of the H3K27me3 samples could give us the same list of HC bivalent promoters. Taking the top high peaks of each H3K27me3 sample, as many as the HC bivalent peaks of the same sample, we checked the degree of overlap between them.

3.3.6 Functional enrichment analysis

We conducted gene ontology functional analyses for the bivalent promoters using DAVID (Dennis et al., 2003) and AmiGO (Carbon et al., 2009).

3.3.7 Overlap between Species

To obtain a list of common HC bivalent, H3K4me3-only, H3K27me3-only and latent genes between the species, we used the one2one orthologous regions between human and mouse (16,639 genes from ensembl BioMart) (Guberman et al., 2011). We calculated the percentage of conservation for each species individually taking into account the corresponding orthologous regions and their chromatin state for the other species.

3.3.8 Clustering using published ChIP-seq data

We downloaded ChIP-seq data from published studies and used them for further classification of our HC bivalent promoters in mouse embryonic stem cells. We gathered four different forms of RNAPII, RNAPIIS5P, RNAPIIS7P and 8WG16 (Brookes et al., 2012a), PRC2 component, Suz12 (Morey et al., 2013), PRC1 subunits, Cbx7 and Ringb (Morey et al., 2013) in murine ESCs. We also downloaded Jarid2 (Tee et al., 2014), H3K27ac (Yu et al., 2013), Utf1 (Jia et al., 2012a) and Ring1b (Ku et al., 2008a). We used seqMINER (Ye et al., 2011) to integrate the multiple TFs and histone modifications and visualize the patterns that are formed genome wide at the HC bivalent promoters.

3.3.9 CpG overlap for the HC promoters

We calculated the overlap of the HC bivalent, H3K4me3-only, H3K27me3-only and latent regions with the CpG island regions as given from the UCSC tracks CpG islands for hg19 and mm10 (Karolchik et al., 2014). We calculated the percentage of overlap with the total number of genes for Gencode 19 and Gencode M2, with the protein coding genes and with the all the HC groups we have detected previously in our analysis.

3.3.10 CpG density and H3K27me3 read density across species

We calculated the CpG density as the ratio of observed to expected CpG counts (Gardiner-Garden and Frommer, 1987) for -5Kb, +5kb around the TSS for 100 bp window. The regions we have used were the bivalent regions for each species and their corresponding regions in other species (human/mouse) using the UCSC liftOver tool (Karolchik et al., 2014). We created heatmaps using custom R scripts for the visualization of the CpG density and H3K27me3 read density ordered by the CpG density of the targeted species.

3.3.11 Transcription and Epigenetic Factors' enrichment using published ChIP-Seq data

We have used data from 49 and 99 ChIP-seq experiments for several transcription factors (TFs), chromatin remodellers and methyltransferases in human and mouse ESCs respectively (Sánchez-Castillo et al., 2015). Initially, we intersected the peak files of all the factors with the promoter regions we have created for the Gencode gene sets. The resulting files were finally intersected with the HC promoters for all the categories and we found the levels of enrichment. For each promoter region we also counted the total number of factors binding significantly at the region. We calculated the numbers of factors binding across the different promoter categories.

3.3.12 RNA sequencing levels

We downloaded a mouse RNA-seq experiment (Yu et al., 2013) in fastq format. After aligning the reads to the mouse reference genome (mm10) using Bowtie 0.12.9 (Langmead et al., 2009), we found the FPKM values using cufflinks 2.2.1 (Trapnell et al., 2012). For human we used RNA-Seq (FPKM values – transformed in natural logarithm-ln scale) data for H1-hESCs from (Djebali et al., 2012) . We created three different classes according to the expression level. We defined as highly expressed genes with expression greater than log (FPKM) > 4. Low expression was defined as 0 < log (FPKM) < 4. Finally, genes with expression equal to zero belonged to the no expression category.

3.3.13 Single cell RNA Sequencing

Using single cell RNA sequencing data (Streets et al., 2014), we inspected the number of genes that had zero levels of expression. We selected genes that belonged to the low expression class (0 < log (FPKM) < 4) for all the HC promoter categories. We then intersected the low expression sets with the FPKM values of the corresponding genes from single RNA-seq. For each gene we counted the number of occurrences of zero expression along the 63 single cell RNA sequencing experiments.

3.3.14 Gain and loss of function perturbation

We collected differentially expressed gene lists (both up- and down-regulated) after over expression of 54 transcription factors and deletion of 37 transcription factors individually in murine ESCs (Xu et al., 2013). We then intersected the gene lists for both gain and loss of function with our HC promoters for all the categories. We checked the levels of perturbation among the promoter types and also which were the genes that were over-perturbed for the majority of the TFs.

3.3.15 Motif enrichment using HOMER

We used the gene based analysis with the command findMotifs.pl from HOMER (Heinz et al., 2010). We performed the analysis for the HC bivalent and H3K4me3 marked promoters using them both as main and background files to each other. Similarly, the same was applied to all peak list from Najafabadi et al. (2015) C2H2 ChIP-seq experiments.

3.3.16 P-value calculation

Covered in section 2.3.6 of Chapter 2. In this chapter we corrected all P-values for multiple hypotheses testing using FDR correction.

3.3.17 Robustness of High Confidence definition

The number of samples we chose as a cut-off for the detection of HC promoters was 6 for mouse and 8 for human. To validate that the results did not depend on this choice of cut off, we conducted key steps of the analysis for one less and one more samples for both human and mouse. Firstly, we checked the overlap of CpG islands with the HC bivalent, H3K4me3-only, H3K27me3-only and latent promoters. Then, we checked the number of factors binding across the various HC promoter categories and which are the enriched factors for each category. Lastly, we performed de-novo motif discovery using HOMER (Heinz et al., 2010). All the analysis is shown in Figure 3.2 demonstrating the robustness of our findings.

3.4 Results

3.4.1 High-confidence bivalent promoters in human and mouse ESCs are enriched for developmental regulators

Bivalent promoters are distinguished by the presence of both H3K4me3 and H3K27me3 modifications and are thought to mark developmental regulators in ESCs. To determine a robust set of bivalent promoters, we collected 11 pairs (i.e., generated by the same lab using same ES cell samples) of H3K4me3 and H3K27me3 ChIP sequencing (ChIP-seq) datasets for human ESCs and 8 pairs for mouse ESCs from the Gene Expression Omnibus (GEO) database and the Roadmap Epigenomics Project (Tables 2.1, 2.2 and Methods). After aligning reads to the respective genomes, peaks were called in each dataset using SICER (Zang et al., 2009) and were overlapped with

57,818 human promoters from GENCODE 19 (Harrow et al., 2012b) and 38,922 murine promoters from GENCODE M2 (Harrow et al., 2012b).

The number of H3K4me3 marked promoters across data sets was highly consistent (human: mean 18,632.55 relative SD 2.8%, mouse: mean 17,554.25 relative SD 11%), in contrast to the number of H3K27me3 marked promoters (human: mean 7,523.45 relative SD 37%, mouse: mean 6,128.75 relative SD 35%) (Tables 3.1 and 3.2). Moreover, the same promoters were consistently identified as H3K4me3 marked across samples, as demonstrated by incrementally intersecting the peaks from multiple datasets (Figure 3.1A, green curve). In contrast, the H3K27me3 marked promoters (Figure 3.1A, purple curve) varied across datasets, strongly influencing the number of bivalent promoters detected (Figure 3.1A, yellow curve). Assigning a bivalent status to a promoter is therefore largely subject to H3K27me3 peak identification on the promoter. Over 85% of H3K27me3 marked promoters in both human and mouse were bivalent promoters (Figure 3.1A, Tables 3.3 and 3.4). Thus, we reconfirm that bivalency at the H3K27me3 marked promoters is rather a rule than an exception (Pan et al., 2007).

HUMAN						
Samples	**Reads K27**	**Peaks K27**	**Reads K4**	**Peaks K4**	**Peaks K27 at promoters**	**Peaks K4 at promoters**
GSE13084_1	21,672,220	10,904	22,955,497	30,373	6,284	19,719
GSE16256_1	17,173,545	8,353	7,763,232	19,338	5,354	18,027
GSE16256_2	19,825,041	24,326	22,053,477	19,423	5,266	17,894
GSE16256_3	57,540,895	13,058	20,360,983	22,756	10,588	19,005
GSE16256_4	19,847,708	14,941	17,950,910	21,622	7,274	18,540
GSE16256_5	60,534,166	12,309	37,036,236	23,600	4,202	18,732
GSE17312_1	12,682,151	9,355	13,730,204	24,133	6,641	18,799
GSE17312_2	12,946,346	11,058	14,190,726	21,873	7,889	18,004
GSE24447_1	15,925,532	39,528	17,211,980	22,607	14,221	18,626
GSE29422_1	53,468,516	12,822	21,930,525	22,917	7,670	18,896
GSE39912_1	28,364,855	12,061	39,595,351	21,042	7,369	18,716

Table 3.1 Total reads and peaks detected at promoters for 11 samples of H3K27me3 and H3K4me3 histone modifications in human ES cells

MOUSE						
Samples	**Reads K27**	**Peaks K27**	**Reads K4**	**Peaks K4**	**Peaks K27 at promoters**	**Peaks K4 at promoters**
GSE12241_1	17,748,450	17,406	19,618,320	33,732	7,334	17,487
GSE31039_1	106,488,411	22,326	42,655,393	32,304	8,807	17,759
GSE38596_1	14,552,549	6,534	7,474,539	16,904	3,346	15,092
GSE39513_1	95,910,319	20,674	208,490,886	65,538	4,475	19,735
GSE46134_1	29,913,075	7,038	20,167,316	20,605	4,366	15,899
GSE46134_2	31,256,236	8,739	30,157,341	19,243	5,246	15,375
GSE47949_1	99,957,560	26,234	88,198,720	62,081	9,097	19,303
GSE47949_2	107,076,939	14,727	105,069,233	66,243	6,359	19,784

Table 3.2 Total reads and peaks detected at promoters for 8 samples of H3K27me3 and H3K4me3 histone modifications in mouse ESCs

HUMAN							
Samples	H3K4me3 only	H3K27me3 only	H3K4me3 marked	H3K27me3 marked	Bivalent	Latent	Bivalent/H3K27me3 marked
1	7254	5497	20732	18975	12881	31589	0.67884058
2	6778	2122	19753	12036	9542	31589	0.792788302
3	6581	1349	19318	9698	8007	31589	0.825634151
4	6470	1001	19039	8445	7111	31589	0.842036708
5	6355	772	18773	7638	6528	31589	0.854673998
6	6257	617	18523	6966	6001	31589	0.861469997
7	6199	506	18333	6358	5511	31589	0.866782007
8	6135	397	18120	5708	4979	31589	0.872284513
9	6049	330	17842	5048	4374	31589	0.866481775
10	5959	234	17526	3998	3400	31589	0.850425213
11	5833	151	16999	2732	2164	31589	0.792093704

Table 3.3 Number of identified promoters in each category as the samples taken into account increase in human ESCs

MOUSE							
Samples	H3K4me3 only	H3K27me3 only	H3K4me3 marked	H3K27me3 marked	Bivalent	Latent	Bivalent/H3K27me3 marked
1	11797	2912	20615	11730	8480	15395	0.722932651
2	11114	1585	19690	8981	7190	15395	0.800579
3	10685	916	19023	7407	6276	15395	0.847306602
4	10130	514	18049	6162	5369	15395	0.871308017
5	9722	270	17287	5137	4600	15395	0.895464279
6	9336	152	16262	4184	3659	15395	0.874521989
7	9015	81	15422	3222	2706	15395	0.839851024
8	8478	52	14086	2207	1534	15395	0.695061169

Table 3.4 Number of identified promoters in each category as the samples taken into account increase in mouse ESCs

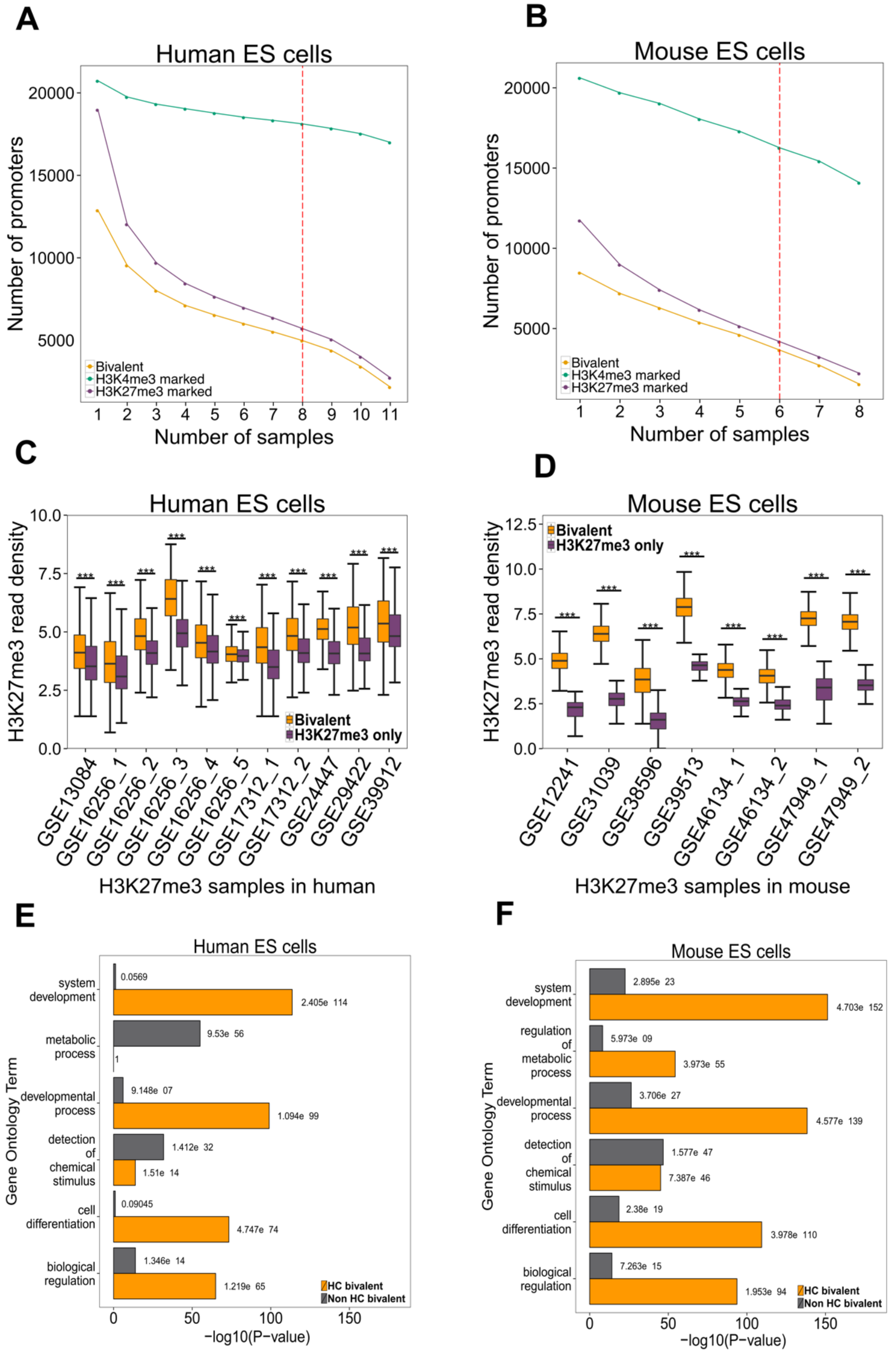


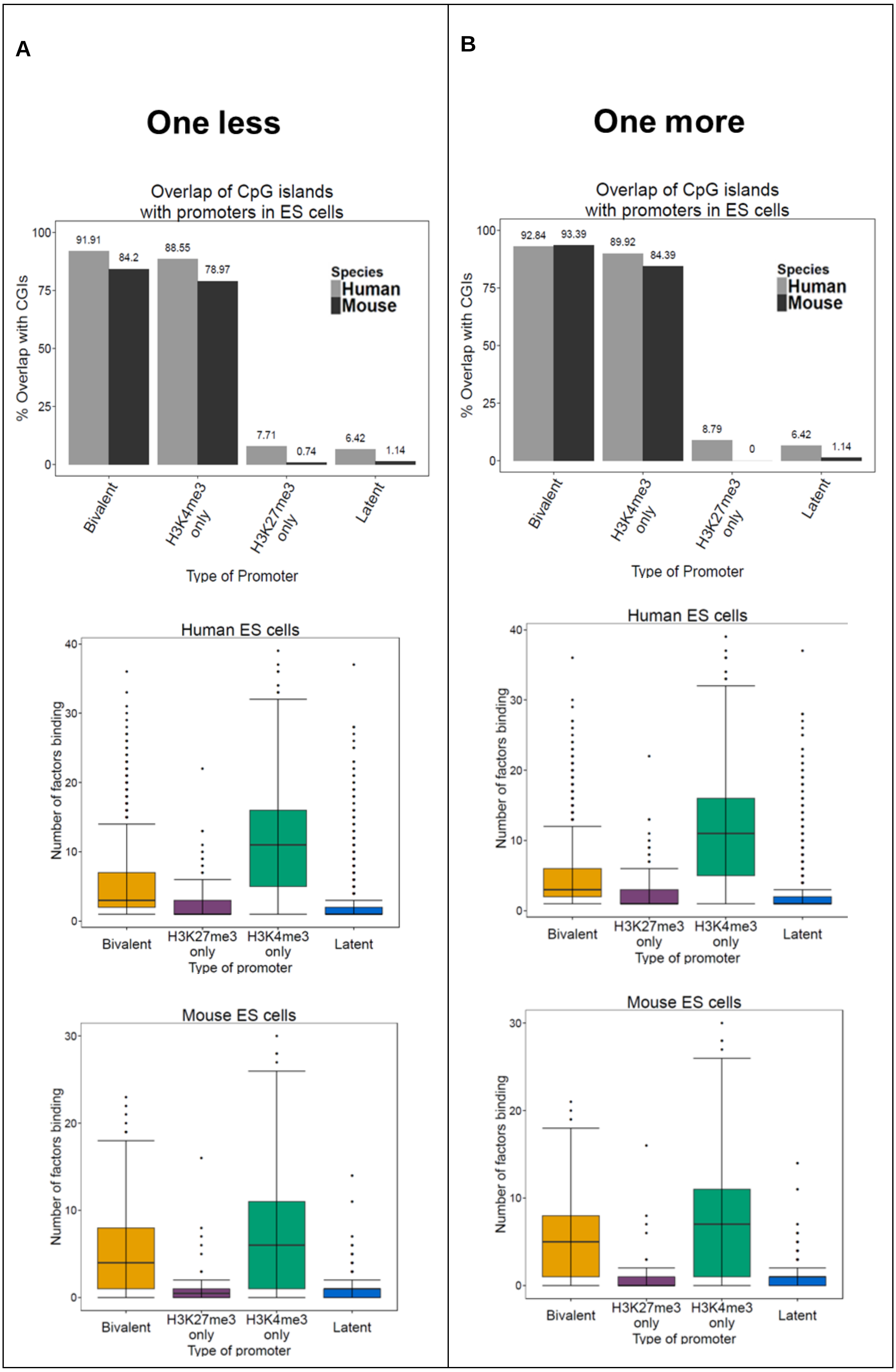
Figure 3.1 Identification of high confidence bivalent promoters in human and mouse ESCs. A. The number of H3K4me3 (green), H3K27me3 (purple) and bivalent (yellow) promoters detected in 'n' or more samples (x axis) in human (left) and mouse ESCs (right). The red dotted line represents cut off used to define high-confidence bivalent promoters. B. H3K27me3 read density (in log scale-natural logarithm-ln) at bivalent and

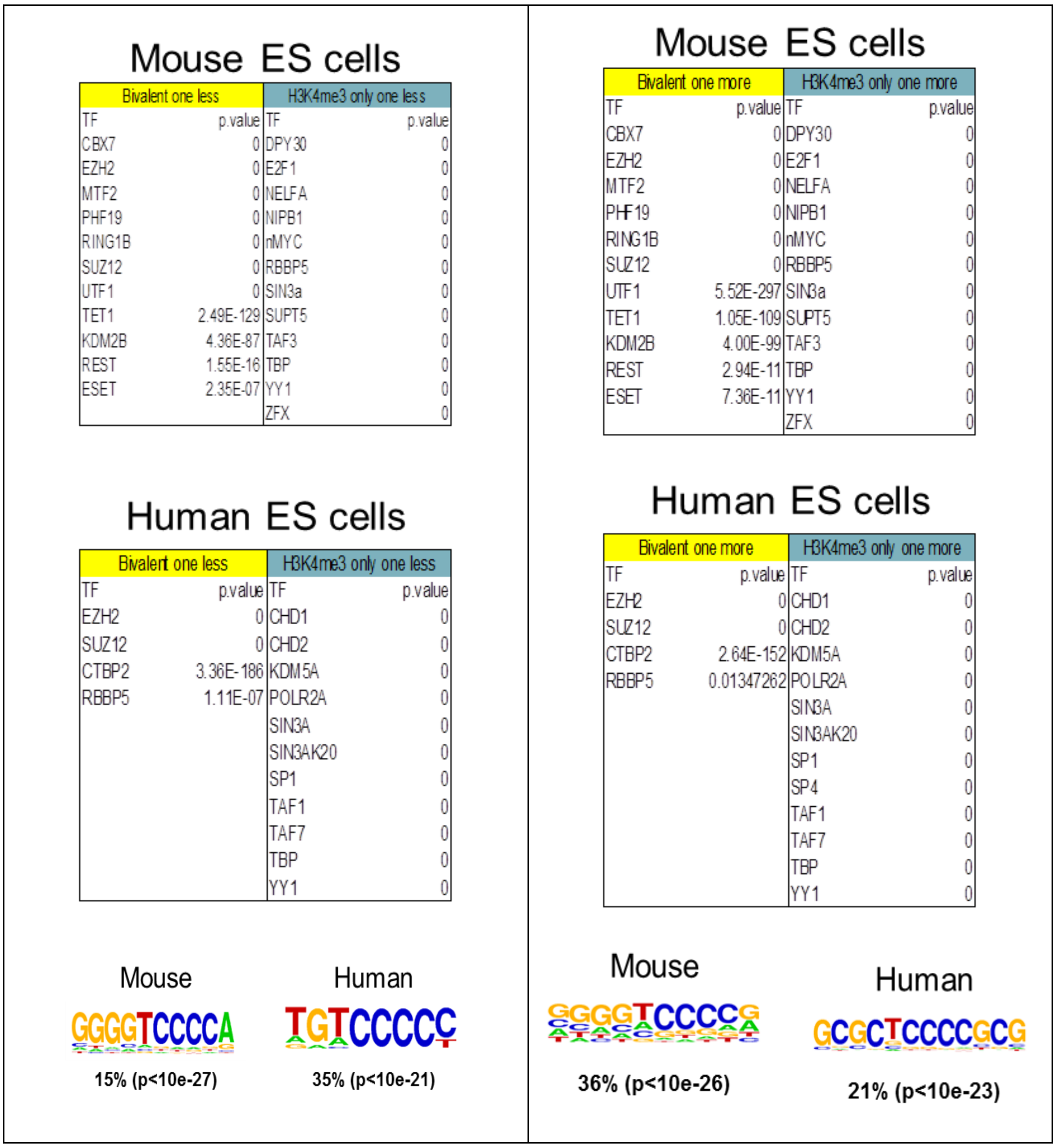
H3K27me3 only promoters in each sample designated by their GEO accession number (x axis) in human (left) and mouse (right) ESCs (* P-value<10^{-4}). C. Gene Ontology terms enriched in HC bivalent promoter list (yellow) or non HC bivalent promoter list (grey) in human (left) and mouse (right) ESCs with their corresponding P-value.**

The sequencing depth across samples varied from 14 million to over 100 million which might contribute to the variation of bivalent promoter detection in individual datasets. Indeed, there was a high correlation between the number of reads and number of peaks across murine datasets (for H3K27me3 Pearson's correlation coefficient (***r)*** = 0.75, for H3K4me3 ***r*** = 0.84), but not across human datasets (for H3K27me3 ***r*** = -0.20, for H3K4me3 ***r*** = 0.14). There are other factors contributing to the variation between samples, for example ESCs were grown in diverse culture conditions, and using different cell lines as well as various antibodies across datasets (Tables 2.1 and 2.2). We therefore defined bivalent promoters identified in more than 70% of the datasets (eight or more human datasets and six or more murine datasets) as high confidence (HC), resulting in 4,979 human and 3,659 murine HC bivalent promoters (Figure 3.1A). Eight HC bivalent regions were validated by ChIP qPCR for the presence of H3K27me3 modification (Mikkelsen et al., 2007) (Table 3.5). Adding or removing a sample in defining HC promoters did not change the key findings of the downstream analysis (Figure 3.2). There was no strong correlation between the fraction of HC bivalent promoters detected in a sample and the sequencing depth of that sample for both histone modifications (Pearson's Correlation: Human: r=-0.34 H3K27me3, r=-0.38 H3K4me3, Mouse: r=0.35 H3K27me3, r=-0.112H3K4me3) (Figure 3.3).

Chr Mikkelsen	*Start Mikkelsen*	*End Mikkelsen*	*Histone mark Mikkelsen*	*Chr*	*Start*	*End*	*Strand*	*Gene Name*	*Ensembl ID*	*Marked*
chr2	118702259	118702859	K27	chr2	118701963	118704964	-	Ankrd63	ENSMUSG00000078137	bivalent
chr15	102955841	102956427	K27	chr15	102953426	102956427	+	Hoxc11	ENSMUSG00000001656	bivalent
chr5	139907676	139908276	K27	chr5	139906942	139909943	+	Elfn1	ENSMUSG00000048988	bivalent
chr10	121310040	121310640	K27	chr10	121309189	121312190	-	Tbc1d30	ENSMUSG00000052302	bivalent
chr3	104961044	104961644	K27	chr3	104959709	104962710	-	Wnt2b	ENSMUSG00000027840	bivalent
chr4	115056768	115057368	K27	chr4	115055425	115058426	+	Tal1	ENSMUSG00000028717	bivalent
chr5	140606626	140607226	K27	chr5	140606340	140609341	+	Lfng	ENSMUSG00000029570	bivalent
chr19	10303905	10304505	K27	chr19	10302877	10305878	-	Dagla	ENSMUSG00000035735	bivalent

Table 3.5 Overlap of validated regions with ChIP-PCR from Mikkelsen et al. with our promoters.


A
One less
Overlap of CpG islands with promoters in ES cells
% Overlap with CGIs
100
75
50
25
0
91.91
84.2
88.55
78.97
7.71
0.74
6.42
1.14
Species
Human
Mouse
Bivalent
H3K4me3 only
H3K27me3 only
Latent
Type of Promoter
Human ES cells
Number of factors binding
40
30
20
10
0
Bivalent
H3K27me3 only
H3K4me3 only
Latent
Type of promoter
Mouse ES cells
Number of factors binding
30
20
10
0
Bivalent
H3K27me3 only
H3K4me3 only
Latent
Type of promoter
B
One more
Overlap of CpG islands with promoters in ES cells
% Overlap with CGIs
100
75
50
25
0
92.84
93.39
89.92
84.39
8.79
0
6.42
1.14
Species
Human
Mouse
Bivalent
H3K4me3 only
H3K27me3 only
Latent
Type of Promoter
Human ES cells
Number of factors binding
40
30
20
10
0
Bivalent
H3K27me3 only
H3K4me3 only
Latent
Type of promoter
Mouse ES cells
Number of factors binding
30
20
10
0
Bivalent
H3K27me3 only
H3K4me3 only
Latent
Type of promoter



Mouse ES cells

Bivalent one less		H3K4me3 only one less	
TF	p.value	TF	p.value
CBX7	0	DPY30	0
EZH2	0	E2F1	0
MTF2	0	NELFA	0
PHF19	0	NIPB1	0
RING1B	0	nMYC	0
SUZ12	0	RBBP5	0
UTF1	0	SIN3a	0
TET1	2.49E-129	SUPT5	0
KDM2B	4.36E-87	TAF3	0
REST	1.55E-16	TBP	0
ESET	2.35E-07	YY1	0
		ZFX	0

Human ES cells

Bivalent one less		H3K4me3 only one less	
TF	p.value	TF	p.value
EZH2	0	CHD1	0
SUZ12	0	CHD2	0
CTBP2	3.36E-186	KDM5A	0
RBBP5	1.11E-07	POLR2A	0
		SIN3A	0
		SIN3AK20	0
		SP1	0
		TAF1	0
		TAF7	0
		TBP	0
		YY1	0

Mouse ES cells

Bivalent one more		H3K4me3 only one more	
TF	p.value	TF	p.value
CBX7	0	DPY30	0
EZH2	0	E2F1	0
MTF2	0	NELFA	0
PHF19	0	NIPB1	0
RING1B	0	nMYC	0
SUZ12	0	RBBP5	0
UTF1	5.52E-297	SIN3a	0
TET1	1.05E-109	SUPT5	0
KDM2B	4.00E-99	TAF3	0
REST	2.94E-11	TBP	0
ESET	7.36E-11	YY1	0
		ZFX	0

Human ES cells

Bivalent one more		H3K4me3 only one more	
TF	p.value	TF	p.value
EZH2	0	CHD1	0
SUZ12	0	CHD2	0
CTBP2	2.64E-152	KDM5A	0
RBBP5	0.01347262	POLR2A	0
		SIN3A	0
		SIN3AK20	0
		SP1	0
		SP4	0
		TAF1	0
		TAF7	0
		TBP	0
		YY1	0

Figure 3.2 Results for CpG enrichment, factor occupancy and factor enrichment remain unchanged when we remove or add one sample from the cut-off. The p-values shown to be 0, should be noted as p-value < extremely small value close to zero. For example, p-value < 1e-256.

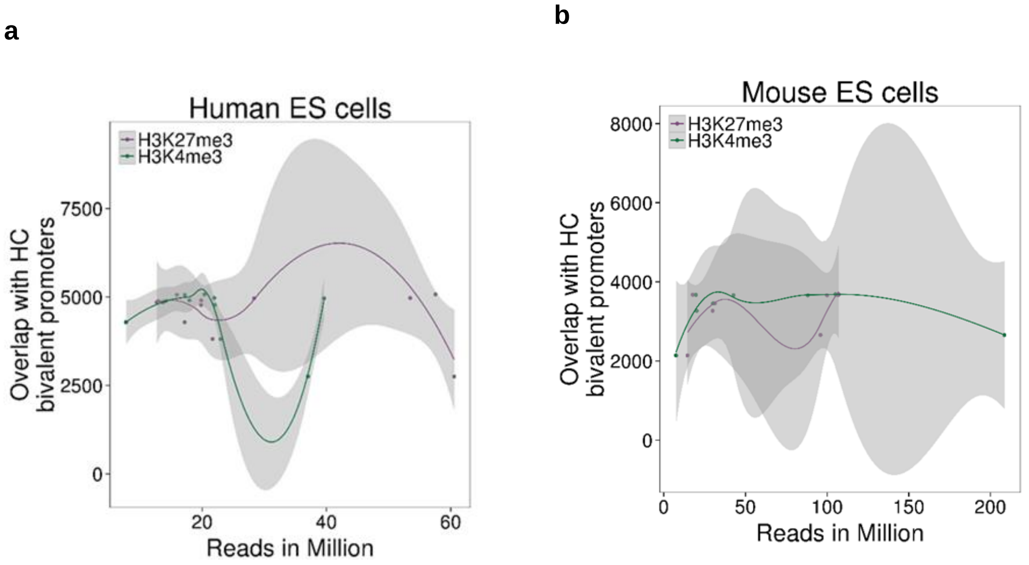
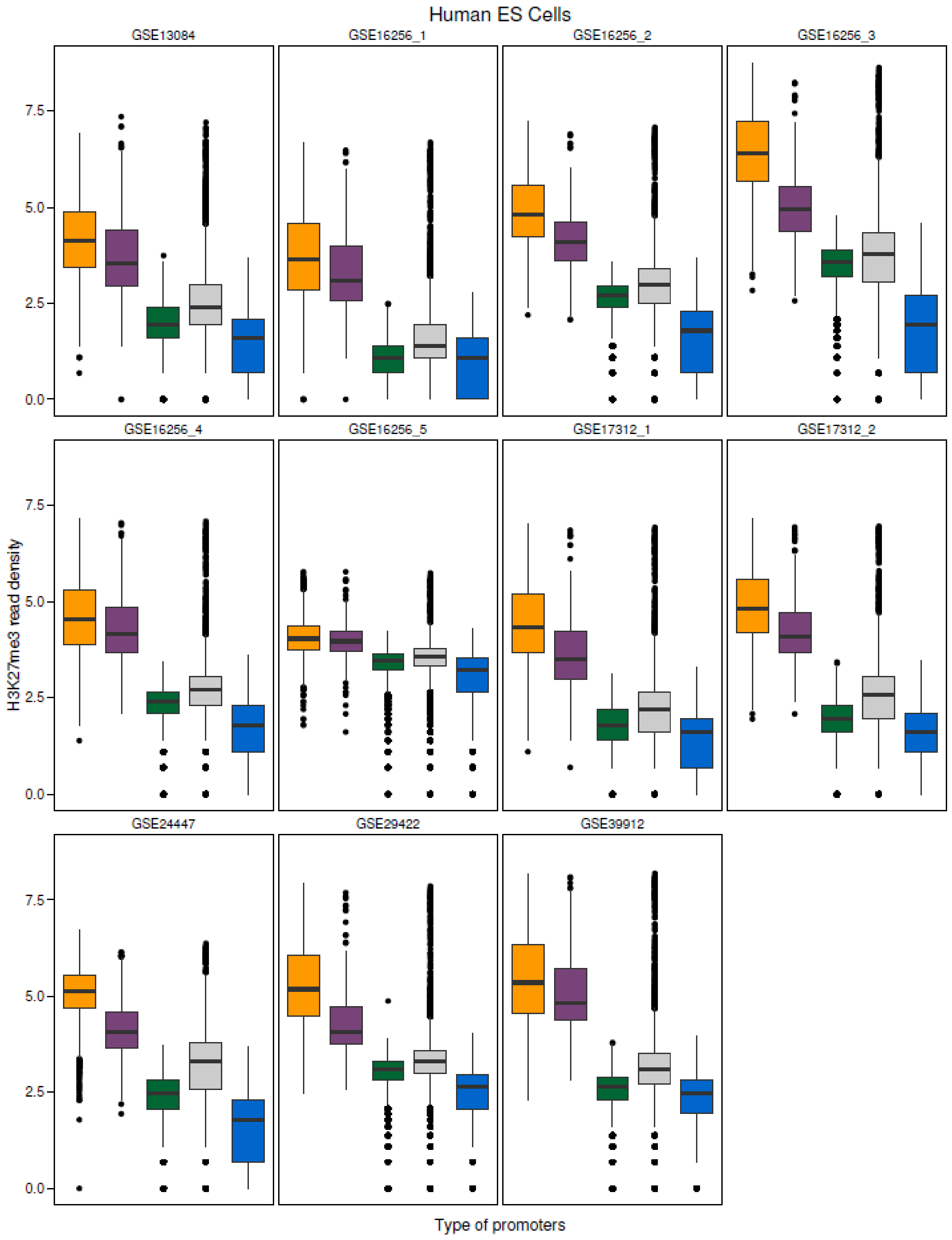


Figure 3.3 Overlap of HC bivalent promoters with bivalent promoters in each sample in a) human and b) mouse ESCs. The correlation of the overlapping HC with the reads of each sample was: For human, r=-0.34 for H3K27me3 samples (purple) and r=-0.38 for H3K4me3. For mouse, r=0.35 for H3K27me3 and r= -0.112 for H3K4me3.

HC bivalent promoters had higher H3K27me3 read density than H3K27me3-only promoters in any individual dataset (Student's t-test, P-value < 0.0001) (Figures 3.1B and 3.4), while H3K4me3 read density at HC bivalent promoters was lower than at H3K4me3-only promoters (Student's t-test, P-value < 0.0001) (Figures 3.5 and 3.6). To test whether integration of multiple samples simply resulted in selecting the peaks with the strongest signal (peak height) from individual H3K27me3 samples, we selected the top (highest H3K27me3 signal) 4,979 human and 3,659 murine bivalent promoter peaks in each dataset and calculated the overlap with HC bivalent promoters. Less than 2/3rd of H3K27me3 top promoters in any individual dataset overlapped with HC bivalent promoters (Figure 3.7).

a


Human ES Cells
GSE13084
GSE16256_1
GSE16256_2
GSE16256_3
GSE16256_4
GSE16256_5
GSE17312_1
GSE17312_2
GSE24447
GSE29422
GSE39912
H3K27me3 read density
0.0
2.5
5.0
7.5
Type of promoters

b

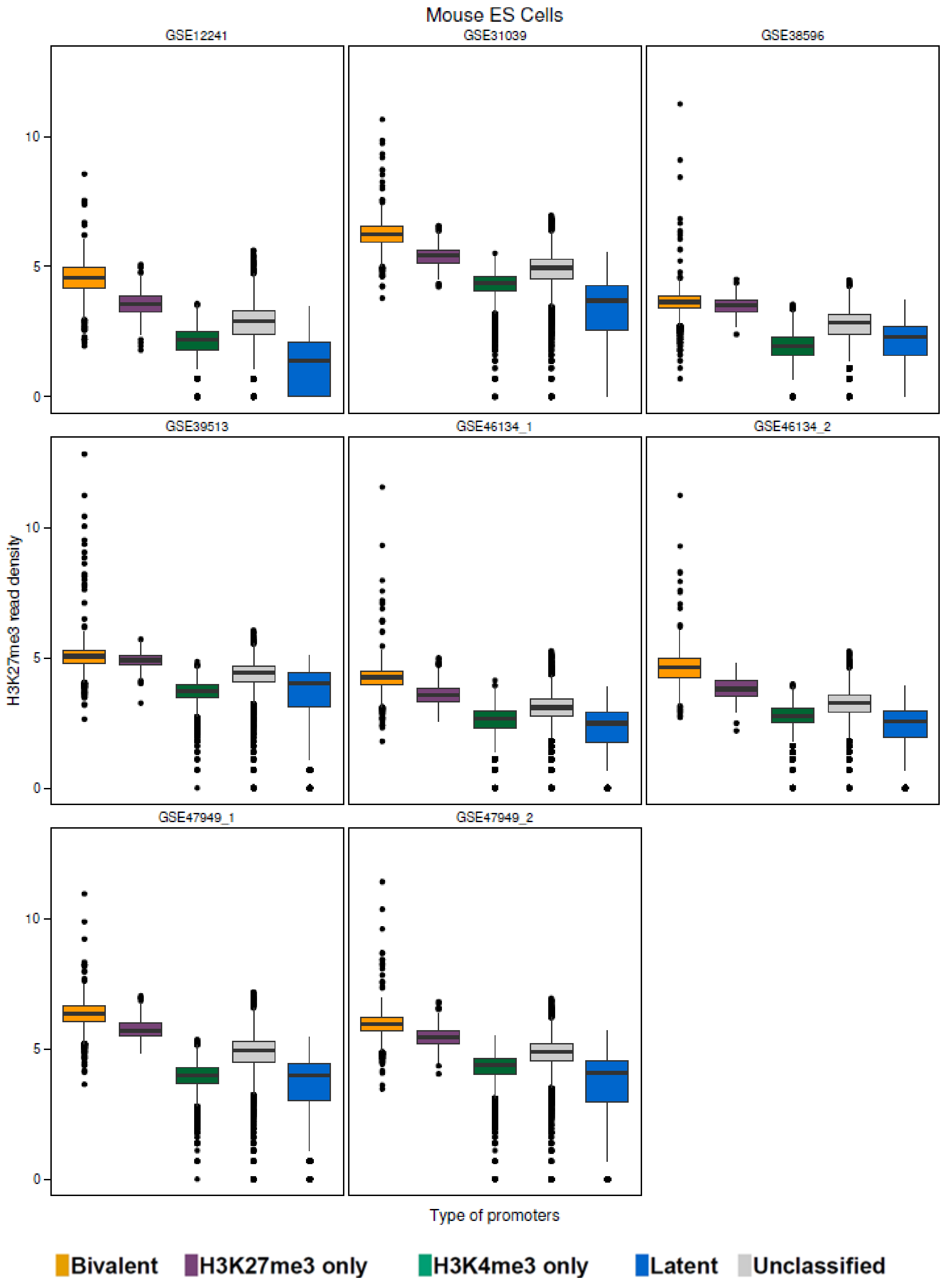
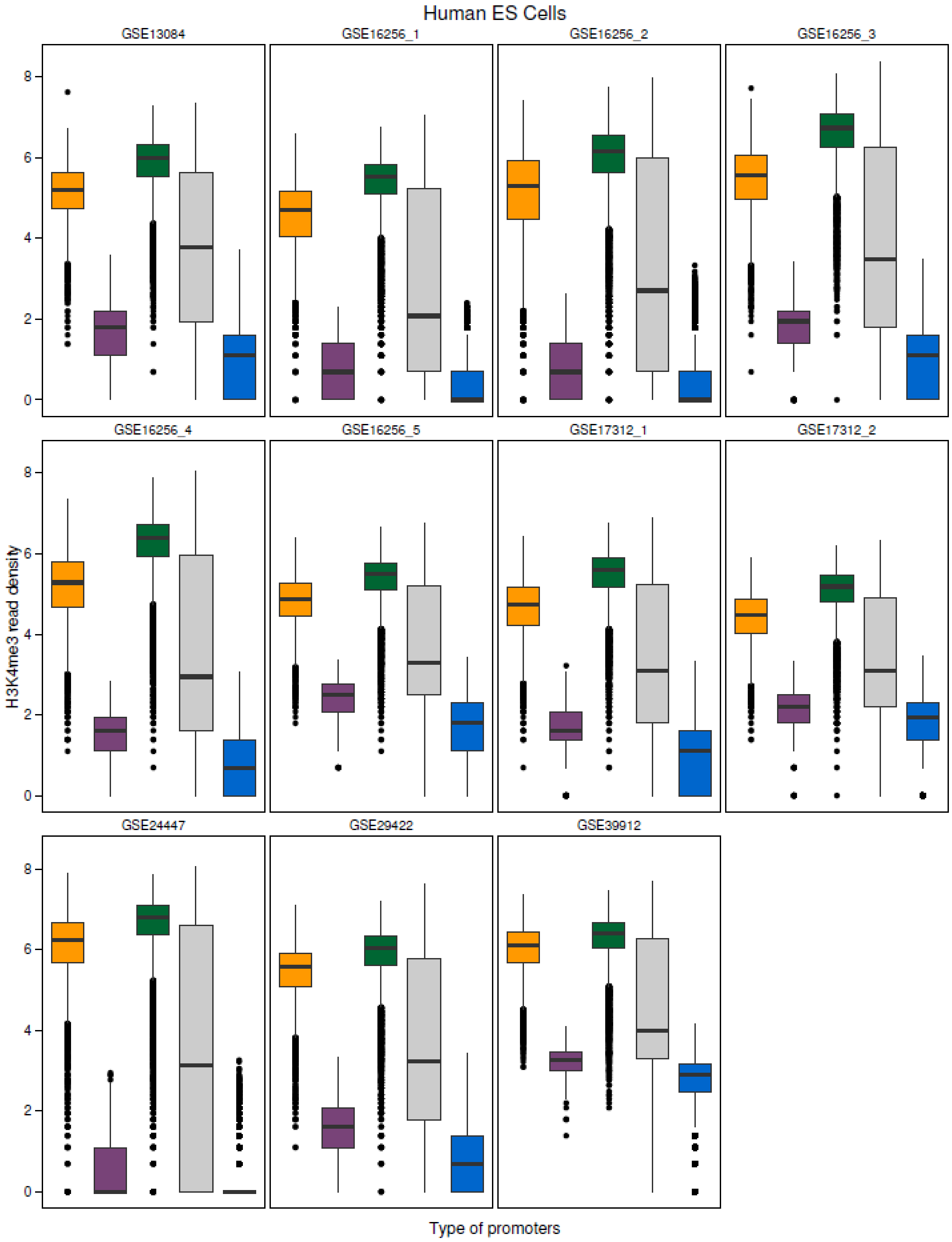


Figure 3.4 Levels of H3K27me3 read density at the promoters according to their classification across samples in a) Human and b) Mouse ESCs

a


Human ES Cells
GSE13084
GSE16256_1
GSE16256_2
GSE16256_3
GSE16256_4
GSE16256_5
GSE17312_1
GSE17312_2
GSE24447
GSE29422
GSE39912
H3K4me3 read density
0
2
4
6
8
Type of promoters

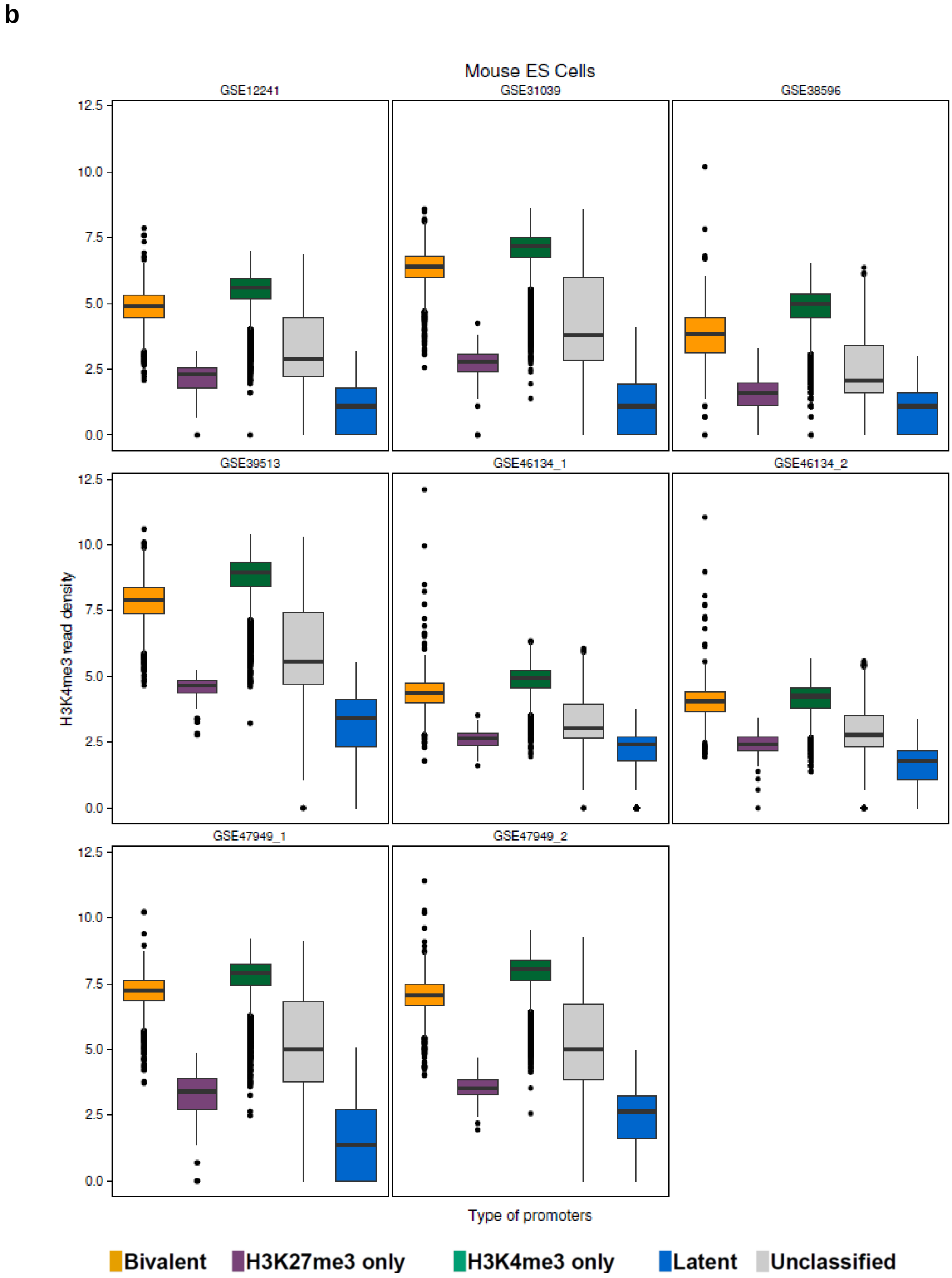


Figure 3.5 Levels of H3K4me3 read density at the promoters according to their classification across samples in a) Human and b) Mouse ESCs

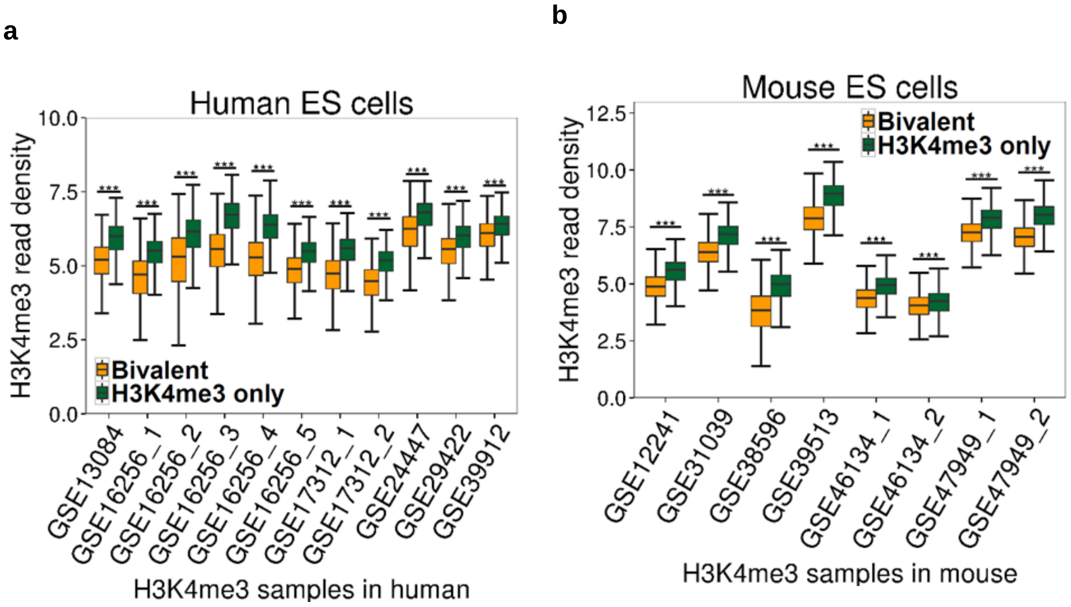


Figure 3.6 H3K4me3 read density at bivalent promoters vs H3K4me3 only promoters in a) human and b) mouse ESCs. (* P-value<10-4)**

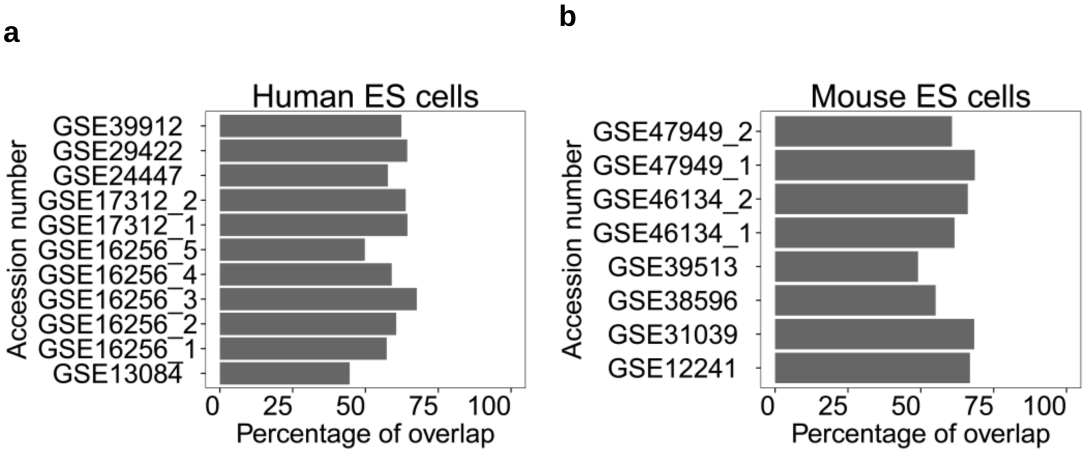


Figure 3.7 Overlap of H3K27me3 top promoters with HC bivalent promoters in any individual dataset in a) Human and b) Mouse ESCs

We also checked whether the peaks of H3K27me3 and H3K4me3 modifications were present at the same genomic location within a promoter region and found that over 95% of H3K27me3 and H3K4me3 peaks overlapped in each pair of samples at HC bivalent promoters. Both chromatin modifications were indeed present at the same genomic location (Figure 3.8). We compared the functional enrichment between high-confidence and non-high-confidence (detected as bivalent in less than 70% of datasets) bivalent promoters and found that only the high-confidence promoters were strongly enriched for processes such as ‘cell differentiation’ and ‘system development’ (Figure 3.1E and 3.1F). Interestingly, metabolic processes were enriched in murine but not human HC bivalent promoters.

In summary, by integrating data from multiple studies we identified HC human and murine bivalent promoters, which could not be identified by simply selecting the top peaks from individual samples. The HC bivalent promoters were highly enriched for developmental regulators compared to non-HC bivalent promoters.

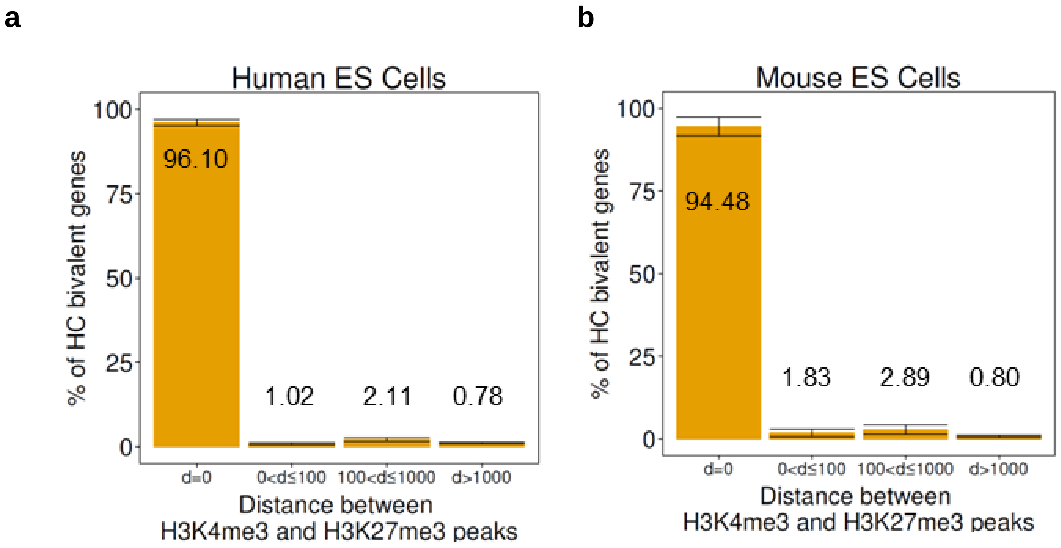


Figure 3.8 Mean distance between H3K27me3 and H3K4me3 peaks in all samples in a) Human and b) Mouse ESCs

3.4.2 High-confidence bivalent promoters are marked by PRC1, PRC2 and RNA Polymerase II

Bivalent promoters are known to show variation in their levels of occupancy by RNA polymerase II (Brookes et al., 2012b) and PRC complexes (Ku et al., 2008a) . To further characterize HC bivalent promoters, we gathered ChIP-seq data in murine ESCs for various forms of RNAPII phosphorylated in different residues (RNAPIIS5P and RNAPIIS7P) as well as RNAPII8WG16 (an antibody that recognizes mostly un-phosphorylated RNAPII) (Brookes et al., 2012b), together with ChIP-seq data for the SUZ12, a subunit of PRC2, responsible for catalysing the histone modification

H3K27me3, the RING1B and CBX7 subunits of PRC1(Morey et al., 2013), responsible for catalysing H2Aub1 and for compacting chromatin, and Jarid2 (Tee et al., 2014). Jarid2 is a co-factor of PRC2 and is methylated by PRC2 which in turn promotes PRC2 activity (Sanulli et al., 2015). All HC bivalent promoters were marked by both PRC1 and PRC2 components albeit at different levels (Figure 3.9A).

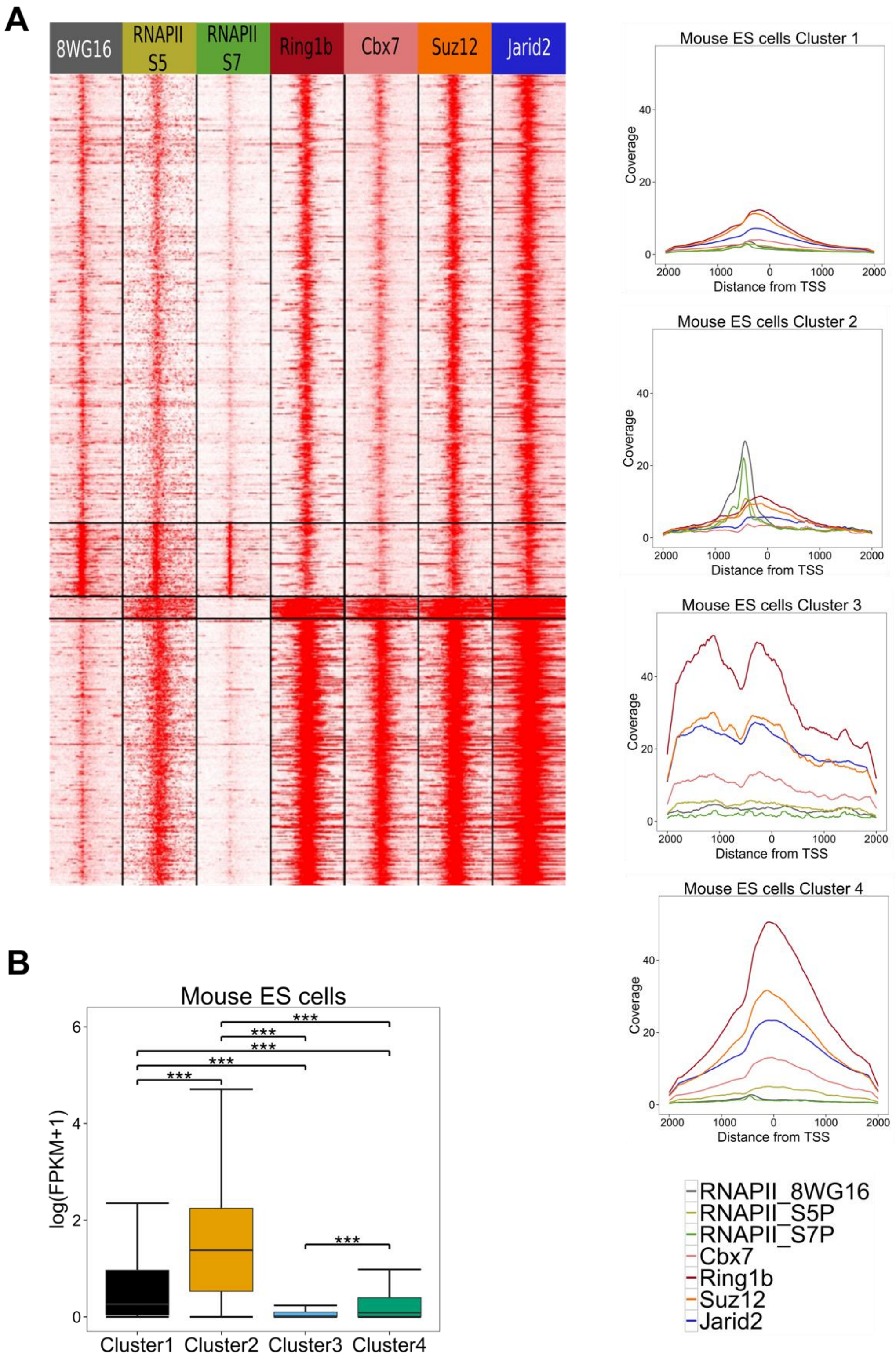


Figure 3.9 Four groups of HC bivalent promoters with distinct biological features A. HC bivalent promoters in murine ESCs classified in four subgroups based on occupancy of PRC1 components (Ring1b, Cbx7), PRC2 (Suz12), Jarid2 and RNA polymerase II (ser7p, ser5p and 8wg16). Each line represents one single promoter while colour code summarized ChIP-seq read densities, from -5kb to +5Kb around TSS. For each cluster,

mean read coverage around TSS is shown on the right. B. Expression levels in mouse ESCs using RNA sequencing data for each of the four clusters. FPKM: Fragment per kilo-base per million (* P-value<10^{-4}).**

HC bivalent promoters could be classified in four distinct clusters based on the presence of PRC1 components and forms of RNAPII (Figure 3.9A). The first two clusters had low PRC1 (Ring1b) levels and high RNAPII (8WG16) levels compared to clusters 3 and 4. The second cluster distinguished from the first cluster by the presence of RNAPII (8WG16 and S5P) modifications as a sharp peak on the promoter. The second cluster consisted of the only group of bivalent promoters marked with RNAPII (S7P). This cluster was enriched for genes involved in metabolic processes. The third and fourth clusters were marked by strong PRC1 (Ring1b), PRC2 (Suz12) and RNAPII (S5P) modifications. Cluster 3 and 4 were distinguished based on the fact that PRC components formed wide domains on cluster 3 and narrow peaks on cluster 4 promoters. Cluster 3 promoters were enriched for regulation of transcription (P value < 10^{-51}) while cluster 4 promoters were enriched for developmental functions such as organ morphogenesis (P value < 10^{-27}). Cluster 3 promoters contained transcription factors important for specific lineages like haematopoiesis factors Gfi1 and Meis1, whereas Cluster 4 contained multiple members of transcription factor families controlling development such as winged helix/forkhead box (Fox) and Hox families.

We noted that bivalent promoters could be distinguished into two groups based on PRC1 occupancy: PRC1 low (cluster 1 & 2) and PRC1 high (cluster 3 & 4). Ku et al. (2008) suggested that PRC1 was absent in our PRC1 low bivalent promoter (Figure 3.10). Ring1b ChIP sequencing at higher sequencing depth confirms that all bivalent promoter are bound by PRC1 albeit at different levels. The PRC1 high group separated into two distinct groups each enriched for a distinct functional category, namely cluster 3 for transcription factors and cluster 4 for developmental controllers. Based on RNAPII occupancy, PRC1 low consisted of two distinct gene sets: RNAPII-low (S7P) and RNAPII-high (S7P). The difference in chromatin signature of these two clusters was also reflected in the expression level namely RNAPII-high (cluster 2) promoters were expressed at higher levels than RNAPII-low or cluster 1 promoters (Kruskal-Wallis test P-value < 0.0001) (Figure 3.9B).

In summary, all HC bivalent promoters are occupied by components of both PRC1 and PRC2. There exists a distinct set of metabolic genes (cluster 2) which though bivalently marked has RNAPII (S7P) and is expressed at a higher level than other bivalent genes.

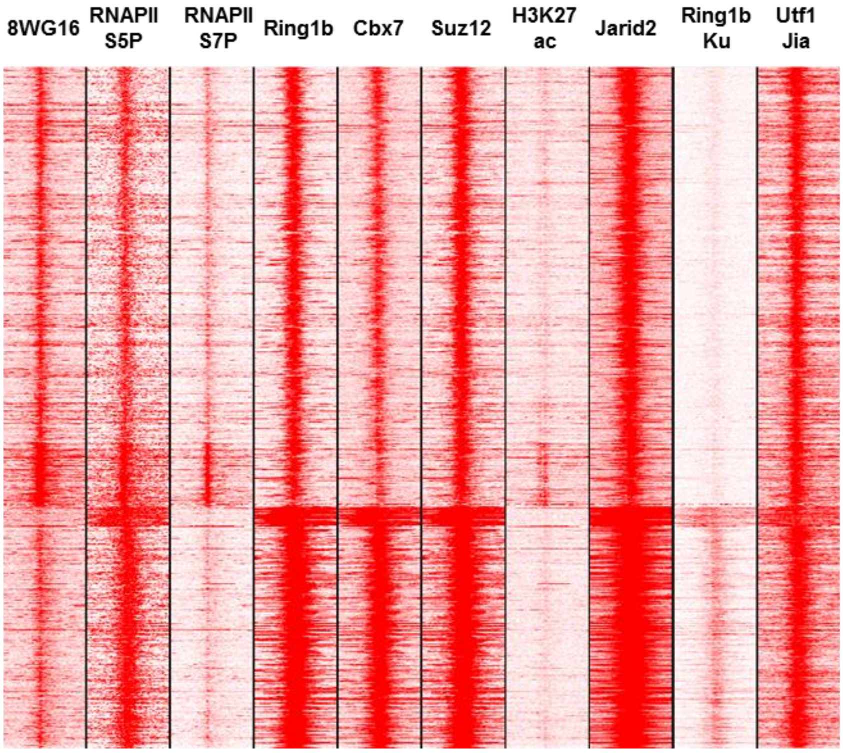


Figure 3.10 Clustering of HC bivalent promoters in mouse ESCs reveals four different groups of bivalent promoters with either low or high levels of PRC1 (Ring1b). Ring1b-Ku (Ku et al., 2008a) **sample shown almost no signal in the first two clusters where Ring1b is low. Utf1** (Jia et al., 2012a) **is present throughout all the clusters.**

3.4.3 Bivalent promoters are lowly expressed and highly sensitive to perturbations in ESCs

RNAPII may be present but stalled at the promoters of bivalent genes and short (abortive) transcripts may be detected at their promoters (De Gobbi et al., 2011). To check whether bivalent genes indeed show a low or leaky expression, we collected RNA sequencing data for murine (Yu et al., 2013) and human (Djebali et al., 2012) ESCs and calculated the mean expression level for the following categories of promoters. We classified promoters into four HC groups (Appendix Tables 1 and 2) depending on the presence or absence of one or both chromatin modifications in over 70% of samples as bivalent promoters, promoters marked only with H3K27me3 (H3K27me3-only), promoters marked only with H3K4me3 (henceforth called ‘active’) and latent promoters (unmarked for H3K27me3 and H3K4me3). Promoters that belonged to any of the previous four categories in less than 70% of the samples, and thus were not considered in that category were marked as unclassified. In human and mouse ESCs, most active promoters were expressed at higher levels than bivalent promoters, and latent promoters were mostly not expressed (FPKM = 0) (Kruskal-Wallis test, P-value < 0.0001) (Figure 3.11A). Low expression can result from two scenarios: either a gene is expressed at low levels in most cells or few cells express a gene while others do not. To determine whether lowly expressed genes in the four groups can be classified into one of the two scenarios, we downloaded single cell RNA sequencing data for 63 mouse ESCs (Streets et al., 2014). Lowly expressed (i.e. FPKM < 4, or log(FPKM) < 1.4) active promoters were expressed in a similar number of single cells as lowly expressed bivalent promoters (Kruskal-Wallis test, P-value > 0.05) (Figure 3.11B) demonstrating that single cell gene expression data cannot distinguish between bivalent and active lowly expressed genes.

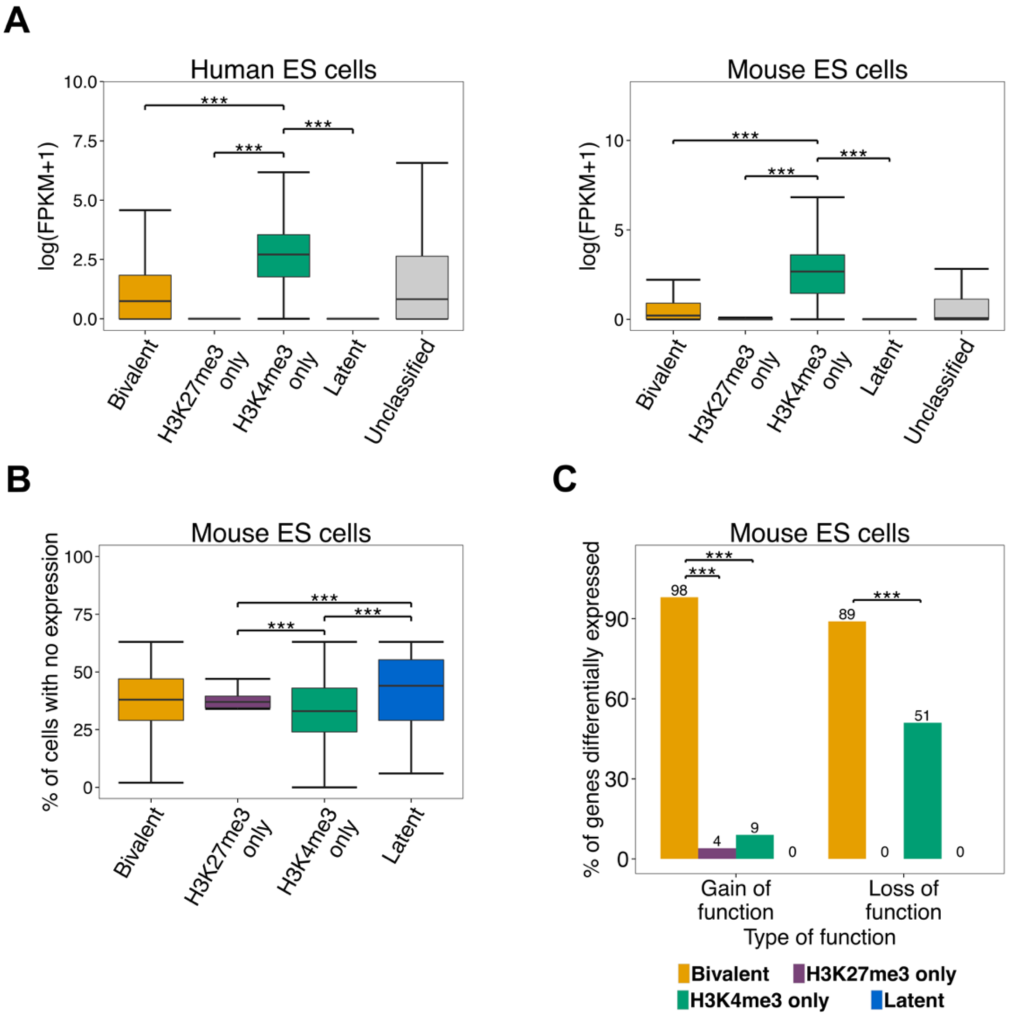


Figure 3.11 Bivalent promoters are lowly expressed in ESCs, are more likely to be differentially expressed upon perturbation. A. Expression levels according to human (left) or mouse (right) ESCs RNA-seq for HC bivalent promoters (yellow), promoters marked with H3K27me3 only (purple), promoters marked with H3K4me3 only (green), latent promoter (blue), or unclassified promoters (grey, see text) (*P-value<10-4). B. From single cell RNA-seq data of mouse ESCs, percentage of cells non expressing the lowly expressed genes (i.e. FPKM < 4) was computed for different classes of promoters (bivalent, H3K27me3 only, H3K4me3 only and latent) (*** P-value <10-4). C. HC bivalent promoters are hypersensitive to changes in the transcription network perturbation. Differentially expressed gene lists were collected from studies overexpressing one of 54 factors (gain of function) or down-regulation of one of 37 factors in ESCs. Percentage of significantly overlapping (P value < 1e-3) bivalent, H3K27me3 only, H3K4me3 only and latent genes with differentially expressed in at least one of the experiments is represented (*** P-value<10-4).**

As bivalent genes are thought to be poised for activation or repression, we hypothesised that these genes might be more likely to be differentially expressed upon perturbation of ESCs. We therefore used a collection of differentially expressed genes upon deletion or over-expression of 91 transcription and epigenetic factors in mouse ESCs, and found that 98% of differentially expressed gene sets by the overexpression of at least one TF significantly overlapped (Hypergeometric test, P value < 1e-3) with bivalent genes, and 89% differentially expressed gene sets by the down-regulation of at least one TF (Figure 3.11C). To check whether this is a property of bivalent genes or lowly expressed genes in general, we also calculated the overlap of active and latent lowly expressed genes with the differentially expressed gene sets upon transcription and epigenetic factor perturbation. We confirmed that bivalent genes are highly susceptible to perturbations compared to active or latent lowly expressed genes (Kruskal-Wallis test, P-value < 0.001) (Figure 3.12).

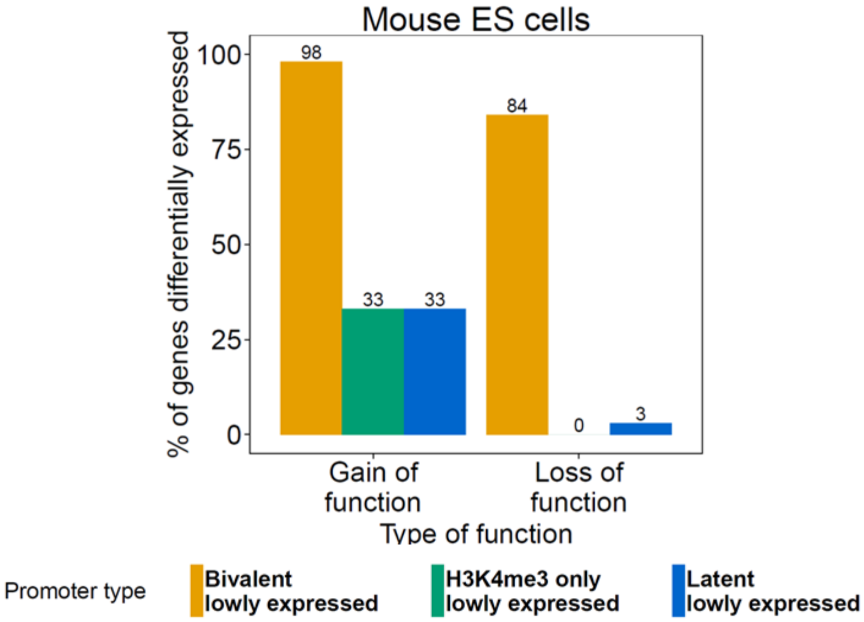
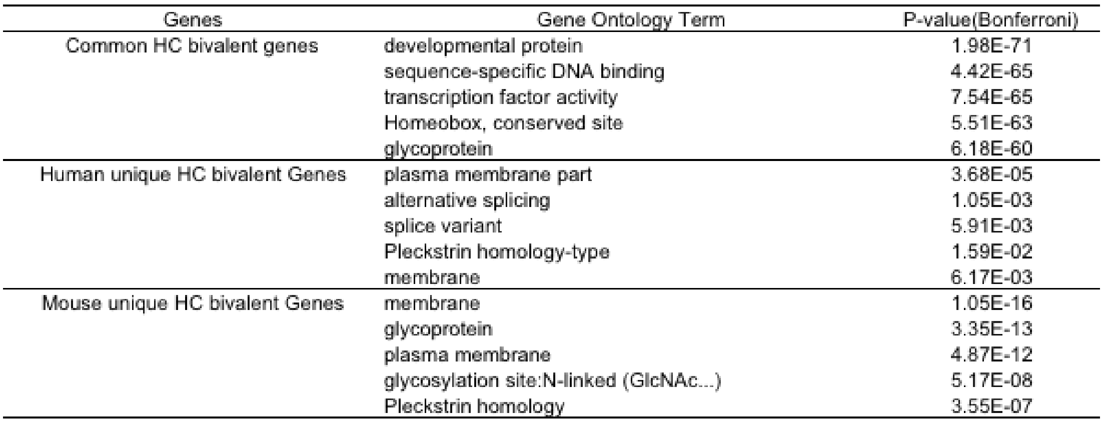


Figure 3.12 Perturbation of lowly expressed HC bivalent, lowly expressed HC H3K4me3 only and lowly expressed HC latent genes when there is gain or loss of function.

3.4.4 Over 50% of bivalent promoters maintain their chromatin status as well as gene expression profile across species

To perform a systematic comparison of chromatin status between human and mouse promoters in ESCs, we used 16,639 one-to-one orthologous genes between the two species (Guberman et al., 2011). We classified orthologous promoters into four HC groups – active (H3K4me3-only), H3K27me3-only, bivalent and latent. Promoters that did not belong to any of the previously mentioned groups were designated as 'unclassified'. We confirmed that HC H3K27me3-only and active promoters indeed had low or no other chromatin modification (Figures 3.4 and 3.5). We then calculated the overlap of the five groups across species (Figure 3.13A). Over 40% of murine orthologous promoters (n=6964) contain an activating mark (H3K4me3-only), in contrast to only 24% of human orthologous promoters (n=3961). There was a 47% overlap of murine active promoters with human active promoters; while 84% of human active promoters overlapped with murine active promoters i.e. most active promoters in human are also active in mouse but not vice versa. Bivalent promoters constitute 17% (n=2854) and 20% (n=3342) of mouse and human orthologous genes respectively. 66% of murine bivalent promoters are also bivalent in human and 56% of human bivalent promoters are bivalent in mouse. The promoters with the H3K27me3-only modification form a very small fraction of orthologous promoters reaching merely 0.2% (n=45) and 0.3% (n=66) in mouse and human respectively. About 20% of H3K27me3-only promoters in one species are bivalent in the other species. Conserved bivalent promoters were enriched for functional categories developmental protein (P value < 10^{-71}) and transcription factor activity (P value < 10^{-65}); whereas species-specific promoters were not enriched for the two above terms (Table 3.6). Specifically, the mouse-specific bivalent promoters were enriched for membrane (P value < 10^{-16}) and glycoprotein (P value < 10^{-13}) and the human-specific for plasma membrane part (P value < 10^{-5}) and alternative splicing (P value < 10^{-3}).



Genes	Gene Ontology Term	P-value(Bonferroni)
Common HC bivalent genes	developmental protein	1.98E-71
	sequence-specific DNA binding	4.42E-65
	transcription factor activity	7.54E-65
	Homeobox, conserved site	5.51E-63
	glycoprotein	6.18E-60
Human unique HC bivalent Genes	plasma membrane part	3.68E-05
	alternative splicing	1.05E-03
	splice variant	5.91E-03
	Pleckstrin homology-type	1.59E-02
	membrane	6.17E-03
Mouse unique HC bivalent Genes	membrane	1.05E-16
	glycoprotein	3.35E-13
	plasma membrane	4.87E-12
	glycosylation site:N-linked (GlcNAc...)	5.17E-08
	Pleckstrin homology	3.55E-07

Table 3.6 Gene ontology terms for the conserved and unique to species HC bivalent genes in Human and Mouse ESCs

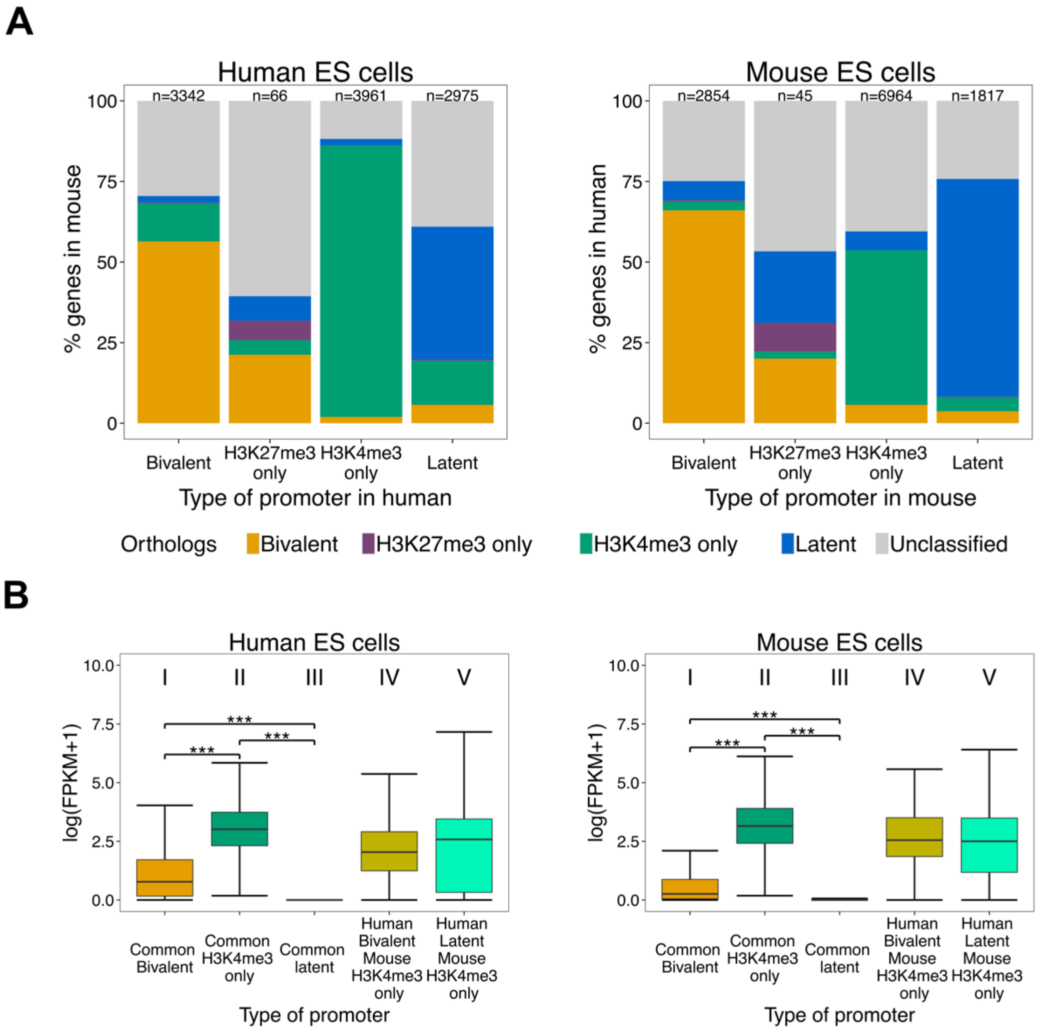


Figure 3.13 Over 50% of bivalent promoters maintain their chromatin status as well as gene expression profile across species A. Overlap of high confidence (HC) H3K4me3 only (green), H3K7me3 only (purple), bivalent (yellow) and latent (blue, absence of both H3K4me3 and H3K7me3 modifications) in human ESCs with the corresponding

categories in mouse ESCs (left), and vice versa (right). Grey: Unclassified promoters (see text). B. Expression levels in human (right) and mouse (left) ESCs using RNA sequencing data for each of the five groups of orthologous genes identified in A (* P-value<10^{-4}).**

To check whether the chromatin status across species is reflected in the gene expression status, we focused on five groups of promoters (Figure 3.13B): three groups (I, II and III) with conserved chromatin status and two groups with divergent chromatin status (IV and V) across species. The gene expression profiles of conserved chromatin groups across species were also conserved. Specifically, active promoters (II) were expressed at higher level than bivalent promoters (I) which in turn were expressed at higher level than latent promoters (III) in both human and mouse ESCs (Kruskal-Wallis test, P-value < 0.0001) (Figure 3.13B). The divergence of chromatin status promoters across species was not reflected in the gene expression level. For example, the orthologous promoters with bivalent status in human and active status in mouse (IV) were expressed at intermediate levels between active (II) and bivalent promoters (I) in both species (Figure 3.13B).

3.4.5 Bivalent promoters are CpG-rich while H3K27me3-only promoters are CpG-poor

As shown in the first section, the bivalent status of promoters is primarily determined by the detection of an H3K27me3 modification (Figure 3.1A). CpG islands (CGIs) have been implicated in Polycomb recruitment and therefore H3K27me3 modification (Deaton and Bird, 2011; Farcas et al., 2012; Riising et al., 2014). CGIs are CpG-rich genomic regions and are sites of transcription initiation (Saxonov et al., 2006). CGI promoters are silenced by either DNA methylation or Polycomb group proteins with approximately a fifth of CGI promoters accounting for bivalent promoters in ESCs (Ku et al., 2008a). About 35% of all GENCODE genes in both human and mouse overlapped with at least one CGI. When only protein coding genes were considered, this overlap increased to 67% for human and 54% for mouse (Figure 3.14A). Mouse promoters in most categories showed lower overlap with CGIs than human promoters (Figure 3.14A). 89% of human active (H3K4me3-only) as well as 82% of murine active promoters contained at least one CGI (Figure 3.14A).

Over 90% of our HC bivalent promoters in ESCs in both species overlap with at least one CGI region, whereas only 8% (37 of 397) of human H3K27me3 only promoters contained a CGI and no mouse H3K27me3 only promoters (none of 152) contained a CGI (Figure 3.14A). Previously CGIs have been associated with H3K27me3 modification in mammalian ESCs (Lynch et al., 2012; Mendenhall et al., 2010), but our results show that this is the case for bivalent promoters but not for H3K27me3 only promoters. We confirmed that the lack of CGIs on active promoters is not due to the CGI detection threshold and that the CpG density at repressed promoters is indeed significantly lower than at CGIs (Kruskal-Wallis test, P-value < 0.0001) (Figure 3.14B). It has been proposed that a high density of un-methylated CpG is sufficient for vertebrate Polycomb recruitment (Mendenhall et al., 2010). The fact that H3K27me3-only promoters are specifically CpG-poor (Figure 3.14A and 3.14B), suggests that, although highly unmethylated CpG islands might be sufficient for Polycomb recruitment, they might not be necessary.

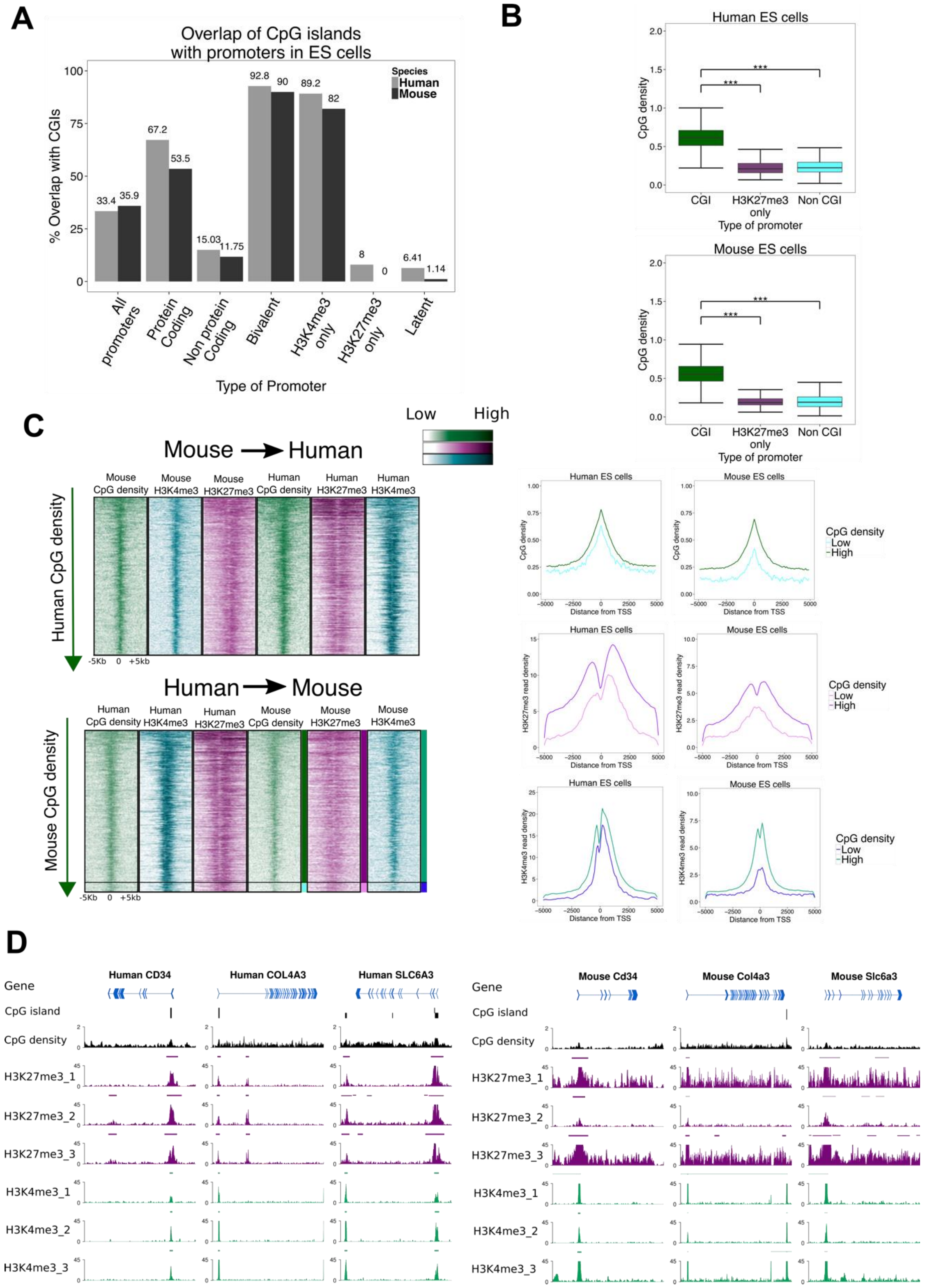


Figure 3.14 Bivalent promoters are CpG island rich while H3K27me3 only are CGI poor. A. Percentage of promoters overlapping with one or more CpG island in human (grey) or mouse (black). B. CpG ratio at H3K27me3 only promoters is similar to non-CGI promoters in human (top) and mouse (bottom) ESCs (* P-value<10^{-4}). C. Relationship between CpG density, H3K27me3 modification and H3K4me3 modification in human and mouse ESCs. There is a loss of human CGI promoters in mouse (bottom, below**

marked black line) but no loss of mouse CGI promoters in human (top). This loss is linked with decreasing H3K4me3 and H3K27me3 in mouse as compare with human. Left panels indicate mean CpG densities, mean H3K27me3 read densities and mean H3K4me3 densities in human and mouse. D. Exemplar murine promoters where CGI loss on promoters does not correspond to the loss H3K27me3 modification. These promoters despite losing CGI keep bivalent promoter status in murine ESCs.

The loss of H3K27me3 in rodents (mouse and rat) compared to human ESCs at many developmental genes has been associated with depletion of CGIs; mouse CGI erosion has been characterised at MYO1G, CLEC4G and MYF6 gene loci with corresponding H3K27me3 loss(Lynch et al., 2012). We performed a cross-species comparison of CpG density, H3K4me3 and H3K27me3 profiles of bivalent promoters (Figure 3.14C). Indeed, about 5% of bivalent human promoters lost CGIs in mouse but not vice versa (indicated by black horizontal line). There was a high correlation between CpG density and H3K4me3 as well as H3K27me3 profiles within each species as well as across species (Figure 3.14C), but the concordance between loss/gain of CGIs and H3K4me3 and/or H3K27me3 mark does not always hold true. Of 70 orthologous CpG-rich bivalent promoters in human where CGI was lost in mouse and analysed their chromatin status, only 18% of these promoters had clearly lost their H3K27me3 mark in mouse ESCs, of which half were classified as H3K4me3-only and the rest as latent in murine ESCs (Figure 3.15). Despite losing CGI on murine promoters, 20% of these orthologous promoters maintained a bivalent chromatin status including Col4a3, Cd34 and Slc6a3 (Figure 3.14D).

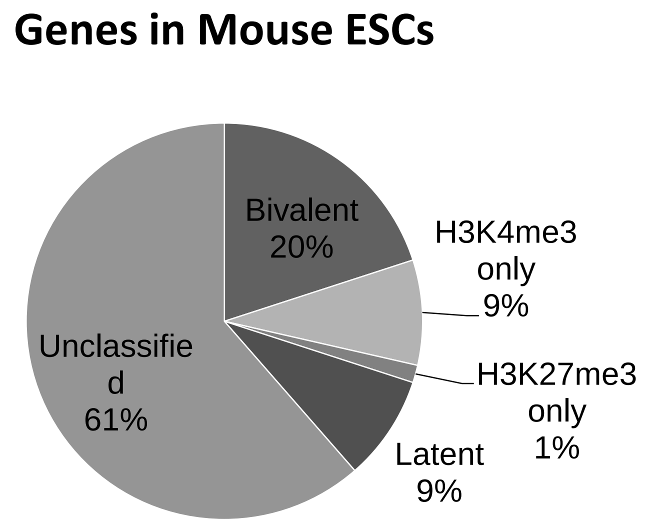


Figure 3.15 Chromatin status low CpG density promoters in mouse ESCs where corresponding human promoters are CpG-rich and bivalent.

In summary, the H3K27me3-only CpG-poor promoters demonstrate that Polycomb recruitment does not only depend on CpG density. Although the CpG density largely correlates with H3K4me3 and H3K27me3 profiles across promoters, the loss of CGI on a promoter does not always imply a corresponding loss of the H3K4me3 and/or H3K27me3 modification on that promoter.

3.4.6 Bivalent promoters are occupied by fewer transcription factors than active promoters and are specifically enriched in a 'TCCCC' sequence motif

As both active (H3K4me3-only) and bivalent promoters are CpG-rich, we investigated possible modes of distinction between the two in ESCs. Voigt, Tee, and Reinberg (2013) proposed a model where the density of transcription factors at the promoters determines establishment of bivalent domains. Specifically, the model suggests that PcG proteins are inhibited from binding at active promoters by an abundance of transcription factors, while at promoter sites with a low occupancy of transcription factors, PcG proteins can easily be recruited at CpG islands to establish the H3K27me3 modification. To test this model, we used publicly available genome-wide TF and epigenetic modifier binding profiles (ChIP-seq data) in murine and human ESCs (Pooley et al., 2014) and calculated the number of transcription factors bound (TF density) at the four classes of promoters. Indeed, the TF density decreases from active to bivalent to H3K27me3-only promoters in both human and mouse ESCs (Kruskal-Wallis test, P-value < 0.0001) (Figure 3.16A).

To identify factors preferentially binding to bivalent promoters, we calculated the overlap between transcription and epigenetic factor binding sites (peaks) and bivalent promoters. Four out of 49 and eleven out of 99 factors characterised by ChIP-seq preferred bivalent promoters in human and mouse respectively (Figure 3.16B). As expected, members of the PcG family were enriched at both human and mouse bivalent promoters (P value $< 10^{-256}$). Moreover, the co-repressor c-terminal binding protein 2 (CTBP2), required for PcG recruitment in Drosophila (Srinivasan and Atchison, 2004), and the RBBP5 (MLL subunit) were enriched at human bivalent promoters (P value < 0.005). The components of both PRC2 (Ezh2, Suz12) and PRC1 (Cbx7, Ring1b) together with two Polycomb-like proteins (Mtf2, Phf9) were enriched at mouse bivalent promoters. Mtf2 and Phf19 recruit the PRC2 complex and are thought to silence transcriptionally active loci (H3K36me3) by recruiting H3K36me3 histone demethylases such as Kdm2b to further recruit PRC2 components for H3K27me3 (Ballaré et al., 2012; Brien et al., 2012; Musselman et al., 2012). Accordingly, Kdm2b was also enriched at mouse bivalent promoters (P value $< 10^{-3}$). Four other epigenetic regulators, Utf1, Tet1, Rest and Setdb1 were highly enriched at mouse bivalent regions. Utf1 (P value $< 10^{-256}$) was recently identified as a component of bivalent chromatin by acting as a buffer against full activation of bivalent genes (Jia et al., 2012a).

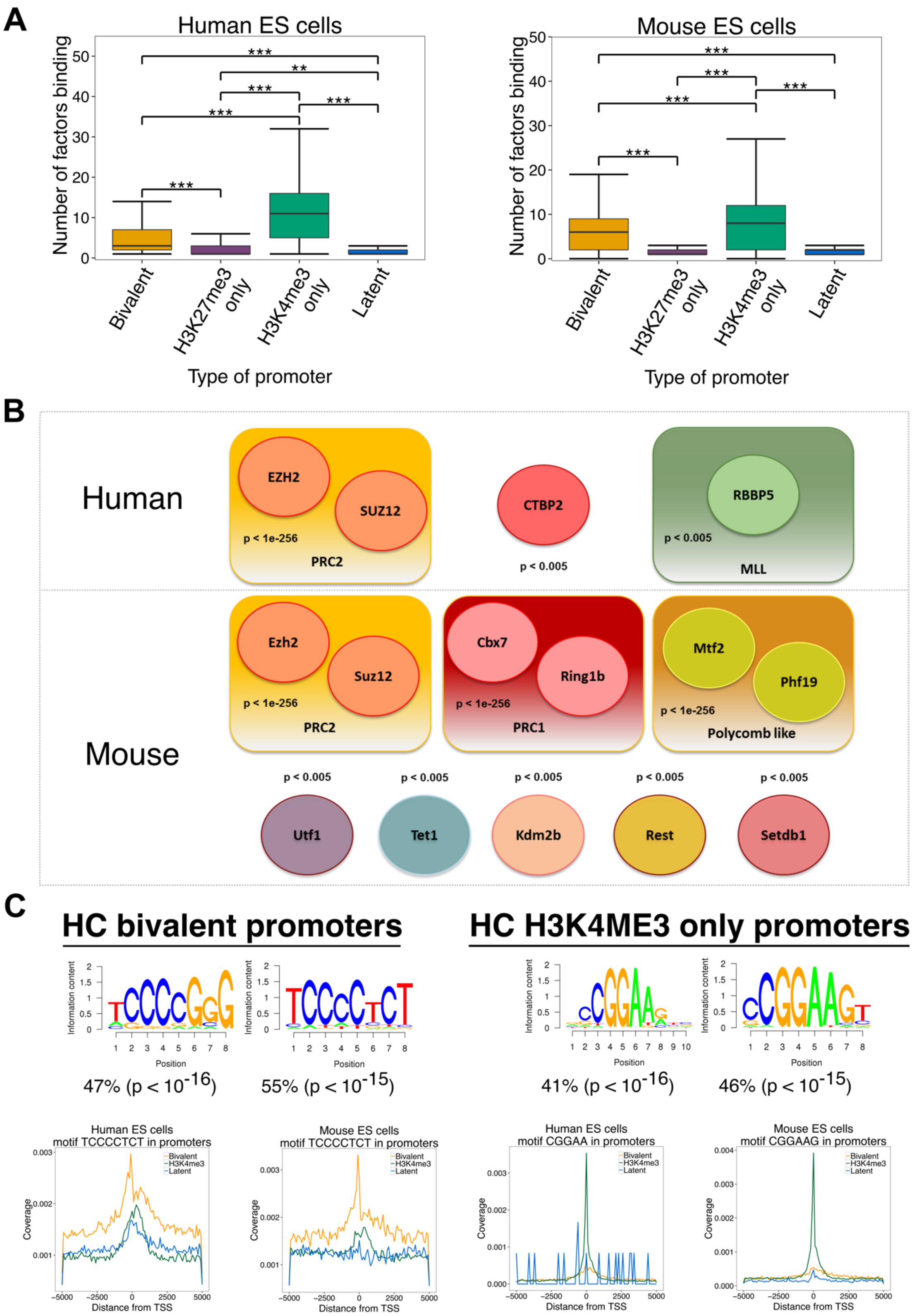
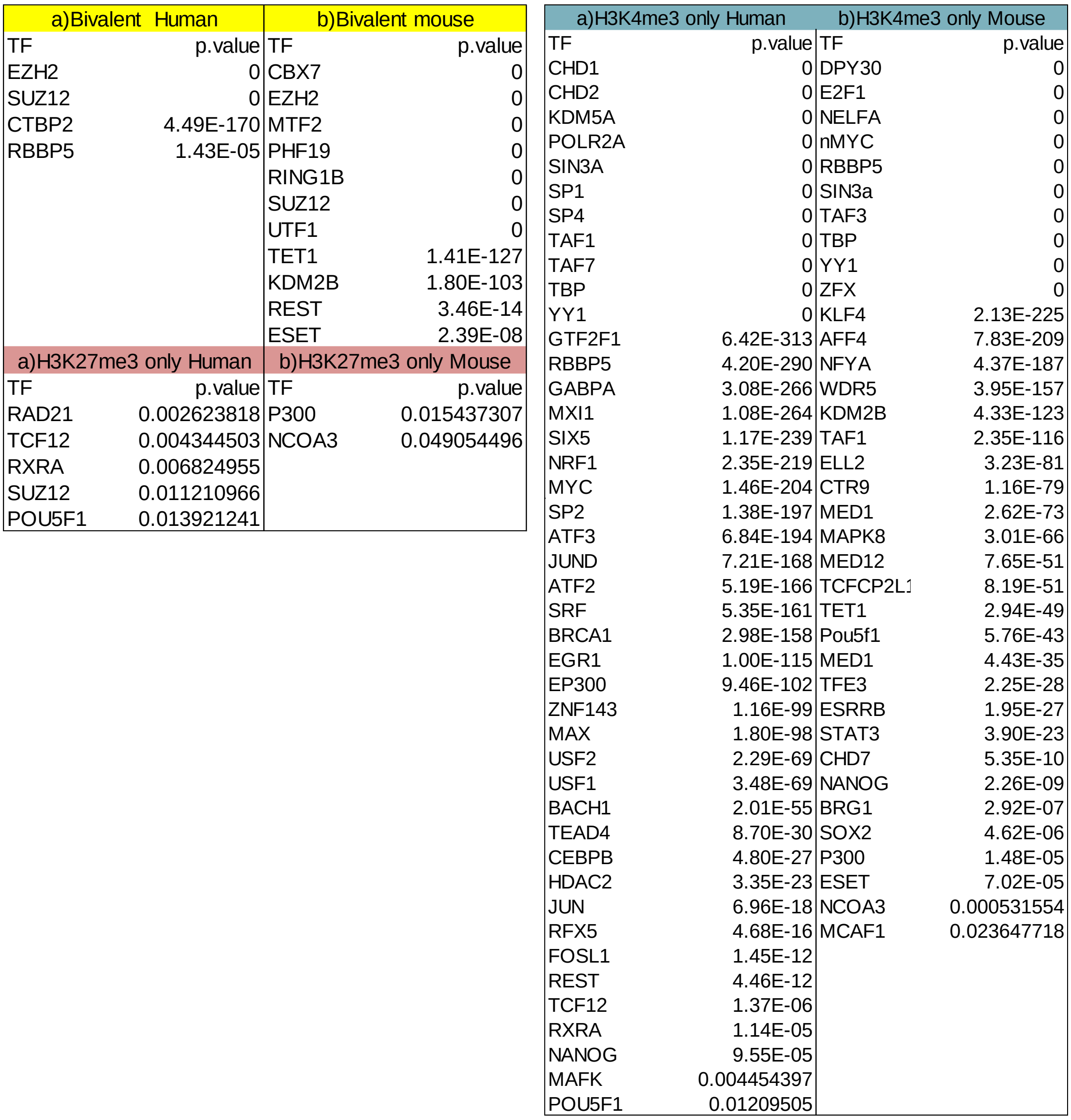


Figure 3.16 A 'TCCCC' sequence motif is specifically enriched in bivalent promoters. A. The average occupancy of factors at HC H3K4me3-only promoters (green) is higher than at HC bivalent promoters (yellow) which is higher than at HC H3K27me3 only promoters (red) and unmarked promoters (blue) in human (left) and mouse (right) ESCs (* equals to <10^{-4}). B. Transcription and epigenetic factors with statistically significant**

overlap with HC bivalent promoters from ChIP sequencing data for 49 in human (up) and 99 factors in mouse (down) ESCs. C. A 'TCCCC' sequence motif is specifically enriched in HC bivalent promoters in both human and mouse ESCs. Similarly a 'CGGAA' motif is enriched HC H3K4me3 promoters in both human and mouse ESCs. Each motif was then mapped to the genome, and motif densities around TSSs of bivalent (black), H3K4me3-only (yellow) and latent (blue) promoters are shown in the left (human) and right (mouse) panels.

As expected, many TFs (33 out of 49 factors in human and 39 out of 99 factors in mouse) were enriched at active (H3K4me3-only) promoters. This included known regulators of pluripotency in ESCs such as Klf4, Esrrb, Oct4, Sox2, and Nanog (Table 3.7). Only two factors enriched in bivalent promoters, Kdm2b and Tet1, were also enriched at active promoters. All other factors showed preference to either bivalent promoters or active but not both. For example, C-Myc can stimulate Pol II elongation (Brien et al., 2012) and was enriched in active promoters in both human and mouse ESCs but not in bivalent promoters.

The observation that some factors are enriched specifically at bivalent promoters suggests that sequence motifs specific to bivalent promoters may determine their binding. We performed de novo motif identification on bivalent promoters by providing active promoter sequences as background in HOMER software (Heinz et al., 2010) and found several AG-rich and GC-rich motifs specific to bivalent promoters (Figure 3.17). These resemble the sequence motifs of Jarid2 (Peng et al., 2009) and Utf1 (Jia et al., 2012a) identified from ChIP-seq data. Interestingly, a 'TCCCC' sequence motif was enriched and found in about 50% of bivalent promoters in both human and mouse (Figure 3.16C). This motif was not enriched in active promoters in either of the species (the number of repressed promoters was not large enough to perform a reliable de novo motif discovery). The 'TCCCC' motif was most similar to the known binding sequence of the Mzf1 transcription factor (Morris et al., 1994). The Mzf1 promoter both in mouse and human ESCs is characterized as HC H3K4me3 only and belonged to the low expressed genes in our analysis. However, in recent Mzf1 ChIP-seq experiment performed in HEK293 cell line (Najafabadi et al., 2015), the "TCCCC" motif was not enriched in Mzf1 peak list (Table 3.8). When de novo motif enrichment was performed on active human and mouse promoters using bivalent promoter sequences as background, they were enriched for a 'CGGAA' motif found in 40% of the active promoter sequences, which was not enriched in bivalent promoters. This motif is the most similar to the known motif for Elk1 transcription factor (Figure 3.16C).



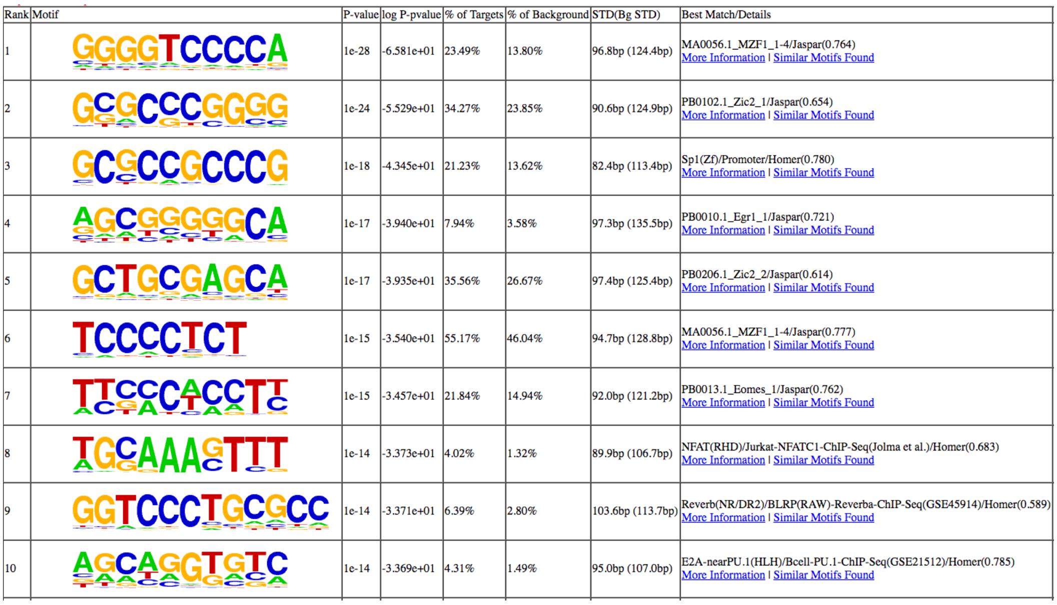
a)Bivalent Human		b)Bivalent mouse	
TF	p.value	TF	p.value
EZH2	0	CBX7	0
SUZ12	0	EZH2	0
CTBP2	4.49E-170	MTF2	0
RBBP5	1.43E-05	PHF19	0
		RING1B	0
		SUZ12	0
		UTF1	0
		TET1	1.41E-127
		KDM2B	1.80E-103
		REST	3.46E-14
		ESET	2.39E-08

a)H3K27me3 only Human		b)H3K27me3 only Mouse	
TF	p.value	TF	p.value
RAD21	0.002623818	P300	0.015437307
TCF12	0.004344503	NCOA3	0.049054496
RXRA	0.006824955		
SUZ12	0.011210966		
POU5F1	0.013921241		

a)H3K4me3 only Human		b)H3K4me3 only Mouse	
TF	p.value	TF	p.value
CHD1	0	DPY30	0
CHD2	0	E2F1	0
KDM5A	0	NELFA	0
POLR2A	0	nMYC	0
SIN3A	0	RBBP5	0
SP1	0	SIN3a	0
SP4	0	TAF3	0
TAF1	0	TBP	0
TAF7	0	YY1	0
TBP	0	ZFX	0
YY1	0	KLF4	2.13E-225
GTF2F1	6.42E-313	AFF4	7.83E-209
RBBP5	4.20E-290	NFYA	4.37E-187
GABPA	3.08E-266	WDR5	3.95E-157
MXI1	1.08E-264	KDM2B	4.33E-123
SIX5	1.17E-239	TAF1	2.35E-116
NRF1	2.35E-219	ELL2	3.23E-81
MYC	1.46E-204	CTR9	1.16E-79
SP2	1.38E-197	MED1	2.62E-73
ATF3	6.84E-194	MAPK8	3.01E-66
JUND	7.21E-168	MED12	7.65E-51
ATF2	5.19E-166	TCFCP2L1	8.19E-51
SRF	5.35E-161	TET1	2.94E-49
BRCA1	2.98E-158	Pou5f1	5.76E-43
EGR1	1.00E-115	MED1	4.43E-35
EP300	9.46E-102	TFE3	2.25E-28
ZNF143	1.16E-99	ESRRB	1.95E-27
MAX	1.80E-98	STAT3	3.90E-23
USF2	2.29E-69	CHD7	5.35E-10
USF1	3.48E-69	NANOG	2.26E-09
BACH1	2.01E-55	BRG1	2.92E-07
TEAD4	8.70E-30	SOX2	4.62E-06
CEBPB	4.80E-27	P300	1.48E-05
HDAC2	3.35E-23	ESET	7.02E-05
JUN	6.96E-18	NCOA3	0.000531554
RFX5	4.68E-16	MCAF1	0.023647718
FOSL1	1.45E-12		
REST	4.46E-12		
TCF12	1.37E-06		
RXRA	1.14E-05		
NANOG	9.55E-05		
MAFK	0.004454397		
POU5F1	0.01209505		

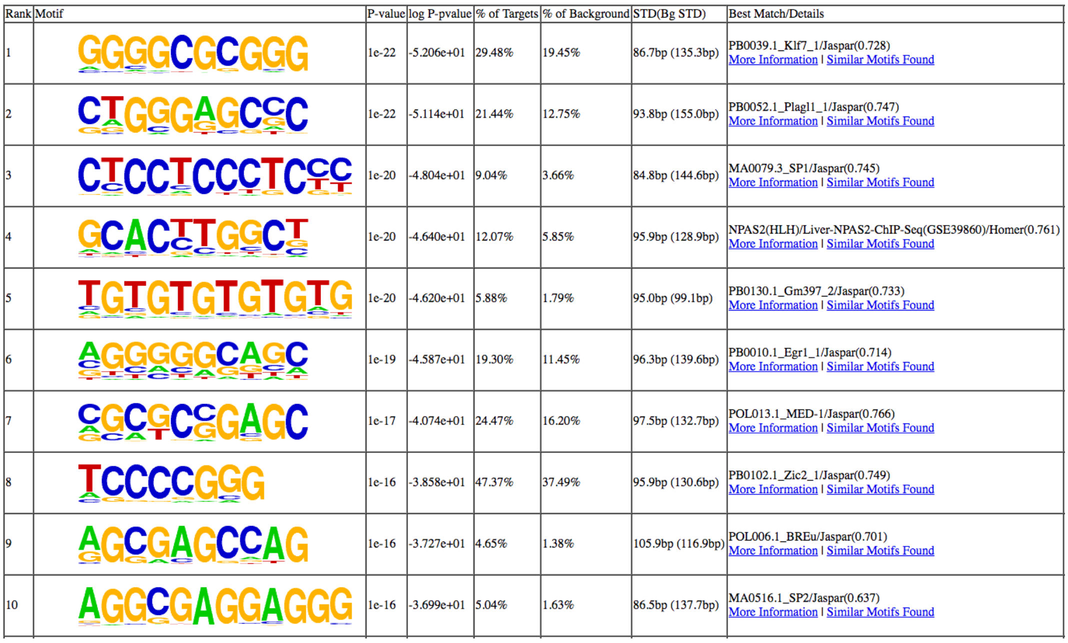
Table 3.7 Factors binding at bivalent, H3K4me3-only and H3K27me3-only promoters in a) human and b) mouse ESCs. The p-values shown to be 0, should be noted as p-value < extremely small value close to zero. For example, p-value < 1e-256.

A Bivalent mouse with K4 background



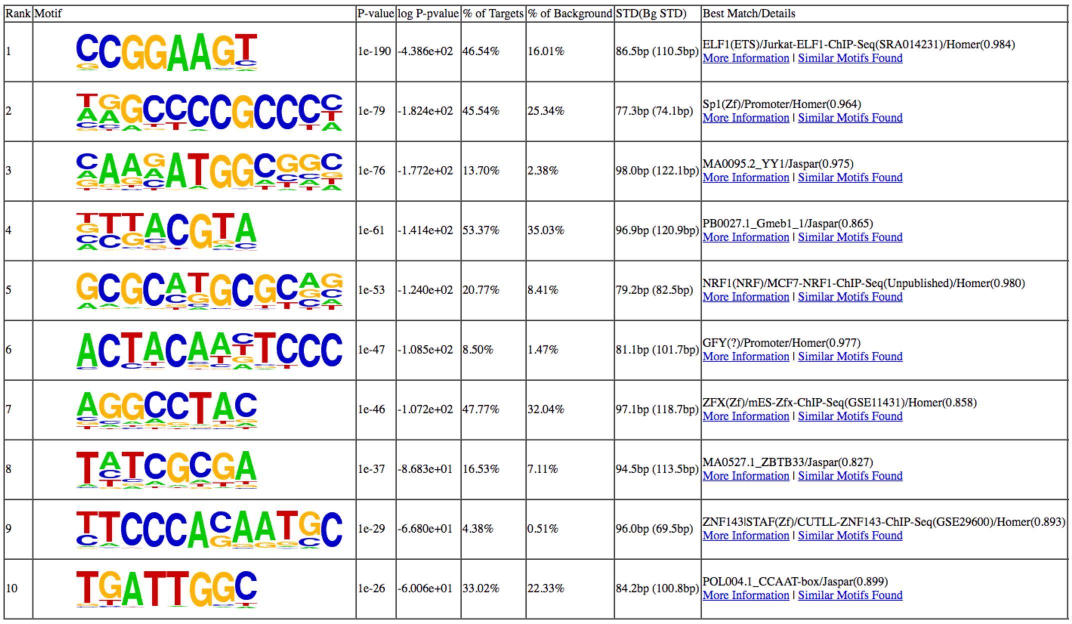
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1	GGGGTCCCCA	1e-28	-6.581e+01	23.49%	13.80%	96.8bp (124.4bp)	MA0056.1_MZF1_1-4/Jaspar(0.764) More Information | Similar Motifs Found
2	GCGCCCGGGG	1e-24	-5.529e+01	34.27%	23.85%	90.6bp (124.9bp)	PB0102.1_Zic2_1/Jaspar(0.654) More Information | Similar Motifs Found
3	GCGCCGCCCG	1e-18	-4.345e+01	21.23%	13.62%	82.4bp (113.4bp)	Sp1(Zf)/Promoter/Homer(0.780) More Information | Similar Motifs Found
4	AGCGGGGGCA	1e-17	-3.940e+01	7.94%	3.58%	97.3bp (135.5bp)	PB0010.1_Egr1_1/Jaspar(0.721) More Information | Similar Motifs Found
5	GCTGCGAGCA	1e-17	-3.935e+01	35.56%	26.67%	97.4bp (125.4bp)	PB0206.1_Zic2_2/Jaspar(0.614) More Information | Similar Motifs Found
6	TCCCCTCT	1e-15	-3.540e+01	55.17%	46.04%	94.7bp (128.8bp)	MA0056.1_MZF1_1-4/Jaspar(0.777) More Information | Similar Motifs Found
7	TTCCCACCTT	1e-15	-3.457e+01	21.84%	14.94%	92.0bp (121.2bp)	PB0013.1_Eomes_1/Jaspar(0.762) More Information | Similar Motifs Found
8	TGCAAAGTTT	1e-14	-3.373e+01	4.02%	1.32%	89.9bp (106.7bp)	NFAT(RHD)/Jurkat-NFATC1-ChIP-Seq(Jolma et al.)/Homer(0.683) More Information | Similar Motifs Found
9	GGTCCCTGCGCC	1e-14	-3.371e+01	6.39%	2.80%	103.6bp (113.7bp)	Reverb(NR/DR2)/BLRP(RAW)-Reverba-ChIP-Seq(GSE45914)/Homer(0.589) More Information | Similar Motifs Found
10	AGCAGGTGTC	1e-14	-3.369e+01	4.31%	1.49%	95.0bp (107.0bp)	E2A-nearPU.1(HLH)/Bcell-PU.1-ChIP-Seq(GSE21512)/Homer(0.785) More Information | Similar Motifs Found

B Bivalent human with K4 background



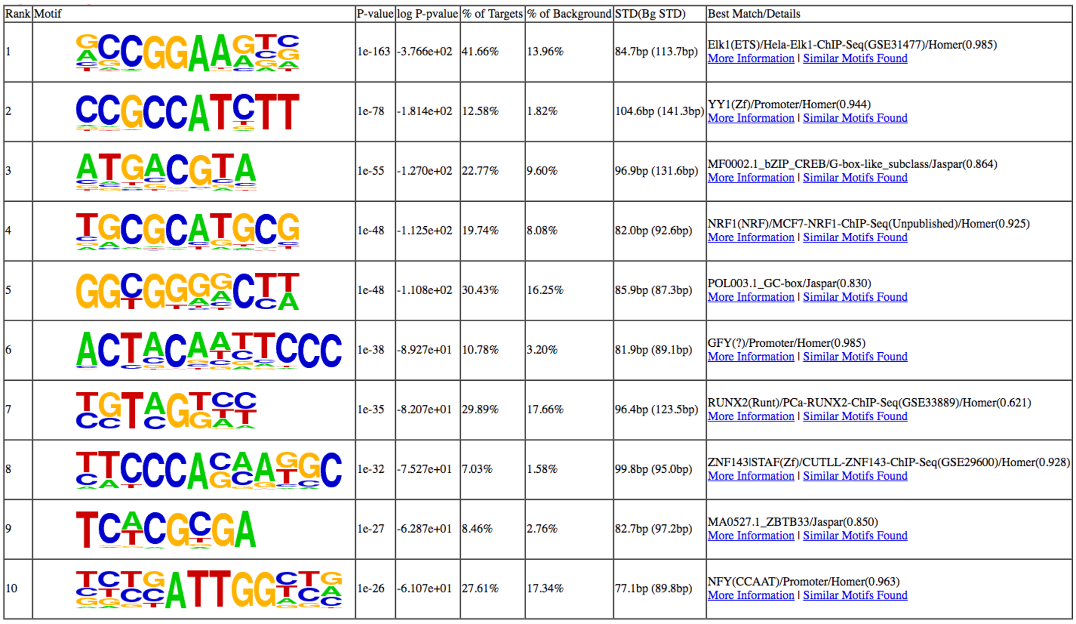
Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1	GGGGCGCGGG	1e-22	-5.206e+01	29.48%	19.45%	86.7bp (135.3bp)	PB0039.1_Klf7_1/Jaspar(0.728) More Information | Similar Motifs Found
2	CTGGGAGCCC	1e-22	-5.114e+01	21.44%	12.75%	93.8bp (155.0bp)	PB0052.1_Plagl1_1/Jaspar(0.747) More Information | Similar Motifs Found
3	CTCCTCCCTCCC	1e-20	-4.804e+01	9.04%	3.66%	84.8bp (144.6bp)	MA0079.3_SP1/Jaspar(0.745) More Information | Similar Motifs Found
4	GCACTTGGCT	1e-20	-4.640e+01	12.07%	5.85%	95.9bp (128.9bp)	NPAS2(HLH)/Liver-NPAS2-ChIP-Seq(GSE39860)/Homer(0.761) More Information | Similar Motifs Found
5	TGTGTGTGTGTG	1e-20	-4.620e+01	5.88%	1.79%	95.0bp (99.1bp)	PB0130.1_Gm397_2/Jaspar(0.733) More Information | Similar Motifs Found
6	AGGGGGCAGC	1e-19	-4.587e+01	19.30%	11.45%	96.3bp (139.6bp)	PB0010.1_Egr1_1/Jaspar(0.714) More Information | Similar Motifs Found
7	CGCGCCGAGC	1e-17	-4.074e+01	24.47%	16.20%	97.5bp (132.7bp)	POL013.1_MED-1/Jaspar(0.766) More Information | Similar Motifs Found
8	TCCCCGGG	1e-16	-3.858e+01	47.37%	37.49%	95.9bp (130.6bp)	PB0102.1_Zic2_1/Jaspar(0.749) More Information | Similar Motifs Found
9	AGCGAGCCAG	1e-16	-3.727e+01	4.65%	1.38%	105.9bp (116.9bp)	POL006.1_BREu/Jaspar(0.701) More Information | Similar Motifs Found
10	AGGCGAGGAGGG	1e-16	-3.699e+01	5.04%	1.63%	86.5bp (137.7bp)	MA0516.1_SP2/Jaspar(0.637) More Information | Similar Motifs Found

C H3K4me3 only mouse with bivalent background



Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-190	-4.386e+02	46.54%	16.01%	86.5bp (110.5bp)	ELF1(ETS)/Jurkat-ELF1-ChIP-Seq(SRA014231)/Homer(0.984) More Information | Similar Motifs Found
2		1e-79	-1.824e+02	45.54%	25.34%	77.3bp (74.1bp)	Sp1(Zf)/Promoter/Homer(0.964) More Information | Similar Motifs Found
3		1e-76	-1.772e+02	13.70%	2.38%	98.0bp (122.1bp)	MA0095.2_YY1/Jaspar(0.975) More Information | Similar Motifs Found
4		1e-61	-1.414e+02	53.37%	35.03%	96.9bp (120.9bp)	PB0027.1_Gmeb1_1/Jaspar(0.865) More Information | Similar Motifs Found
5		1e-53	-1.240e+02	20.77%	8.41%	79.2bp (82.5bp)	NRF1(NRF)/MCF7-NRF1-ChIP-Seq(Unpublished)/Homer(0.980) More Information | Similar Motifs Found
6		1e-47	-1.085e+02	8.50%	1.47%	81.1bp (101.7bp)	GFY(?)/Promoter/Homer(0.977) More Information | Similar Motifs Found
7		1e-46	-1.072e+02	47.77%	32.04%	97.1bp (118.7bp)	ZFX(Zf)/mES-Zfx-ChIP-Seq(GSE11431)/Homer(0.858) More Information | Similar Motifs Found
8		1e-37	-8.683e+01	16.53%	7.11%	94.5bp (113.5bp)	MA0527.1_ZBTB33/Jaspar(0.827) More Information | Similar Motifs Found
9		1e-29	-6.680e+01	4.38%	0.51%	96.0bp (69.5bp)	ZNF143|STAF(Zf)/CUTLL-ZNF143-ChIP-Seq(GSE29600)/Homer(0.893) More Information | Similar Motifs Found
10		1e-26	-6.006e+01	33.02%	22.33%	84.2bp (100.8bp)	POL004.1_CCAAT-box/Jaspar(0.899) More Information | Similar Motifs Found

D H3K4me3 only human with bivalent background



Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-163	-3.766e+02	41.66%	13.96%	84.7bp (113.7bp)	Elk1(ETS)/Hela-Elk1-ChIP-Seq(GSE31477)/Homer(0.985) More Information | Similar Motifs Found
2		1e-78	-1.814e+02	12.58%	1.82%	104.6bp (141.3bp)	YY1(Zf)/Promoter/Homer(0.944) More Information | Similar Motifs Found
3		1e-55	-1.270e+02	22.77%	9.60%	96.9bp (131.6bp)	MF0002.1_bZIP_CREB/G-box-like_subclass/Jaspar(0.864) More Information | Similar Motifs Found
4		1e-48	-1.125e+02	19.74%	8.08%	82.0bp (92.6bp)	NRF1(NRF)/MCF7-NRF1-ChIP-Seq(Unpublished)/Homer(0.925) More Information | Similar Motifs Found
5		1e-48	-1.108e+02	30.43%	16.25%	85.9bp (87.3bp)	POL003.1_GC-box/Jaspar(0.830) More Information | Similar Motifs Found
6		1e-38	-8.927e+01	10.78%	3.20%	81.9bp (89.1bp)	GFY(?)/Promoter/Homer(0.985) More Information | Similar Motifs Found
7		1e-35	-8.207e+01	29.89%	17.66%	96.4bp (123.5bp)	RUNX2(Runt)/PCa-RUNX2-ChIP-Seq(GSE33889)/Homer(0.621) More Information | Similar Motifs Found
8		1e-32	-7.527e+01	7.03%	1.58%	99.8bp (95.0bp)	ZNF143|STAF(Zf)/CUTLL-ZNF143-ChIP-Seq(GSE29600)/Homer(0.928) More Information | Similar Motifs Found
9		1e-27	-6.287e+01	8.46%	2.76%	82.7bp (97.2bp)	MA0527.1_ZBTB33/Jaspar(0.850) More Information | Similar Motifs Found
10		1e-26	-6.107e+01	27.61%	17.34%	77.1bp (89.8bp)	NFY(CCAAT)/Promoter/Homer(0.963) More Information | Similar Motifs Found

Figure 3.17 De novo motif enrichment for a) bivalent promoters in mouse ESCs, b) bivalent promoters in human ESCs, c) H3K4me3 only promoters in mouse ESCs and d) H3K4me3 only promoters in human ESCs.

C2H2 TF	P-value	% of peaks with motifs	Rank
CTCF	-	-	-
KLF10	-	-	-
KLF14	-	-	-
MZF1	-	-	-
YY1	1.00E-1603	9.87%	13
ZBTB12	-	-	-
ZBTB18	-	-	-
ZFP3	1.00E-46	9.16%	4
ZIC2	-	-	-
ZNF136	-	-	-
ZNF16	-	-	-
ZNF189	1.00E-39	0.24%	16
ZNF200	-	-	-
ZNF250	-	-	-
ZNF264	-	-	-
ZNF273	-	-	-
ZNF317	-	-	-
ZNF322	1.00E-45	4.19%	16
ZNF33A	-	-	-
ZNF35	1.00E-25	17.63%	23
ZNF382	-	-	-
ZNF415	-	-	-
ZNF416	-	-	-
ZNF41	-	-	-
ZNF454	-	-	-
ZNF45	1.00E-140	2.46%	2
ZNF467	-	-	-
ZNF519	-	-	-
ZNF528	1.00E-196	0.88%	7
ZNF574	-	-	-
ZNF621	-	-	-
ZNF653	-	-	-
ZNF669	1.00E-209	8.73%	4
ZNF675-2	-	-	-
ZNF675	-	-	-
ZNF684	1.00E-04	3.98%	10
ZNF692	1.00E-12	0.55%	7
ZNF71	1.00E-25	43.73%	7
ZSCAN22	-	-	-
ZSCAN31	-	-	-

Table 3.8 TCCCC motif enrichment in C2H2 ChIP-sequencing peaks from Najafabadi et al. in Human ESCs. Rank N means that N-1 non TCCCC motif were more enriched.

In summary, bivalent promoters are bound by fewer transcription factors than active (H3K4me3-only) promoters, but more than H3K27me3 only and latent promoters. Active promoters were preferentially occupied by pluripotency factors. On the other hand, bivalent promoters were enriched for Polycomb factors as well as other chromatin modifiers. The factors enriched at bivalent promoters show very little overlap with the ones enriched at active promoters. These finding are consistent with the observed spatial segregation of transcriptional networks in ESCs where Nanog and Polycomb proteins were shown to occupy distinct nuclear spaces (Denholtz et al., 2013). Finally, we identified a ‘TCCCC’ sequence motif specifically at bivalent promoters and a ‘CGGAA’ sequence motif at active promoters.

3.5 Discussion

Bivalent chromatin domains bearing both H3K4me3 and H3K27me3 modifications have been shown to be a key feature of developmentally regulated genes in ESCs (B. E. Bernstein et al. 2006; Mikkelsen et al. 2007; Jia et al. 2012; Xiao Dong Zhao et al. 2007; Pan et al. 2007b). These domains are thought to be ‘poised’, with an ability to quickly become active (losing H3K27me3) or inactive (losing H3K4me3) during differentiation (Mikkelsen et al., 2007; Mohn et al., 2008). While many studies have produced ChIP-seq data for both H3K4me3 and H3K27me3 in ESCs in both humans (Pan et al., 2007; Zhao et al., 2007) and mice (Jia et al., 2012a; Mikkelsen et al., 2007), differences in species, ES growth conditions, ChIP protocols (shearing, cross link, antibodies used) and high throughput sequencing setup (with or without replicate, with or without input) have rendered a comparison across studies challenging. By systematic integration of available data, we identified robust lists of 4,979 and 3,659 high confidence bivalent promoters in human and mouse respectively. Since our work is using the data of previous studies using H3K4me3 and H3K27me3 ChIP-seq to define bivalency in ESCs, we are biased toward a confirmation of the original studies, as their data is integrated in our dataset. However, our integrative approach (see methods) renders this analysis resistant to any outlier experiments. By cumulatively integrating the samples, it became evident that the detection of bivalency on promoters is dependent on the reliable detection of the H3K27me3 modification. Over 85% of H3K27me3 promoters were bivalent, i.e. they also had the H3K4me3 mark. This confirms that bivalency in ESCs is rather the rule than the exception. The three main chromatin states on promoters in ESCs are thus active, bivalent and latent (no mark). Correspondingly, active promoters were expressed, bivalent were lowly expressed and latent were mostly not expressed.

Bivalent promoters are thought to be poised for rapid activation or inactivation during differentiation (Jia et al., 2012a; Voigt et al., 2013). To tease out whether the low expression at bivalent promoters is a result of some cells expressing the genes while others not, or the genes are expressed at low levels in most cells, we used single cell gene expression data. Bivalent genes were expressed in a similar number of single cells as lowly expressed active genes. It is therefore unlikely that bivalency is a result of mixture of cell populations in ESCs. Similarly, H3K27me3 read density was higher at HC bivalent promoters than at H3K27me3-only promoters, again arguing in disfavour of a mix-population model. The low transcription level can be interpreted as a "leaking" transcription rate, in the absence of a strong repressive chromatin environment. During development, these poised domains have been shown to resolve as either active (by losing the H3K27me3 mark) or inactive (by losing the H3K4me3 mark), and in some cases gaining DNA methylation (Deaton and Bird, 2011), depending on the cellular lineage. In agreement with this model, we have found that >90% of differentially expressed (either up-regulated or down-regulated) gene sets when any one of a set of 91 transcription factors was either overexpressed or knocked down in mouse ESCs were enriched for bivalent genes. This finding suggests that bivalent genes are hypersensitive to most perturbations of the regulatory network in ESCs.

We computed binding profiles of PRC components (PRC1 and PRC2) and various forms of RNA polymerase II at bivalent promoters in murine ESCs. All HC bivalent promoters were marked by Suz12, Jarid2, Ring1b and Cbx7. To note, the PRC2-only group defined by (Ku et al., 2008a) overlapped with PRC1-low clusters, the PRC1 signal detected due to higher sequencing depth in latter case (Figure 3.10). Thus all bivalent promoters were occupied by both PRC1 and PRC2. Accordingly, H2Aub showed enrichment at HC bivalent promoters (Figure 3.18). Recent studies have suggested that true bivalency is better associated with H2Aub than H3K27me3 (Brookes et al., 2012b). We note that H2Aub predominantly but not exclusively marks bivalent promoters (Table 3.9) as it also marks a fraction of H3K4me3-only expressed gene promoters (Figure 3.19). Based on PRC1 and RNAPII occupancy, bivalent promoters grouped into four clusters. Clusters 1 and 2 had low PRC1 occupancy and high RNAPII (8WG16) levels while Cluster 3 and 4 were PRC-rich with low RNAPII

(8WG16) levels. Cluster 2 was enriched for metabolic genes and marked with RNAPII (S7P) and cluster 2 genes were expressed at higher levels than the other three clusters. The bivalent promoters therefore consist of sub-groups of genes which at functional, epigenetic and transcriptional level are quite different from each other.

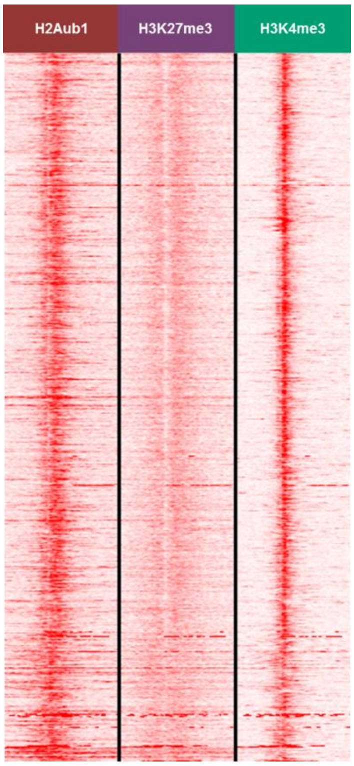


Figure 3.18 Signal of H2Aub1, H3K27me3 and H4K4me3 histone modifications at the HC bivalent promoters in mouse ESCs.

Marked in HC	Number of HC promoters	Overlap with H2Aub1/H3K4me3 promoters (4028)	Percentage
Bivalent	3659	2916	72.39325
H3K27me3 onl	152	0	0
H3K4me3 only	9336	451	11.19662
Unclassified	10354	661	16.41013
Latent	15391	0	0

Table 3.9 Overlap of HC promoters in all categories with H2Aub1/H3K4me3 promoters. (In total 4518 H2Aub1 peaks where found in promoters and the majority of them were accompanied by H3K4me3)

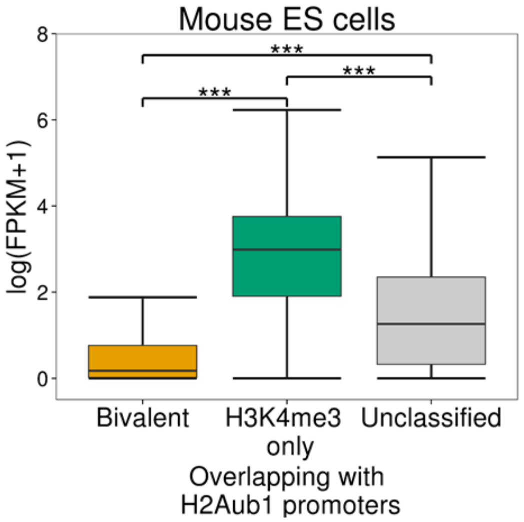


Figure 3.19 Expression levels of HC bivalent, H3K4me3 only and Unclassified promoters overlapping with the H2Aub1 bivalent promoters.

More than half of high-confidence bivalent promoters were conserved between human and mouse, suggesting the existence of a set of genes bivalently marked across most mammalian ESCs (Appendix tables 1 and 2). These genes were very highly enriched for transcription regulators and developmental factors, compared to the species specific bivalent promoters. On the other hand, divergence of epigenetic status across species did not imply divergence of gene expression i.e. promoters with bivalent chromatin status in human and active chromatin status in mouse did not have gene expression profiles similar to bivalent genes in human and active genes in mouse. Further analysis is necessary to understand whether the differences between mouse and human ESCs are indeed species-specific or developmental stage specific as human ESCs do not share the same developmental state as mouse ESCs (Takashima et al., 2014; Tesar et al., 2007).

Since a high density of un-methylated CpG is sufficient for vertebrate Polycomb recruitment (Farcas et al., 2012; Mendenhall et al., 2010; Riising et al., 2014), it is assumed that the presence of CpG islands determines H3K27me3 modification. Over 90% of bivalent promoters contained a CpG island while few to none of the H3K27me3-only promoters had a CpG island. Wachter et al. (2014) recently suggested that bivalency is the default chromatin structure for CpG-rich, G+C-rich DNA

(Wachter et al., 2014). The presence of H3K27me3 on CpG-poor promoters without H3K4me3 modification in ESCs (Figures 3.10 and 3.21) suggests mechanisms other than CpG islands for Polycomb recruitment.

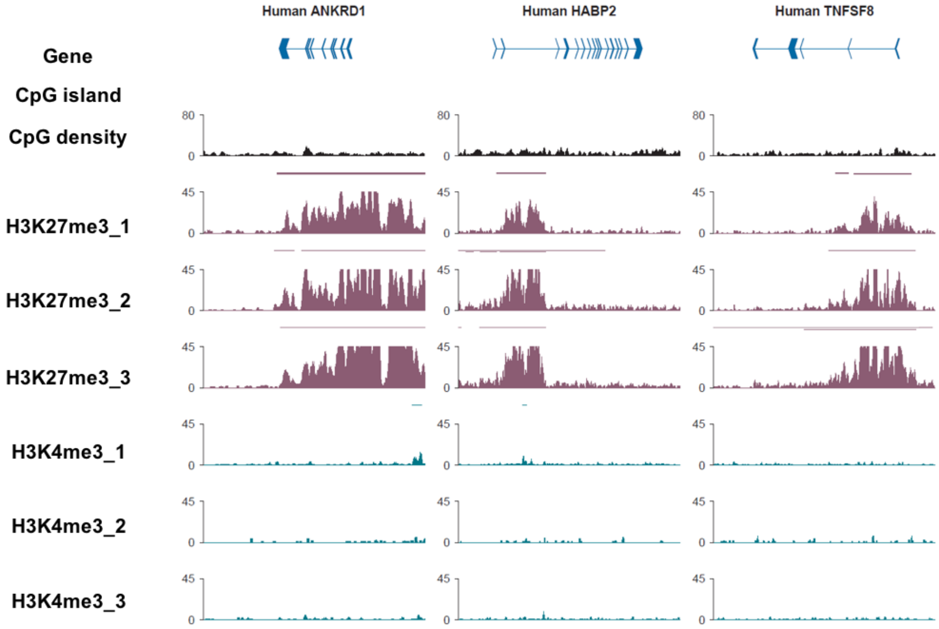


Figure 3.20 Examples of H3K27me3-only promoters in human ESCs

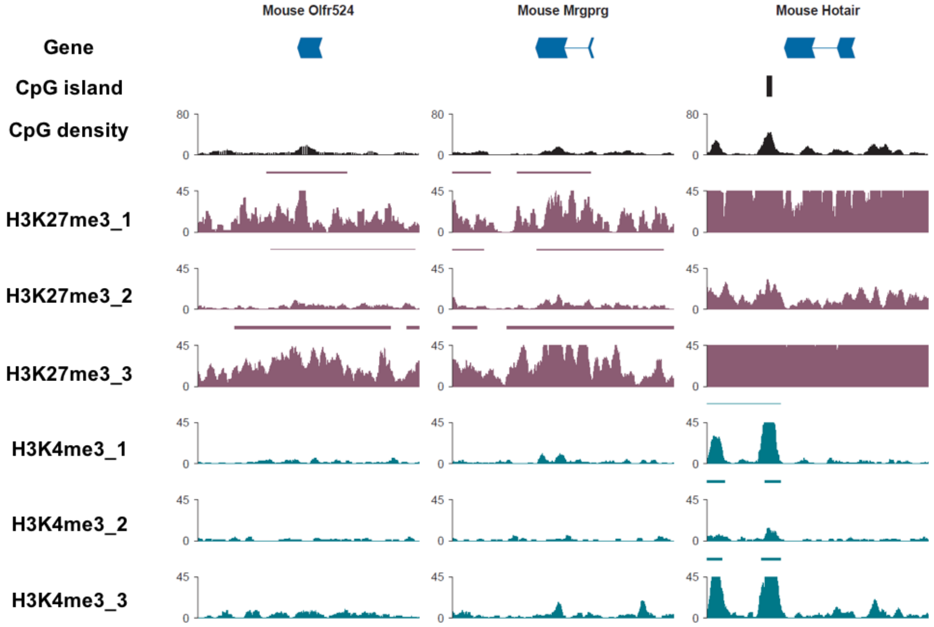


Figure 3.21 Examples of H3K27me3-only promoters in mouse ESCs

On bivalent promoters, the CpG density and H3K27me3 modification are highly correlated. By performing a cross-species comparison, a small fraction (~5%) of human CpG-rich HC bivalent promoters has the corresponding CpG eroded in the mouse genome, while no CpG-rich bivalent promoters in mouse are eroded in human. This erosion of CpG density was correlated with the loss of H3K27me3 and H3K4me3 (Lynch et al., 2012). However, in about 20% of the cases, the CpG density loss in mouse compared to human did not correspond to a loss of H3K27me3. This reiterates the finding that CpG density might be sufficient but not necessary for H3K27me3 modification.

It is intriguing how bivalent domains are established in ESCs. Voigt et. al (2013) proposed a model where H3K4me3 marked promoters occupied by a low number of transcription factors allowed the establishment of H3K27me3 modification. Indeed, HC bivalent promoters were bound by fewer factors than active promoters in human and mouse ESCs. HC bivalent promoters were specifically enriched in ChIP-seq peaks for many members of the PRC1, PRC2 and MLL complexes as expected. We also found enrichment for several additional proteins known to be involved in recruiting these complexes, including CTBP2, Mtf2 and Phf19. Other factors frequently binding to HC bivalent promoters included Kdm2b, Utf1, Tet1, Rest and Stedb1. These factors are involved in establishing diverse epigenetic modifications suggesting the complex epigenetic regulation of these regions.

As active (H3K4me3-only) and bivalent promoters are both CpG rich, it is key to unravel the distinguishing factors between these two groups. De novo motif discovery at HC bivalent promoters identified a ‘TCCCC’ motif in both human and mouse ESCs which was not enriched at active promoters. This motif was present in about half of the HC bivalent promoters and is similar to the sequence motif of MZF1 (Morris et al., 1994), although this was not confirmed in recent MZF1 ChIP-seq experiment in HEK293 cell line (Najafabadi et al., 2015) (Table 3.8). Similarly, a ‘CGGAA’ motif was enriched specifically at active promoters and is similar to the sequence motif of ELK1. Further experiments are mandate to establish whether these sequence motifs indeed play a role at bivalent and active promoters, and if yes, through which factors?

Characterising factors associated with these motifs will be the first step to study their functional relevance.

In summary, this meta-analysis revealed several novel aspects of bivalency in mammalian ESCs and will serve as a resource for future studies to further understand transcriptional regulation during embryonic development. Further work will be aimed at understanding how the HC bivalent promoters identified here are resolved in different cellular lineages during differentiation.

Chapter 4 Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data

4.1 Chapter Introduction

This chapter was published in 2016 in Computational Biology and Chemistry (for Asia Pacific Bioinformatics Conference 2016, San Francisco) under DOI: http://dx.doi.org/10.1016/j.compbiolchem.2016.02.004. Here we assess in detail the gene expression heterogeneity observed in ESC populations, with a specific interest in the expression patterns of bivalent genes that we detected in Chapter 3.

Abstract

Gene expression heterogeneity contributes to development as well as disease progression. Due to technological limitations, most studies to date have focused on differences in mean expression across experimental conditions, rather than differences in gene expression variance. The advent of single cell RNA sequencing has now made it feasible to study gene expression heterogeneity and to characterise genes based on their coefficient of variation. We collected single cell gene expression profiles for 32 human and 39 mouse embryonic stem cells and studied correlation between diverse characteristics such as network connectivity and coefficient of variation (CV) across single cells. We further systematically characterised properties unique to High CV genes. Highly expressed genes tended to have a low CV and were enriched for cell cycle genes. In contrast, High CV genes were co-expressed with other High CV genes, were enriched for bivalent (H3K4me3 and H3K27me3) marked promoters and showed enrichment for response to DNA damage and DNA repair. Taken together, this analysis demonstrates the divergent characteristics of genes based on their CV. High CV genes tend to form co-expression clusters and they explain bivalency at least in part.

4.2 Introduction

Transcription control is fundamental to mammalian system in defining gene expression programs that establish and maintain specific cell states during development. Any aberration to this process can result into disease phenotype. Microarray technology enables a genome-wide snapshot of the transcription landscape during development and disease by parallel quantification of large numbers of messenger RNA transcripts from different cell types and tissues (Schulze and Downward, 2001). This technology is widely used for differential gene expression analysis where studies are performed on a pool of hundreds of thousands of cells with an assumption that the variation across multiple samples from a cell population is largely due to experimental noise. Difference between mean values of gene expression is therefore the focus of such analyses and rarely the variability across the samples (Mar et al., 2011).

The breakthroughs in sequencing technology have now made it feasible to generate gene expression data for hundreds of individual cells from a cell population (Pan, 2014) providing new insights into early development (Tang et al., 2010) and differentiation (Shalek et al., 2013). Single cell RNA-seq sequencing is used for characterisation of hidden subpopulations of rare cell types, as closely related cells with the same phenotype can be discriminated to distinguish functionally each subgroup (Buettner et al., 2015). Importantly, the gene expression quantification by single-cell RNA-seq is consistent with the existing gold standards (Wu et al., 2013). The single cell gene expression data is variable between individual cells in contrast to the high concordance across replicates of populations of cells (Shalek et al., 2013). Though part of variation across individual cells is attributed to various confounding factors such as random technical noise mainly due to transcription bursts (Brennecke et al., 2013), protein fluctuations (Karwacki-Neisius et al., 2013) or mRNA fluctuations in response to cell cycle (Singh et al., 2013), there is no doubt about the biological relevance of variation in development (Xue et al., 2013), evolutionary adaptation, and disease (Feinberg and Irizarry, 2010).

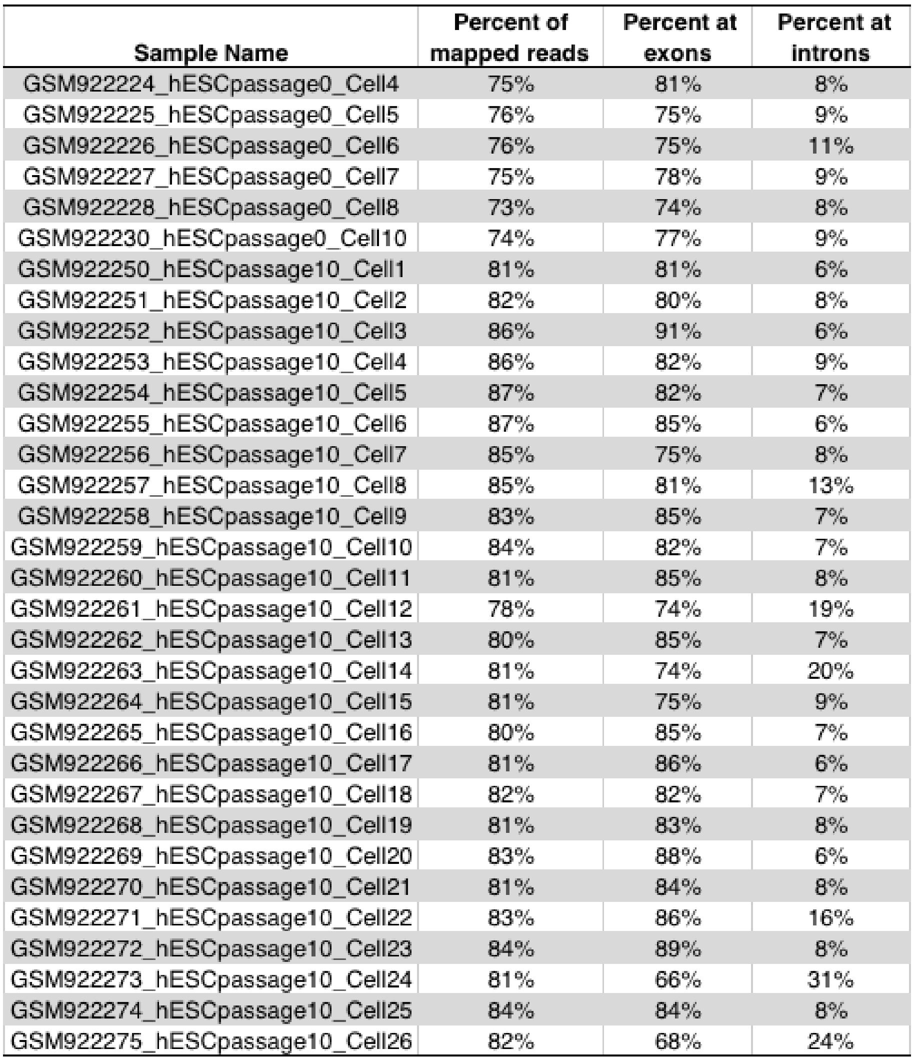
Importantly, variation at a single cell level in genetically identical organisms in homogeneous environments indicates its role in generating diversity (Raj et al., 2010).

Achieving such diversity is particularly important in the context of stem cells. The pluripotent state is a delicate equilibrium between the ability of self-renewal and differentiation, hence an imbalance (the variation of key pluripotency factors) could lead tipping the scale in favour of differentiation (Karwacki-Neisius et al., 2013). Accordingly, a high concordance was noted between global gene expression variability and heterogeneity of human pluripotency states (Mason et al., 2014). The differences between gene sets at the two ends of the spectrum of variation demonstrated that low variance genes were highly connected in the regulatory networks providing a causal hypothesis for their low variance (Mar et al., 2011). Highly variable genes, on the other hand, are thought to represent elements which fluctuate as the stem cell population moves between self-renewal and differentiation-potential (Mason et al., 2014). We collected single cell RNA sequencing data in human (Streets et al., 2014) and mouse (Yan et al., 2013) embryonic stem cells and identified ‘High CV’ (CV: Coefficient of Variation) gene sets. The multi-facetted bioinformatic analysis was based on CV enabled systematic characterisation of differences between the stable and variable gene sets.

4.3 Methods

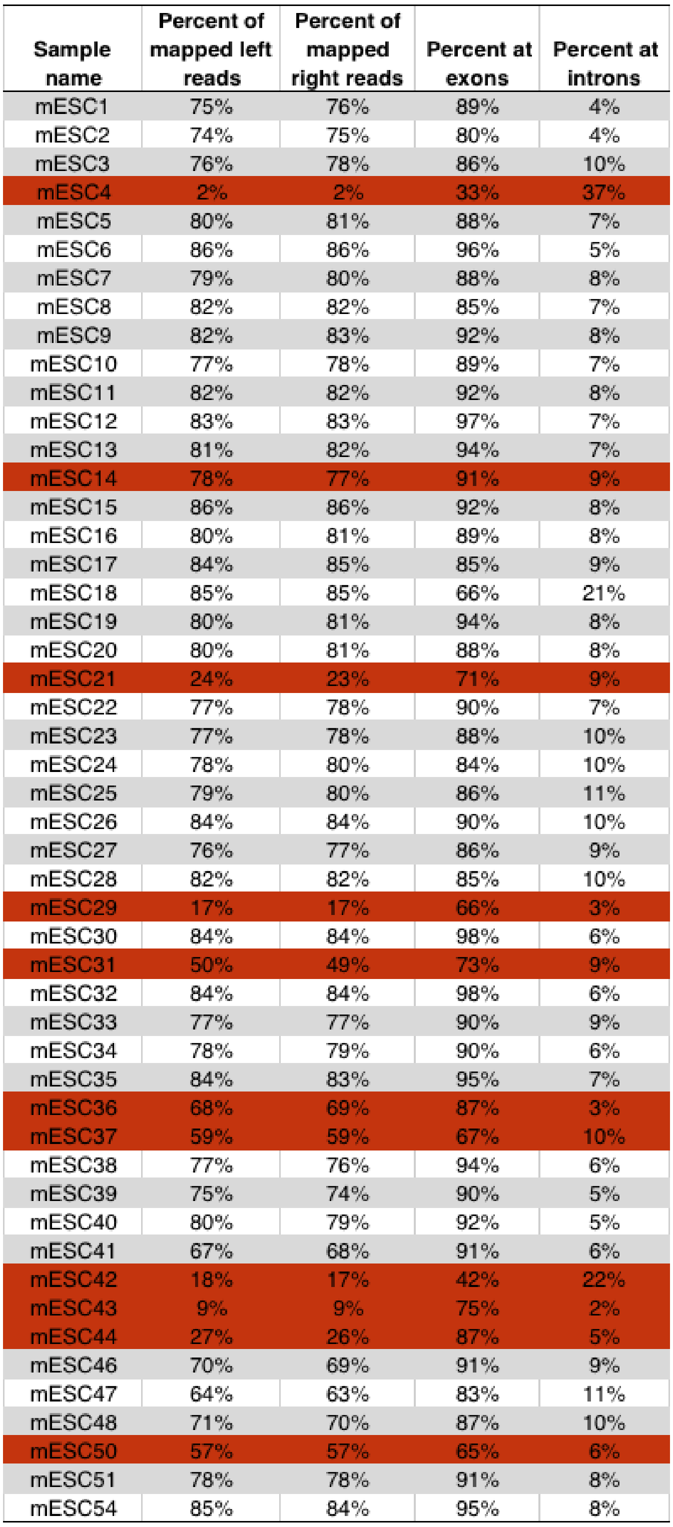
4.3.1 Data collection and processing

Single cell RNA-seq data was obtained from Gene Expression Omnibus (GEO) database (Barrett et al., 2013) in fastq format. We downloaded 63 mouse single ES cell RNA-seq data (paired end) (GSE47835, SRP025171) (Streets et al., 2014) and 32 human single ES cell RNA-seq data (single end) (GSE36552, SRP011546) (Yan et al., 2013). After quality control using FastQC 0.11.2, alignment was done with TopHat 2.0.9 (Trapnell et al., 2009) using mm10 and hg38 as reference genomes and the GENCODE(Harrow et al., 2012b) annotations (M4 and 22) for mouse and human respectively. Expression values for each single cell were calculated following the Cufflinks 2.2.1(Trapnell et al., 2010) pipeline. The aligned reads were converted to expression values using the cuffquant command. Gene expression values for all single cell libraries were generated using the cuffnorm command with the default library normalization method (geometric). 39 mouse ESCs were selected for final analysis after discarding 24 cells due to low read quality, poor alignment scores or failure to map to the reference genome. The quality of the samples was further assessed using the 'read_distribution.py' module from the RSeQC package, which is used as an RNA-seq QC package (Wang et al., 2012). The percentages of mapped reads in total, to exons and to introns are shown in Tables 4.1 and 4.2 for human and mouse ESCs respectively.



Sample Name	Percent of mapped reads	Percent at exons	Percent at introns
GSM922224_hESCpassage0_Cell4	75%	81%	8%
GSM922225_hESCpassage0_Cell5	76%	75%	9%
GSM922226_hESCpassage0_Cell6	76%	75%	11%
GSM922227_hESCpassage0_Cell7	75%	78%	9%
GSM922228_hESCpassage0_Cell8	73%	74%	8%
GSM922230_hESCpassage0_Cell10	74%	77%	9%
GSM922250_hESCpassage10_Cell1	81%	81%	6%
GSM922251_hESCpassage10_Cell2	82%	80%	8%
GSM922252_hESCpassage10_Cell3	86%	91%	6%
GSM922253_hESCpassage10_Cell4	86%	82%	9%
GSM922254_hESCpassage10_Cell5	87%	82%	7%
GSM922255_hESCpassage10_Cell6	87%	85%	6%
GSM922256_hESCpassage10_Cell7	85%	75%	8%
GSM922257_hESCpassage10_Cell8	85%	81%	13%
GSM922258_hESCpassage10_Cell9	83%	85%	7%
GSM922259_hESCpassage10_Cell10	84%	82%	7%
GSM922260_hESCpassage10_Cell11	81%	85%	8%
GSM922261_hESCpassage10_Cell12	78%	74%	19%
GSM922262_hESCpassage10_Cell13	80%	85%	7%
GSM922263_hESCpassage10_Cell14	81%	74%	20%
GSM922264_hESCpassage10_Cell15	81%	75%	9%
GSM922265_hESCpassage10_Cell16	80%	85%	7%
GSM922266_hESCpassage10_Cell17	81%	86%	6%
GSM922267_hESCpassage10_Cell18	82%	82%	7%
GSM922268_hESCpassage10_Cell19	81%	83%	8%
GSM922269_hESCpassage10_Cell20	83%	88%	6%
GSM922270_hESCpassage10_Cell21	81%	84%	8%
GSM922271_hESCpassage10_Cell22	83%	86%	16%
GSM922272_hESCpassage10_Cell23	84%	89%	8%
GSM922273_hESCpassage10_Cell24	81%	66%	31%
GSM922274_hESCpassage10_Cell25	84%	84%	8%
GSM922275_hESCpassage10_Cell26	82%	68%	24%

Table 4.1 Percentages of total mapped reads, reads mapping to exons and reads mapping to introns for the human ESCs. No single cells were discarded, since they presented a satisfactory percentage of mapping reads back to the reference genome.



Sample name	Percent of mapped left reads	Percent of mapped right reads	Percent at exons	Percent at introns
mESC1	75%	76%	89%	4%
mESC2	74%	75%	80%	4%
mESC3	76%	78%	86%	10%
mESC4	2%	2%	33%	37%
mESC5	80%	81%	88%	7%
mESC6	86%	86%	96%	5%
mESC7	79%	80%	88%	8%
mESC8	82%	82%	85%	7%
mESC9	82%	83%	92%	8%
mESC10	77%	78%	89%	7%
mESC11	82%	82%	92%	8%
mESC12	83%	83%	97%	7%
mESC13	81%	82%	94%	7%
mESC14	78%	77%	91%	9%
mESC15	86%	86%	92%	8%
mESC16	80%	81%	89%	8%
mESC17	84%	85%	85%	9%
mESC18	85%	85%	66%	21%
mESC19	80%	81%	94%	8%
mESC20	80%	81%	88%	8%
mESC21	24%	23%	71%	9%
mESC22	77%	78%	90%	7%
mESC23	77%	78%	88%	10%
mESC24	78%	80%	84%	10%
mESC25	79%	80%	86%	11%
mESC26	84%	84%	90%	10%
mESC27	76%	77%	86%	9%
mESC28	82%	82%	85%	10%
mESC29	17%	17%	66%	3%
mESC30	84%	84%	98%	6%
mESC31	50%	49%	73%	9%
mESC32	84%	84%	98%	6%
mESC33	77%	77%	90%	9%
mESC34	78%	79%	90%	6%
mESC35	84%	83%	95%	7%
mESC36	68%	69%	87%	3%
mESC37	59%	59%	67%	10%
mESC38	77%	76%	94%	6%
mESC39	75%	74%	90%	5%
mESC40	80%	79%	92%	5%
mESC41	67%	68%	91%	6%
mESC42	18%	17%	42%	22%
mESC43	9%	9%	75%	2%
mESC44	27%	26%	87%	5%
mESC46	70%	69%	91%	9%
mESC47	64%	63%	83%	11%
mESC48	71%	70%	87%	10%
mESC50	57%	57%	65%	6%
mESC51	78%	78%	91%	8%
mESC54	85%	84%	95%	8%

Table 4.2 Percentages of total mapped reads, reads mapping to exons and reads mapping to introns for the mouse ESCs. In red we show the single cells that were discarded at the second step of quality check, since they presented a non-satisfactory percentage of mapping reads back to the reference genome.

4.3.2 Biological over technical variation threshold

From the initial normalized FPKM value matrix, we discarded the genes with 35 or more, zero expression values for mouse and 28 or more, zero expression values for human. We calculated the mean FPKM values (mean expression) across all cells for each of the remaining genes. We selected 229 (mESCs) and 217 (hESCs) highly expressed genes (> 150 FPKM is each single cell) as highly confident sets. The remaining genes were sorted according to their mean expression levels and divided in windows of 1,000 genes each (16 windows mouse, 19 windows human). The lowest windows (1,259 genes in mouse, 1,025 genes in human) were comprised of genes with the lowest mean expression levels, hence suffering from high levels of technical variation. We calculated the Pearson correlation coefficient for each pair of highly expressed genes with each gene in each window. For each window, (except the lowest one) we compared the distribution of correlation of all the gene pairs with the distribution of correlation of the lowest window using a t-test. We kept the genes with significantly higher correlation (probability distribution shifted to the right) compared to the lowest window (comparable to random noise). CV was determined as the ratio of standard deviation to mean for each gene across single cells.

4.3.3 Transcription factor enrichment

We used data from 49 and 99 ChIP-seq experiments for transcription factors and chromatin remodellers in human and mouse embryonic stem cells respectively (Pooley et al., 2014). We selected peaks in promoter regions (+/- 1kb from the TSS) of the two groups (High CV and Non High CV). For each promoter region, we also counted the total number of factors binding at the region.

4.3.4 miRNA target interactions

Data of miRNA target interactions in ESCs were retrieved from the ESCAPE database (Xu et al., 2013). From 693,552 interactions, we kept only the interactions that their target genes were in our one-to-one orthologs list and divided the number of miRNA interactions per gene in 3 bins (1-50, 51-100, >100).

4.3.5 Protein-protein interactions

Data of protein-protein interactions were retrieved from the ESCAPE database (Xu et al., 2013). One-to-one orthologs were used to map the genes for each category and for the total list of interactions. The number of proteins interacting with each gene were divided in four bins (1, 2, 3, >3).

4.3.6 Overlap with bivalent and active genes

We overlapped our genes with genes that were classified as bivalent or active (H3K4me3 marked) in human and mouse ESCs using previous work from our lab (Mantsoki et al., 2015) and studied their differences at the level of CV.

4.3.7 Overlap with CpG islands and TATA box promoters

We calculated the overlap of the promoters of the genes with the CpG island regions as given from the UCSC tracks unmasked CpG islands for hg38 and mm10 (Karolchik et al., 2014). 2,742 murine and 2,010 human TATA-box motif promoters were retrieved from the Eukaryotic Promoter Database(Dreos et al., 2015).

4.3.8 Gene type classification

We calculated the fraction of genes that belonged to a specific gene type (from GENCODE annotation files). We selected only the types of genes with at least 30 genes in all the groups and plotted the CV for each category.

4.3.9 High variation threshold

For the sets of genes that were above the threshold of technical noise we calculated the coefficient of variation (CV) using the standard definition of ratio of the standard deviation to the mean, and divided them in four groups (quartiles) according to their CV. The High variation (High CV) genes were the ones that were falling in the fourth quartile of the CV. The rest of the genes were defined as Non High CV. Gene ontology enrichment was performed using DAVID (Dennis et al., 2003).

4.3.10 Correlation co-expression analysis

We calculated the Pearson correlation coefficient between all the pairs of High CV genes using FPKM values. We randomly permutated the FPKM values between cells for each gene to generate random data. The correlation distributions of High CV genes were significantly different (Wilcoxon test) than the random ones and we investigated their co-expression patterns by hierarchical clustering (flashClust package in R) visualised with heatmaps (heatmap.2 in R).

4.3.11 Conservation analysis

17,009 one-to-one orthologs from ensembl BioMart (Guberman et al., 2011) were used to calculate CV values in each species. After intersecting the orthologs with the 4,000 genes (for both mouse and human) we end up with a gene set containing 2,363 orthologous genes.

4.3.12 Topological associated domains

A lists of topological associated domains (TADs) for mouse and human ESCs (Dixon et al., 2012) was used to calculate the number of genes per TAD for the High CV and Non High CV genes in our analysis.

4.3.13 Bulk expression data

For the bulk RNA analysis we used 3 biological replicates of Microarray data from mouse ESCs (GSM1326660-2) (Zhang et al., 2014) and 4 biological replicates of RNA-seq data from hESCs (GSE33480) (Djebali et al., 2012).

4.3.14 Sequence conservation

The sequence conservation scores where obtained from PhyloP100way (Human) and PhyloP60way (Mouse) tracks available at UCSC.

4.4 Results

4.4.1 Correlation based approach to identify genes with significant biological variation in mammalian Embryonic Stem cell single RNA-seq data

To study the gene expression variability across individual cells, we collected RNA sequencing data for 32 human and 39 mouse single ESCs. After normalising the data across cells, we calculated FPKM values for 43,345 mouse and 60,468 human GENCODE (Harrow et al., 2012b) genes in each single cell. Single cell sequencing data suffers from low genome coverage and high amplification bias. These biases contribute to technical variation (noise) which hinders capturing biological variation across individual single cells. To distinguish the genes with significantly higher biological variation over technical variation, we developed a correlation-based approach. As highly expressed genes tend to have lower technical noise, we selected top 229 (mouse) and 217 (human) highly expressed genes (see Methods) across single cells. We then binned the genes based on their mean expression level. We calculated the correlation of genes in each bin with the highly expressed genes. We noted that technical noise was inversely related to the mean expression of gene sets i.e. higher the gene expression, lower the technical noise. We selected a threshold on expression value where the correlation with highly expressed genes was statistically significant over correlation with gene sets with technical noise (see Methods). This procedure resulted in selection of 4229 genes over 2.9 mean expression (natural logarithm transformation- ln) threshold (log(FPKM+1)) in murine ESCs (Figure 4.1A and 4.2) and 4217 genes over log mean expression threshold of 3.1 in human ESCs (Figure 4.1B and 4.3) with significantly higher biological noise than technical noise.

Gene expression variability was negatively correlated with the mean expression level i.e. highly expressed genes had low CV while lowly expressed genes spanned a wide spectrum on CV range (Figure 4.1A and 4.1B). The functional enrichment of low CV genes resulted in enrichment for cell cycle functional category specifically the ‘M phase' of mitotic cell cycle for both human and mouse ESCs. We further calculated the functional enrichment for highly expressed genes irrespective of CV values. They were also enriched for cell cycle functional category in both human and mouse ESCs. We therefore inferred that highly expressed genes tend to have low CV and are involved in cellular functions such as cell cycle.

We further checked if different gene categories provided by GENCODE (Harrow et al., 2012b) demonstrate variability comparable to protein coding genes (Figure 4.1C and 4.1D). The lincRNAs had higher CV values in both human (t-test, P-value < 0.01) and mouse ESCs (t-test, P-value < 0.05). An overwhelming fraction of murine processed pseudogenes had low CV (t-test, P-value < 0.05). In contrast, a significant fraction of human processed pseudogenes had CV higher than protein-coding genes (t-test, P-value < 0.001). Processed transcripts and antisense transcripts on the other hand show no significant difference, possibly due to low sample numbers.

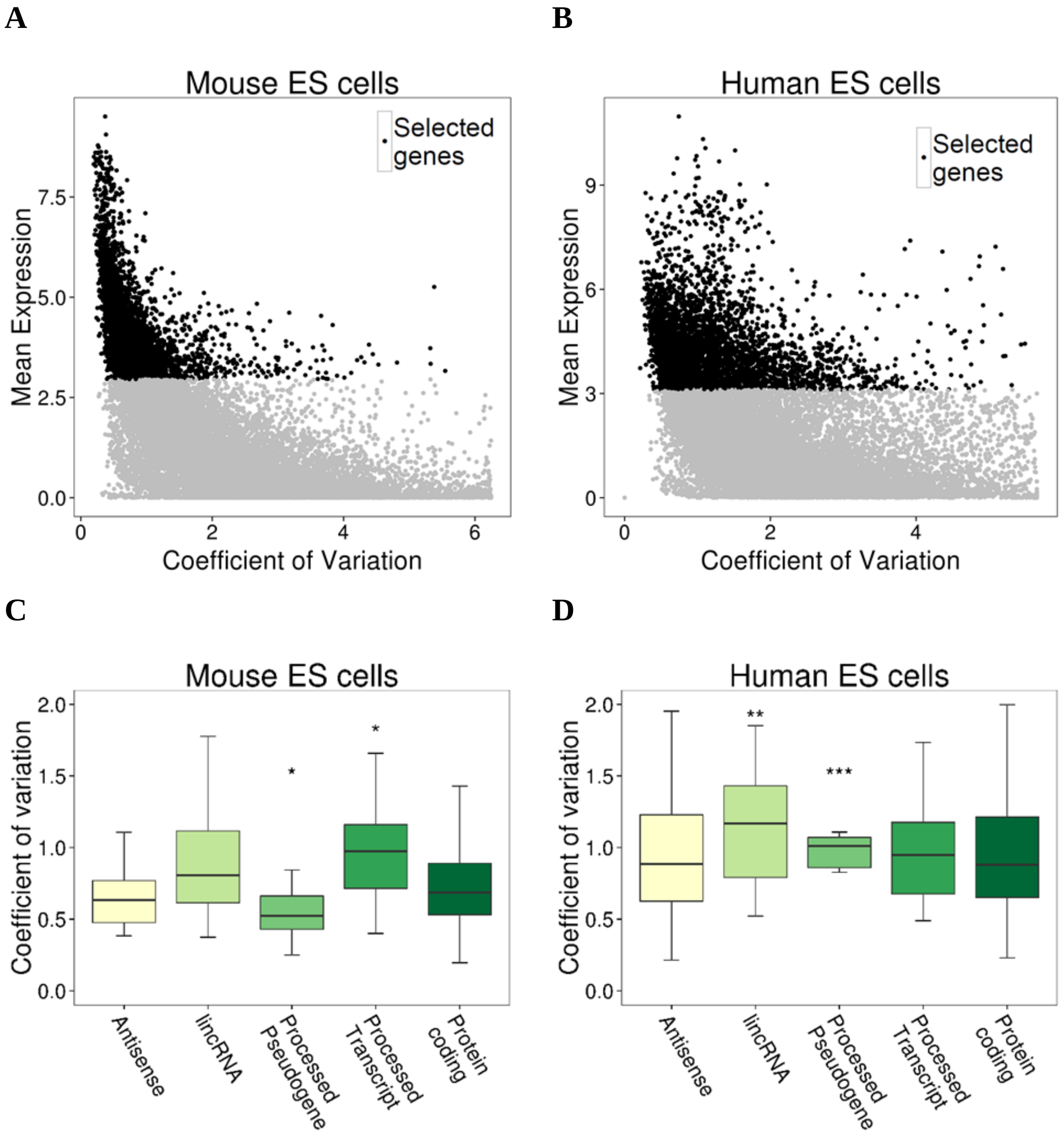
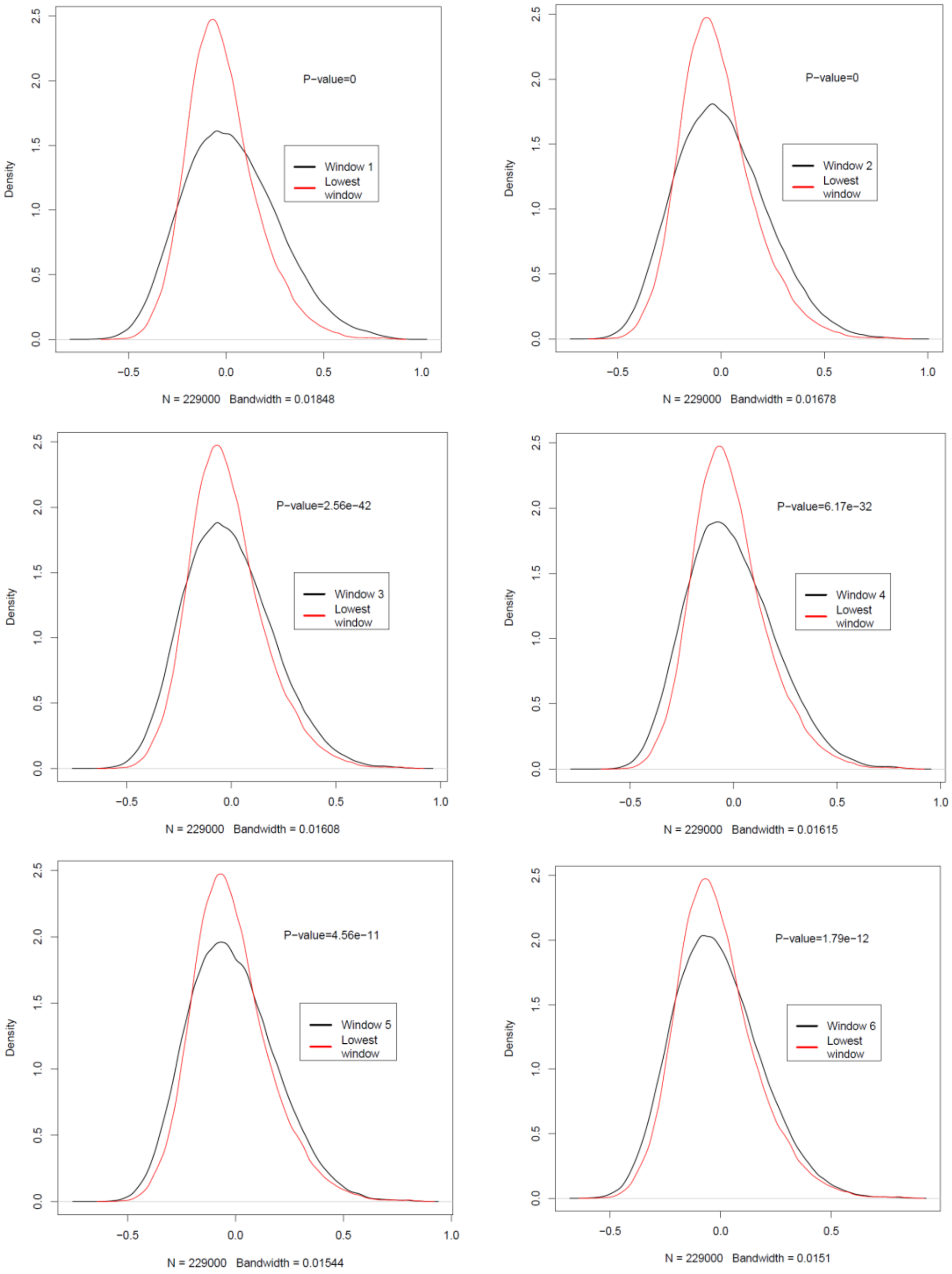
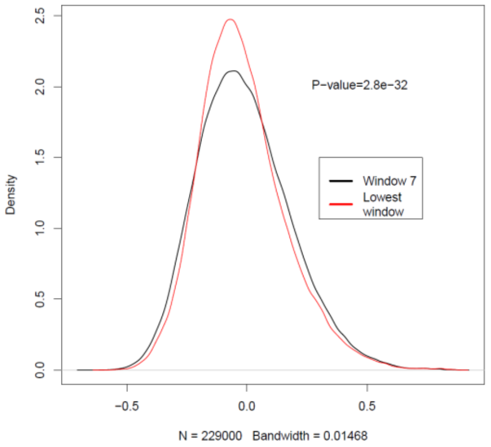
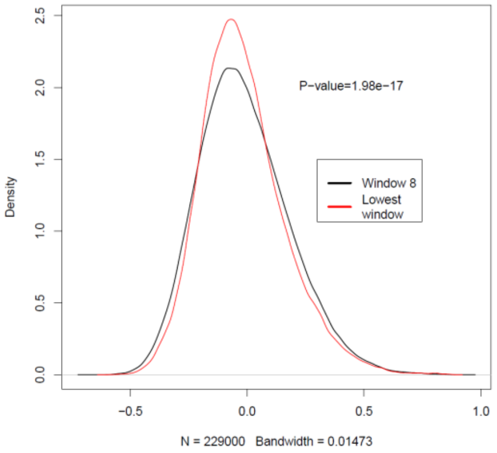


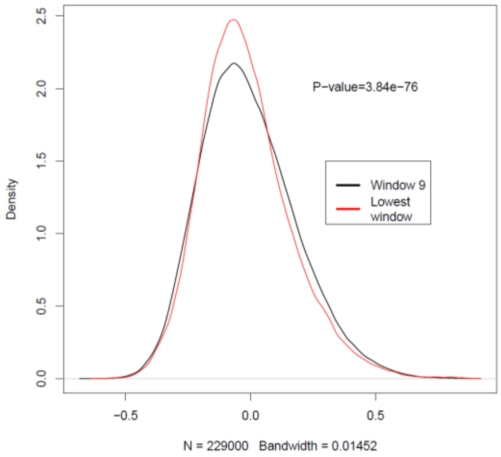
Figure 4.1 Correlation based approach for the identification of genes above the threshold of technical variation (A, B) Scatterplots showing genes according to their mean expression (log (mean FPKM+1)) and coefficient of variation in Mouse and Human ESCs. The genes highlighted in black were chosen for the analysis, since they were more correlated with the highly expressed genes. (C, D) Gene types in Mouse and Human ESCs and their respective CV levels (shown only the genes types that were found in 30 genes or more).

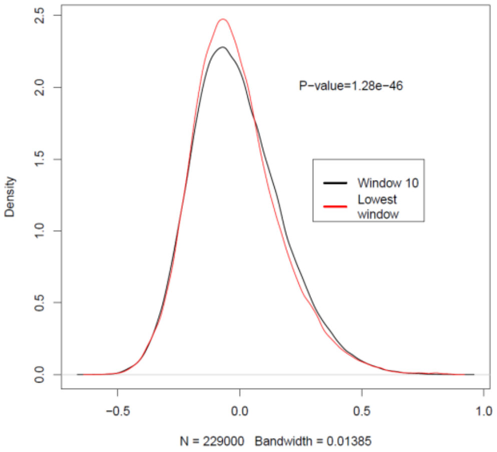
Mouse ESCs

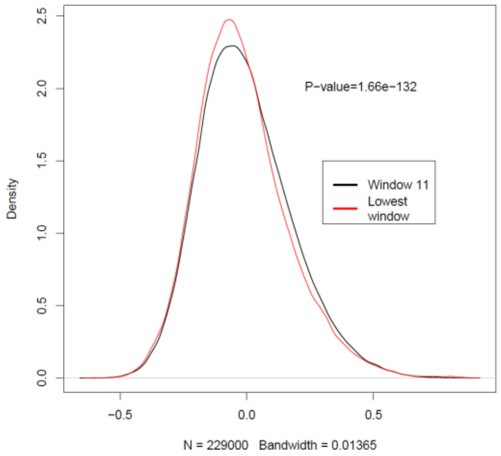

Density
0.0
0.5
1.0
1.5
2.0
2.5
-0.5
0.0
0.5
1.0
P-value=0
Window 1
Lowest window
N = 229000 Bandwidth = 0.01848
P-value=0
Window 2
Lowest window
N = 229000 Bandwidth = 0.01678
P-value=2.56e-42
Window 3
Lowest window
N = 229000 Bandwidth = 0.01608
P-value=6.17e-32
Window 4
Lowest window
N = 229000 Bandwidth = 0.01615
P-value=4.56e-11
Window 5
Lowest window
N = 229000 Bandwidth = 0.01544
P-value=1.79e-12
Window 6
Lowest window
N = 229000 Bandwidth = 0.0151

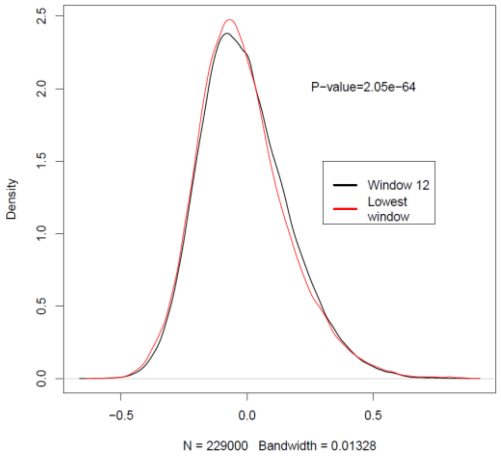

P−value=2.8e−32
Window 7
Lowest window
Density
0.0
0.5
1.0
1.5
2.0
2.5
−0.5
0.0
0.5
N = 229000 Bandwidth = 0.01468


P−value=1.98e−17
Window 8
Lowest window
Density
0.0
0.5
1.0
1.5
2.0
2.5
−0.5
0.0
0.5
1.0
N = 229000 Bandwidth = 0.01473


P−value=3.84e−76
Window 9
Lowest window
Density
0.0
0.5
1.0
1.5
2.0
2.5
−0.5
0.0
0.5
N = 229000 Bandwidth = 0.01452


P−value=1.28e−46
Window 10
Lowest window
Density
0.0
0.5
1.0
1.5
2.0
2.5
−0.5
0.0
0.5
1.0
N = 229000 Bandwidth = 0.01385


P−value=1.66e−132
Window 11
Lowest window
Density
0.0
0.5
1.0
1.5
2.0
2.5
−0.5
0.0
0.5
N = 229000 Bandwidth = 0.01365


P−value=2.05e−64
Window 12
Lowest window
Density
0.0
0.5
1.0
1.5
2.0
2.5
−0.5
0.0
0.5
N = 229000 Bandwidth = 0.01328

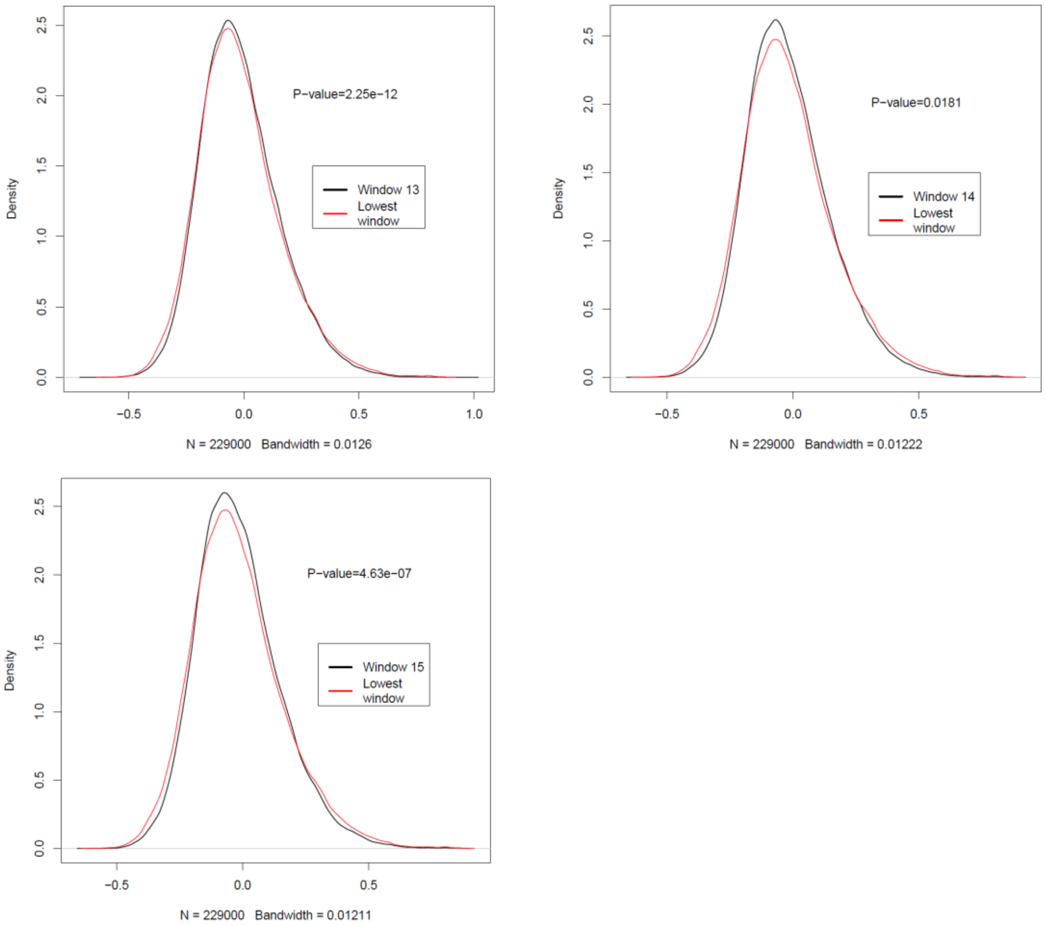
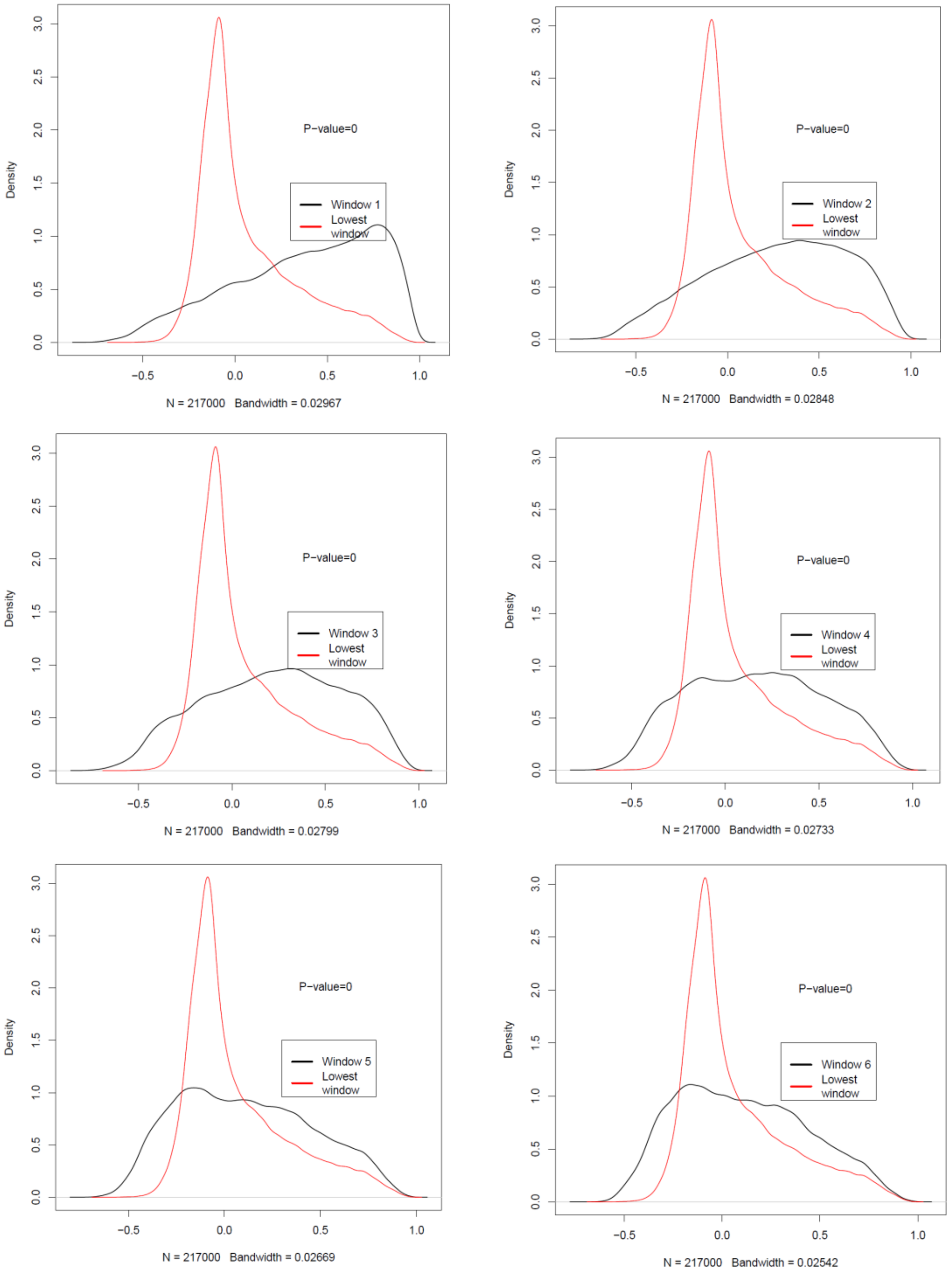
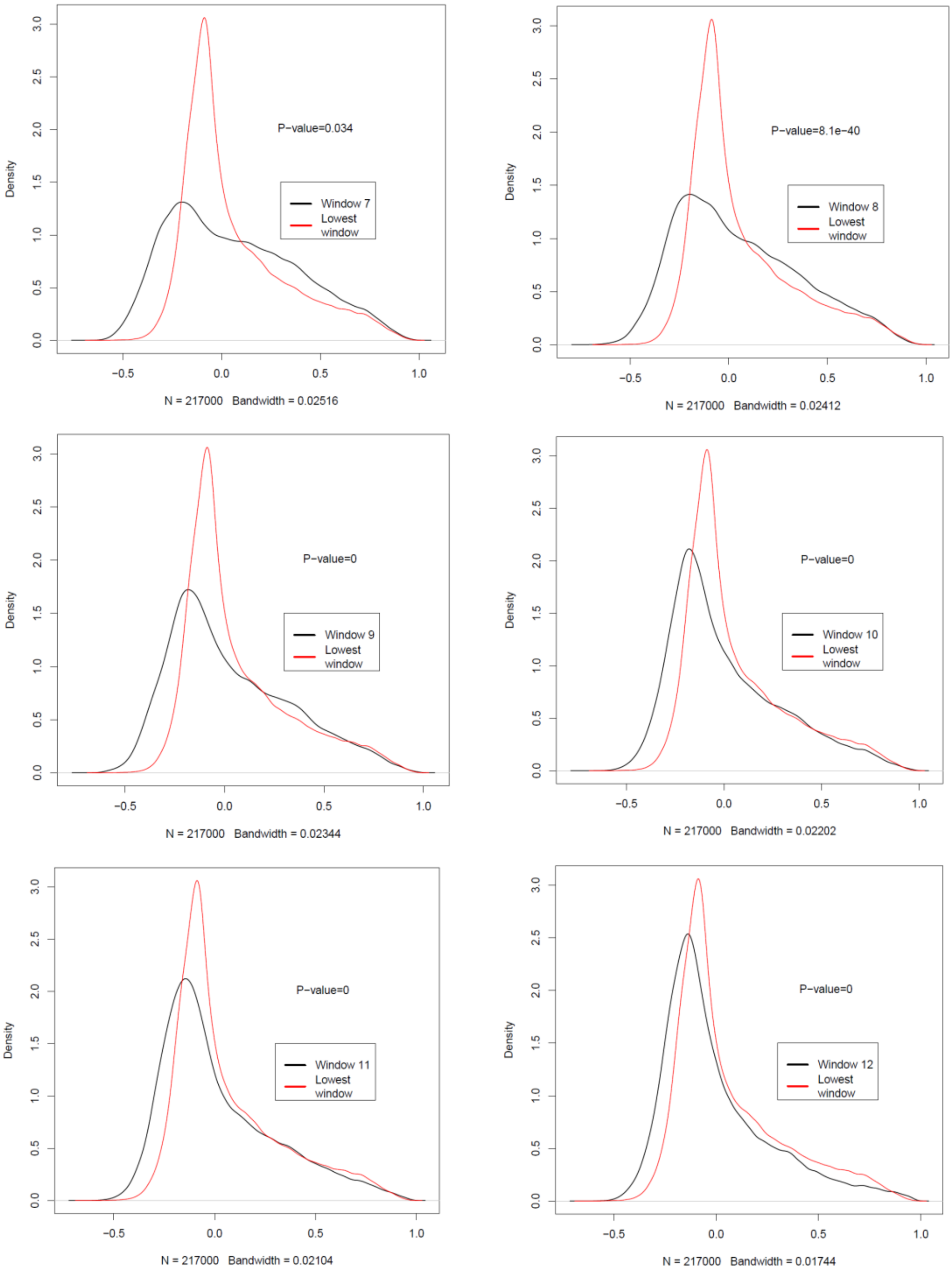


Figure 4.2 Distributions of correlation of high expression genes with genes in 15 windows of mean expression compared with the lowest mean expression window, in Mouse ESCs. The genes of Windows 1 to 4 were chosen as the ones that are above the threshold of technical variation.

Human ESCs


Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-0.5
0.0
0.5
1.0
P-value=0
Window 1
Lowest window
N = 217000 Bandwidth = 0.02967
Window 2
Lowest window
N = 217000 Bandwidth = 0.02848
Window 3
Lowest window
N = 217000 Bandwidth = 0.02799
Window 4
Lowest window
N = 217000 Bandwidth = 0.02733
Window 5
Lowest window
N = 217000 Bandwidth = 0.02669
Window 6
Lowest window
N = 217000 Bandwidth = 0.02542


Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-0.5
0.0
0.5
1.0
P-value=0.034
Window 7
Lowest window
N = 217000 Bandwidth = 0.02516
P-value=8.1e-40
Window 8
Lowest window
N = 217000 Bandwidth = 0.02412
P-value=0
Window 9
Lowest window
N = 217000 Bandwidth = 0.02344
P-value=0
Window 10
Lowest window
N = 217000 Bandwidth = 0.02202
P-value=0
Window 11
Lowest window
N = 217000 Bandwidth = 0.02104
P-value=0
Window 12
Lowest window
N = 217000 Bandwidth = 0.01744

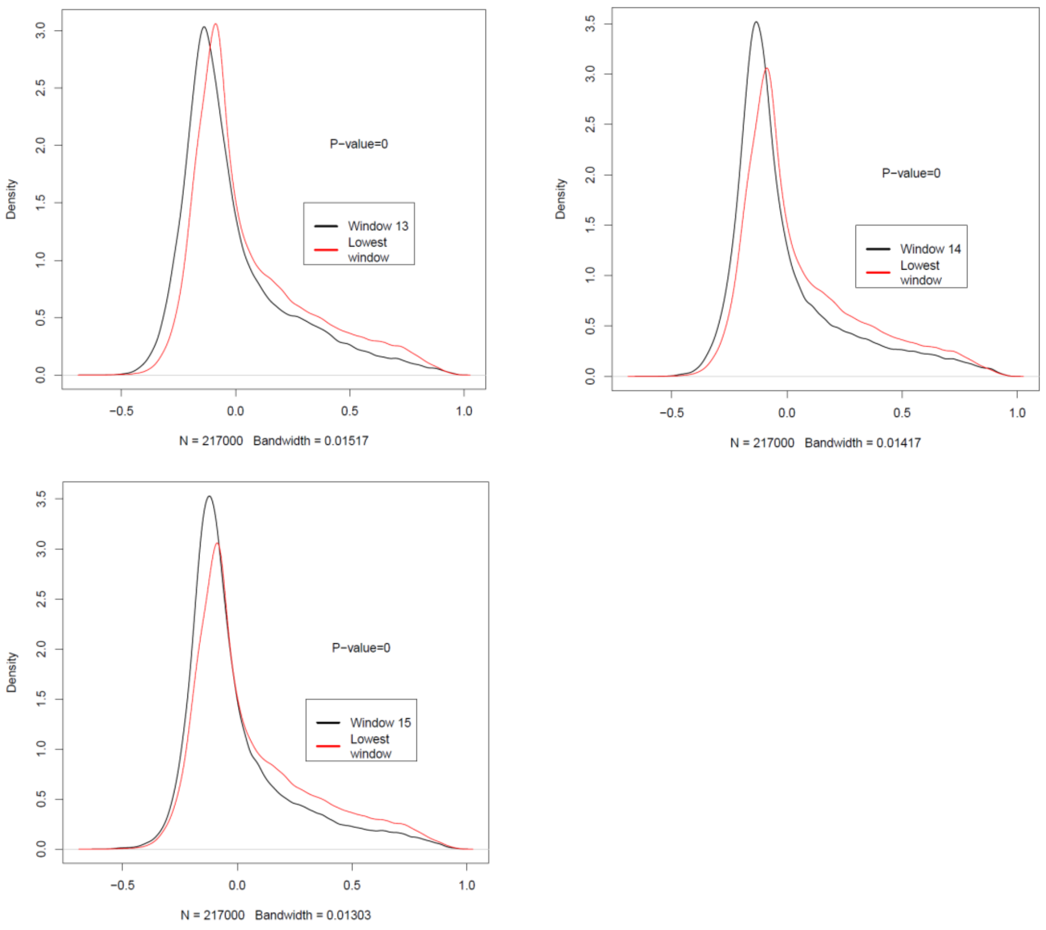


Figure 4.3 Distributions of correlation of high expression genes with genes in 18 windows of mean expression compared with the lowest mean expression window, in Human ESCs. The genes of Windows 1 to 4 were chosen as the ones that are above the threshold of technical variation.

4.4.2 Genes occupied by many transcription factors have a lower CV

In order to study the level of transcription control among three groups of promoters, we calculated the number of factors binding at each promoter using ChIP sequencing compendia for transcription and epigenetic factors in human and mouse ESCs (Pooley et al., 2014). The mean CV for genes bound by less than 10 factors was significantly higher than the mean CV for genes bound by more than 10 factors in both human (t-test, P-value < 0.001) and mouse (t-test, P-value < 0.001) ESCs (Figure 4.4A and 4.4B).

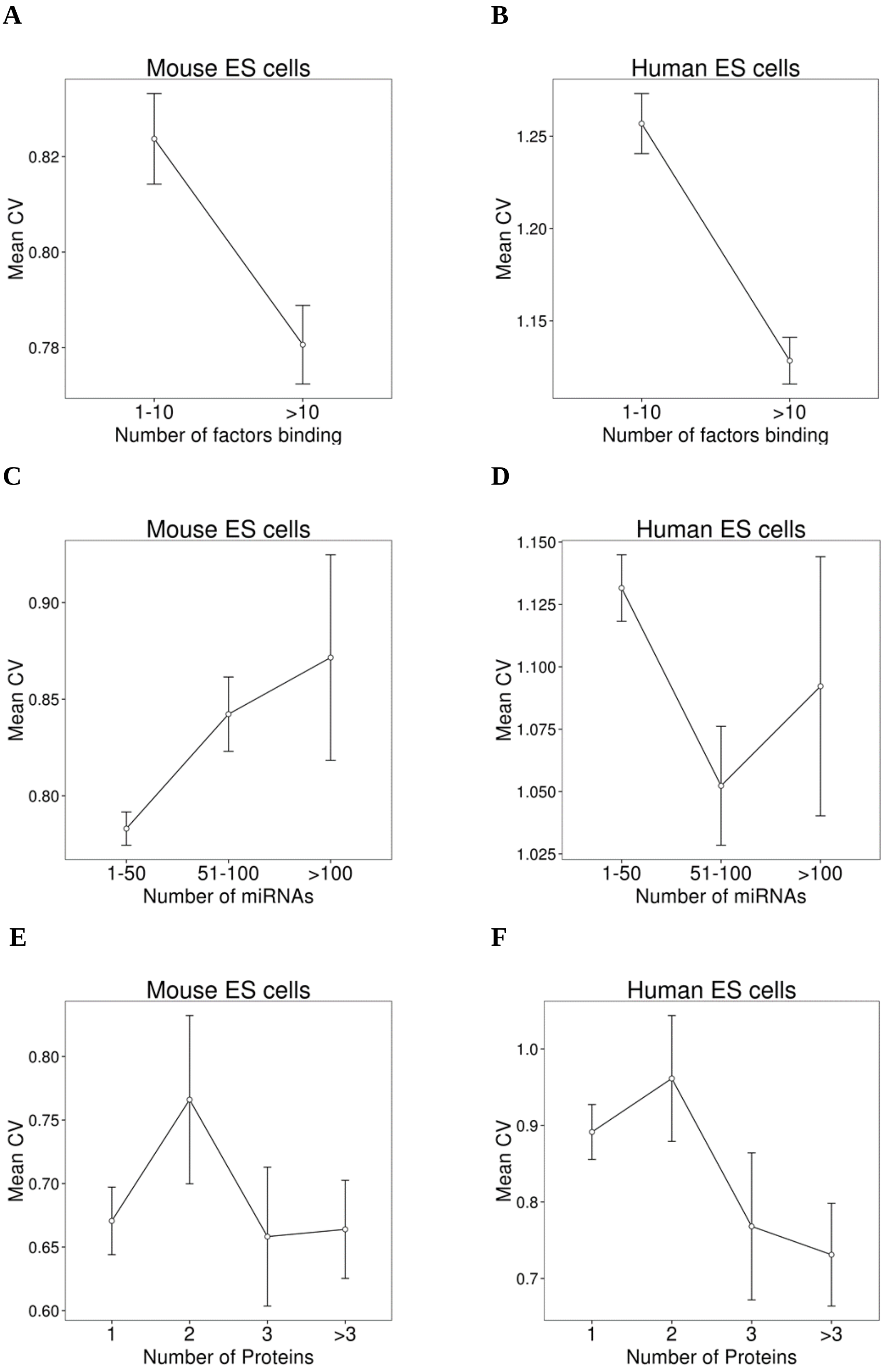


Figure 4.4 Mean CV levels according to quantification of transcription factors, miRNA targets and protein-protein interactions. (A, B) Transcription and epigenetic factor

occupancy (number of factors binding) at the promoters of genes is inversely correlated with their Mean CV in Mouse (99 ChIP-seq TFs) and Human (49 ChIP-seq TFs) ESCs. (C, D) Bins of miRNAs targeting each gene and their responding Mean CV levels (only interactions with genes in orthologs one2one list have been used) in Mouse and Human ESCs. (E, F) Genes (only interactions with genes in orthologs one2one list have been used) with known protein-protein interactions for Mouse and Human ESCs and their responding Mean CV levels.

This result was consistent when average binding of individual factors was tested as well i.e. genes more likely to be bound by more factors tended to have low CV. We obtained the number of putative binding sites of transcription factors in gene promoters from UCSC. Again, number of putative binding sites varied inversely with the CV value (Figure 4.5).

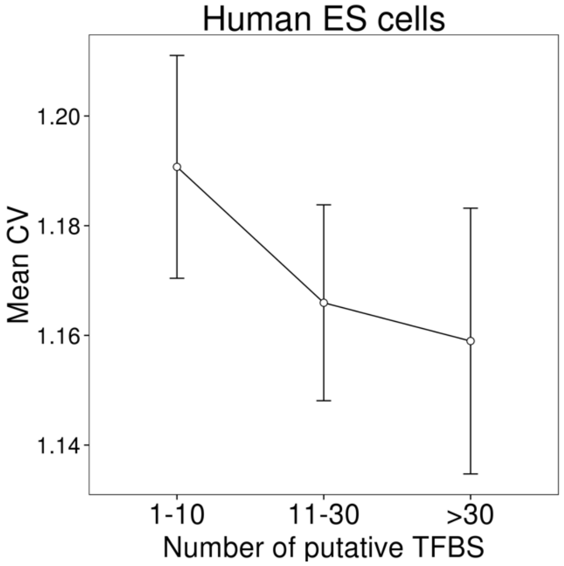


Figure 4.5 Number of putative Transcription Factor Binding Sites (TFBS) per gene (shown in 3 bins) and their corresponding Mean CV values. There was no statistically significant difference between means

To test the regulation at post-transcriptional level, we collected putative miRNA targets predicted by four miRNA prediction methods(Xu et al., 2013). Unlike TF targets, there was no bias towards the number of miRNA targets with respect to their mean CV, either in human or mouse ESCs (Figure 4.4C and 4.4D).

Finally we collected known protein-protein interactions (PPI) in mouse and human ESCs(Xu et al., 2013) and calculated the number of known interacting partners for each of the genes. Similarly, to miRNA targets, there was no statistically significant difference between the mean CV values based on the number of interacting partners at protein level in either human or mouse ESCs (Figure 4.4E and 4.4F).

4.4.3 High expression variability genes correlate with DNA repair and bivalency

The activity of signalling pathways such as TGF-β-related signalling pathways are thought to prime cells for differentiation contributing to the heterogeneity between cells in ESCs (Galvin-Burgess et al., 2013). The CV value did not distinguish any particular signalling pathway. The differences in micro-environments sensed by the signalling pathway can manifest in large expression changes of its downstream target genes. We therefore tested whether transcription factor and chromatin remodeller binding prefers or avoids gene promoters based on their CV measure using the ChIP sequencing data compendium for 49 and 99 factors in mouse and human ESCs respectively(Pooley et al., 2014). Unsurprisingly, many promoter specific factors such as E2F1, TAF1, and YY1 did not show any bias for the CV. High CV genes in mouse ESCs showed an exclusive binding preference of the following four factors: NCOA3 (Hypergeometric test, P-value < 0.0001), p300 (Hypergeometric test, P-value < 0.0001), MCAF1 (Hypergeometric test, P-value < 0.01) and p53(Hypergeometric test, P-value < 0.05).

NCOA3 is a nuclear receptor activator with a histone acetyltransferase activity, recruiting the chromatin modifying proteins p300, CARM1 and CBP at the *Nanog* locus (Wu et al., 2012). NCOA3 is thought to be critical for both the induction and maintenance of pluripotency, acting as an essential Esrrb coactivator (Percharde et al., 2012). ESRRB is downstream of NANOG which is a direct target of TGF-β mediated SMAD signalling(Xu et al., 2008). NANOG targets did not show any bias with respect to CV.

MCAF1 is a nuclear protein associated with heterochromatin, shown to colocalize with SETDB1 in PML bodies (Sasai et al., 2013). PML is a protein involved in the senescence pathway through the p53 signalling, and its overexpression leads to premature senescence (Pearson et al., 2000). p53 is a sequence specific transcription factor with tumour suppressor activity, regulating cell cycle arrest, apoptosis, senescence and stem cell differentiation, acting as an activator or suppressor of its downstream targets (Vousden and Prives, 2009). Upon DNA damage, p53 activates differentiation associated genes and represses self-renewal genes, affecting the status of ESCs (Li et al., 2012).

Accordingly, high CV genes showed enrichment for biological processes such as cellular response to stress (Fisher's exact test, adjusted P-value < 10^{-4}), response to DNA damage stimulus (Fisher's exact test, adjusted P-value < 10^{-3}) and DNA repair (Fisher's exact test, adjusted P-value < 10^{-3}) in both murine and human ESCs.

The genes overlapping with bivalent promoters had statistically significant higher CV values than the ones overlapping with the active promoters (presence of H3K4me3 and absence of H3K27me3 modifications) in both human (Hypergeometric test, P-value < 0.001) and mouse (Hypergeometric test, P-value < 0.001) ESCs (Figure 4.6A and 4.6B). Genes with high CV showed a weak functional enrichment for embryonic development and transcription control; the functional categories associated with bivalent genes (Bernstein et al., 2006b).

As specific promoter structures such as presence of TATA boxes have been previously associated with genes with highly fluctuating single-cell levels within populations(Choi and Kim, 2009), we calculated TATA and CpG island fraction for all human and mouse promoters (-/+ 1Kb from TSS). The CpG-rich promoters showed lower CV values than the CpG-poor promoters and the difference was statistically significant in both human and mouse ESCs (t-test P-value<0.001) (Figure 4.6C and 4.6D). Unlike CpG promoters, TATA box promoters could not be distinguished based on the CV value (Figure 4.6E and 4.6F).

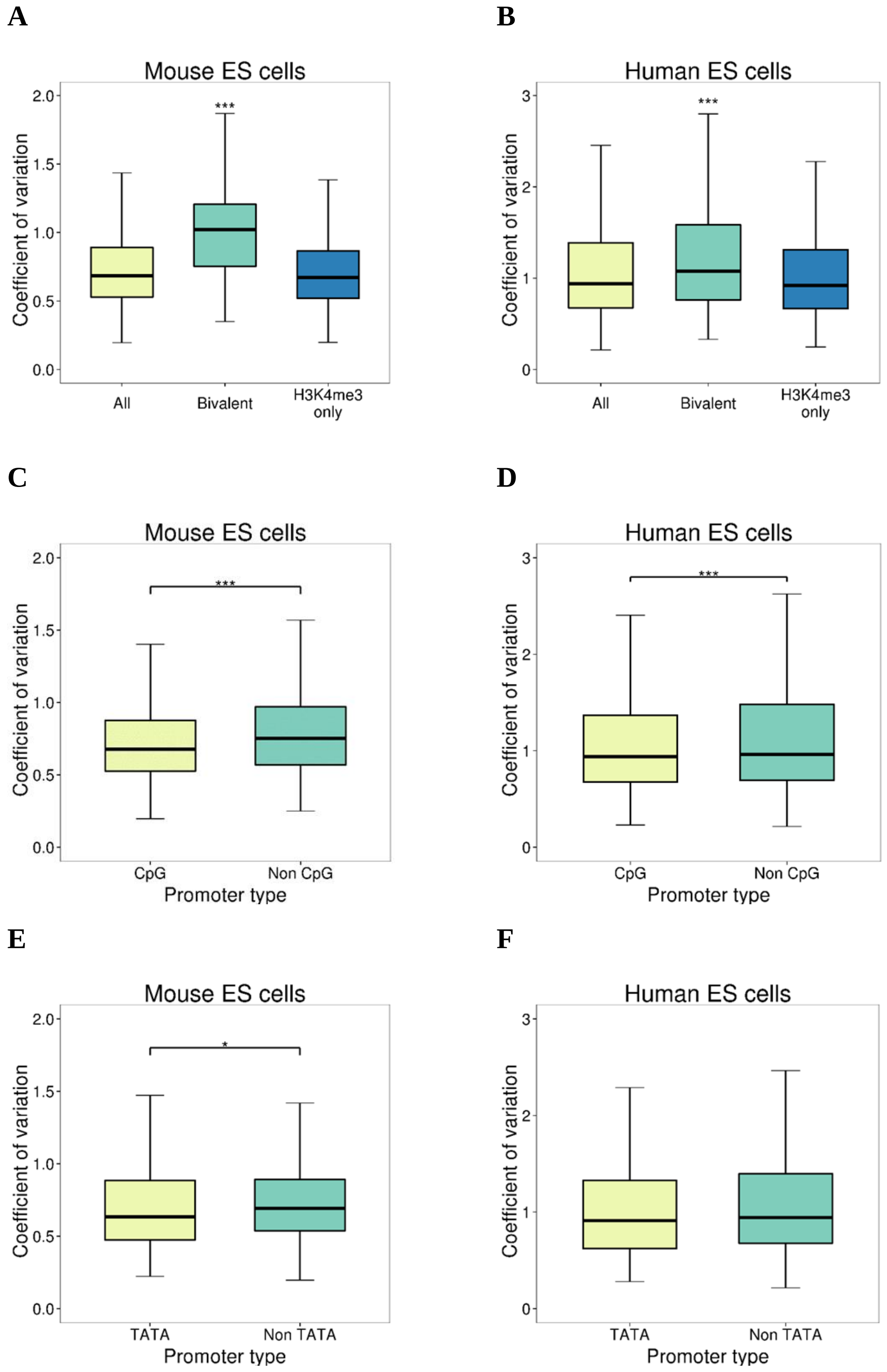


Figure 4.6 Chromatin modifications and sequence features of genes and their corresponding coefficient of variation. (A, B) Overlapping genes with bivalent and active (H3K4me3 marked) gene promoters in response to their CV, in Mouse and Human ESCs. Bivalent genes show significantly higher CV levels than all the promoters (irrespective of overlap) and the active promoters (pairwise t-test, P-value < 0.001) (C)

CV levels of genes having a CpG island and a non- CpG island promoter. (D) CV levels of genes having a TATA box and a non-TATA box promoter.

4.4.4 High CV genes form dense highly co-expressed clusters

In order to study the characteristics of genes with high variability, we defined genes with CV value greater than 0.92 (3rd quartile value) as High CV in mouse (Figure 4.7A) and genes with CV value greater than 1.45 (3rd quartile value) in human ESCs (Figure 4.7B). We then checked whether the expression of High CV genes varies concordantly across single cells by calculating Pearson's correlation coefficient between all pairs of High CV genes. A subset of High CV genes was significantly more correlated with each other compared to expected from a random permutation (Figures 4.7C (mouse) and 4D (human)).

The highly correlated network (Pearson's correlation coefficient > 0.95) of High CV genes grouped them mainly into only few tightly co-expressed clusters in both human and mouse ESCs (Figure 4.8 and 4.9). Interestingly, the genes in each cluster were highly expressed only in one individual cell (Figure 4.7E (mouse) and 4.7F (human)). We firstly confirmed that these single cells (e.g. single cell 24 and 26 in humans) did not suffer from poor technical quality of samples (Figure 4.10). We also removed these two cells and redefined the High CV gene set (Figure 4.11) to find a similar result. This assured that the significant co-expression among High CV genes is not an artefact of few aberrant single cells.

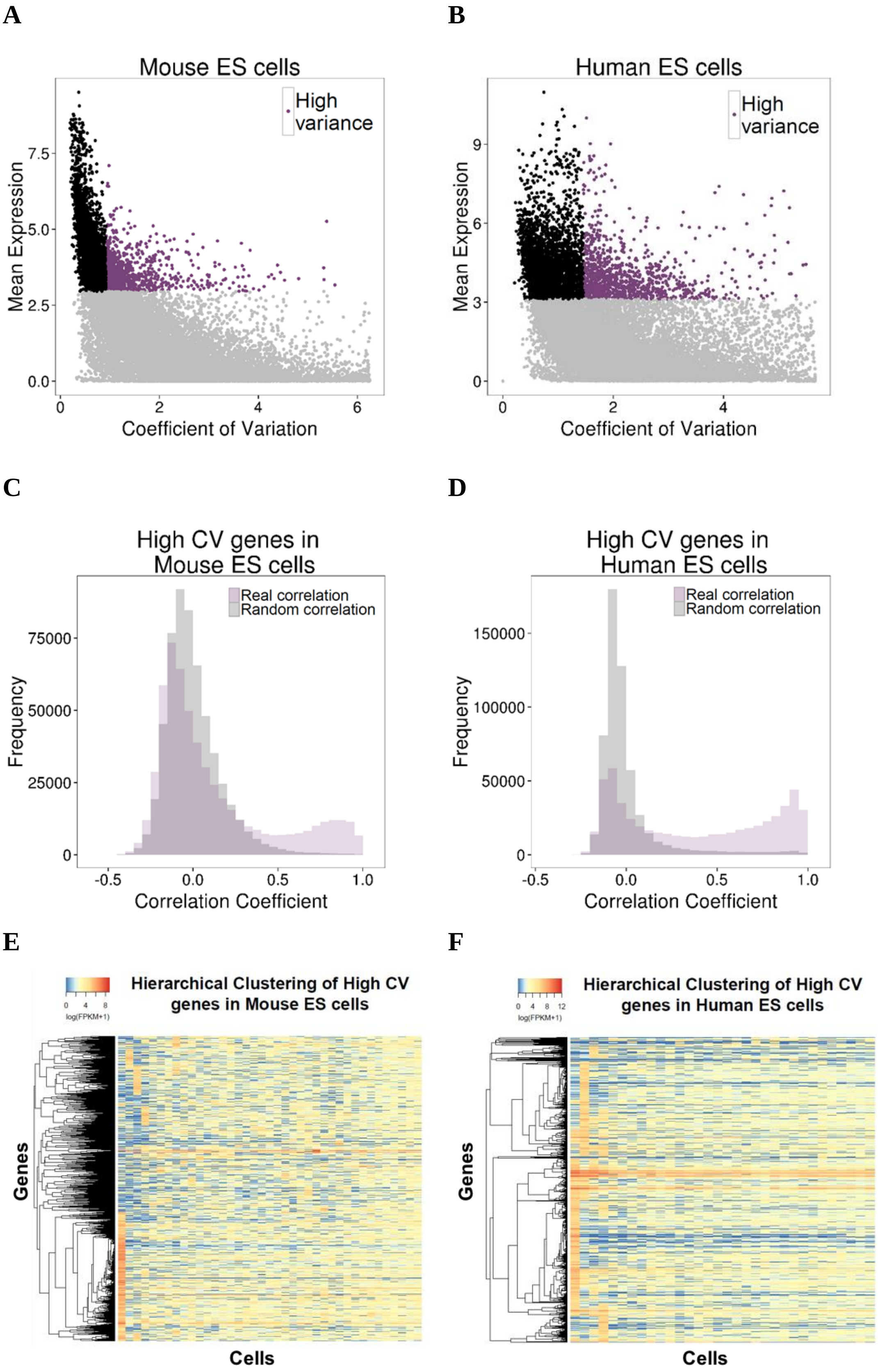

A
Mouse ES cells
High variance
Mean Expression
7.5
5.0
2.5
0.0
0
2
4
6
Coefficient of Variation
B
Human ES cells
High variance
Mean Expression
9
6
3
0
0
2
4
Coefficient of Variation
C
High CV genes in Mouse ES cells
Real correlation
Random correlation
Frequency
75000
50000
25000
0
-0.5
0.0
0.5
1.0
Correlation Coefficient
D
High CV genes in Human ES cells
Real correlation
Random correlation
Frequency
150000
100000
50000
0
-0.5
0.0
0.5
1.0
Correlation Coefficient
E
Hierarchical Clustering of High CV genes in Mouse ES cells
0 4 8
log(FPKM+1)
Genes
Cells
F
Hierarchical Clustering of High CV genes in Human ES cells
0 4 8 12
log(FPKM+1)
Genes
Cells

Figure 4.7 High variance genes are more correlated than expected by chance (A, B) Scatterplot of genes in response to their CV and mean expression. Highlighted in purple are the High variance genes, selected based on their CV (CV value greater than the third quartile of the distribution). (C, D) Correlation coefficient distributions for the High variance (High CV) genes in Mouse and Human ESCs (statistically significant difference (p<0.001, Wilcoxon test) between the real and random distributions). (E, F) Heatmaps of gene expression (in log(FPKM+1) values) for the High variance genes (High CV) in Mouse and Human ESCs.

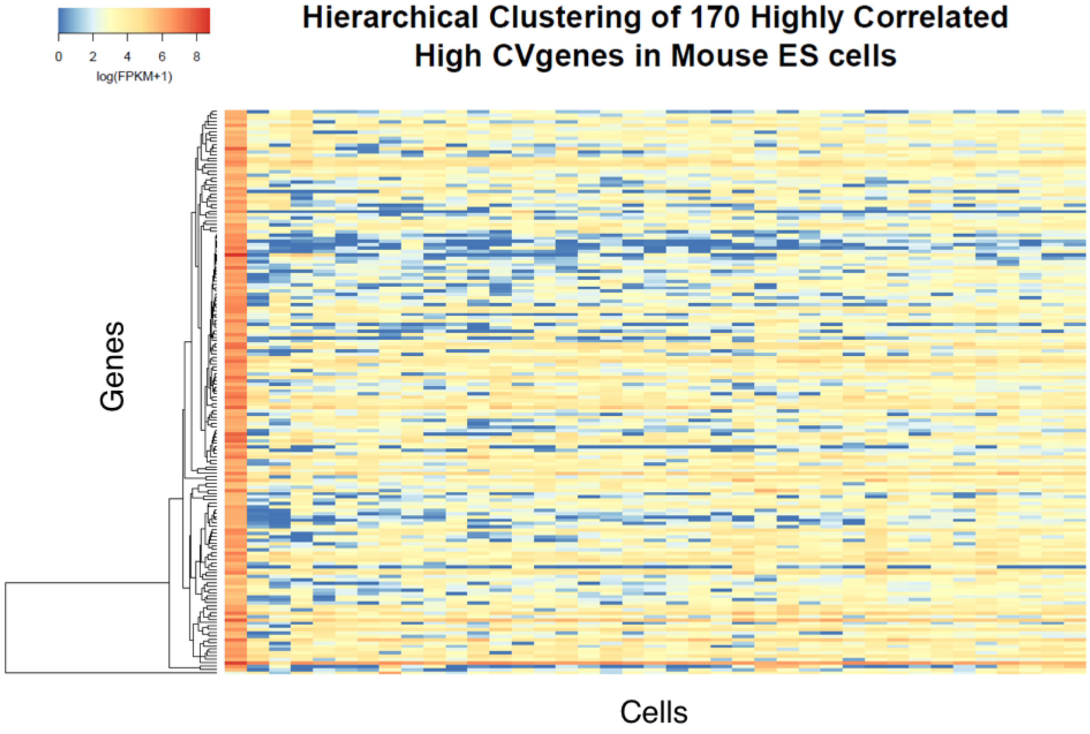


Figure 4.8 Heatmap of gene expression values (in log(FPKM+1)) of 170 highly correlated and highly variable (High CV) genes in Mouse ESCs.

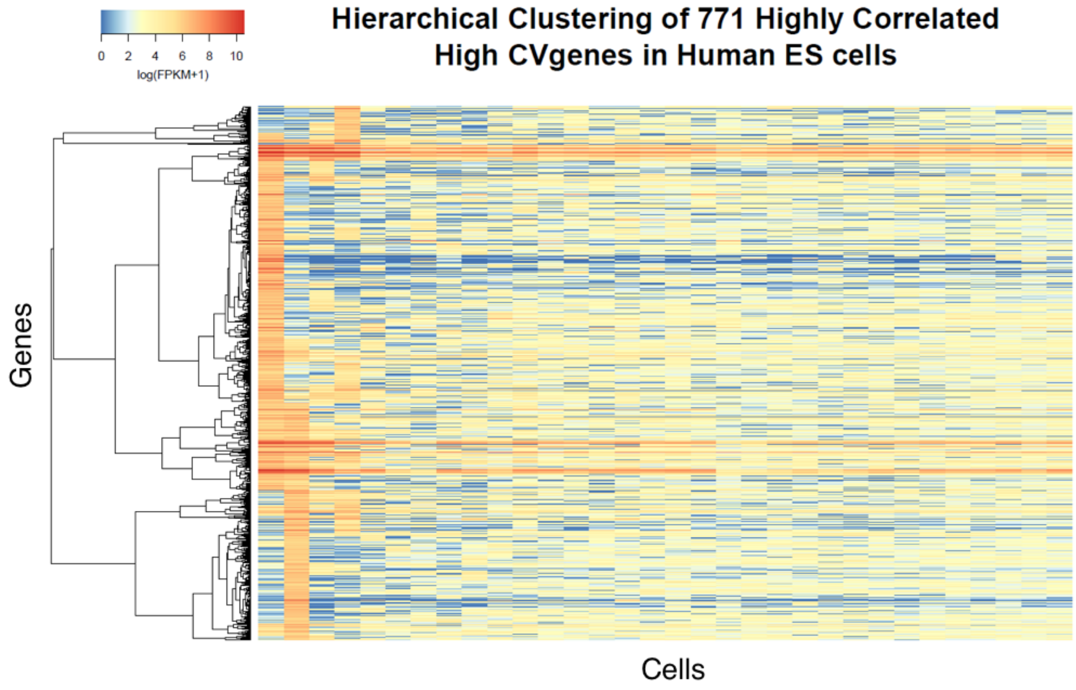


Figure 4.9 Heatmap of gene expression values (in log(FPKM+1)) of 771 highly correlated and highly variable (High CV) genes in Human ESCs.

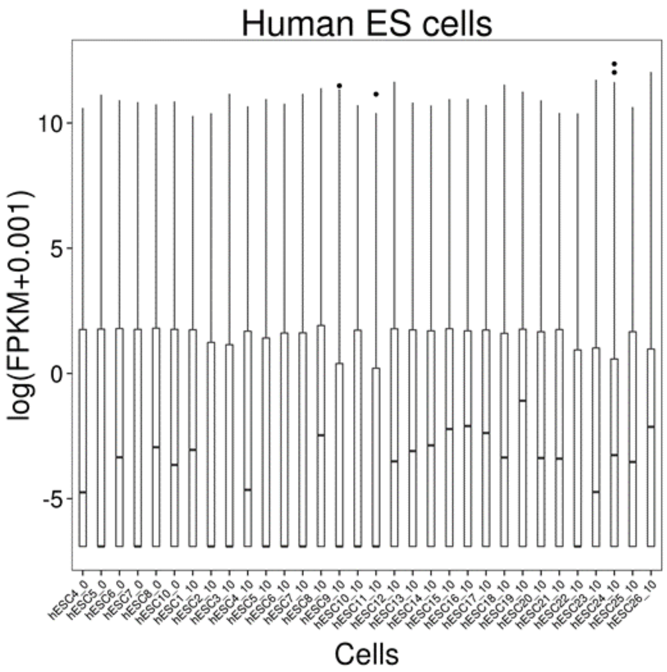


Figure 4.10 Boxplot of FPKM values for all cells in human

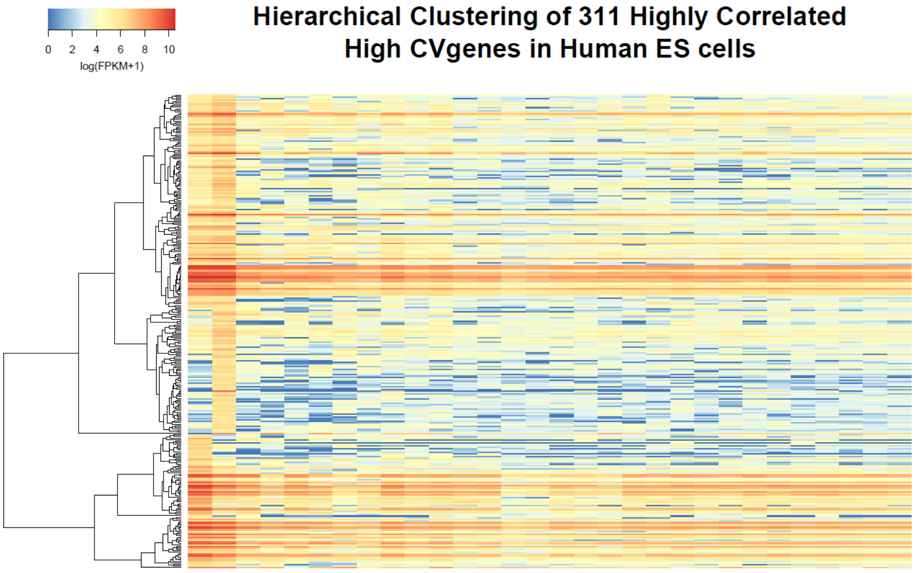


Figure 4.11 Heatmap of High CV genes in Human ESCs after discarding cells 24 and 26

The co-expressed genes derived from large-scale analyses of mammalian expression data have demonstrated that neighbouring genes tend to have similar expression profiles(Lercher et al., 2002). As high CV genes formed tight co-expression clusters, we checked whether they tend to be in gene neighbourhoods with each other compared to other genes. We did not observe any tendency of genes clustering based on CV value. We also checked whether there was any bias towards similar CV genes co-existing in topological associated domains (TADS) inferred from Hi-C chromatin capture data in human and mouse ESCs (Dixon et al., 2012). There was no bias towards associating similar CV value genes with same TADS. Also, tightly co-expressed High CV genes in each cluster were not specifically enriched for any biological process nor primed for specific lineage.

4.4.5 CV values are conserved across species

In order to check whether the CV values are conserved between bulk and single cell experiments, we obtained gene expression values for bulk RNA in human and mouse ESCs. The CV values of genes from single cells and bulk RNA showed no correlation in both human (Pearson’s correlation coefficient *r*=0.09) and mouse (Pearson’s correlation coefficient *r*=0.06) ESCs (Figure 4.12A and 4.12B).

To test whether gene expression variability from single and bulk RNA-seq is conserved across species, we collected one-to-one orthologs between human and mouse (Guberman et al., 2011). The gene expression tends to be conserved across species for single (Pearson’s correlation coefficient *r*=0.23) (Figure 4.12C) i.e. orthologs of genes with lower CV in mouse are more likely to have lower expression variance across human single ESCs and vice versa. We confirmed that the distribution of CV values for orthologous genes in mouse was not significantly different from mouse-specific genes (Figure 4.12D). We further checked whether the expression conservation goes hand-in-hand with the conservation at the sequence level. Indeed, sequence conservation showed a negative correlation with the CV values in both human and mouse ESCs in their 5’UTR, their 3’UTR and their exons (Figure 4.12E and 4.12F). Thus tight regulation of gene expression level is a feature that appears to be conserved and selected during evolution.

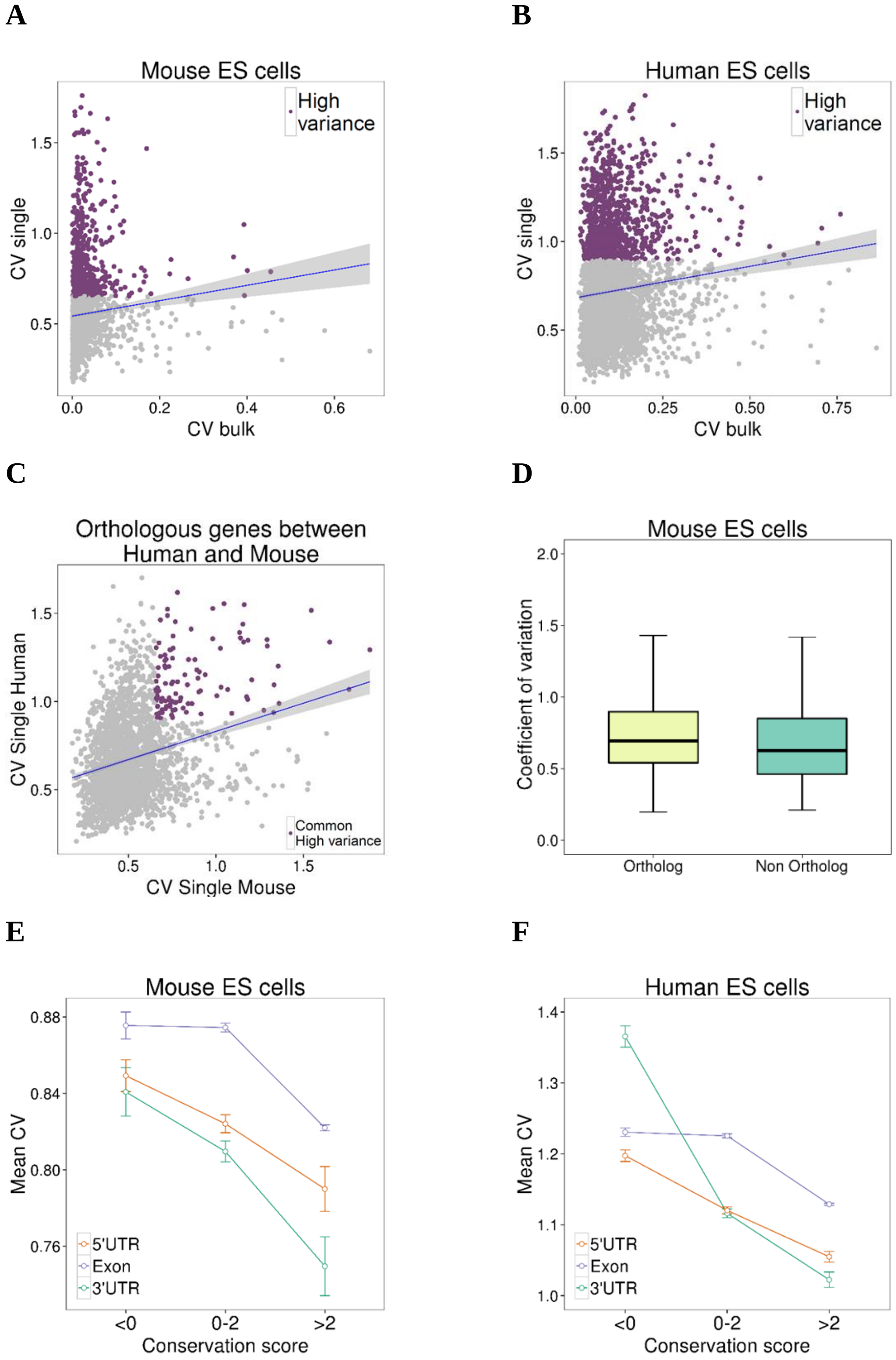


Figure 4.12 Conservation of expression variability across technologies and species. (A, B) Scatterplot of CV values in a bulk expression study against CV values in a single cell RNA–seq study in Mouse and Human ESCs. There is a positive correlation between the CV values of the two technologies (Pearson's r=0.06 for mouse, r=0.09 for human). (C)

Scatterplot of CV values of orthologous genes between human and mouse from single RNA-seq studies in ESCs. There is a positive correlation of CV values between species (Pearson's r=0.23) and 10% of High CV genes (highlighted in purple) are conserved as highly variant between species (D) Boxplot of CV values of orthologous and non-orthologous genes between human and mouse in ESCs (3,675 orthologs and 554 non-orthologs out of 4,229 genes in our analysis). (E, F) Sequence conservation scores and their corresponding Mean CV values for 5'UTR, Exons and 3'UTRs in Mouse and Human ESCs.

4.5 Discussion

Single cell RNA-seq data holds a great promise for studying variability across individual cells with the hindrance of large technical noise inherent to these data. Though availability of data from a limited number of cells (32 in human, 39 in mouse) could influence the results, it has been recently shown that 30 cells is the lower limit of sample size to sufficiently converge to the complexity of large cell populations (Marinov et al., 2014). We used a correlation based approach to define a set of genes with biological variation significantly higher than technical variation across single cells. We then studied the characteristics of expression variability for 4,217 genes in human and 4,229 genes in mouse single ESCs, where the estimated biological variability was significantly greater than the technical variability. We noted that highly expressed genes tended to have lower CV (Figure 4.1A & 4.1B). Since ESCs are not synchronized in their cell cycle and can belong to different development stages, we specifically looked whether genes with high CV were developmental stage specific or involved in specific function, but did not find a strong evidence for it.

High CV genes form co-expression clusters. Tightly co-expressed High CV genes in each cluster were highly expressed only in one or a few single cell(s) and genes in each cluster were not specifically enriched for any biological process. This fits with the notion of pluripotent cells to alternate between different transient and reversible cell states where transient states do not show any functional bias or lineage priming. High CV genes showed enrichment for response to DNA damage and DNA repair and were exclusively bound by regulators of DNA damage and senescence pathways like MCAF1 and p53. They also showed significant overlap with bivalent genes in human and mouse ESCs. Indeed, it has been previously shown that genes whose promoters are bound by Polycomb regulators can produce highly variant levels of transcripts per cell (scRNA-seq data used) despite the co-incidence of H3K27me3 at the locus (Kumar et al., 2014). More specifically, Polycomb bound genes expressed at higher levels, showed weaker H3K27me3 signatures than the ones with transcripts detected in fewer cells. This confirms that at least a subset of bivalent genes can indeed be attributed to heterogeneity in ESCs.

Though many characteristics of CV genes are conserved across species, there are some differences. Interestingly the vast majority of murine processed pseudogenes have lower CV than protein-coding genes while human processed pseudogenes have higher CV than protein-coding genes. Processed pseudogenes have recently been demonstrated to play a regulatory role by competing with other genes for the binding of small RNAs (Poliseno et al., 2010). This potential species specific regulatory aspect needs to be explored in detail.

Taken together, genes with lower CV tend to be highly expressed, tightly regulated at transcriptional level as they are likely to be central to many cellular processes. High CV genes, on the other hand, are highly expressed only in individual single cells which possibly partly explains the bivalent genes (with both active and inactive chromatin status) observed in bulk studies.

Chapter 5 Chromatin dynamics and RNAPII pausing at murine promoters in eight cell lineages

5.1 Chapter Introduction

This chapter contains unpublished work and it was written in a manuscript format following the general format of the previous chapters. Here we have integrated epigenetic and transcriptomic data in ESCs and differentiated lineages with a focus on promoter chromatin state dynamics and more particularly the resolution of bivalency after ESC differentiation. Data was collected only for mouse cell types.

5.2 Introduction

The vast versatility of mammalian cell types is generated with the assistance of epigenetic features and transcription factors which control the transcription of genes that reside at the shared genome across all the cell types of an organism (Reik, 2007; Rivera and Ren, 2013). Large collaborative efforts including ENCODE (ENCODE Project Consortium, 2012) and Roadmap Epigenomics (Bernstein et al. 2010) have generated genome-wide profiles of epigenetic features, chromatin accessibility, transcription factor (TF) DNA binding and gene expression (Mortazavi et al., 2008; Park, 2009; Wang et al., 2009; Zhou et al., 2011), across hundreds of cell types and tissues. This publicly available data constituted the building blocks of many studies which assessed the variability of the epigenetic landscape during development or disease onset (Maurano et al., 2012; Thurman et al., 2012; Xie et al., 2013; Zhu et al., 2013).

Multivariate statistical models have been instrumental to classify the chromatin into biologically meaningful states that correspond not only to coding regions of the genome but also to non-coding regulatory elements such as promoters and enhancers (Birney et al., 2007; Guenther et al., 2007; Heintzman et al., 2007). Furthermore, combinatorial patterns of multiple epigenetic datasets have been evaluated in individual (Ernst and Kellis, 2010) or multiple cell types (Ernst et al., 2011) and tissues (Kundaje et al., 2015) giving rise to a plethora of distinct chromatin states. The combinatorial approach of multiple epigenetic data sets has particularly focussed on the study of poised promoters (Ernst and Kellis, 2010; Ernst et al., 2011), a set of regulatory regions that exhibited both activating (H3K4me3) and silencing (H3K27me3) histone marks and were previously described as bivalent promoters (Bernstein et al., 2006b; Mikkelsen et al., 2007). Many developmental factors showed bivalent promoter status in embryonic stem cells (ESCs) as well as in other non-pluripotent cell types (Mikkelsen et al. 2007; Mohn et al. 2008; Roh et al. 2006; Barski et al. 2007; Cui et al. 2009; Adli et al. 2010). Bivalency is thought to safeguard genes from terminal silencing (DNA methylation), therefore allowing ESCs to retain their plasticity (Williams et al. 2011; Voigt et al. 2013).

Bivalent chromatin has been closely associated with specific variants of phosphorylated RNA polymerase II (RNAPII) that is engaged at the promoter of the genes but not proceeding to productive elongation (Brookes and Pombo, 2009; Brookes et al., 2012b). This phenomenon, known as RNAPII promoter proximal pausing (Guenther et al. 2007; Muse et al. 2007; Zeitlinger et al. 2007; Krumm et al. 1995), was subsequently found at promoters of genes with a wide range of expression and biological function (Guenther et al., 2007; Min et al., 2011; Williams et al., 2015). Bivalent promoters in ESCs cultured in serum were divided in groups that featured distinct variants of phosphorylated RNAPII (S5P only, S5P and S2P) and variable levels of gene expression (Brookes et al., 2012b). However, follow up studies have shown that naïve ESCs exhibit increased levels of RNAPII at promoter sites (Marks et al., 2012), combined with extreme pausing at cell cycle genes and signalling factors rather than at developmental genes (Williams et al., 2015). These findings suggest that RNAPII pausing plays an important role in the expression fine-tuning of gene sets with diverse functional output, affecting the differentiation potential of the cell. While the implications of RNAPII pausing have been studied mainly in ESCs and *Drosophila* (Muse et al., 2007; Zeitlinger et al., 2007), its role in mammalian differentiation and cell lineage commitment is essentially unknown.

The specific combinations of bivalency marks with RNAPII and their respective biological functions are largely undetermined in ESCs as well as lineage committed cell types. Here we perform an integrative analysis of histone modification (H3K4me3, H3K27me3), RNAPII (8WG16) binding and expression profiling data, at the promoter regions of genes in eight different murine cell types. The promoter profiles across eight cell types clustered in nine major profile subgroups with distinct functional characteristics and divergent roles during mammalian development.

5.3 Methods

5.3.1 ChIP-seq data collection and processing

Murine ChIP-seq datasets for H3K4me3, H3K27me3 and RNAPII (8WG16) were collected in fastq format from Gene Expression Omnibus (GEO) (Barrett et al., 2013) for eight cell types (ESCs, PMNs, MEFs, BMDMs, DCs, B cell, MBs, MTs) (Table 5.1). Alignment of reads was done using Bowtie 2 using the mm10 reference genome and the default parameters (Langmead and Salzberg, 2012). SAM to BAM conversion of the aligned files was done using the SAMtools pipeline (Li et al., 2009). Total number of reads aligned to the genome is shown at Table 5.2. The bam files that belonged to the same experiment (technical replicates) were merged into a single bam file in order to proceed with the further analysis.

Name	Library Layout	Sample Name	Cell type	Experiment	Antibody
H3K27me3_1_ESC_GSE46134	SINGLE	GSM1124780	Embryonic Stem Cells	SRX266816	07-449
H3K4me3_1_ESC_GSE46134	SINGLE	GSM1124778	Embryonic Stem Cells	SRX266814	07-473
8WG16_ESC_GSE34518_all	SINGLE	GSM850469	Embryonic Stem Cells	SRX112178	8WG16(MMS-126R)
PMN_H3K27me3	SINGLE	GSM968714	Progenitor motor neurons	SRX160499	ab6002
PMN_H3K4me3	SINGLE	GSM968715	Progenitor motor neurons	SRX160500	ab8580
PMN_PolII_8WG16	SINGLE	GSM968717	Progenitor motor neurons	SRX160502	8WG16(MMS-126R)
MEF_H3K27me3_2	SINGLE	GSM1382349	Mouse Embryonic Fibroblasts	SRX535294	07-449
MEF_H3K4me3_all	SINGLE	GSM769029	Mouse Embryonic Fibroblasts	SRX085452	07-473
MEF_PolII_8WG16_all	SINGLE	GSM918761	Mouse Embryonic Fibroblasts	SRX143853	8WG16(MMS-126R)
BMDM_H3K27me3	SINGLE	GSM1314676	Bone marrow derived macrophages	SRX450325	07-449
BMDM_H3K4me3_all	SINGLE	GSM1000065	Bone marrow derived macrophages	SRX185786	07-473
BMDM_PolII_8WG16_all	SINGLE	GSM918720	Bone marrow derived macrophages	SRX143812	8WG16(MMS-126R)
DC_H3K27me3	SINGLE	GSM1384949	classical Dendritic Cells	SRX835927	NA
DC_H3K4me3	SINGLE	GSM1384941	classical Dendritic Cells	SRX835919	17-614
DC_Pol2	SINGLE	GSM1199835	mouse dendritic cells	SRX330713	NA
Bcell_H3K27me3	SINGLE	GSM1048209	CD43 negative splenic B cells	SRX208211	ab6002
Bcell_H3K4me3	SINGLE	GSM1509364	CD43 negative splenic B cells	SRX707681	ab8580
Bcell_PolII_8WG16	SINGLE	GSM1048213	CD43 negative mouse resting B cells	SRX208215	8WG16(MMS-126R)
MB_H3K27me3_all	SINGLE	GSM628007	Proliferating Myoblasts	SRX031886	17-622
MB_H3K4me3_all	SINGLE	GSM628005	Proliferating Myoblasts	SRX031884	07-473
MB_PolII_8WG16_all	SINGLE	GSM628011	Proliferating Myoblasts	SRX031890	8WG16(MMS-126R)
MT_H3K27me3_all	SINGLE	GSM628008	48-hours differentiated Myotubes	SRX031887	17-622
MT_H3K4me3_all	SINGLE	GSM628006	48-hours differentiated Myotubes	SRX031885	07-473
MT_PolII_8WG16_all	SINGLE	GSM628012	48-hours differentiated Myotubes	SRX031891	8WG16(MMS-126R)

Table 5.1 ChIP-seq samples, accession numbers, cell types, names given and antibodies used

Name	Total reads	Reads mapping to the genome	% reads mapping to the genome	Number of peaks
H3K27me3_1_ESC_GSE46134	29913075	13525738	45.22%	7038
H3K4me3_1_ESC_GSE46134	20167316	13437685	66.63%	20605
8WG16_ESC_GSE34518_all	32465308	23166679	71.36%	25120
PMN_H3K27me3	22964988	11374610	49.53%	4584
PMN_H3K4me3	21072673	9256509	43.93%	18867
PMN_PolII_8WG16	6843975	5244475	76.63%	19113
MEF_H3K27me3_2	33178460	32509163	97.98%	22472
MEF_H3K4me3_all	41470370	17110146	41.26%	20214
MEF_PolII_8WG16_all	71713121	51805888	72.24%	71969
BMDM_H3K27me3	13682117	12623558	92.26%	20310
BMDM_H3K4me3_all	26150869	24787557	94.79%	28651
BMDM_PolII_8WG16_all	61935086	55287733	89.27%	115667
DC_H3K27me3	205643964	201922110	98.19%	47861
DC_H3K4me3	196128930	193145895	98.48%	28563
DC_Pol2	15414031	15005191	97.35%	22754
Bcell_H3K27me3	76454312	73329993	95.91%	16002
Bcell_H3K4me3	31800729	30468857	95.81%	17971
Bcell_PolII_8WG16	47662490	36098123	75.74%	29673
MB_H3K27me3_all	23133748	21765658	94.09%	10774
MB_H3K4me3_all	34143979	33074892	96.87%	23699
MB_PolII_8WG16_all	48387677	46362196	95.81%	47118
MT_H3K27me3_all	21278916	19870534	93.38%	9755
MT_H3K4me3_all	30967336	29877440	96.48%	21639
MT_PolII_8WG16_all	51400978	49311337	95.93%	43711

Table 5.2 ChIP-seq data, number of reads, reads mapping back to the genome and number of peaks called

5.3.2 Peak calling

SICER (Xu et al., 2014) was used to detect peaks for the histone marks (H3K4me3 and H3K27me3). Input controls were not used for any of the samples in any of the cell types. Specific parameters were defined for H3K4me3, such as *window=200* and *gap size=200*. For H3K27me3 on the other hand, *window=200* and *gap size=2x300*, since this mark covers wider chromatin domains. The rest of the parameters (same for H3K4me3 and H3K27me3) were *effective genome fraction=0.7*, *redundancy threshold=1*, *fragment size=150* and *E-value=100*. MACS (Zhang et al., 2008) was used for the detection of peaks for the RNAPII samples, using the default parameters and no control. The total number of peaks detected for each sample are shown at Table 5.2.

5.3.3 RNA-seq data collection and processing

Murine RNA-seq datasets were collected from GEO (Barrett et al., 2013) in fastq format for all of the eight cell types mentioned previously (Table 5.3). Alignment was done with TopHat 2.0.9 (Trapnell et al., 2009) using mm10 as reference genome and the GENCODE.vM4 (Harrow et al., 2012) as annotation file. Expression values for each cell type were calculated following the Cufflinks 2.2.1 (Trapnell et al., 2010) pipeline. The aligned reads were converted to expression values using the *cuffquant* command with *library-type=fr-unstranded*. A bam file with FPKM values was created for each of the samples. Also, a file with gene expression values for all cell types was generated using the *cuffnorm* command with the default library normalization method (*geometric*), creating a matrix where each row was representing a gene and each column a cell type.

Name	LibraryLayout	Run	Sample Name	Cell type	Experiment
ES cells_run1	SINGLE	SRR391028	GSM850476	Undifferentiated ES cells (ES-OS25)	SRX112175
ES cells_run2	SINGLE	SRR391029	GSM850476	Undifferentiated ES cells (ES-OS25)	SRX112175
ES cells_run3	SINGLE	SRR391030	GSM850476	Undifferentiated ES cells (ES-OS25)	SRX112175
ES cells_run4	SINGLE	SRR391031	GSM850476	Undifferentiated ES cells (ES-OS25)	SRX112175
PMN_RNASeq	PAIRED	SRR1610573	GSM1524263	ESC-derived motor neuron progenitors	SRX731072
MEF_RNASeq_run1	SINGLE	SRR496251	GSM929719	Mouse Embryonic Fibroblast	SRX147592
MEF_RNASeq_run2	SINGLE	SRR496252	GSM929719	Mouse Embryonic Fibroblast	SRX147592
BMDM_RNASeq_run1	SINGLE	SRR496223	GSM929705	Bone marrow derived macrophage	SRX147578
BMDM_RNASeq_run2	SINGLE	SRR496224	GSM929705	Bone marrow derived macrophage	SRX147578
DC_RNASeq	SINGLE	SRR2040609	GSM1696234	immature DC	SRX1038966
Bcell_RNASeq_3	PAIRED	SRR628317	GSM1048203	CD43 negative mouse resting B cells	SRX208221
MB_RNASeq_run1	SINGLE	SRR074113	GSM628028	Proliferating Myoblasts	SRX032210
MB_RNASeq_run2	SINGLE	SRR074114	GSM628028	Proliferating Myoblasts	SRX032210
MT_RNASeq_run1	SINGLE	SRR074115	GSM628029	Confluent Myoblasts	SRX032211
MT_RNASeq_run2	SINGLE	SRR074116	GSM628029	Confluent Myoblasts	SRX032211

Table 5.3 RNA-seq samples, accession numbers, cell types and names given

5.3.4 Hierarchical trees for histone marks and gene expression

GENCODE.vM4 (Harrow et al., 2012a) was the chosen annotation for the creation of custom promoter regions (22,179 unique genes with gene length > 300 bp). The promoter BED file was created by taking the -5 kb, +5 kb area around the TSSs of GENCODE genes as promoter. Peak BED files were intersected with the custom promoter file using the *intersectBED* command from the BEDtools suite (Quinlan and Hall, 2010). The intersected peak-promoter files from all cell types were merged into one file, where each row was representing a gene promoter. In the columns, binary values of 0 or 1, would represent the absence or existence respectively, of a peak at that promoter for that cell type. For the hierarchical tree of gene expression, the output matrix from cuffnorm command was used. Hierarchical clustering was performed using the *hclust* function from the *fastcluster* package in R (Müllner, 2013). The *Euclidean* distance of the columns of each matrix (cell types) was used as a dissimilarity matrix and the method chosen was *complete*.

5.3.5 Clustering of gene promoters across cell types

We loaded the peak files in R for all the ChIP-seq datasets and converted them to GRanges objects. The bam files for each RNA-seq sample were also loaded in R creating custom coverage files with the *GRcoverageInbins* function (as object we used the promoter file (32,840 regions) converting it to GRanges, *Nnorm*=TRUE, *Snorm*=FALSE, *Nbins*=20) from the *compEpiTools* package (Kishore et al., 2015) in Bioconductor (Huber et al., 2015). For each of the cell types, we subsequently created a combined matrix of histone marks, RNAPII, RNA-seq coverage (normalized by library size in each cell type), CpG island regions (from UCSC) and gene annotation for sense and antisense transcripts (Gencode.vM4). Using the *heatmapData* function from *compEpiTools* (Kishore et al., 2015) we created a 140 column matrix (20 bins for each of the features) where the first 20 columns were representing the H3K27me3 peaks, columns 21-40 were H3K4me3 peaks, columns 41-60 were RNAPII peaks, columns 61-80 were RNA-seq coverage, columns 81-100 were CpG islands, columns 100-120 were sense transcript annotation and columns were 120-140 antisense transcript annotation. RNA-seq coverage was log2-scaled and transformed (values only in the (0,1) range) for each cell type separately.

We combined the matrices from all the cell types to acquire an initial super-matrix of 32,840*8=262,720 rows. Each row had a distinctive name of the ensembl gene id and the cell type it belonged to. We subsequently discarded the rows where more than 80% of the columns (only columns 1 to 80 were considered) had a zero value. This resulted in a matrix of 117,438 rows where each gene promoter was found in at least one cell type. This would essentially mean that we kept only the gene promoters that were presenting a signal in at least one of the histone marks, RNAPII or expression.

Hierarchical clustering of the gene promoter matrix was performed using the *hclust* function from the *fastcluster* package in R (Müllner, 2013). As a dissimilarity matrix we used the *Euclidean* distance of the rows of the matrix only for histone marks, RNAPII and expression values. The method used in *hclust* function was *complete*. After inspection of the initial clustering, through heatmap visualisation, we cut the resulting tree in groups using the *cutree* function from *stats* package in R (Team, 2016) and specifying *k*=60. The high number of groups specified facilitated the detection of groups that even though they had small number of genes, they presented a highly unique pattern of marks or expression.

We created a custom function in R to merge the clusters presenting highly similar patterns. The central function incorporated in our function was *clusterSim* (method=*"centroid"*) from the *flexclust* package in R (Leisch and Friedrich, 2006). *clusterSim* computed the pairwise distances between all centroids of the 60 groups and scaled them between (0,1). The similarity value was then given by subtracting the distance from 1. In our function we merged clusters whose similarity values were over the 99th quantile of the similarity values distribution for all pairwise comparisons. The newly merged clusters along with the ones that were not similar with any other cluster were renamed. Clusters having less than 100 genes were discarded.

The final clusters were visualised with the *heatmap.2* function from the *gplots* package in R (Gregory R. et al. 2016). The final clustered matrix contained 116,741 gene promoters which translated in 22,179 unique genes across all cell types.

5.3.6 Over and under representation of cell types in clusters

Significance of over or under representation of each cell type in each cluster was assessed using the hypergeometric test in R (*phyper*) from *stats* package. Since the cell types were not equally represented in the total population, we calculated the test using normalised values for the number of genes across clusters. The number of genes in any one cluster for any one cell type were divided by the total number of genes for that cell type and then multiplied by 10,000, resulting in a matrix where virtually the total number of genes would be 8 (cell types) *10,000=80,000.

5.3.7 Functional enrichment analysis

Gene Ontology (GO) enrichment analysis was done using the *topGO* package (Alexa and Rahnenführer, 2016) in Bioconductor (Huber et al., 2015) and the statistical test used to quantify the significance of the GO terms was *fisher's exact test*. GO term enrichment was done for the total number of genes in each cluster, but also separately for cell type specific genes in each cluster.

5.3.8 Maximum parsimony trees

After the clustering of gene promoters, we acquired lists of genes that belonged in each cluster, specified by their cell type. For each cluster gene list we created a matrix where the rows were the unique ensembl gene ids of that cluster and the columns were the names of the 8 cell types. If a gene was found in that cluster for a particular cell type a value of 1 was put at that specific cell, else the value was 0. The downstream analysis was conducted for each cluster separately and the binary matrices were the inputs for the next step where we used the *ape* package in R (Paradis et al., 2004). First we calculated the pairwise distances between the genes in the matrix using the *dist.gene* function. The neighbor joining tree estimation (Saitou and Nei, 1987; Studier and Keppler, 1988) was performed with the *nj* function using as input the result from the right previous step. Finally, the reconstruction of the most parsimonious ancestral states (Hanazawa et al., 1995; Narushima and Hanazawa, 1997) was done using the *MPR* function were we used as inputs the initial binary matrix, the resulting tree from the previous step, ESCs as *outgroup* and we kept only the lower values of the reconstructed sets for each ancestral node.

5.3.9 Average profiles and profile similarities

We calculated the average values for the columns of the matrix representing the 10kb region around the TSS, for the histone marks, RNAPII and RNA-seq of each cluster. This resulted in one row matrices containing, mean profiles for all the variables used for clustering of the promoters. To assess the general profile patterns shared between groups of clusters, we performed a hierarchical clustering for all the mean profile matrices. We used the *hclust* function from the *fastcluster* package in R (Müllner, 2013). The *Euclidean* distance of the rows was used as a dissimilarity matrix and the method chosen was *complete*.

5.3.10 Gene overlap between clusters and cell types

We calculated the overlap of genes between pairwise combinations of "Cell type - Cluster" sets of genes. For example, the overlap of genes between the genes that belonged in B cells in Cluster 1 with the genes of ESCs in Cluster 3.

Only movements of genes moving between clusters were taken into account and not movements within clusters. The significance of the overlap was assessed by the hypergeometric test in R (*phyper*) from *stats* package. As mentioned previously, the genes were not equally represented in each cell type, thus the normalized number of genes was used. Hence, in phyper, *q* was the number of genes overlapping, *m* was the number of genes in the "Cell type - Cluster" of origin, *n* was the total number of genes in the cell of origin minus *m* and *k* was the total number of genes in the "Cell type - Cluster" of arrival.

For subsequent analysis we kept only the gene movements within clusters for which *p-value* $<10^{-10}$ and the number of overlapping genes was larger than 50. To visualize the interactions between the "Cell type - Cluster" of origin and arrival, we used the *chordDiagram* function from the *circlize* package (Gu et al., 2014) in R. The size of the links is defined by the number of genes moving and the colour is a mix between the colours defining the clusters of interaction.

We also calculated the movement of genes across triplets of "Cell type - Cluster" sets of genes. We did not assess for the significance of the overlap but we kept only the interactions where more than 50 genes were overlapping.

5.3.11 Transcription factor enrichment

We downloaded data from 683 ChIP-seq experiments of TFs in multiple murine cell types from the CODEX database (Sánchez-Castillo et al., 2015). We calculated the overlap of the TF binding regions with the regions 1 kb around the TSS of each gene in each cluster. We used the function *countOverlaps* from the *GenomicRanges* package (Lawrence et al., 2013) in Bioconductor (Huber et al., 2015). To assess the significance of the overlaps we used the hypergeometric test in R (*phyper*) from *stats* package. Since TFs tend to bind more regularly at promoter regions, this would bias the p-value calculation. Hence, we narrowed down the total number of regions to only the regions that overlapped with a promoter.

5.3.12 Motif enrichment

Using the unique ensembl gene ids from the gene promoters in each cluster we used the *findMotifs.pl* command from the HOMER suite (Heinz et al., 2010) and searched for known and de-novo motifs at the 1kb areas flanking the TSSs. We systematically discarded the putative results of de-novo motifs that were represented in less than 15% of the regions and had a *p-value* > 10^{-10}. Similarly, known motifs were discarded when found in less than 15% of the regions and had a *p-value* > 10^{-5}.

5.3.13 RNAPII pausing index calculation

The formula we used to calculate the RNAPII pausing index (travelling ratio) is given by Muse et al. 2007 and is the following:

$$S = log_2\left(d\left(RNAPII_{promoter}\right)\right) - log_2(d(RNAPII_{genebody}))$$

which essentially is the ratio of RNAPII read density at the promoter area to the RNAPII read density in the gene body. *d* stands for the number of reads per nucleotide (nt) in the given region. The difference between the densities in log2 units equals to the ratio of fold enrichment in these regions, meaning a value of 1 would represent a 2-fold greater enrichment of RNAPII signal at the promoter region rather than in the gene body (Muse et al. 2007). We created two GRanges objects: 1) The promoter area ranging 600 bp around the TSS of the gene and 2) the gene body area ranging +600 from the TSS until the Transcription Ending Site (TES) of the gene. Using the *GRcoverage* function (as objects we used the previously mentioned promoter and gene body files (22,179 regions), *Nnorm*=FALSE, *Snorm*=TRUE) from the *compEpiTools* package (Kishore et al., 2015) in Bioconductor (Huber et al., 2015) we computed the read coverage at those regions for each cell type and gene in our clusters. This resulted in 2 column matrices (separately for each cell type) where the first column was the normalized (by region width in bp) read density at the promoters and the second column the normalized read density at the gene body. Finally, we calculated the ratio of the read densities using the pausing index formula.

Among other groups of genes, we calculated pausing indices for four more categories, namely the developmental, cell cycle, pro-pluripotency, pro-differentiation and ES signalling genes. To get a list of ensembl gene ids for the above groups of genes we used the same GO terms that were used by Williams et al. 2015 for developmental and cell cycle genes, whereas the genes involved in the ES pluripotency network were given by the Signaling pathways regulating pluripotency of stem cells - Mus musculus (mmu04550 entry) from KEGG pathways database (Kanehisa et al., 2016). For developmental genes the GO terms were : GO:0045165, GO:0048864, GO:0007498 and for cell cycle genes were: GO:0007049. We narrowed down the selection of genes in the ES pluripotency network to pro-pluripotency (stem cell population maintenance – GO:0019827, negative regulation of cell differentiation – GO:0045596), pro-differentiation (positive regulation of cell differentiation – GO:0045597) and ES-signalling genes (cytokine activity – GO:0005125, regulation of MAPK cascade – GO:0043408).

5.4 Results

5.4.1 H3K27me3, H3K4me3, RNAPII and expression at promoters in eight murine cell types

To study the dynamics of epigenetic and transcription control at promoters during development, we collected ChIP-sequencing data for two chromatin modifications (activating – H3K4me3 and silencing – H3K27me3) and RNA polymerase II (8WG16) as well as expression data (RNA sequencing) across murine embryonic stem cells (ESCs), progenitor motor neurons (PMNs), embryonic fibroblasts (MEFs), bone marrow derived macrophages (BMDMs), dendritic cells (DCs), B cells, myoblasts (MBs) and myotubes (MTs) (see Methods). The gene expression quantified at 22,179 (see Methods) GENCODE.vM4 (Harrow et al., 2012b) promoters using RNA sequencing (RNA-seq) reflects low variation of expressed promoters (FPKM>1, Figure 5.1) across cell types (11178±396) with PMNs showing the highest number of expressed promoters (11,604) and BMDMs the lowest (10,582). The hierarchical clustering of cell types using expression data matched the known developmental relationships across cell types (Figure 5.2A) with the three hematopoietic cell types (B cells, BMDMs and DCs) clustering together and the two progenitor cell types (PMNs and ESCs) forming a separate cluster. BMDMs and DCs clustered tightly together despite large technical variation between samples (Figure 5.2B). MEFs clustered with MBs and MTs, albeit close to the progenitor cells.

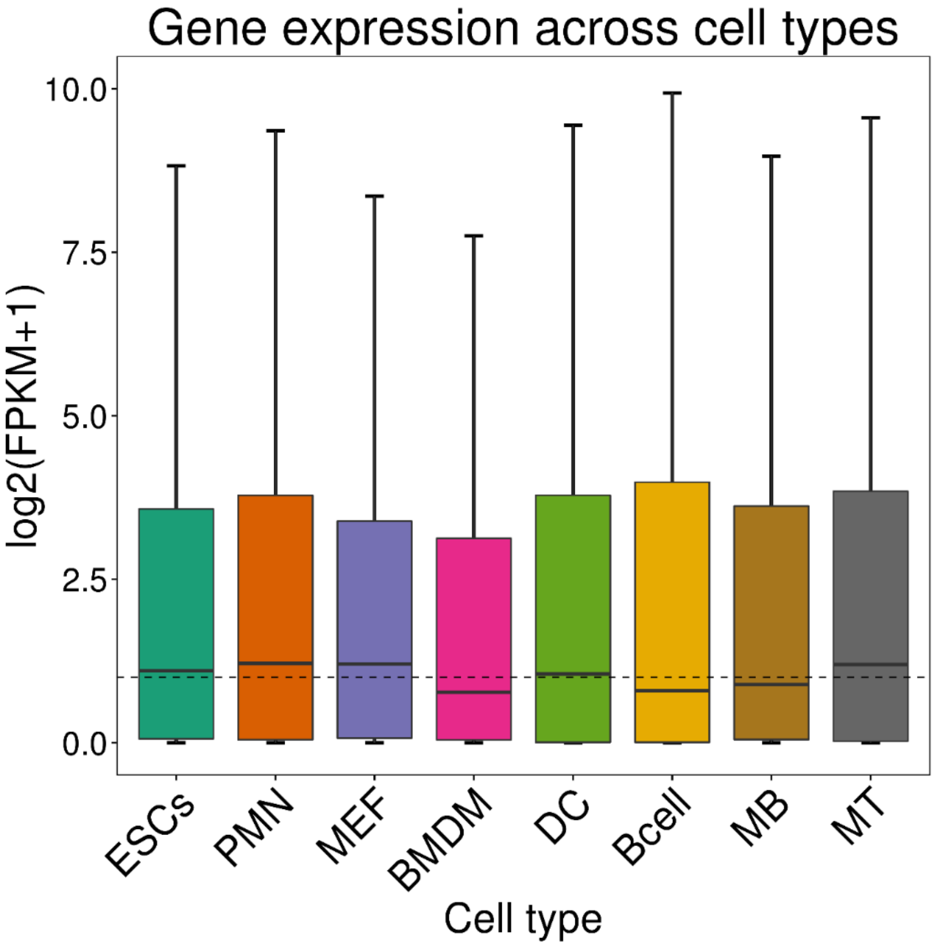


Figure 5.1 Gene expression distribution for all the genes used in the study (22,179) across cell types. The dashed line at log2 (FPKM+1) = 1 **represents the threshold imposed to classify genes as expressed and not-expressed.**

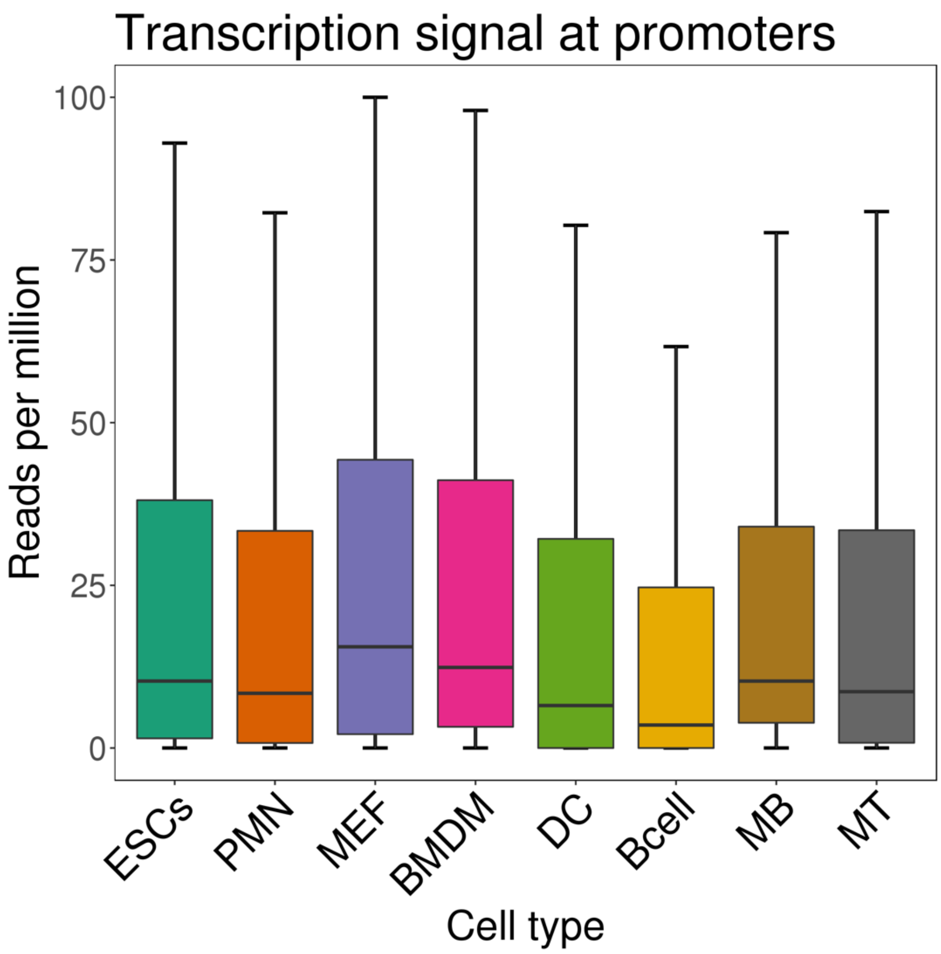


Figure 5.2 Distribution of the RNA-seq reads (reads per million) at the promoter regions (-5KB, +5KB from the TSS) shows comparable levels across all cell types.

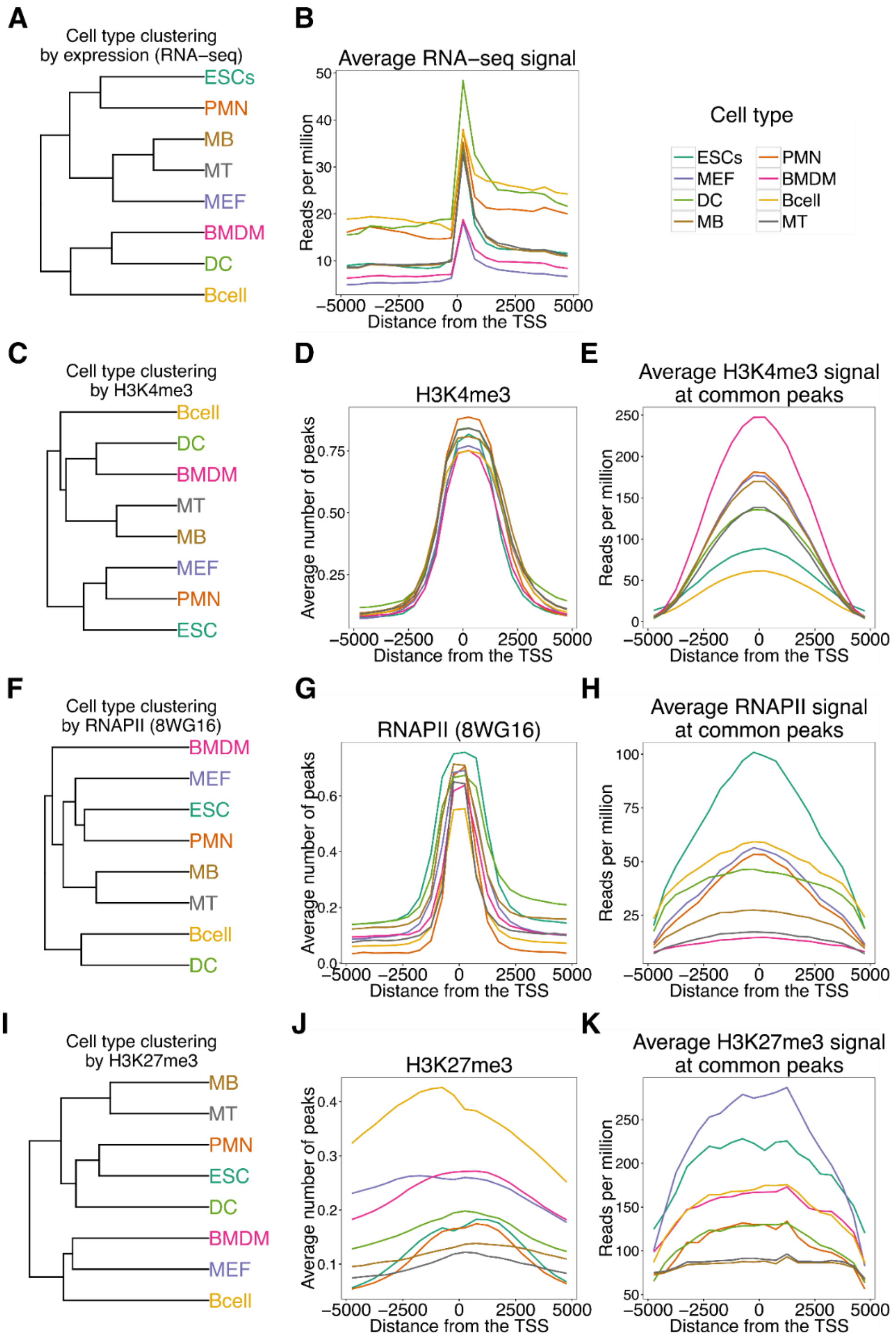


Figure 5.3 Expression, H3K4me3, H3K27me3 and RNAPII (8WG16) signatures at promoters of 22,179 genes in eight murine cell types. A) Hierarchical clustering of normalized expression values (see Methods) across eight cell types results in a tree where relationships between cell types are largely reconstituted. B) Average normalized RNA-seq signal (reads per million -RPM) across the gene promoters (±5kb) per cell type displays signal variability across cell types. C) Hierarchical clustering of H3K4me3

marked promoters across all cell types is positively correlated with the expression data and results in a tree in agreement with the known developmental relationships between cell types. D) The average number of H3K4me3 detected peaks at the promoters is highly consistent across all the cell types. E) The average H3K4me3 signal at common peaks across all cell types is highly variable, with BMDMs showing the strongest signal. F) Hierarchical clustering of RNAPII (8WG16) binding is closely correlated with the H3K4me3 tree, rather than the expression tree. G) The average number of RNAPII peaks at the promoters is consistent across cell types, however less than in H3K4me3 marked promoters. H) The average RNAPII signal at common peaks at the promoters is highly variable with ESCs displaying the strongest signal. I) Hierarchical clustering of H3K27me3 marked promoters across all cell types, results in a tree where only the relationships of MBs and MTs are reconstituted. J) The average number of H3K27me3 peaks at the promoters is variable across the cell types, with B cells showing the largest number of detected peaks in all cell types. K) The average H3K27me3 signal at common peaks is highly variable across cell types with MEFS showing the strongest signal.

We then assessed the presence or absence of H3K4me3, H3K27me3 modifications and RNA polymerase II (RNAPII) binding at the promoters across eight cell types by peak calling in each sample using SICER (Xu et al., 2014) and focussed on the peaks at the 10 Kilobase (KB) region flanking the Transcription Start Site (TSS) of the 22,179 genes. Though the variation in the number of H3K4me3 marked promoters across cell types (15686 ±804) was larger compared to the expression data, H3K4me3 modification was consistent with gene expression (Pearson’s correlation coefficient: 0.34, Table 5.4) at the promoters across 8 cell types.

Accordingly, the hierarchical tree of H3K4me3 peaks at promoters across 8 cell types (Figure 5.2C) was largely in agreement with the one obtained using expression data (Figure 5.2A). Interestingly, H3K4me3 profiles closely associated PMNs with MEFs rather than ESCs. We confirmed this was not due to technical issues such as over detection of peaks in one sample rather than the other (Figure 5.2D). BMDMs and B cells clustered together despite a high signal variability at common H3K4me3 peaks found across all cell types (Figure 5.2E). To study the dynamics of H3K4me3 modification between cell types, we used a maximum parsimony based approach (see Methods). Maximum parsimony approach predicts the chromatin modification status at each intermediate node of a tree by allowing minimum number of epigenetic changes within the tree (Hanazawa et al., 1995; Narushima and Hanazawa, 1997). Over 80% of promoters (15690 out of 19022 promoters with H3K4me3 mark across all 8 cell types) retained H3K4me3 modification across cell types (Figure 5.4).

Signature	ESCs	PMN	MEF	BMDM	DC	B cell	MB	MT
Expressed	11372	11604	11575	10582	11227	10723	10894	11451
Not expressed	10807	10575	10604	11597	10952	11456	11285	10728
H3K4me3 marked	15980	16207	16231	16295	16393	14184	15260	14937
Not H3K4me3 marked	6199	5972	5948	5884	5786	7995	6919	7242
H3K4me3 only marked	11904	13077	10455	11467	12451	8047	13225	12895
Not H3K4me3 only marked	10275	9102	11724	10712	9728	14132	8954	9284
H3K27me3 marked	4578	3359	7918	7107	5510	9793	3573	3342
Not H3K27me3 marked	17601	18820	14261	15072	16669	12386	18606	18837
H3K27me3 only marked	502	229	2142	2279	1568	3656	1538	1300
Not H3K27me3 only marked	21677	21950	20037	19900	20611	18523	20641	20879
Bivalent	4076	3130	5776	4828	3942	6137	2035	2042
Not bivalent	18103	19049	16403	17351	18237	16042	20144	20137
RNAPII bound	14296	12546	15953	16512	10503	11235	13255	11625
Not RNAPII bound	7883	9633	6226	5667	11676	10944	8924	10554

Table 5.4 Classification of 22,179 gene promoters in each cell type according to 1) expression levels (Expressed when log2 (FPKM+1) >1), 2) H3K4me3 marks, 3) H3K4me3 only – H3K4me3 marks that do not overlap with H3K27me3 marks, 4) H3K27me3 marks, 5) H3K27me3 only – H3K27me3 marks that do not overlap with H3K4me3 marks, 6) Bivalent marks – H3K4me3 and H3K27me3 peaks overlapping at the region, 7) RNAPII bound promoters

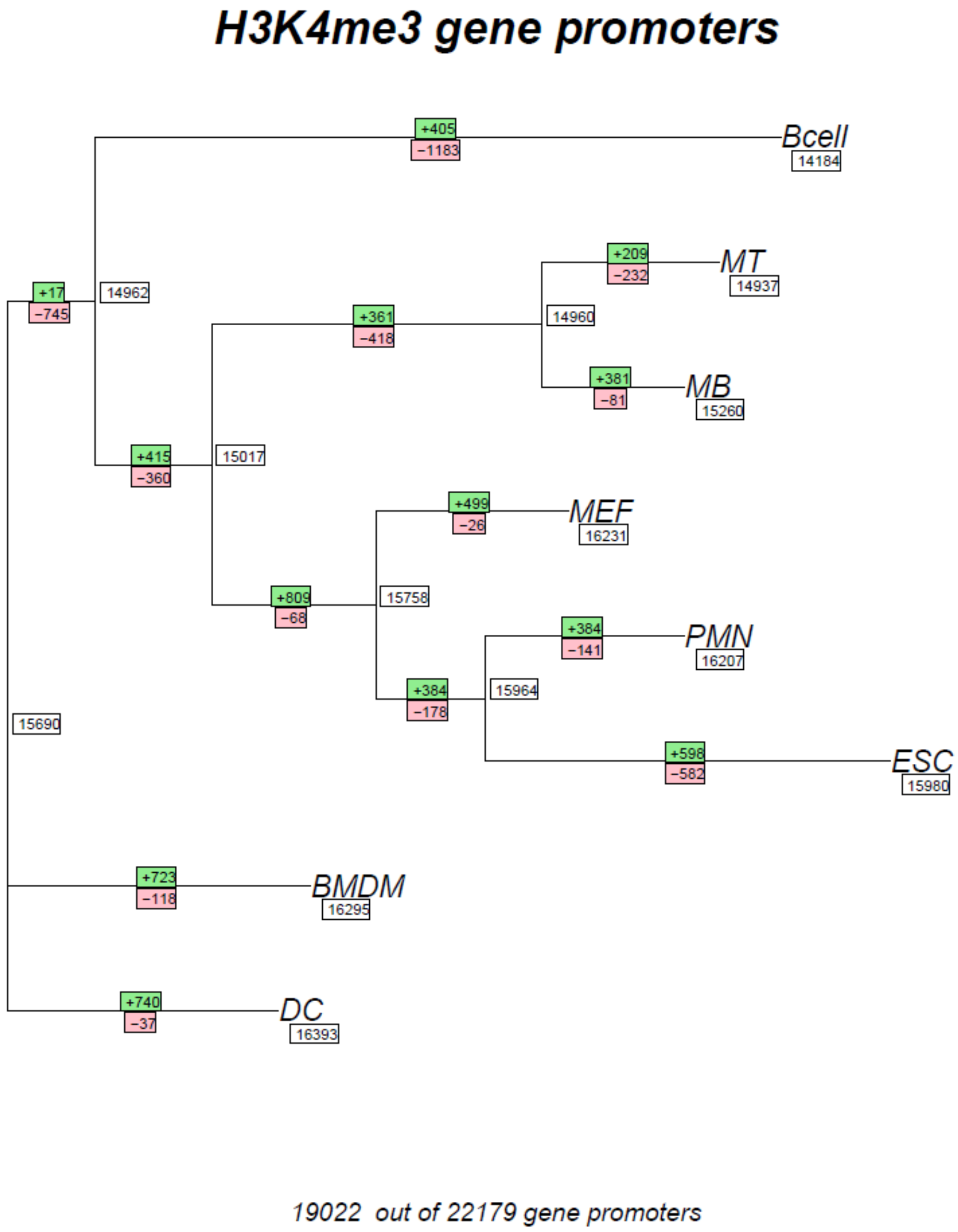


Figure 5.4 H3K4me3 dynamics at the promoter regions (-5Kb, +5Kb) across the cell types. More than 80% promoters (15,690 shared out of 19,022 promoters with H3K4me3 mark across all 8 cell types) retain H3K4me3 modification across cell types. This is a maximum parsimony tree representing the minimum number of changes necessary (gain of genes with a H3K4me3 peak shown in green, loss of genes without a H3K4me3 peak shown in red) for the reconstruction of the tree of relationships between the cell types.

The hierarchical clustering of RNAPII (8WG16) modification at the promoters was not as consistent with the expression (Pearson's correlation coefficient: -0.10) as the H3K4me3 tree (Figure 5.2F). Notably, the number of RNAPII marked promoters varied highly across cell types (13240±2197) where BMDMs had over 16,000 RNAPII occupied promoters while DCs had only about 10,000 RNAPII occupied promoters (Figure 5.5). Accordingly, DCs and B cells clustered together in the RNAPII hierarchical tree (Figure 5.2F). The parsimony tree using RNAPII peaks at promoters showed that RNAPII peaks were shared to a lesser extent (about 66%, 13,192 of 19758 promoters with RNAPII peaks across all 8 cell types) than H3K4me3 between cell types (Figure 5.5). Similarly, we tested the technical variability between samples by calculating average RNAPII peak strength in each cell type across common peaks (Figure 5.2H). The RNAPII peak strength showed low correlation with the number of RNAPII peaks (Figure 5.2G).

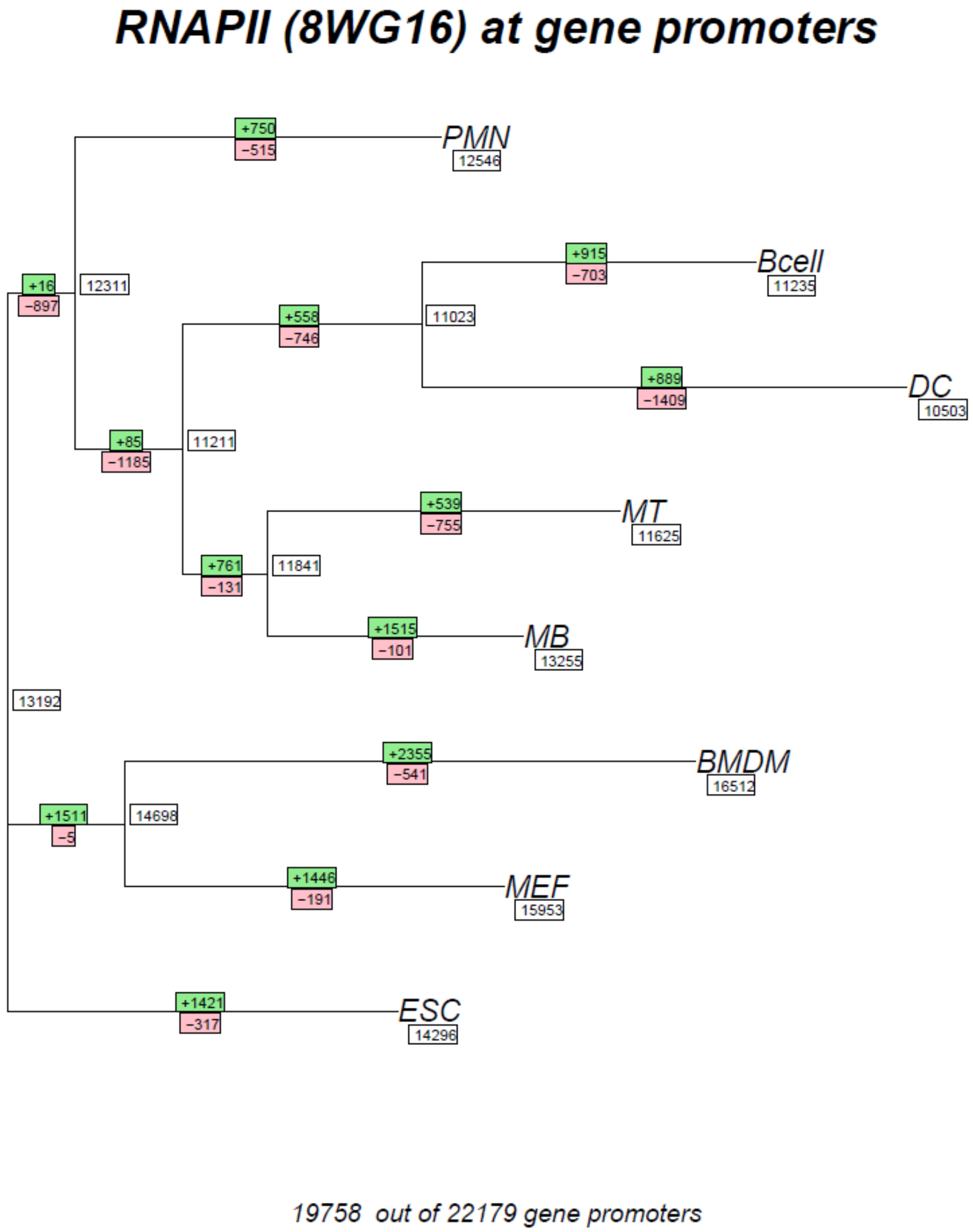


Figure 5.5 RNAPII (8WG16) dynamics at the promoter regions (-5Kb, +5Kb) across the cell types. More than 66% promoters (13,192 shared out of 19,758 promoters with RNAPII binding across all 8 cell types) retain RNAPII binding across cell types. This is a maximum parsimony tree representing the minimum number of changes necessary (gain of genes with a RNAPII peak shown in green, loss of genes without a RNAPII peak shown in red) for the reconstruction of the tree of relationships between the cell types.

The number of H3K27me3 marked promoters were highly dynamic across cell types (5647±2405) with B cells, BMDMs and MEFs marked with H3K27me3 at a large number of promoters (7,000 to 10,000). PMNs, MBs and MTs on the other hand, were the cell types with the smallest number of H3K27me3 marked promoters (about 3,000) (Table 5.4). The hierarchical tree of H3K27me3 promoter peaks across cell types also did not agree with expression data where MEFs clustered with B cells and BMDMs, while DCs clustered with progenitor type cells (Figure 5.2I). We verified that this variability is not solely due to technical reasons by calculating average number of detected peaks across all promoters (Figure 5.2J) and average H3K27me3 signal at common peaks in each cell type (Figure 5.2K). In the H3K27me3 parsimony tree, only about 16% (2,523 of 15,005 H3K27me3 marked promoters) of H3K27me3 promoters were shared across all cell types (Figure 5.6).

Bivalent promoters are defined as those marked with both repressing (H3K27me3) and activating (H3K4me3) modifications and are enriched for developmental regulators (Bernstein et al., 2006b; Mikkelsen et al., 2007). We sub-classified H3K27me3 promoters into bivalent and H3K27me3-only promoters depending on presence or absence of H3K4me3 modifications at the same promoter in the same cell type. Over 90% of H3K27me3 promoters in ESCs and PMNs were bivalent while only about 60% of H3K27me3 promoters in B cells were bivalent (Table 5.4). Interestingly, around 21% of bivalent promoters (2,598 out 11,919 bivalent promoters in all 8 cell types) were shared across cell types (Figure 5.7) and were enriched for pattern specification process (Fisher's exact test, P-value < 10^{-8}) and developmental protein (Fisher's exact test, P-value < 10^{-15}). Complementarily, ESCs and PMNs had the lowest (6-10%) and MBs, MTs and B cells had the highest (37-43%) numbers of H3K27me3-only promoters (Table 5.4). MBs and MTs clustered tightly together for all epigenetic modifications.

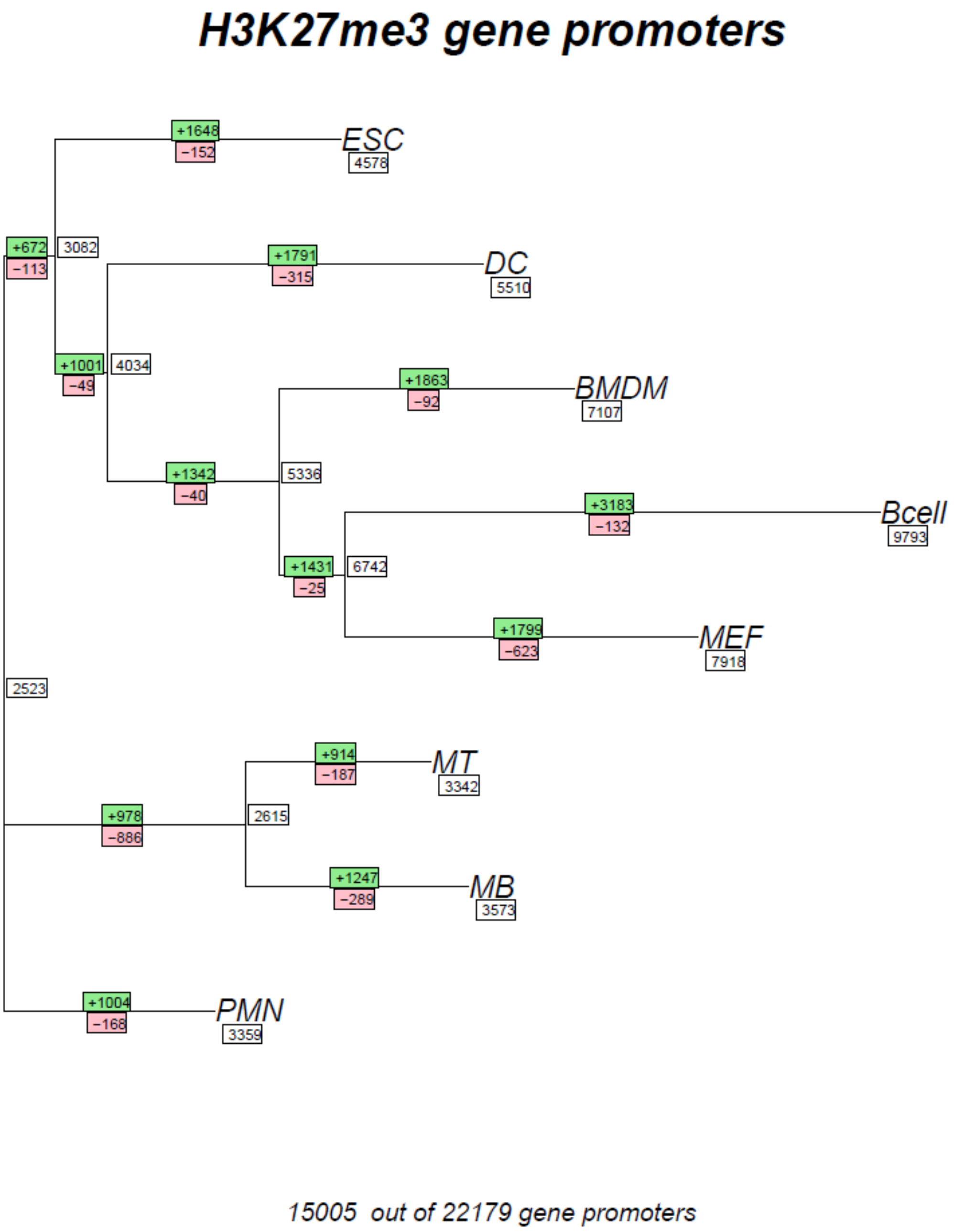


Figure 5.6 H3K27me3 dynamics at the promoter regions (-5Kb, +5Kb) across the cell types. Only around 17% promoters (2,523 shared out of 3,359 promoters with H3K27me3 mark across all 8 cell types) retain H3K27me3 modification across cell types. This is a maximum parsimony tree representing the minimum number of changes necessary (gain of genes with a H3K27me3 peak shown in green, loss of genes without a H3K27me3 peak shown in red) for the reconstruction of the tree of relationships between the cell types.

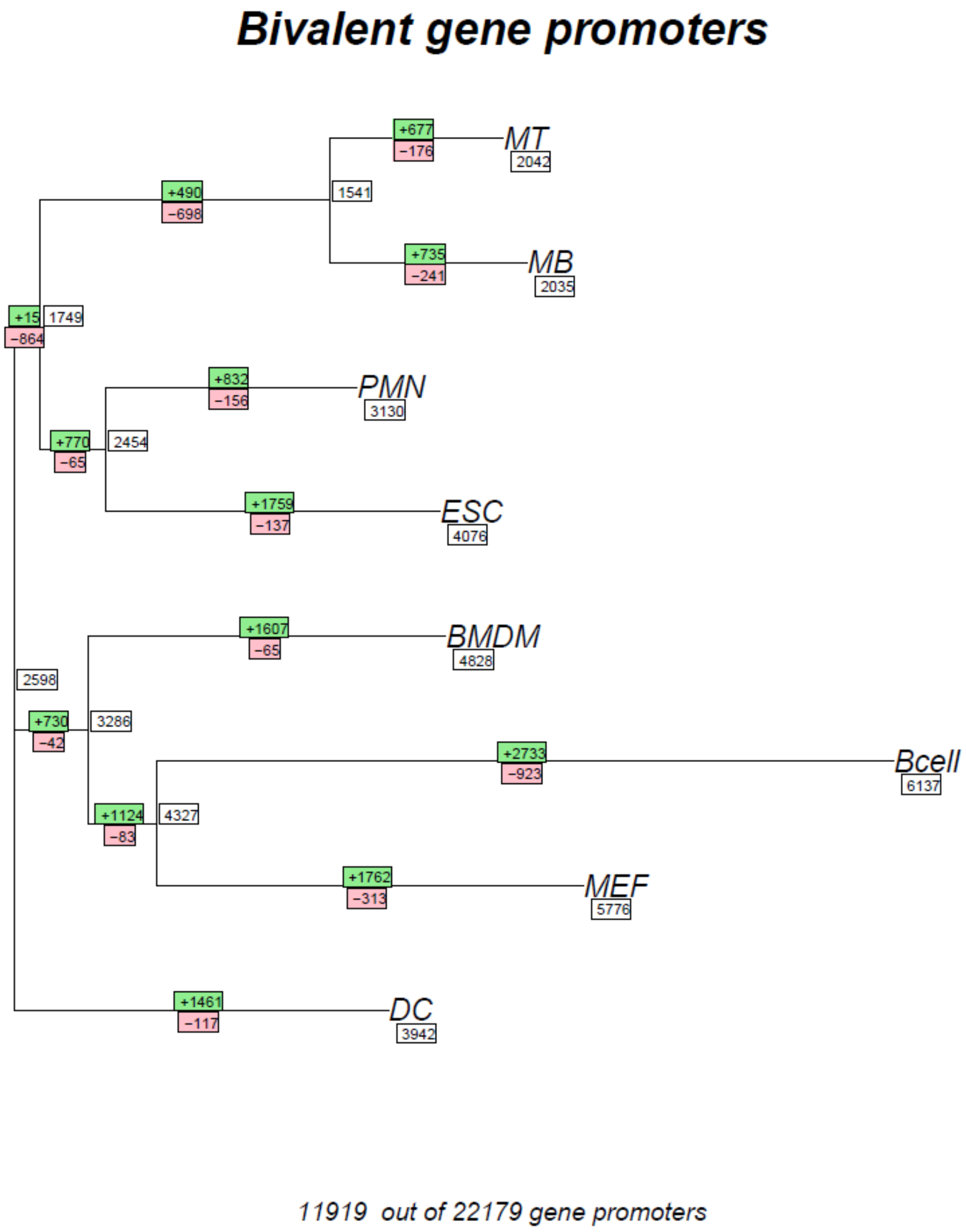


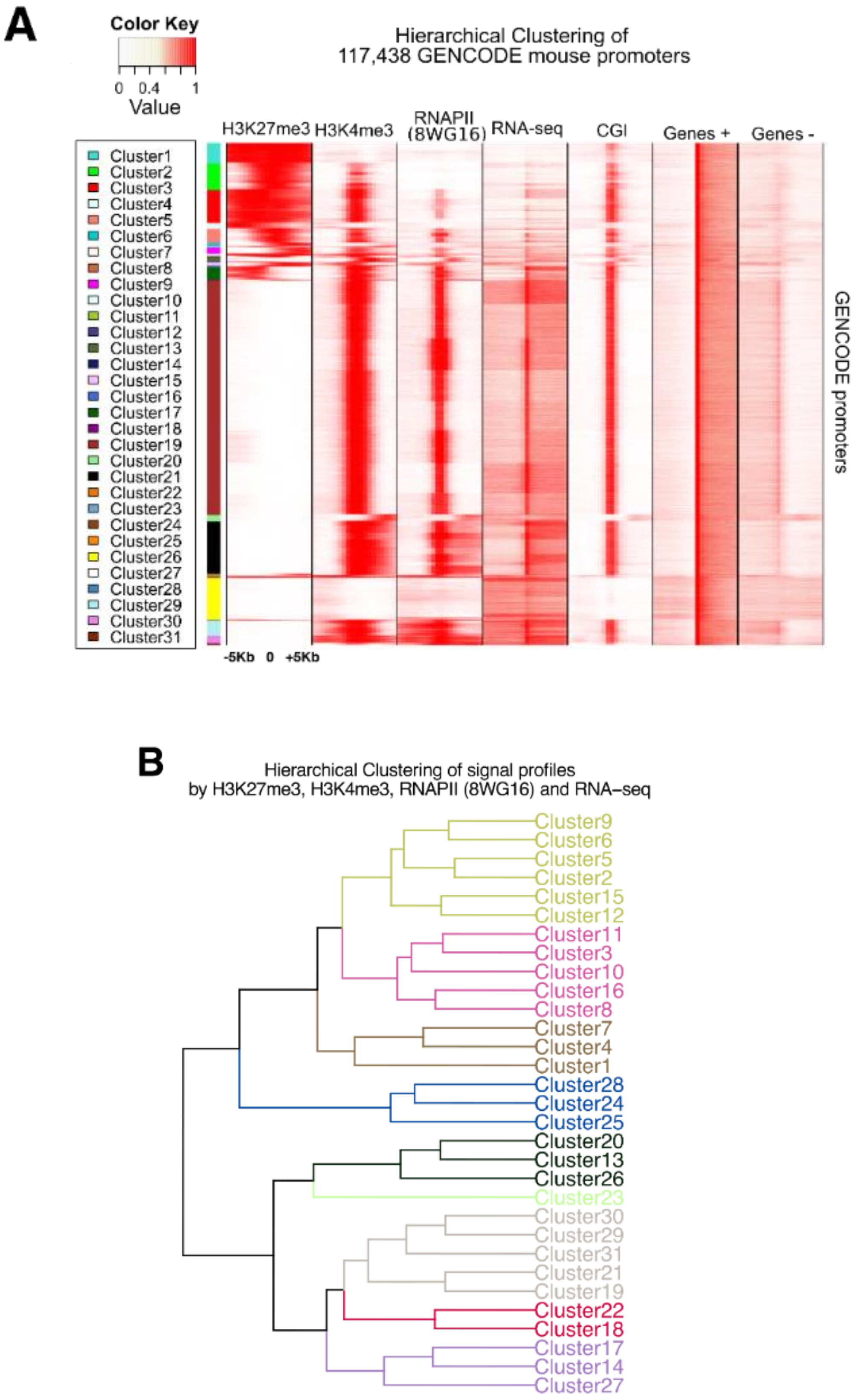
Figure 5.7 Bivalency dynamics at the promoter regions (-5Kb, +5Kb) across the cell types. Only around 22% promoters (2,598 shared out of 11,919 promoters with H3K27me3 mark across all 8 cell types) retain H3K27me3 modification across cell types. This is a maximum parsimony tree representing the minimum number of changes necessary (gain of genes with a bivalent peak shown in green, loss of genes without a bivalent peak shown in red) for the reconstruction of the tree of relationships between the cell types.

Taken together, the cell type relationships established using RNAPII binding and H3K27me3 modifications at promoters did not agree with those using expression data across eight cell types.

5.4.2 Nine major epigenetic and expression profiles at promoters across cell types

In the previous section, we noted that the chromatin modifications at promoters across cell types does not fully agree with expression dynamics. To further systematically analyse the patterns of chromatin and expression dynamics at promoters across cell types, we clustered H3K4me3, H3K27me3 and RNAPII peaks as well as RNA-seq signal at 22,179 GENCODE.vM4 gene promoters in eight cell types. Promoters occupied by RNAPII can either be active or paused depending upon whether or not the RNAPII signal is more enriched at the core promoter than in the gene body (Brookes and Pombo, 2009; Brookes et al., 2012b). To capture such relevant features of chromatin modifications, we defined a wide window (±5kb) around the TSS to characterise each modification at a promoter level in a given cell type (see Methods).

We obtained epigenetic and transcription profiles (4 data types) at each promoter in a given cell type for a total of 117,438 promoter-cell types (each gene promoter was found in at least one cell type, see Methods). We clustered promoter-cell types by hierarchical clustering using the Euclidean distance as a distance measure (see Methods), resulting in 31 clusters with distinct patterns across four data types (Figure 5.8A). The number of promoter-cell types in each cluster varied largely across clusters. Cluster 19 consisted over 54,000 promoter-cell types while cluster 8 consisted of only 105 promoter-cell types (Table 5.5). As expected H3K4me3 and RNAPII modifications largely overlapped with expressed promoters; most of which belonged to cluster 19 (Figure 5.8A).


A
Color Key
0 0.4 1
Value
Hierarchical Clustering of
117,438 GENCODE mouse promoters
H3K27me3
H3K4me3
RNAPII (8WG16)
RNA-seq
CGI
Genes +
Genes -
Cluster1
Cluster2
Cluster3
Cluster4
Cluster5
Cluster6
Cluster7
Cluster8
Cluster9
Cluster10
Cluster11
Cluster12
Cluster13
Cluster14
Cluster15
Cluster16
Cluster17
Cluster18
Cluster19
Cluster20
Cluster21
Cluster22
Cluster23
Cluster24
Cluster25
Cluster26
Cluster27
Cluster28
Cluster29
Cluster30
Cluster31
-5Kb 0 +5Kb
GENCODE promoters
B
Hierarchical Clustering of signal profiles
by H3K27me3, H3K4me3, RNAPII (8WG16) and RNA–seq
Cluster9
Cluster6
Cluster5
Cluster2
Cluster15
Cluster12
Cluster11
Cluster3
Cluster10
Cluster16
Cluster8
Cluster7
Cluster4
Cluster1
Cluster28
Cluster24
Cluster25
Cluster20
Cluster13
Cluster26
Cluster23
Cluster30
Cluster29
Cluster31
Cluster21
Cluster19
Cluster22
Cluster18
Cluster17
Cluster14
Cluster27

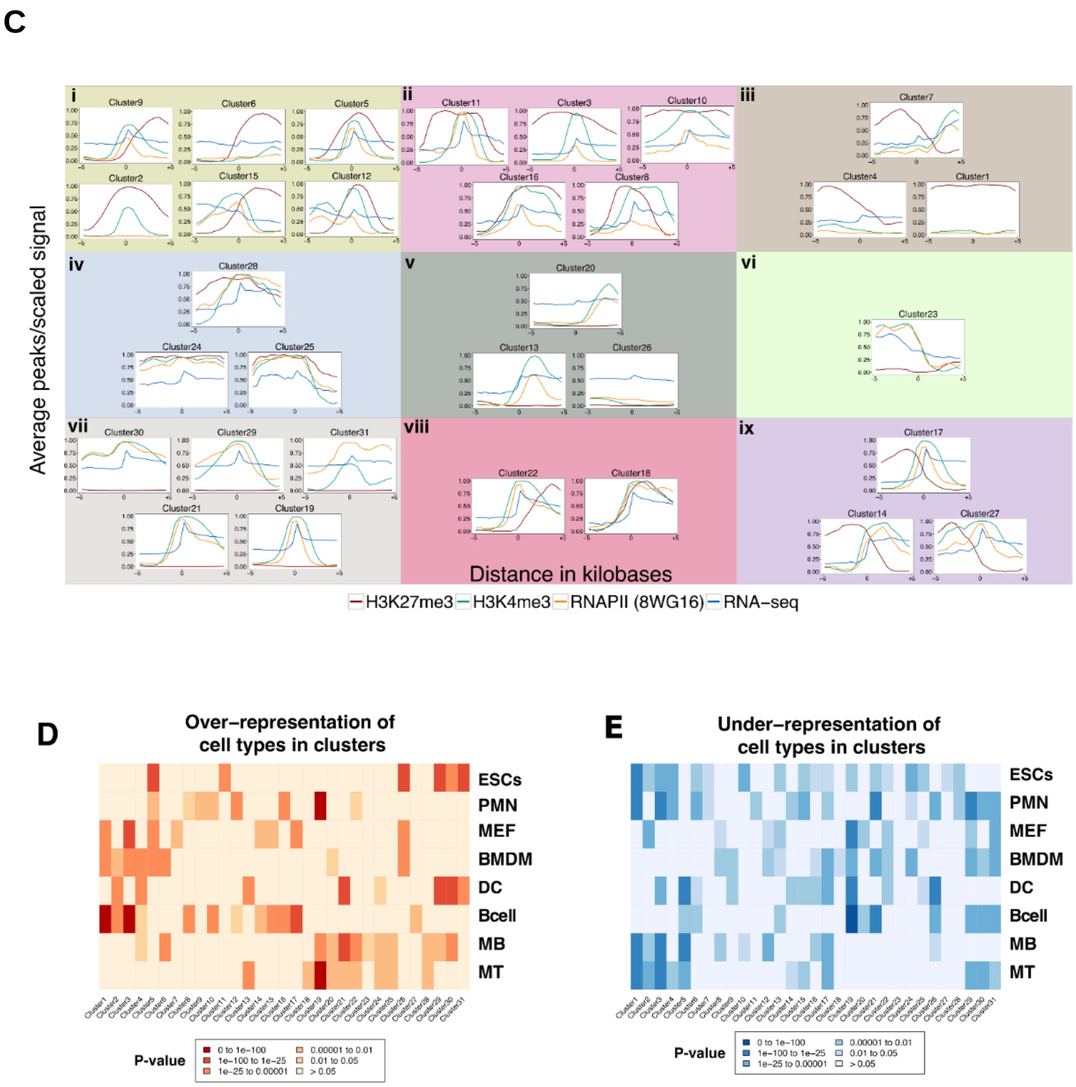


Figure 5.8 Identification of nine major epigenetic and expression profiles, comprised of 31 distinct clusters. A) Hierarchical clustering of the combined profiles of H3K27me3 (peaks), H3K4me3 (peaks), RNAPII (peaks) and expression signal (reads per million) across 117,438 distinct gene promoters-cell type. 31 clusters of distinct signatures were detected. B) Hierarchical clustering of the average profile signals across clusters results in the identification of 9 major profile sub-groups. C) Average number of peaks/Average RNA-seq signal at all 31 clusters, which belong in the following sub-groups: i) bivalent-narrow-H3K27me3, ii) bivalent-wide-H3K27me3, iii) H3K27me3-only, iv) bivalent-wide-active, v) wide-active, vi) antisense-active, vii) highly-active, viii) bivalent-highly-active, ix) boundary-H3K27me3-active. D) Enrichment of cell types in each cluster (significance was assessed with hypergeometric test) E) Under-enrichment of cell types per cluster (significance was assessed with hypergeometric test).

Cluster	Total number of genes	Unique genes	Ratio Unique/total number of gene promoters
1	4543	3124	0.6876513
2	6185	4328	0.6997575
3	**7679**	**4564**	**0.5943482**
4	1400	1140	0.8142857
5	3275	2649	0.808855
6	823	679	0.8250304
7	158	118	0.7468354
8	105	100	0.952381
9	1464	1236	0.8442623
10	486	393	0.808642
11	137	134	0.9781022
12	289	254	0.8788927
13	924	615	0.6655844
14	196	166	0.8469388
15	687	514	0.7481805
16	304	262	0.8618421
17	3002	2396	0.7981346
18	212	161	0.759434
19	**54580**	**13185**	**0.241572**
20	**1520**	**663**	**0.4361842**
21	**12250**	**5854**	**0.4778776**
22	361	309	0.8559557
23	166	109	0.6566265
24	**285**	**149**	**0.522807**
25	115	72	0.626087
26	**9746**	**3750**	**0.3847732**
27	124	103	0.8306452
28	118	97	0.8220339
29	3428	2212	0.6452742
30	1705	1047	0.6140762
31	474	406	0.8565401

Table 5.5 Total clusters detected (31) ordered by how they are displayed in the heatmap in Figure 2A. Total number of gene promoters-cell type in each cluster, unique number of gene promoters in each cluster independently of cell type information, and ratio of unique to total number of genes. The clusters in black background displayed a ratio lower than 0.6, indicating clusters with gene promoter signatures conserved across cell types.

To facilitate biological interpretation of the 31 clusters, we further grouped them into 9 major sub-groups namely: i) bivalent-narrow-H3K27me3, ii) bivalent-wide-H3K27me3, iii) H3K27me3-only, iv) bivalent-wide-active, v) wide-active, vi) antisense-active, vii) highly-active, viii) bivalent-highly-active, ix) boundary-H3K27me3-active, based on the patterns of the four data types across promoters (Figure 5.8B and 5.8C). For example, boundary-H3K27me3-active cluster 17 (Figure 5.8C (ix)) showed H3K27me3 modification upstream of TSS and H3K4me3, RNAPII signal and weak transcription at the TSS and downstream. Bivalent clusters marked simultaneously with H3K27me3 and H3K4me3 modifications were divided into four groups. They were grouped as bivalent-narrow-H3K27me3 (Figure 5.8C (i)) and bivalent-wide-H3K27me3 (Figure 5.8C (ii)) based on the H3K27me3 pattern at the promoter and were grouped into bivalent-wide-active (Figure 5.8C (iv)) and bivalent-highly-active (Figure 5.8C (viii)) according to the RNA-seq signal at the promoter. Bivalent-wide-H3K27me3 cluster 10 was enriched for ‘pattern specification process’ (Fisher’s exact test , P-value < 10^{-30}) and ‘Embryonic morphogenesis’ (Fisher’s exact test, P-value < 10^{-30}) and cluster 3 was enriched for ‘nervous system development’ (Fisher’s exact test , P-value < 10^{-30}). On the other hand, bivalent-narrow-H3K27me3 cluster 2 was highly enriched for ‘cell-cell signalling’ (Fisher’s exact test , P-value < 10^{-30}) and cluster 5 was enriched for genes involved in ‘cell development’ (Fisher’s exact test , P-value < 10^{-18}).

We have previously noted that H3K27me3-only promoters lacked CpG islands in mouse ESCs (Mantsoki et al., 2015). We therefore calculated CpG density at promoters in all 9 major groups and clusters. The H3K27me3-only clusters were indeed mostly CpG poor (only ~25% to 50% overlapped with a CpG island) (Figure 5.9C) and were among the clusters with the lowest mean CpG density (group mean CpG density=0.39, see Figure 5.9A and 5.9B). Active and bivalent clusters were enriched for CpG islands (Figure 5.9C) across all cell types studied. The subgroups with mean CpG densities higher than 0.70 were the highly-active (vii) and Boundary-H3K27me3-active (ix) (Figure 5.9A).

We then investigated if particular cell types were over or under-represented in the clusters by hypergeometric testing after correcting for cell type specific differences (see Methods). In over half of the clusters all cell types were equally represented

(Figure 5.8D). ESCs were underrepresented while B cells were over-represented in 'H3K27me3-only' clusters (Figure 5.8D and 5.7E). Bivalent-wide-H3K27me3 cluster 3 had MEFs, B cells and BMDMs over-represented, whilst the rest of the cell types were under-represented. Surprisingly, ESCs were over-represented only in bivalent cluster 5 (bivalent-narrow-H3K27me3) and cluster 11 (bivalent-wide-H3K27me3) (Figure 5.8D and 5.8E). PMNs, MEFs and BMDMs were also represented in bivalent cluster 5, whereas DCs, B cells, MBs and MTs were under-represented.

To identify clusters with promoters shared across cell types, we calculated the ratio of unique number of promoters to total number of promoters in each cluster (Table 5.5). This ratio was the lowest for cluster 19 (0.24, Table 5.5) with expressed genes enriched for 'cellular macromolecule catabolic process' (Fisher's exact test , P-value < 10^{-30}). To understand in greater detail, the gene gains and losses in each cluster across cell types, we reconstructed the maximum parsimony trees for each cluster (see Methods, Figures 5.9.1-5.9.31).

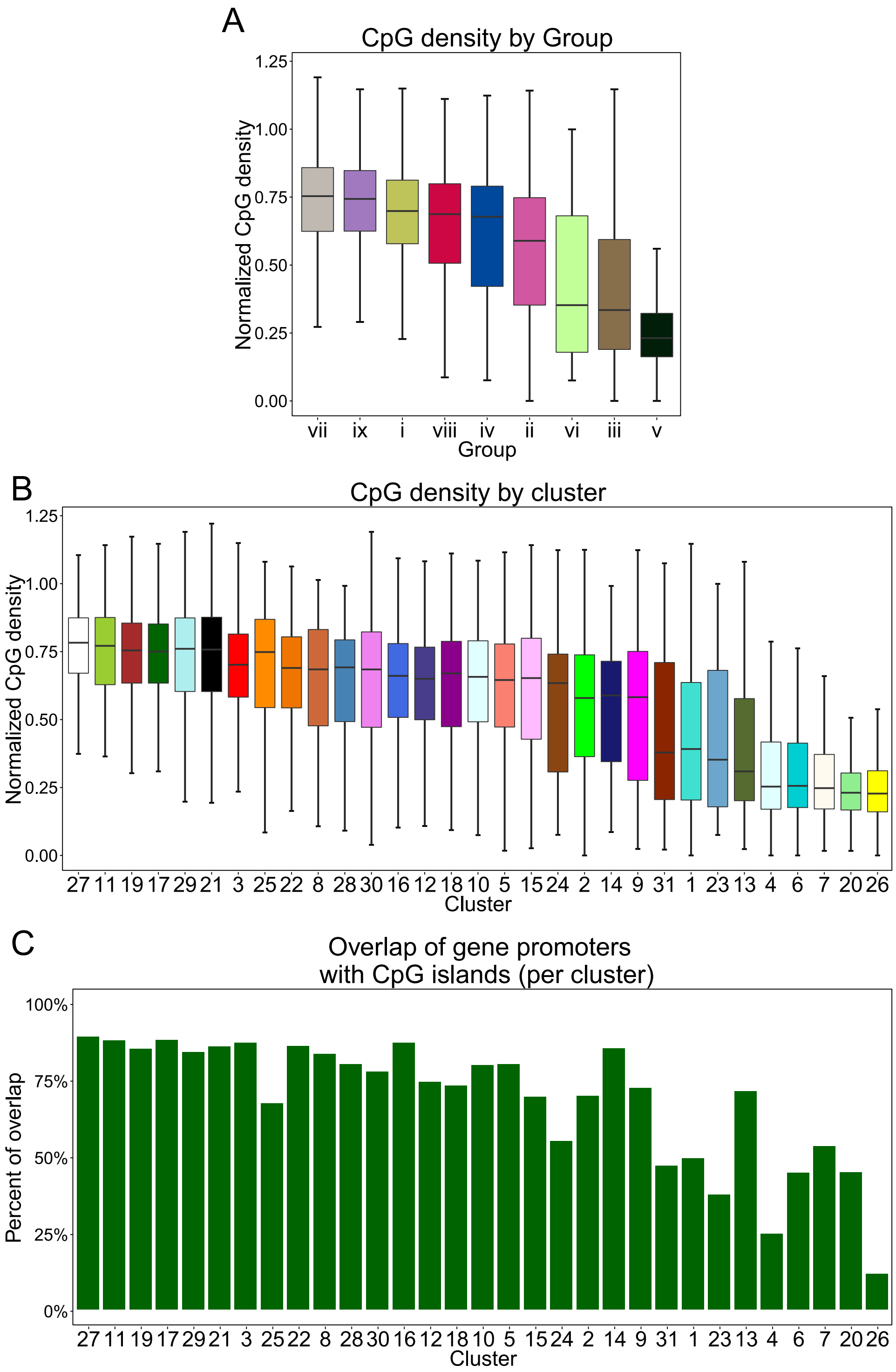
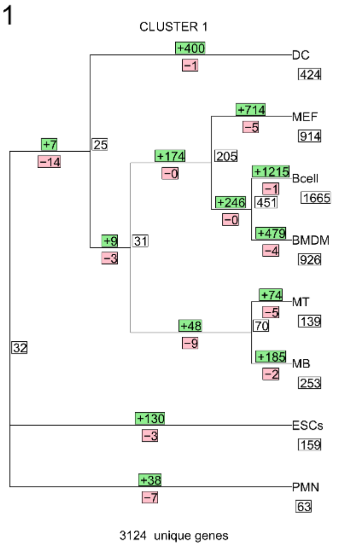


Figure 5.9 A) Boxplot of normalised CpG density across promoters for each of the 9 major profile sub-groups. The groups are ordered from the highest to the lowest according to the mean CpG density across the clusters belonging in that sub-group (the line at the middle of the boxplots denotes the median of the distribution). B) Boxplot of normalised CpG density across promoters for each of the 31 clusters. The clusters are ordered from the highest to the lowest according to the mean CpG density across

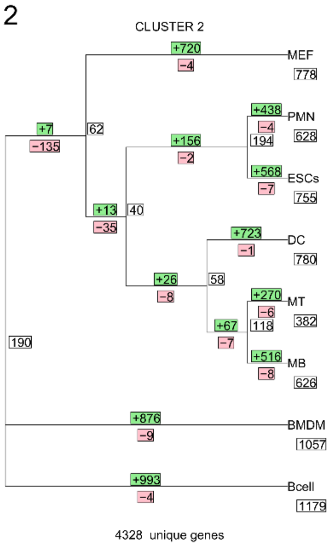
the promoters belonging in that cluster (the line at the middle of the boxplots denotes the median of the distribution. C) Percent of overlap of the promoter regions (± 500 bp from the TSS) in each cluster with CpG islands in the mouse genome. Clusters are ordered according to the ordering of the clusters in Figure S6B.

Six clusters with a ratio lower than 0.6 (Table 5.5) contained genes conserved across cell types. Genes in active clusters 19, 20, and 21, bivalent clusters 3 and 24, and wide-active cluster 26 showed high gene overlap across cell types (Figure 5.11). Clusters 19, 20 and 21 with high levels of H3K4me3 and RNAPII, accompanied by high expression signals (Figure 5.11A) were enriched for protein coding genes (P-value < 0.001, hypergeometric test, Figure 5.11B). In bivalent-wide-H3K27me3 (ii) cluster 3, 15% of genes in ESCs were shared across other seven cell types (Figure 5.10.3). In contrast, bivalent-narrow-H3K27me3 cluster 5 contained highly cell type specific genes with only 3% of genes in ESCs shared across other seven cell types and very few common genes shared at the later states (Figure 5.10.5). Finally, cluster 26 was amongst the highly conserved clusters with over 20% of genes in ESCs shared with other seven cell types (Figure 5.10.26). Cluster 26 consisted largely of processed pseudogenes and sense intronic RNA genes (P-value < 0.0001, hypergeometric test, Figure 5.11B) and showed functional enrichment for G-protein coupled receptor signalling pattern (Fisher's exact test, P-value < 10^{-10}) and sensory perception of chemical stimulus (Fisher's exact test, P-value < 10^{-7}).

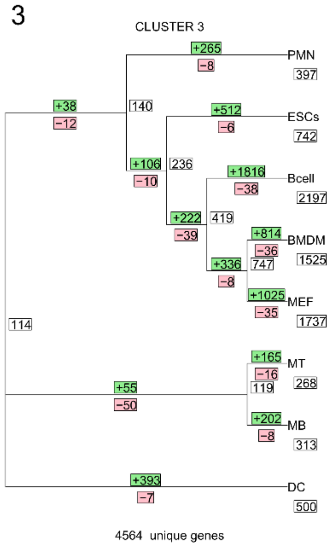
1


CLUSTER 1
+400
−1
DC
424
+7
−14
25
+714
−5
MEF
914
+174
−0
205
+1215
−1
Bcell
1665
+246
−0
451
+479
−4
BMDM
926
+9
−3
31
+74
−5
MT
139
+48
−9
70
+185
−2
MB
253
32
+130
−3
ESCs
159
+38
−7
PMN
63
3124 unique genes

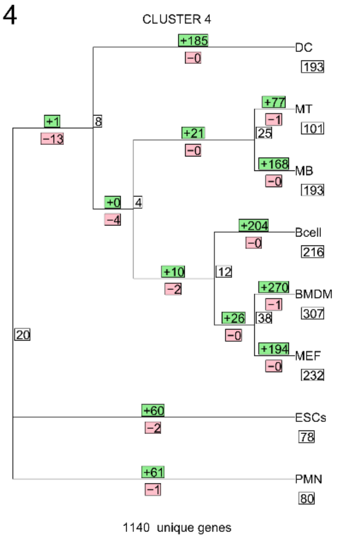
2


CLUSTER 2
+720
−4
MEF
778
+7
−135
62
+438
−4
194
PMN
628
+156
−2
+568
−7
ESCs
755
+13
−35
40
+723
−1
DC
780
+26
−8
58
+270
−6
MT
382
+67
−7
118
+516
−8
MB
626
190
+876
−9
BMDM
1057
+993
−4
Bcell
1179
4328 unique genes

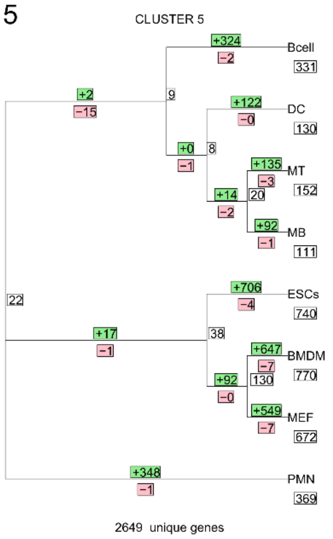
3


CLUSTER 3
+265
−8
PMN
397
+38
−12
140
+512
−6
ESCs
742
+106
−10
236
+1816
−38
Bcell
2197
+222
−39
419
+814
−36
BMDM
1525
+336
−8
747
+1025
−35
MEF
1737
114
+165
−16
MT
268
+55
−50
119
+202
−8
MB
313
+393
−7
DC
500
4564 unique genes

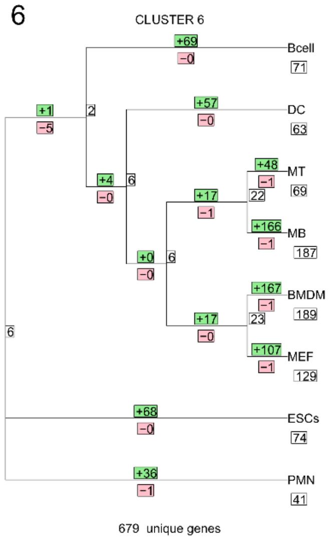
4


CLUSTER 4
+185
−0
DC
193
+1
−13
8
+77
−1
MT
101
+21
−0
25
+168
−0
MB
193
+0
−4
4
+204
−0
Bcell
216
+10
−2
12
+270
−1
BMDM
307
+26
−0
38
+194
−0
MEF
232
20
+60
−2
ESCs
78
+61
−1
PMN
80
1140 unique genes

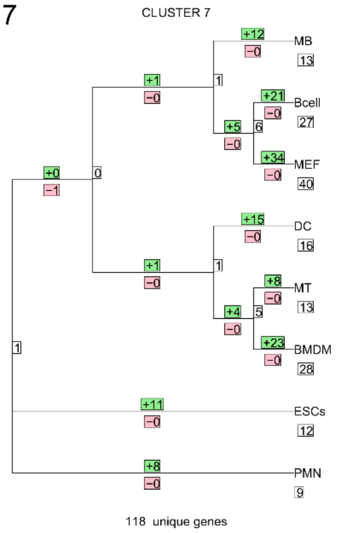
5


CLUSTER 5
+324
−2
Bcell
331
+2
−15
9
+122
−0
DC
130
+0
−1
8
+135
−3
MT
152
+14
−2
20
+92
−1
MB
111
22
+706
−4
ESCs
740
+17
−1
38
+647
−7
BMDM
770
+92
−0
130
+549
−7
MEF
672
+348
−1
PMN
369
2649 unique genes

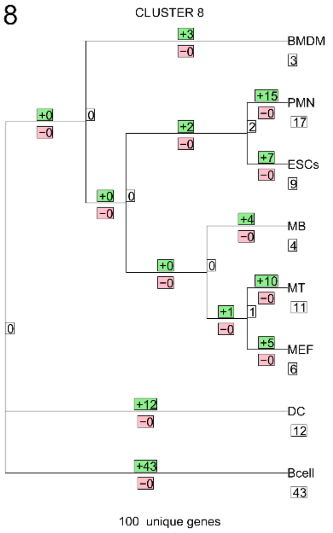
6


CLUSTER 6
+69
−0
Bcell
71
+1
−5
2
+57
−0
DC
63
+4
−0
6
+48
−1
MT
69
+17
−1
22
+166
−1
MB
187
+0
−0
6
+167
−1
BMDM
189
+17
−0
23
+107
−1
MEF
129
6
+68
−0
ESCs
74
+36
−1
PMN
41
679 unique genes

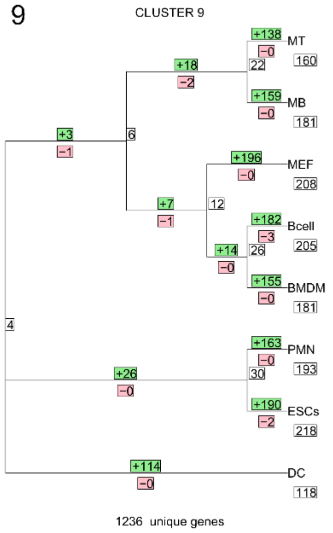
7


CLUSTER 7
+12
−0
MB
13
+1
−0
1
+21
−0
Bcell
27
+5
−0
6
+34
−0
MEF
40
+0
−1
0
+15
−0
DC
16
+1
−0
1
+8
−0
MT
13
+4
−0
5
+23
−0
BMDM
28
1
+11
−0
ESCs
12
+8
−0
PMN
9
118 unique genes

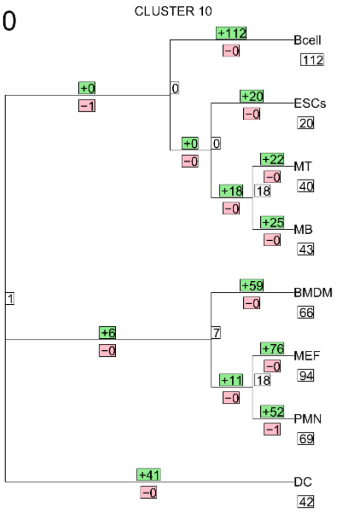
8


CLUSTER 8
+3
−0
BMDM
3
+0
−0
0
+15
−0
PMN
17
+2
−0
2
+7
−0
ESCs
9
+0
−0
0
+4
−0
MB
4
+0
−0
0
+10
−0
MT
11
+1
−0
1
+5
−0
MEF
6
0
+12
−0
DC
12
+43
−0
Bcell
43
100 unique genes

9

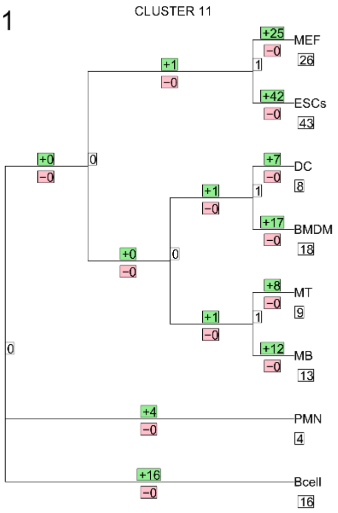

CLUSTER 9
+138
−0
MT
160
+18
−2
22
+159
−0
MB
181
+3
−1
6
+196
−0
MEF
208
+7
−1
12
+182
−3
Bcell
205
+14
−0
26
+155
−0
BMDM
181
4
+163
−0
PMN
193
+26
−0
30
+190
−2
ESCs
218
+114
−0
DC
118
1236 unique genes

10



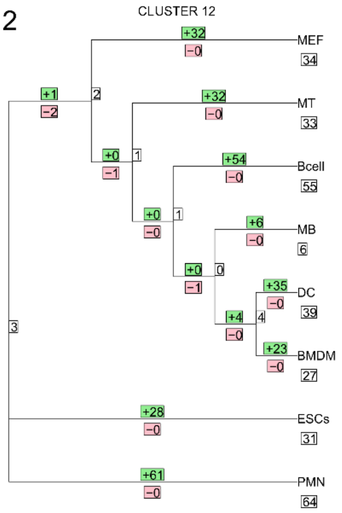
393 unique genes

11



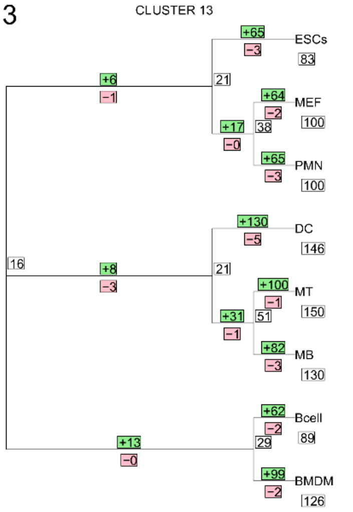
134 unique genes

12



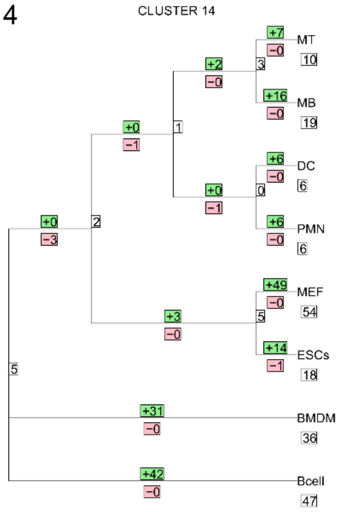
254 unique genes

13



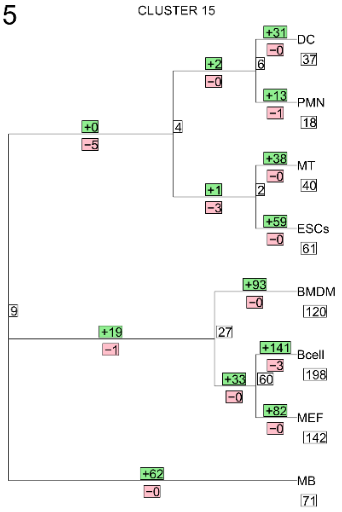
615 unique genes

14



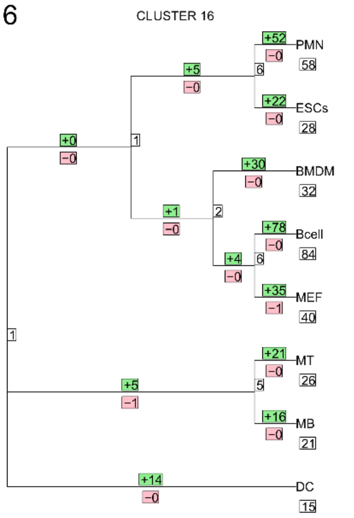
166 unique genes

15



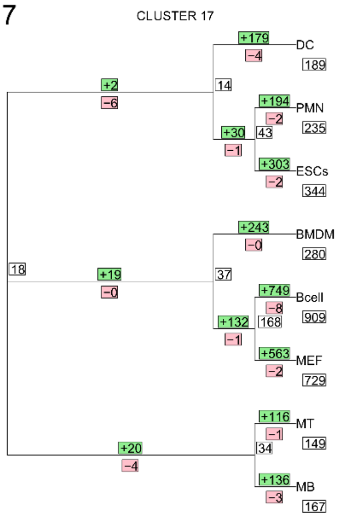
514 unique genes

16



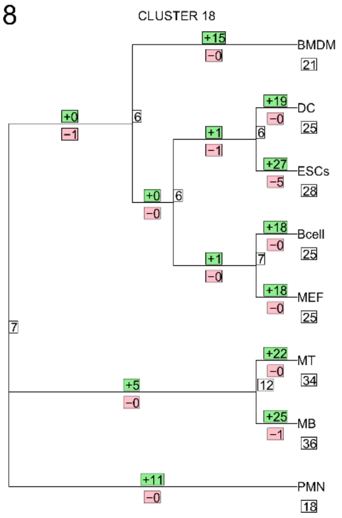
262 unique genes

17



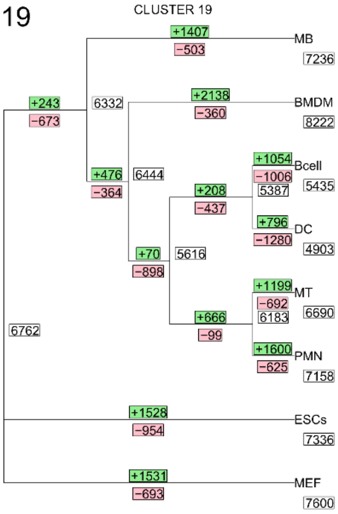
2396 unique genes

18



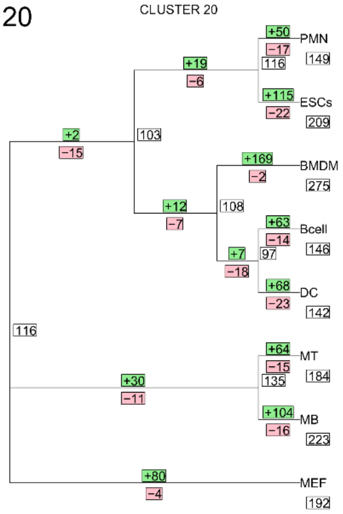
161 unique genes

19



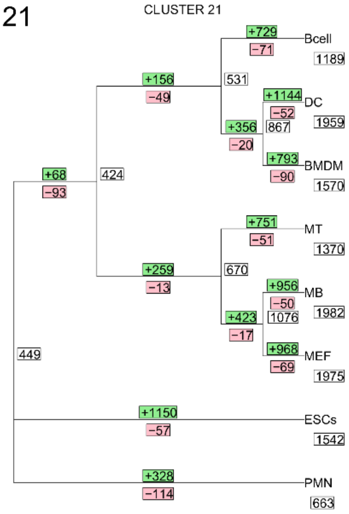
13185 unique genes

20



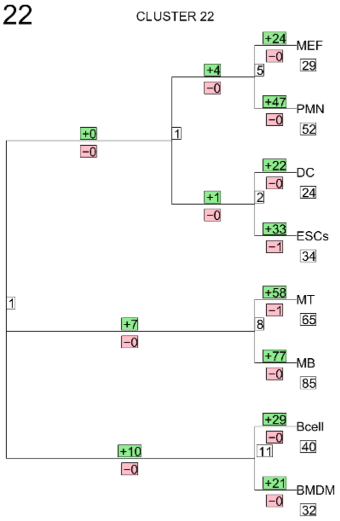
663 unique genes

21



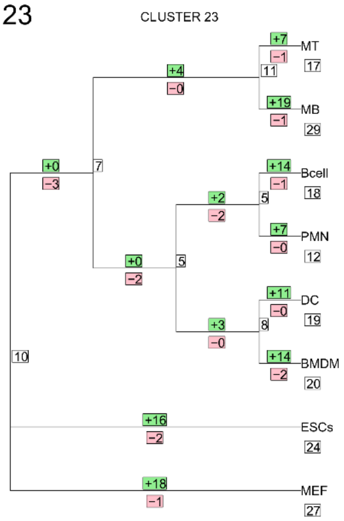
5854 unique genes

22



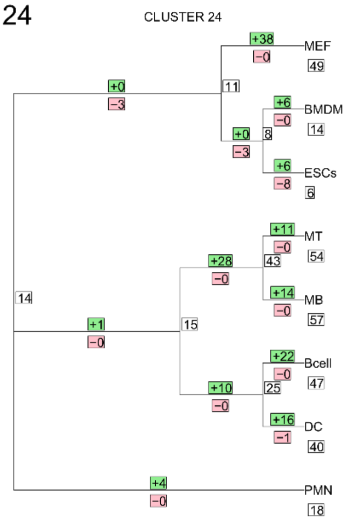
309 unique genes

23



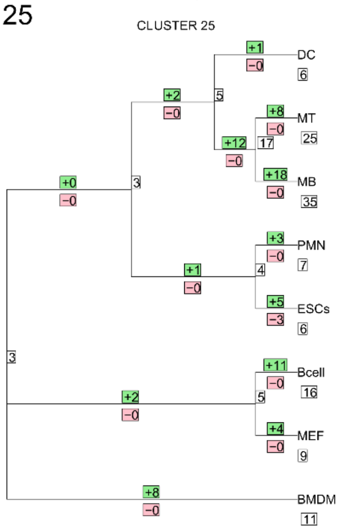
109 unique genes

24



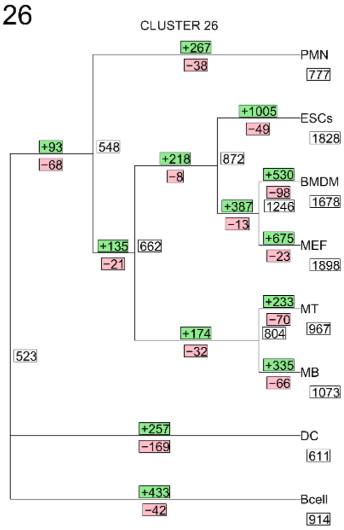
149 unique genes

25



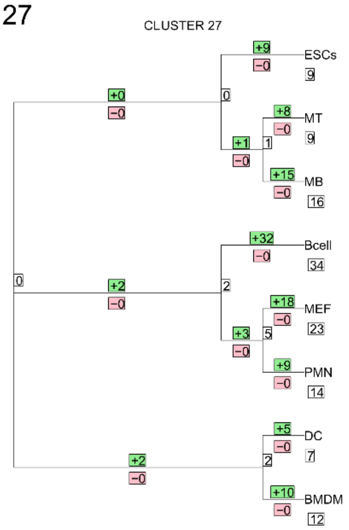
72 unique genes

26



3750 unique genes

27



103 unique genes

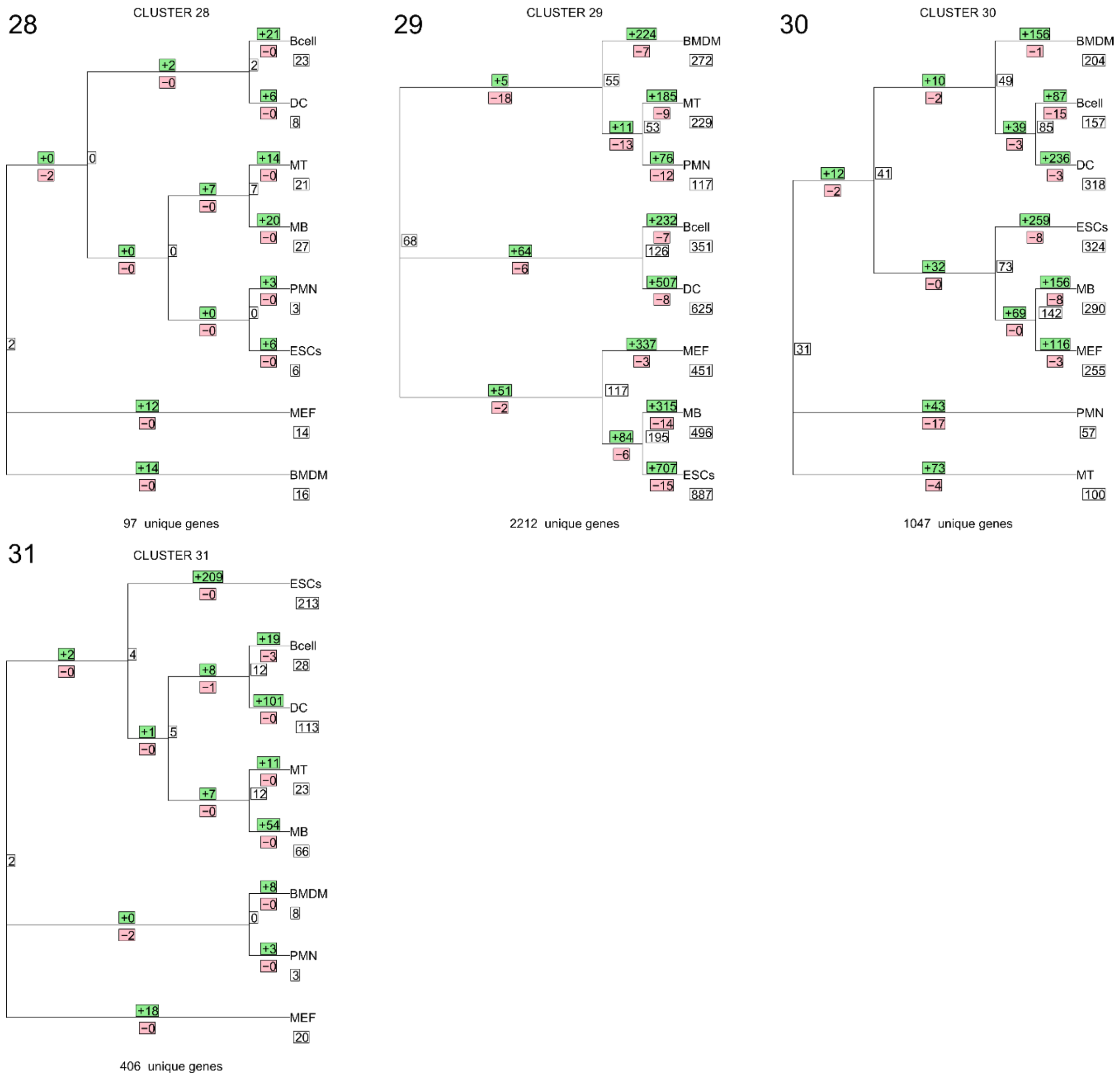


Figure 5.10.1-5.9.31 Maximum parsimony trees for all of the 31 clusters defined in the analysis. The white labels signify the number of genes that are shared across the cell types at the respective ancestral states or the number of genes that are unique in one cell type (if the label is positioned underneath the cell type name). The green and red labels signify the number of genes gained or lost respectively, from one ancestral states to another.

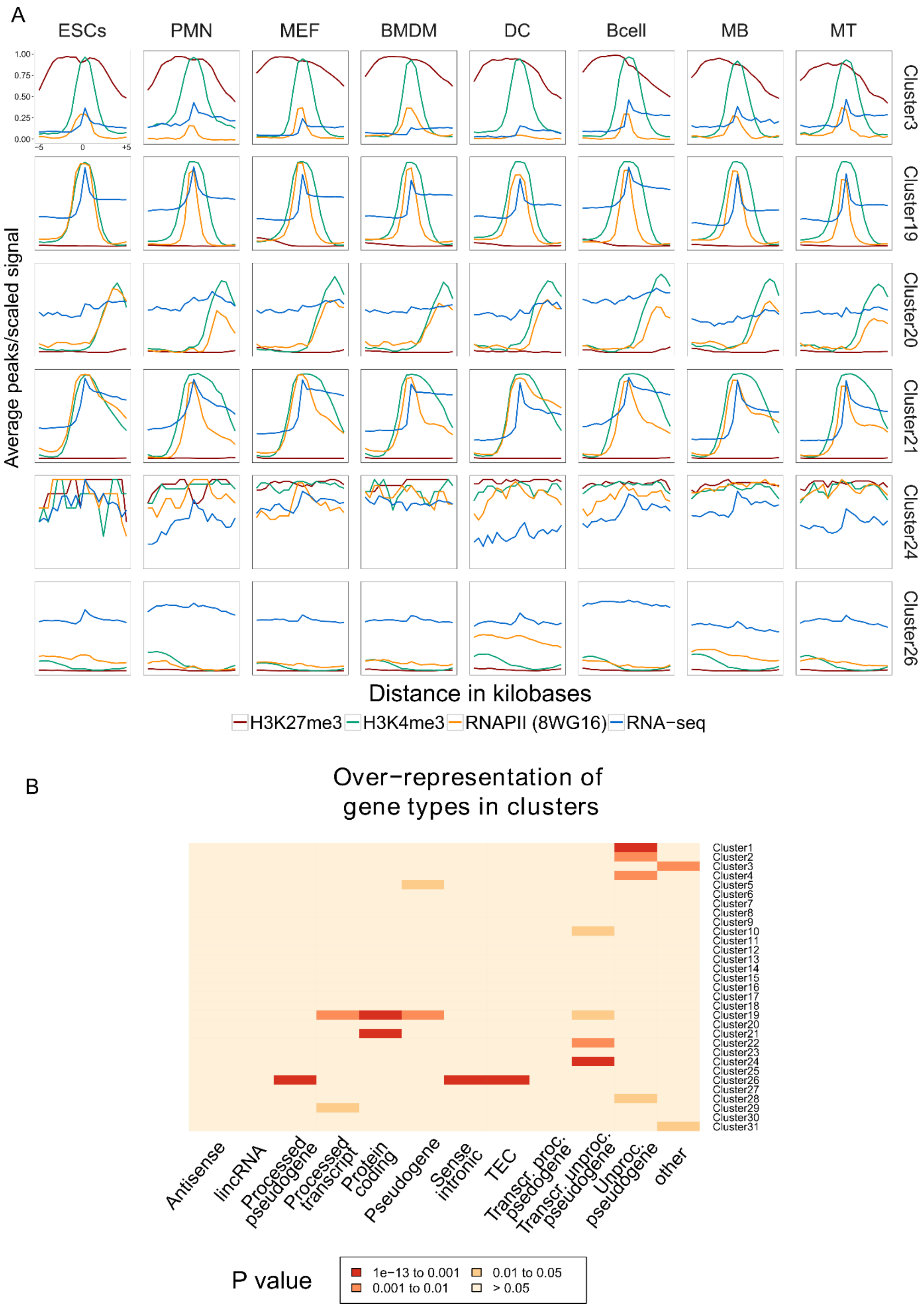


Figure 5.11 A) **Clusters with ratio of unique to total number of genes lower than 0.6 showed similar profiles across the 8 cell types. Clusters 19, 20 and 21 (active) were showing high levels of H3K4me3 and RNAPII, accompanied by high expression signals. Cluster 3 and 24 are bivalent clusters that share few genes at the ancestral states (Figure S9 and S30) and their underlying profile across the 8 cell types. Clusters 26 is showing only an expression signal across the 10kb area flanking the TSS of the**

promoters and the profile is conserved across cell types. B**) Enrichment of various gene types as annotated from ENCODE in all 31 clusters.**

In summary, the chromatin and expression profiles of promoters in 8 cell types showed 9 major groups enriched for specific functional properties and six clusters shared their signal profiles in genes across cell types.

5.4.3 Promoter dynamics across cell types and chromatin states

Bivalent promoters in ESCs are thought to become either active or repressed after differentiating into mature cell types (Mikkelsen et al., 2007; Pan et al., 2007; Zhao et al., 2007). To systematically study major chromatin state transitions across cell types we calculated the number of overlapping genes across cell types in individual clusters. The statistical significance of the number of promoters shared between cell types across clusters was calculated using a hypergeometric test after correcting for the cell type bias of each cluster (see Methods). About 2.2% of all possible promoter overlaps between cell types across clusters were statistically significant (hypergeometric test, P-value <0.01) with the majority of them representing promoters moving across clusters belonging to the same sub-group. For example, 632 genes belonged to cluster 19 in PMNs and cluster 29 in ESCs. Though these genes are expressed in both cluster 19 and 29, they show divergent epigenetic profiles in two cell types namely a wide H3K4me3 and RNAPII signal upstream of TSS in ESCs while a sharp narrow H3K4me3 and RNAPII signal at the promoter in PMNs. We further specifically focussed on the significant overlaps across clusters between sub-groups (Figure 5.12A).

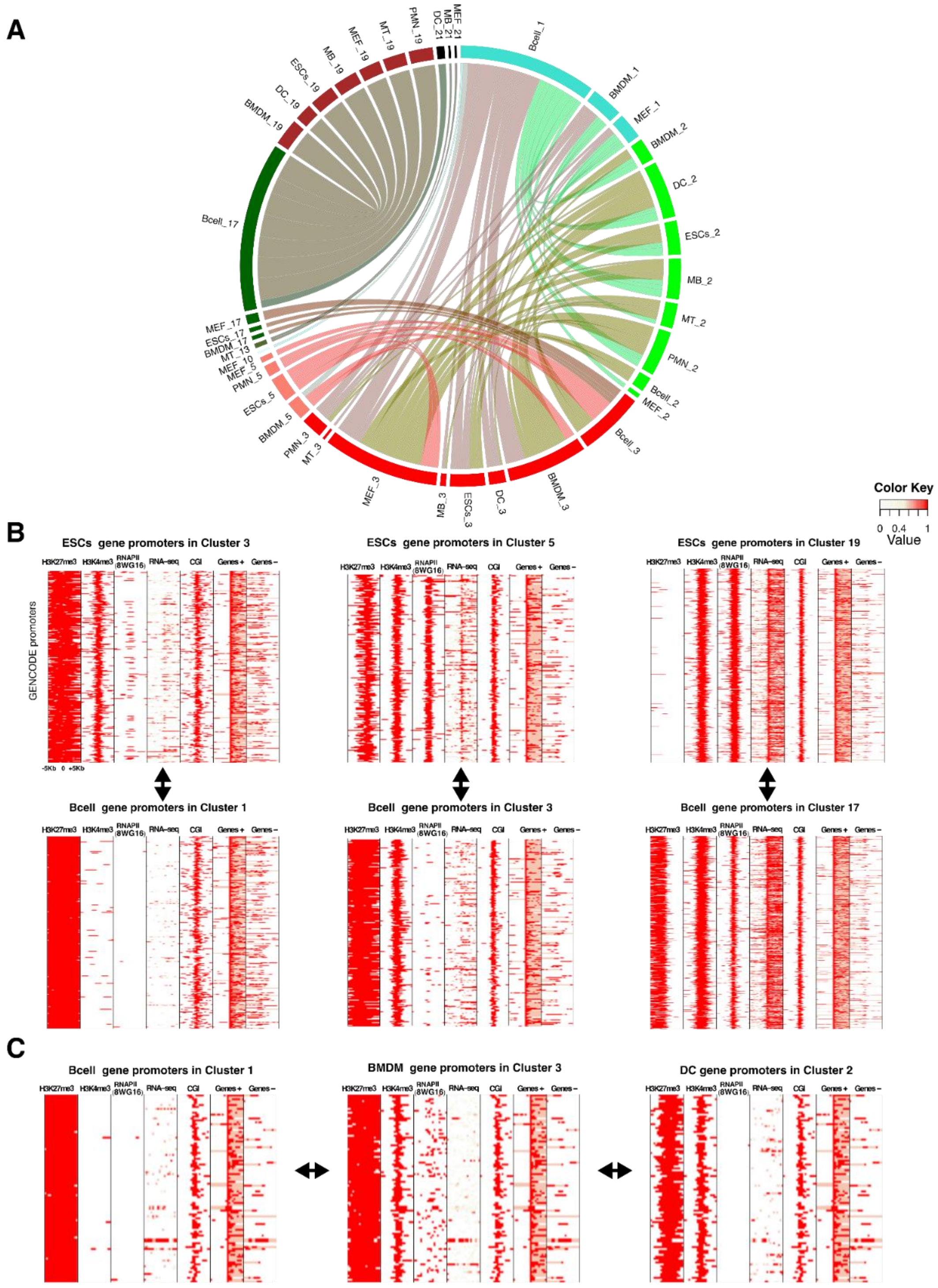


Figure 5.12 Promoter dynamic across cell types and chromatin states. A) Significant chromatin state transitions across cell types and clusters. Four major chromatin state changes across cell type pairs emerged, namely H3K27me3-only <-> bivalent-wide-H3K27me3, H3K27me3-only <-> bivalent-narrow-H3K27me3, bivalent-narrow-H3K27me3 <-> bivalent-active and bivalent-active <-> highly-active. B) Bivalent-wide-

H3K27me3 cluster 3 promoters in ESCs overlapped highly H3K27me3-only cluster 1 promoters in B cells and were enriched for ‘pattern specification process’ (P-value < 10^{-30}). Bivalent-narrow-H3K27me3 cluster 5 promoters in ESCs overlapped highly with bivalent-wide-H3K27me3 cluster 3 promoters in B cells and were enriched for ‘Nervous system development’ (P-value < 10^{-30}). Highly-active cluster 19 promoters in ESCs overlapped highly with boundary-H3K27me3-active cluster 17 promoters in B cells and were enriched for ‘ncRNA metabolic process’ (P-value < 10^{-22}). C) Significant sets of genes overlapping in 3 distinct cell types and clusters. 98 genes enriched for ‘cell fate commitment’ (P-value < 10^{-6}) were present in B cells in cluster 1, in DCs in cluster 2 and in BMDMs in cluster 3.

There were four major chromatin state changes across cell type pairs namely H3K27me3-only ↔ bivalent-wide-H3K27me3, H3K27me3-only ↔ bivalent-narrow-H3K27me3, bivalent-narrow-H3K27me3 ↔ bivalent active and bivalent active ↔ highly-active. H3K27me3-only promoters in B cells were either bivalent-narrow-H3K27me3 or bivalent-wide-H3K27me3 in most other cell types. Similarly, bivalent active promoters in B cells were active in most other cell types. To exclude the possibility that the aberrant H3K27me3 modification at the promoters in B cells compared to the other cell types is due to a technical problem of that sample, we replaced the H3K27me3 sample used in B cells with another H3K27me3 ChIP-seq replicate in B cells from the same study and found the same result despite a small decrease in the H3K27me3 signal (Figure 5.13A and 5.13B).

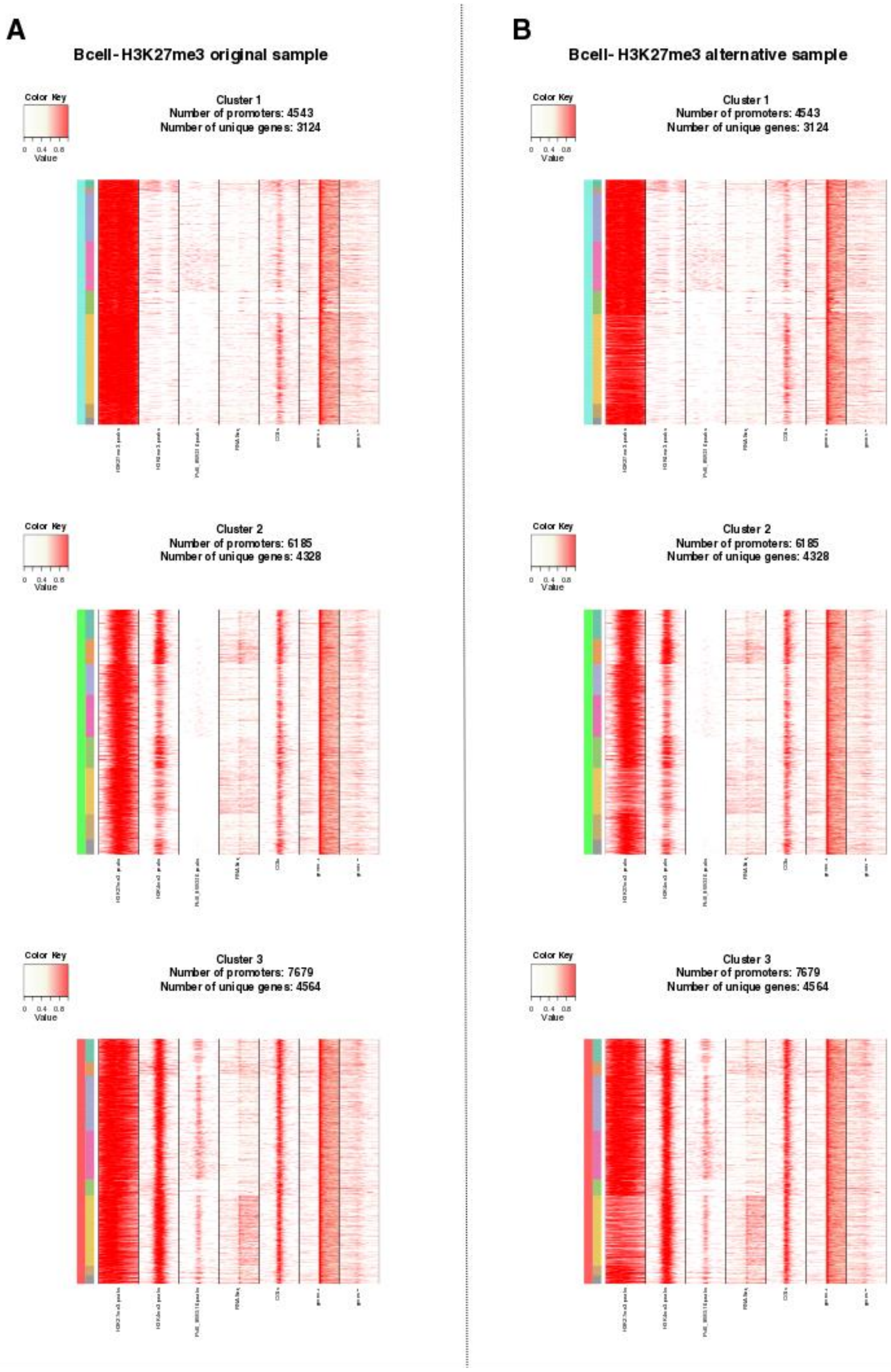


Figure 5.13 A) **Representative heatmaps for clusters 1, 2 and 3, which show extremely high levels of H3K27me3 across the 10 kb regions flanking the TSS of the genes. These profiles are plotted using the H3K27me3 sample from B cells that was originally used throughout the study.** B) **We have replaced the H3K27me3 sample in B cells with**

another replicate to assess if there are extremes differences at the signal. There is a minimal decrease, but not significant enough since the H3K27me3 is still strong.

Bivalent-wide-H3K27me3 cluster 3 promoters in ESCs overlapped highly with H3K27me3-only cluster 1 promoters in B cells and were enriched for ‘pattern specification process’ (Fisher’s exact test, P-value < 10^{-30}) (Figure 5.12B). Bivalent-narrow-H3K27me3 cluster 5 promoters in ESCs overlapped highly with bivalent-wide-H3K27me3 cluster 3 promoters in B cells and were enriched for ‘Nervous system development’ (Fisher’s exact test, P-value < 10-^{30}) (Figure 5.12B). Highly-active cluster 19 promoters in ESCs overlapped highly with boundary-H3K27me3-active cluster 17 promoters in B cells and were enriched for ‘ncRNA metabolic process’ (Fisher’s exact test, P-value < 10^{-22}) (Figure 5.12B).

We further calculated statistical significance for the overlap of the promoters belonging to three different clusters in three cell types. The majority of the significant cluster triplets consisted of promoters in clusters 1 in B cells present in cluster 2 and cluster 3 in other two cell types. For example, 98 genes enriched for ‘cell fate commitment’ (Fisher’s exact test, P-value < 10^{-6}) were present in B cells in cluster 1, in DCs in cluster 2 and in BMDMs in cluster 3 (Figure 5.12C).

Taken together, we noted significant patterns of epigenetic dynamics across cell types predominantly between 6 clusters (clusters 1, 2, 3, 17, 19 and 21). Importantly, the major epigenetic dynamics across cell types were not reflected at the expression level.

5.4.4 Transcription factor binding and motif enrichment across clusters

To investigate whether any of the specific cluster patterns were associated with binding of transcription regulators, we calculated transcription-related factor binding enrichment using the CODEX (Sánchez-Castillo et al., 2015) ChIP sequencing data compendium (see Methods). All clusters were significantly enriched (P-value < 0.01, hypergeometric test) for at least one factor (Figure 5.14A). Highly-active clusters 19, 21 and 29 were enriched for binding of most of the TFs in the analysis while clusters 4, 6, 7, 20 and 26 were not enriched for binding of most of them. H3K27me3-only cluster 1 and bivalent-narrow-H3K27me3 cluster 2 were both highly enriched for Polycomb components binding (Suz12, Ezh2, Rnf2, Mtf2 and Ring1b) as well as Kdm2b, Notch1 and Klf2. Bivalent cluster 2 was additionally enriched for binding of Hdac2, Ldb1 and Foxa2.

The similarities between the epigenetic profiles of clusters were not in full agreement with the similarities in their transcription-related factor binding i.e. the hierarchical clustering of TF enrichment at each cluster promoter (Figure 5.14A) did not result in the same sub-grouping as obtained using similarities between the profiles (Figure 5.8B). For example, bivalent-wide-H3K27me3 cluster 3 and bivalent-narrow-H3K27me3 cluster 5 were enriched for binding of very similar factors despite the differences in H3K27me3 signal and RNAPII occupancy. Nevertheless, many clusters belonging to the same sub-groups showed similar transcription control i.e. clusters 19, 21, 29 and 31 classified as highly-active and showed enrichment for more than half of the TFs in this analysis. Bivalent-wide-H3K27me3 clusters (8, 11, and 16) also clustered very close to each other showing significantly high enrichment for Polycomb complex components. Interestingly, clusters 20 and 26 of the sub-grouping wide-active (v) were significantly enriched for binding of similar TFs. More specifically, eight-twenty-one 2 (Eto2), T-cell acute lymphocytic leukaemia protein 1 (Tal1), LIM Domain Only 2 (Lmo2) and Gata2 were significantly enriched (P-value < 0.05, hypergeometric test), all TFs involved in hematopoietic development (Anguita et al., 2004; Soler et al., 2010; Vicente et al., 2012).

Clusters 3 and 5 were also bound uniquely by some TFs that showed dissimilar gene ontologies (Table 5.6). TFs binding at bivalent-wide-H3K27me3 cluster 3 were associated with positive regulation of transcription (Fisher’s exact test, P-value < 0.01), somatic stem cell population maintenance (Fisher’s exact test, P-value < 10^{-6}) and regulation of transcription for RNAPII promoter in response to stress (Fisher’s exact test, P-value < 10^{-4}). On the other hand, the TFs binding uniquely to bivalent-narrow-H3K27me3 cluster 5 (Table 5.6) were enriched for regulation of cytokine biosynthetic process (Fisher’s exact test, P-value < 0.01), liver development (Fisher’s exact test, P-value < 0.01) and regulation of myeloid cell differentiation (Fisher's exact test, P-value < 0.01).

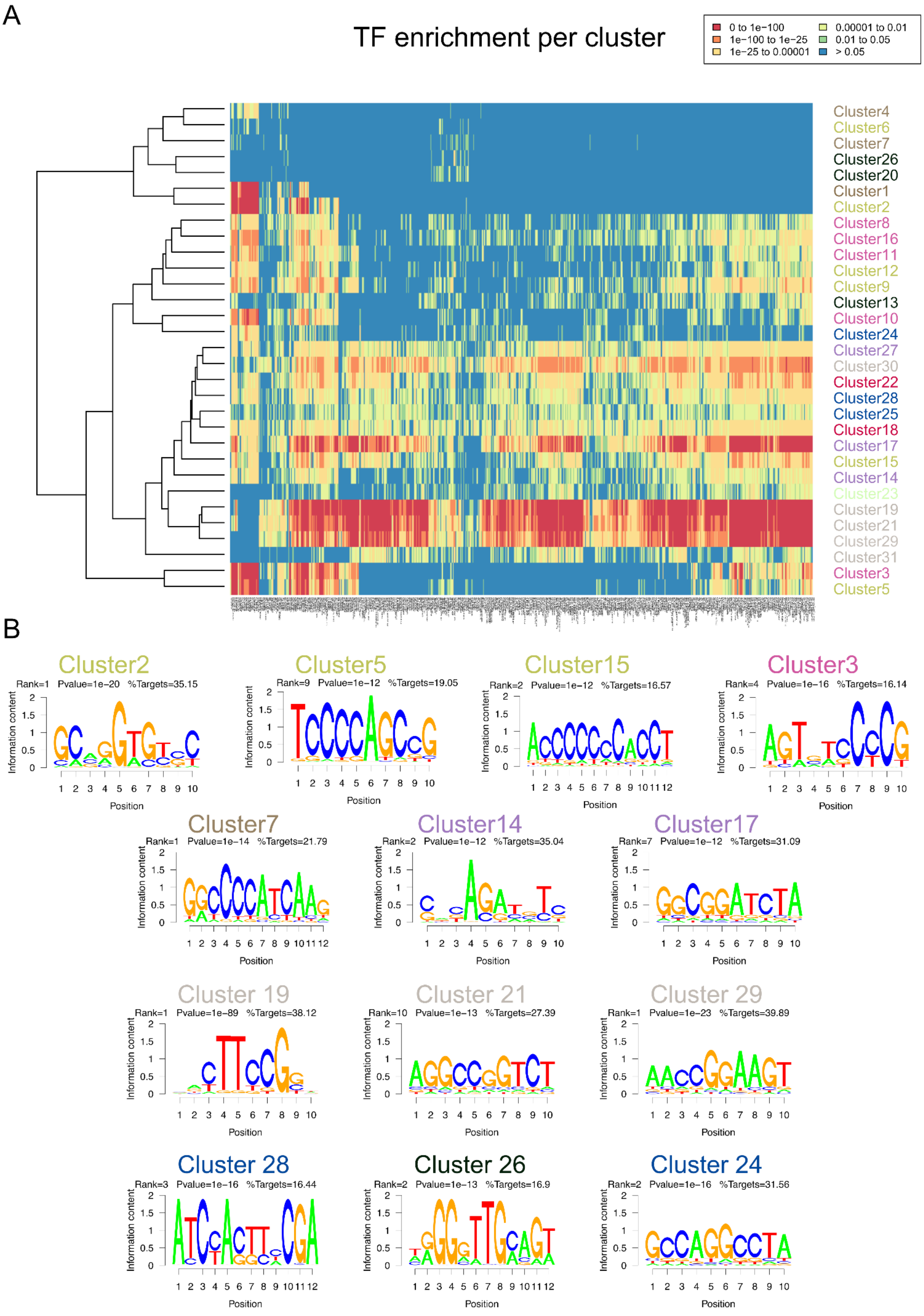


Figure 5.14 Transcription factor binding and motif enrichment across the clusters. A) Transcription-related factor binding enrichment using the CODEX ChIP-seq data compendium (see Methods). All clusters were significantly enriched (hypergeometric test - P value < 0.01) for at least one factor. B) Thirteen clusters showed at least one de-novo motif enrichment with more than 15% of targets and P value < 10^{-10}.

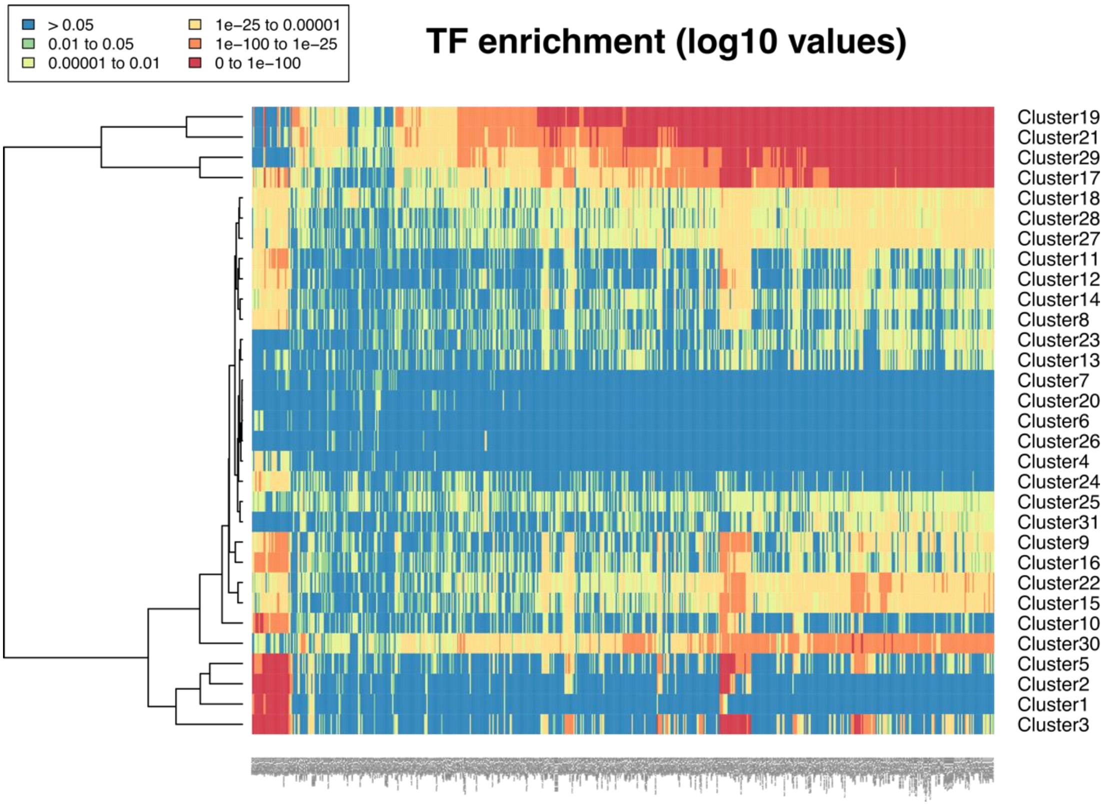


Figure 5.15 Transcription factor binding enrichment using the CODEX ChIP-seq data compendium. The clustering was done on the log10 transformed P-values, derived from a hypergeometric test.

Considering that clustering the raw P-values might give us a non-trustworthy clustering of the TF binding profiles across clusters (the result could be driven solely by the high P-value, due the existence of zero P-values), we used the log10 transformed P-values from the hypergeometric tests. This has resulted in a more expected and cleaner distinction of clusters for the TF binding profiles (Figure 5.15). More specifically, the bivalent (Cluster 3 and 5) and H3K27me3-only cluster (Cluster 1) (at the bottom of the heatmap) had greater distance from the active clusters (Cluster 19, 21,29). This adds more value to our study, confirming that bivalent and active clusters are highly likely to be regulated by very dissimilar TF networks.

We further performed de-novo and known motif discovery for all clusters using HOMER (Heinz et al., 2010). Thirteen clusters showed at least one de-novo motif enrichment with more than 15% of targets and P value < 10^{-10} (Figure 5.14B). Bivalent clusters 2, 3, 5, and 15 were enriched for GC rich motifs. More specifically, bivalent clusters 3 and 5 were enriched for the same ‘TCCCC’ sequence motif, a motif we previously identified enriched at bivalent promoters in ESCs (Mantsoki et al., 2015). H3K27me3-only cluster 7, Boundary-H3K27me3-active clusters 14 and 17 showed enrichment of diverse GC rich motifs. Highly-active clusters 19 and 29 were enriched for GC-rich ETS factor motifs (‘CCGGAA’) while cluster 21 for a zinc-finger matching motif ‘AGGCCGG’. Bivalent-wide-active clusters had motifs enriched in more than 16% of the target sequences with ‘ATCCACTT’ for cluster 24 and ‘GGGTTG’ for cluster 28. Finally, cluster 26 was enriched for a GC-rich motif (‘CCAGGCC’) in more than 31% of the target sequences, which matches to the Zinc Finger Protein, X-Linked (ZFX).

In conclusion, clusters with divergent epigenetic profiles were enriched for binding of similar factors. Nevertheless, they were enriched for specific de-novo sequence motifs.

5.4.5 RNAPII pausing across clusters and cell types

We incorporated RNAPII (8WG16) in our analysis, since it has been established that RNAPII pausing regulates the transcription of many genes involved in development, cell cycle and metabolism (Marks et al., 2012; Min et al., 2011; Tee et al., 2014)**.** Differences in RNAPII occupancy were present across clusters and profile subgroupings (Figure 5.8C), thus we studied RNAPII variability across clusters in greater detail by calculating the RNAPII pausing index (Muse et al. 2007) defined as the ratio of RNAPII signal at the core promoter to RNAPII signal within the gene body (see Methods). When the clusters were ordered according to their respective mean pausing index (Figure 5.16A), they followed a general trend where the active subgroupings (highly-active (vii), boundary-H3K27me3-active (ix)) were followed by the bivalent subgroupings (bivalent-narrow-H3K27me3 (i), bivalent-wide-H3K27me3 (ii)). Active clusters 19, 17, 29 and bivalent cluster 11 presented the highest pausing indices (in dark red background in Figure 5.16A), while active bi-directional (26, 13, 20) and H3K27me3-only (4, 1, 7) clusters presented very low (in yellow background, Figure 5.16A) or no pausing at all (shown in grey background, Figure 5.16A).

There was a moderate correlation between the average pausing index and average mRNA expression levels between clusters (Pearson's correlation coefficient: 0.53). The majority of active clusters had elevated levels of engaged RNAPII at the core promoter, in contrast with the bivalent and H3K27me3-only clusters that showed divergent patterns of pausing ranging from mid to no pausing at all. This is in agreement with a recent publication demonstrating that high pausing index is more associated with highly expressed genes involved in cell cycle regulation rather than bivalent developmental regulators (Williams et al. 2015). The clusters with the highest pausing index were not among the ones with highest expression levels. Clusters with mid-range pausing index including active cluster 30 and bivalent active clusters 22, 18 and 28 were among the clusters with the highest mean expression values (Figure 5.16B). Bivalent clusters 11 (high pausing), 5, 16, 9, 12 (mid-pausing) and 10 (low-pausing) were expressed at low levels. Also, the majority of the low-pausing (3, 2, 26) and no-pausing (6, 1) clusters showed extremely low expression values (Figure 5.16B).

Williams et al. (2015) proposed that RNAPII pausing in ESCs is a regulatory mechanism which facilitates the retaining of their self-renewal properties. We noted that cell types of higher developmental potential (ESCs and PMNs) and MEFs, displayed a high mean pausing index (Figure 5.17A). In contrast, cells with reduced developmental potential (BMDMs, DCs, B cells, MBs, MTs) had a mid-range mean pausing index (Figure 5.17A). We further calculated the pausing indices for genes involved in cell cycle and developmental genes (see Methods), accompanied by groups of different histone modification patterns (bivalent, H3K4me3-only and H3K27me3-only) in our study (Figure 5.17B). Cell cycle and H3K4me3-only genes presented high levels of RNAPII pausing, while developmental and bivalent genes presented mid levels of RNAPII pausing. In all groups, RNAPII pausing reduced from progenitor to committed cell types (Figure 5.17B). Developmental genes belonged to H3K27me3-only cluster 1, bivalent clusters 2 and 3 and active clusters 19 and 21 (Figure 5.17C), while cell cycle genes were mostly found in active clusters 19 and 21 (Figure 5.17D).

To assess in detail the RNAPII pausing in relation to the regulatory networks governing ESCs we used genes involved in the pluripotency network of mouse ESCs from KEGG pathways (Kanehisa et al., 2016) (see Methods) and we kept only the developmental and cell cycle genes that were also annotated in that list. Genes annotated as in favour of self-renewal (Pro Pluripotency) exhibited extremely high RNAPII pausing (pausing index > 4) accompanied by high expression is ESCs (Mapk1, Stat3, Rif1) and MEFs (Mapk1, Stat3) (Figure 5.16C). Map2k1 (or Erk1) is the only Pro differentiation gene that is highly expressed and shows high levels of expression in both ESCs and MEFs (Figure 5.16C).

Jak2 and Wnt9a (Pro differentiation) were highly paused and lowly expressed in ESCs but they showed elevated levels of expression in MEFs retaining their high pausing. Interestingly, the changes in expression of Jak2 and Wnt9a between ESCs and MEFs were also marked by a change in the underlying clustering classification (Figure 5.16D). More specifically, they both convert to active states (Jak2-cluster 21, Wnt9a-cluster 14) in MEFs from bivalent states in ESCs (Jak2-cluster 3, Wnt9a-cluster 5.

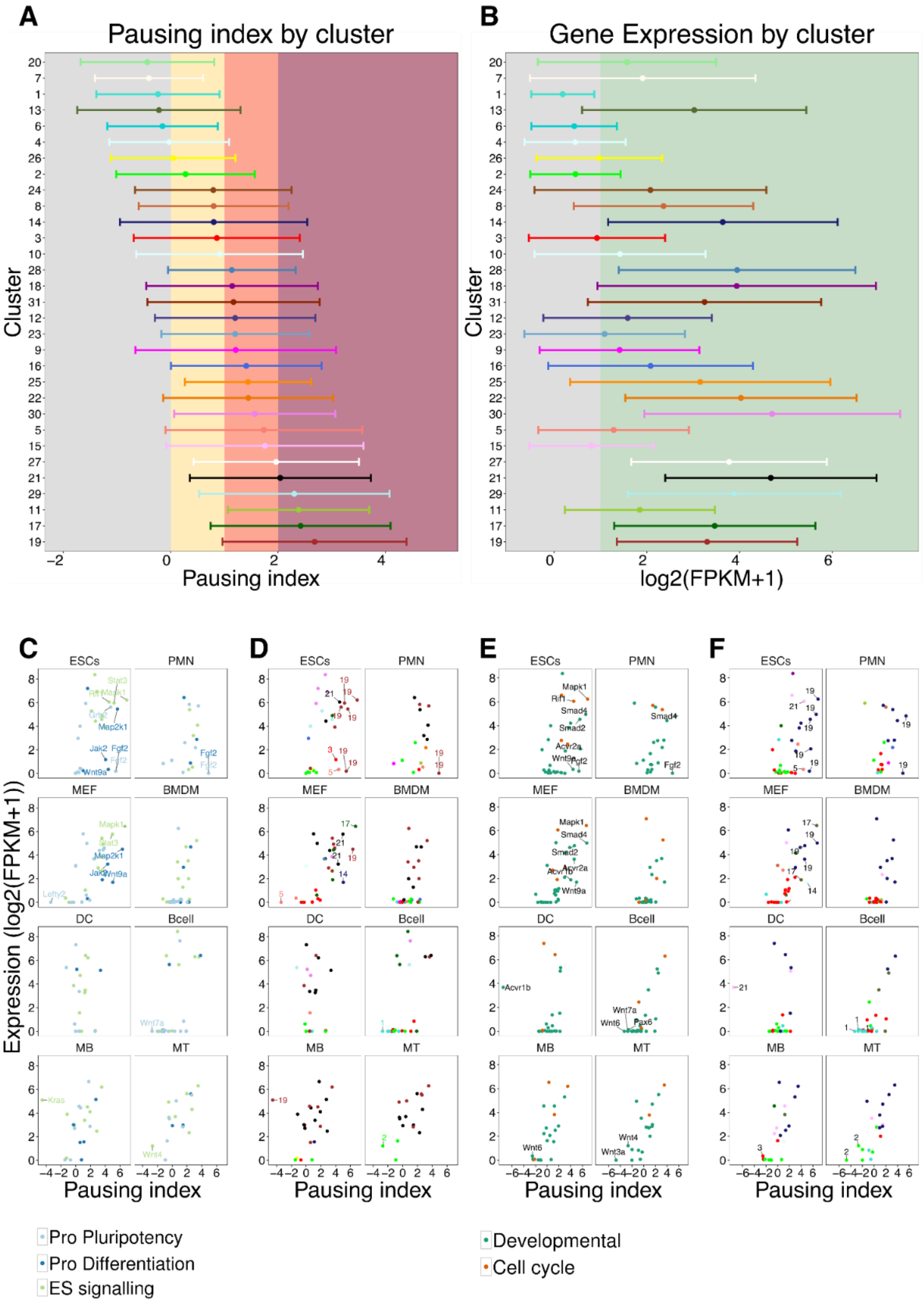


Figure 5.16 RNAPII pausing across clusters and cell types. A) Distribution of pausing indices across gene promoters in all 31 clusters in our study. Clusters are ordered according to their mean pausing index and the point shown in the middle of the distribution denotes the mean, flanked by the error bars of standard deviation. The colours in the background denote the level of RNAPII pausing - grey: no pausing,

yellow: low pausing, red: mid pausing, dark red: high pausing. B) Distribution of expression levels (log2(FPKM+1)) of genes in each cluster. The clusters were in the same order as in Figure 5A and the point in the middle of the distribution denotes the mean, flanked by the error bars of standard deviation. The colour background denotes a threshold on expression (log2 (FPKM+1) =1) as defined in Figure S1 – grey: not expressed, green: expressed. C) RNAPII pausing versus expression (log2(FPKM+1)) for genes annotated as: 1) Pro pluripotent, 2) Pro Differentiation, 3) ES signalling (see Methods). Genes annotated as in favour of self-renewal (Pro Pluripotency) exhibited extremely high RNAPII pausing (pausing index > 4) accompanied by high expression is ESCs and MEFS. D) RNAPII pausing vs expression values genes in Figure 5C. The colours and numbers imposed display the respective cluster colour where each gene belonged. E) RNAPII pausing vs expression for developmental and cell cycle genes (involved in the ESCs pluripotent network, see Methods) across all cell types. F) RNAPII pausing vs expression for genes in Figure 5E. The colours and numbers imposed display the respective cluster colour where each gene belonged.

Next, cell cycle and developmental genes involved in the pluripotency network were assessed in terms of their expression versus RNAPII pausing. We decided to proceed narrowing down these two gene sets, because of the high number of the annotated cell cycle genes and their extreme levels of high RNAPII pausing, which was obscuring the developmental genes’ patterns (Figure 5.17E and 5.17F). Mapk1 and Rif1, were annotated as cell cycle and Pro pluripotency genes and as previously noted, they presented a highly paused-highly epxressed pattern. Developmental genes Smad4/2 were highly expressed and paused across ESCs, PMNs and MEFs and Acvr1b/2a, Wnt9a and Fgf2 were highly paused albeit lowly expressed in the progenitor cells (Figure 514E and Figure 514F).

In conclusion, RNAPII pausing patterns are maintained at high levels for progenitor cells. Pro pluripotency and cell cycle genes present a highly paused-highly expressed pattern across ESCs, PMNs and MEFs, which is in accordance with the gene classification in active clusters and supports strongly the theory that RNAPII pausing assists cells to retain their pluripotent characteristics. Finally, developmental genes involved in pluripotency network present similar RNAPII pausing levels with cell cycle genes and they belong to active chromatin states rather than bivalent, especially for progenitor cells.

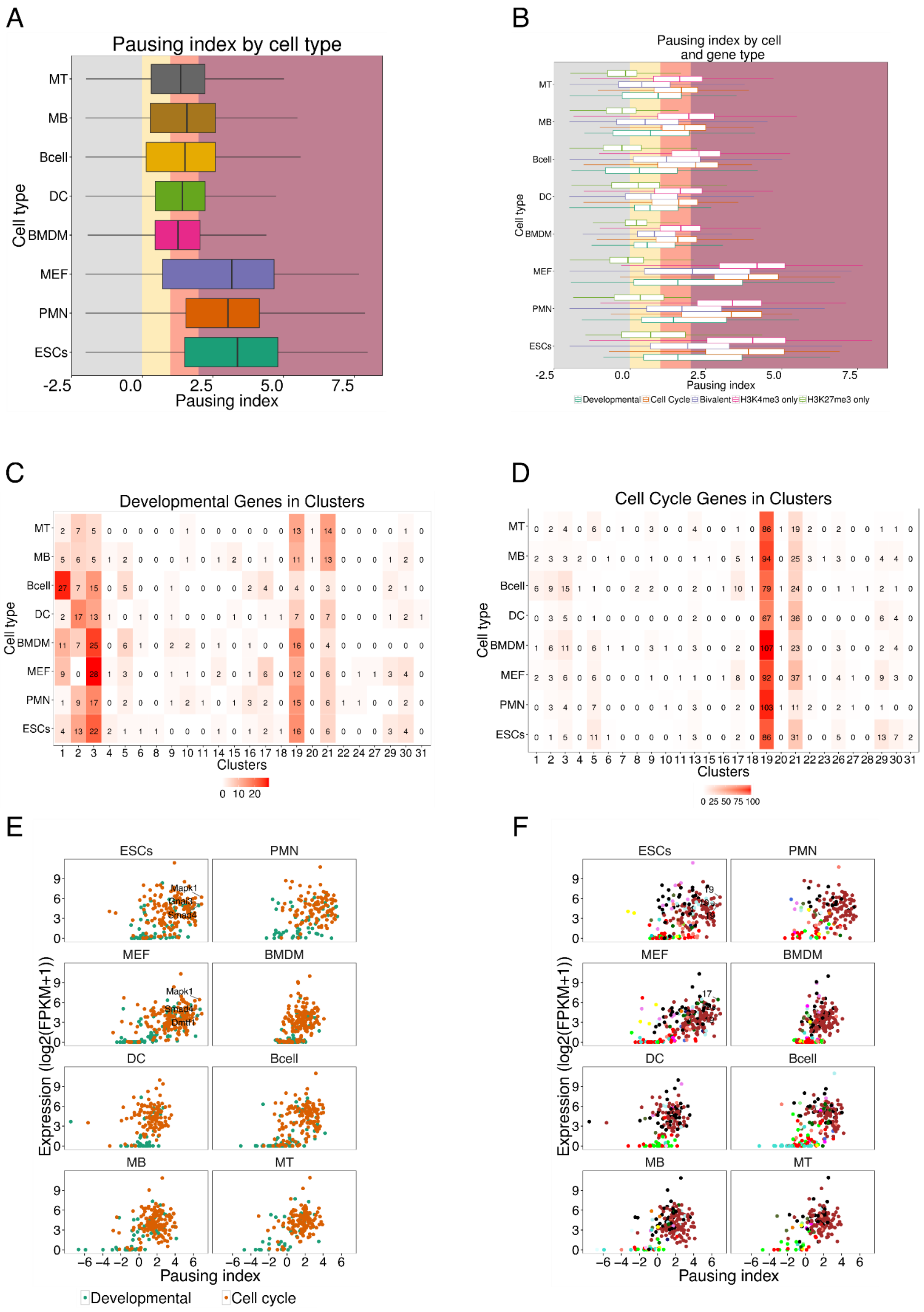


Figure 5.17 A) Boxplot of the pausing index distribution across the 8 cell types used in the study. Cells were ordered according from the highest to lowest mean pausing index value (the line at the middle of the boxplots denotes the median of the distribution). B) Boxplots of the pausing index distribution across all cell types, only for the genes that belonged to the following categories according to gene ontology: 1) Developmental

(GO:0045165, GO:0048864, GO:0007498), 2) Cell cycle (GO:0007049), 3) Bivalent (having both H3K4me3and H3K27me3 in this study), 4) H3K4me3 only and 5) H3K27me3 only. C) Distribution of developmental genes across clusters and cell types. D) Distribution of cell cycle genes across clusters and cell types. E) Scatterplots of RNAPII pausing vs expression (log2 (FPKM+1)) values for developmental and cell cycle genes across all cell types. F) Scatterplots of RNAPII pausing vs expression (log2 (FPKM+1)) values for developmental and cell cycle genes across all cell types. The colours imposed display the respective cluster colour where each gene belonged.

5.4.6 Discussion

In this study, we have conducted integrated analysis of epigenetic marks (H3K4me3, H3K27me3), RNAPII (8WG16) binding and expression (RNA-seq) in eight mouse cell types of variable developmental potential. Hierarchical clustering of cell types for any of the datatypes did not fully overlap with each other demonstrating that all four data types bring non-redundant information into the downstream analysis.

Considering a fixed region around the promoter of each gene would certainly impact the transcriptional signal due to variable numbers of intronic regions among individual genes. This could cause variable levels of average read signal, not necessarily reflecting biological variation. For that reason, it would be wiser if we have used the total FPKM value for each gene, rather than the transcriptional signal at the promoter. But because we were interested in RNAPII dynamics, we decided to go on with the first approach, recognising that the average transcriptional signal for each gene might be obscured by the concentration of intronic regions at the +5 kb region after the promoter. However, there have been cases in the literature, where RNA-seq reads map to intronic sequences, for almost 40% of the total number of reads (Ameur et al., 2011; Gaidatzis et al., 2015). This could make our analysis more robust, given that the percentage of reads mapping to introns versus exons would range in comparable levels across the samples.

The hierarchical tree using H3K27me3 modification at promoters was greatly discordant with the known developmental relationships between cell types. Specifically, B cells had a very large number of H3K27me3 marked promoters. The average number of H3K27me3 peaks was not uniformly distributed (Figure 5.8J) and this was not a result of ChIP sequencing quality, estimated by the signal at the common peaks (Figure 5.2K). This suggests that there might be a real biological difference separating B cells from the rest of the cell types.

We clustered profiles of silencing (H3K27me3) and activating (H3K4me3, RNAPII and RNA-seq) signals across gene promoters for all the cell types in our study. This resulted into nine major profile sub-groups, comprising of 31 distinct clusters. The active expression patterns (clusters 19, 20, 21) were retained to a very high degree among cell types, followed by the bi-directional expression patterns (cluster 26) and lastly the bivalent patterns (cluster 3 and 24) tended to be cell type specific (Figure 5.11 and Table 5.5).

The epigenetic and expression state transitioning between cell types was divided predominantly in two groups. H3K27me3-only cluster genes in one cell type were likely to be bivalent in another cell type and active cluster genes in one cell type were likely to be boundary H3K27me3 in another cell type (Figure 5.12A). There were only a handful of cases transitioning between bivalent and active states (Bcell_3 ↔ ESCs_17, Bcell_3 ↔ MEFS_17, Bcell_3 ↔ BMDMS_17, MTs_13 ↔ MBs_21 and MTs_13 ↔ MEFS_21) implying that bivalency is not the predominant intermediate state for switching on or off transcription during differentiation. B cells, unlike other cell types, were highly enriched for H3K27me3 either at the core promoter or at the boundary upstream the TSS. Further investigation is needed to know whether B cells are more prone to acquire or retain longer Polycomb silencing than other cell types and what are the mechanisms involved.

Hierarchical clustering of TF enrichment for clusters revealed the major profile sub-groups for the highly-active clusters (19, 21, 29, 31), two of the bi-directional expression clusters (20, 26) and some of the bivalent clusters (8, 16, 11, 12, 9) (Figure 5.14A). Clusters 3 and 5 belonged to different sub-groups but had closely associated TF binding patterns. Interestingly, they shared similar binding patterns for most transcription regulators and a common de-novo motif (‘TCCCC’) which was previously recognised at the sequences of high confident (HC) bivalent promoters in human and mouse ESCs (Mantsoki et al., 2015). Of note, they also were bound uniquely by some TFs and bivalent-wide-H3K27me3 cluster 3 was uniquely enriched for binding of Nanog, Oct4 and p300 (Table 5.6). Our results suggest that the bivalent genes in cluster 3 are involved in developmental regulatory functions across multiple cell types and are possibly directly affected by active pluripotency and signalling factors (Figure 5.16C, Table 5.6) that exhibit RNAPII at their promoters (Williams et al., 2015). In contrast, bivalent genes in cluster 5 are more tissue-specific and show higher levels of expression possibly due to transcription leaking (De Gobbi et al., 2011).

Cluster	Unique TFs enriched
Cluster3	Rbbp5, Pou5f1, Sin3A, Nanog, Tcf3, Prep, Chd4, HoxB4, Cbx8, Runx1, Rbpj, Gfi1b, PU.1, Mxi1, Sfpi1, Cebpb, Runx2, Jun, p300
Cluster5	Ebf1, Ldb1, Stat5B, E2f1, Irf1, Maff, Rel, Egr1, Rad21, Stat3, Ncoa3, Tfe3, Sox2, Tal1, Ctcf, Cebpa, Rela, Meis1, Scl, Gata6, Eto2, Ldb1, Mtgr1, Ascl2, Fosl2

Table 5.6 Unique TFs enriched specifically in Cluster 3 or Cluster 5 (P value < 0.05)

We used ChIP-seq data for RNAPII (8WG16) which shows a very high overlap with global run-on sequencing (GRO-seq) (Core et al., 2008) data, used for the calculation of RNAPII pausing (Williams et al., 2015). RNAPII pausing index was not highly correlated with the expression across clusters, however the majority of active clusters exhibited a highly paused-highly expressed configuration (Figure A and B). Bivalent clusters showed mid-pausing levels followed by H3K27me3-only clusters at low or no pausing at all (Figure 5.16A). ESCs, MEFs and PMNs had persistently higher pausing indices than the rest of the cell types independently of gene type (Figure 5.17A). MEFs show high concordance with ESCs both in terms of expression and pausing, indicating that they retain their ES-like characteristics more than other cell types of similar developmental hierarchy (Yusuf et al., 2013). MEFs are frequently used in induced pluripotent stem cells (iPSCs) experiments (Takahashi and Yamanaka, 2006), thus it would be very interesting to assess the utility of pausing index in suggesting candidate cell types for reprogramming research.

In conclusion, we have successfully integrated omics datasets of epigenetic marks, transcription factors and gene expression, conducting a promoter-level analysis for eight cell types of different developmental potentials. We have functionally characterised nine major promoter profile signals and the transcriptional control that governs them, offering a valuable resource for further studies in the regulation of transcription during development. Future work would entail the incorporation of multiple histone modifications and components of complexes involved in RNAPII pausing, investigating the signatures at promoter and enhancer elements as well.

Chapter 6 Discussion

6.1 Overview of data analysis and challenges in data integration from multiple sources

The characteristics of bivalent chromatin (H3K4me3 and H3K27me3) have been assessed by many studies across mouse and human cell types mainly through ChIP-seq protocols (Adli et al., 2010; Barski et al., 2007; Cui et al., 2009; Ku et al., 2008b; Mikkelsen et al., 2007; Pan et al., 2007; Zhao et al., 2007). Most of the studies have generated lists of bivalent genes in isolation and their pairwise comparisons have not yielded high overlap (Marks et al., 2012). Here, we systematically assessed the degree of overlap among available published studies in search of a robust list of bivalent promoters. Direct comparison of bivalent gene sets across studies is challenging, due to different experimental protocols and subsequent data analysis steps followed by each lab. Due to low sample numbers we could not assess systematically the role of different antibodies in the detection of histone mark enriched regions and bivalency per se. Therefore, the choice of a reliable method for the detection of enriched regions bearing the histone marks under investigation is important. The principal aim of our analysis was to detect similarly enriched regions across multiple samples. However, most of the algorithms already developed, employed strenuous quantitative comparisons for the identification of differentially enriched regions between pairs of samples usually after a peak-calling step (Heinig et al., 2015; Schweikert et al., 2013; Shao et al., 2012).

Peak-calling (Kharchenko et al., 2008; Xu et al., 2014; Zhang et al., 2008) is the most widely used method in the literature where enriched regions (with respect to background noise) are assigned with a level of confidence (p-value). In this study we have used a peak-based method, being aware that we might have a large number of false positives due to the great range of read depth across the samples, which incorporates variant signal to noise (S/N) ratios in the results. To adjust for the variability in the read depth across samples we would have to limit the number of reads in all samples similar to the sample with the fewest number of reads for each histone mark. This would lead to undesirable loss of information, thus we developed a complementary cutoff-based method, initially generating the normalised read coverage (reads per million) for each sample which was subsequently quantile normalised across samples. There was a clear signal for promoters highly enriched in H3K4me3, after fitting the bimodal distribution on the normalised H3K4me3 read density distribution. Unfortunately, H3K27me3 normalised read density was not following the same pattern and a set of arbitrary thresholds were chosen.

The limitation of the cut-off approach is that it does not take into consideration the unique pattern of H3K27me3 which, as it became apparent in later stages of our analysis, is found in various forms across the mammalian promoters and cell types. Retrospectively, having confirmed that H3K27me3 is the defining mark for bivalency and the majority of H3K27me3 promoters seem to be bivalent in ESCs, we could have employed a different method uniquely applying to the H3K27me3 characteristics. For instance, after peak calling we would examine the normalised signal of H3K27me3 in common peaks across the samples and divide it into two distributions depending on whether the peaks would be bivalent or H3K27me3-only. Then we would define thresholds as in the cut-off based approach, with the benefit of having two separate distributions with uniform characteristics (H3K27me-bivalent and H3K27me3-only). Finally, we would apply the thresholds on the entire set of reads in each sample after isolating the bivalent peaks from the H3K27me-only peaks. This proposed method is based on the assumption that H3K27me3 in bivalent regions is deposited under the same mechanisms by PRC2, hence the H3K27me3 read distribution should be identical. On the contrary, H3K27me3-only regions show wider and less enriched signals, suggesting a different mechanism from the one at bivalent promoters.

The peak-based method generated better results than the cut-off, where a higher number of bivalent promoters was overlapping with bivalent promoters from previous studies (Mikkelsen et al. 2007) and developmental factors based on gene ontology enrichment. The cumulative intersection of peaks across samples, corrected to a certain degree the bias of outlier peaks in each experiment, deriving either from antibody specificity or read depth sensitivity. Nevertheless, an alternative explanation for the high replication of previous results might be that we use a similar approach to detect histone mark enrichment from the same primary material (ChIP-seq experiment).

We have identified high confidence promoters bearing both marks (HC bivalent), with their expression at low levels in accordance with the literature. The low expression level of genes with bivalent promoters has been attributed to transcription leaking at those genes due to failure of Polycomb repressors in silencing them efficiently (De Gobbi et al. 2011) or due to mixed cell populations, where some cells express them and others do not express them at all (Brookes et al. 2012). To address this issue we used single cell transcriptomics data from human and mouse ESCs. Single cell RNA-seq (scRNA-seq) can capture the variation of expression across individual cells from a cell population (Pan 2014), yet high technical noise, due to low starting material and high amplification bias, overshadows the real biological variation (Marinov et al. 2014; Stegle et al. 2015).

Technical variation is anti-correlated with the mean expression of genes across cells, thus we chose sets of consistently highly expressed genes and measured their correlation with windows of genes of declining mean expression. Our correlation-based approach could separate and discard the genes with extreme technical variation, however it suffers from one major caveat. The characterisation of a portion of bivalent genes in terms of their expression variation is hindered, since the majority of them are lowly expressed, hence not adequately captured by the scRNA-seq techniques and consequently discarded when they do not meet our thresholds. Recent studies have implemented more sophisticated statistical methods taking into account the confounding factor of expression level and increasing the power of the results with larger populations of single cells. A deconvolution approach deploys a distance to the median (DM) as a gene expression variability measure, which does not depend as strongly on the mean expression levels of each gene as the CV (Kolodziejczyk et al., 2015). A noise decomposition method was applied for the assessment of noise in allele stochastic expression, accounting for cell-to cell variability and dividing biological from technical variation with the assistance of spike-in controls (Kim et al., 2015). Our method could potentially evolve and take into account more confounding factors that cause noise. However, no further analysis was done in this project since it was beyond the scope of bivalent chromatin.

To investigate the dynamics of bivalent chromatin in combination with expression patterns and RNAPII pausing we integrated ChIP-seq and bulk RNA-seq data in 8 murine cell types of diverse developmental potential. Similar with the concerns raised for the integrative analysis in ESCs, different read sequencing depths, antibodies used, and protocols across studies, have rendered the accurate division of promoter profiles a challenging task. The clusters were diversely populated both in terms of numbers of genes and cell type representation. This phenomenon could arise either due to real biological difference or because of experimental artefacts such as read coverage bias. Average profiles of binary peaks were compared with the normalised read numbers at the common peaks across all ChIP-seq samples and cell types. The apparent variation in the average number of peaks was not correlated with the signal at the common peaks, even for H3K27me3 which was the most divergent mark in the analysis. The issue raised above could have been prevented if we had detected clusters of chromatin and expression patterns separately for each cell type. Thus, we could have also identified cell type specific states that would not be burdened with the bias of variable read depth and would not be discarded due to limited numbers of genes. On the other hand, integration of similar profiles across cell types would be computationally costly due the large numbers of pairwise comparisons between states.

The meta-analysis has generated results which were in accordance with previously published results, adding value to the reproducibility of the data. We also made new observations improving our understanding of the studied mechanisms, as discussed below.

6.2 Chromatin status and expression signatures across species and cell types

A systematic comparison of chromatin signatures between human and mouse ESCs, using orthologous genes, showed that more than half of HC bivalent promoters retain the same chromatin configuration and are far more enriched for developmental regulators than the species specific bivalent promoters. The conserved set of bivalent promoters across mammalian ESCs implies that bivalent chromatin is an important feature of epigenetic regulation accompanied by the other conserved patterns such as H3K4me3-only and latent. Gene expression levels between species were highly maintained, irrespective of whether their chromatin state was conserved or divergent.

For example, orthologous genes that were detected as latent in human and H3K4me3 only in mouse, maintained equivalent expression profiles similar to active genes. This could mean that expression of actively transcribed genes in human could be regulated by a more complex network of histone modifications (i.e. H3K27ac, H3K9,14Ac, H3K79me3) (Guenther et al., 2007; Kouzarides, 2007) and TFs that were not incorporated in our study, hence they were detected as latent. Also, deposition of activating histone modifications could be happening after, and thus be a consequence of, transcription initiation induced by RNAPII recruitment, and the transcriptional machinery could later preserve those marks at the promoter of the locus for as long as gene transcription takes place (Rybtsova et al., 2007). However, it remains elusive whether these differences in chromatin states constitute species specificity or just differences between the pluripotent states of mouse and human ESCs. Human ESCs seem to be closer to the primed state of epiblast stem cells (EpiSCs) rather than the naïve pluripotency of mouse ESCs (Takashima et al., 2014). Further work using ESCs cultured with media that induce naïve pluripotency features in mouse (Marks et al., 2012) and human (Chan et al., 2013; Gafni et al., 2013; Hanna et al., 2010; Theunissen et al., 2014; Ware et al., 2014), could shed a light on the degree of conservation of the bivalent state across mammalian ESCs.

In the promoter-wise analysis of 8 developmentally distinct cell types, 31 clusters emerged that could be narrowed down to 9 major profile subgroups. The ratio of unique to total number of genes in each cluster, along with the maximum parsimony trees built for each cluster, have yet again pointed to the conservation of three major profiles: the active profile, the bivalent profile and the bi-directional expression profile which lacked any of the two histone marks included in this study, and could be associated with the latent signature. Interestingly, the bi-directional expression profile was enriched for processed pseudogenes but was mainly void of gene type annotation, suggesting that we ought to elucidate the function of this unknown set of genes.

We have also systematically studied the major chromatin transitions across cluster and cell types. Only 2.2% of possible promoter transitions showed a significant number of genes overlapping, mainly representing promoters moving across clusters of very similar profiles. Outside the spectrum of profile similarity, gradual transitions from H3K27me3-only state through to bivalent state and finally to highly active state, were again not reflected at gene expression, where comparable levels of promoter transcriptional signals were noted throughout the comparisons.

The apparent conservation of the three chromatin patterns between cell types of the same species but also across species, supports their high degree of involvement in the transcriptional regulation. However, since the expression profiles seem to be quite unaffected we could postulate that 1) the chromatin marks in our study seem to be a consequence of the transcriptional activity at the promoter locus rather than a direct cause that recruits the transcription initiation machinery, and/or 2) that these chromatin marks recapitulate only a subset of the transcriptional information (which could be assessed by integrating other histone marks and non-histone epigenetic marks as well as TFs).

6.3 Different occupancy levels of Polycomb components and RNAPII could be reflected by the expression variance of bivalent genes

PcG complexes (PRC1 and PRC2) are the main repressive regulators of bivalent genes (Lewis 1978; Kennison 1995; Schuettengruber et al. 2007; Margueron & Reinberg 2011) and are considered to be recruited in a sequential manner at the loci about to be silenced, where PRC2 deposits H3K27me3 (Czermin et al. 2002) which consequently assists PRC1 to be recruited and establish a more robust silenced chromatin structure with H2Aub deposition (Endoh et al., 2012; Stock et al., 2007b). Bivalent promoters have been divided in PRC2-only and PRC1-PRC2, with the ones bound by both complexes considered to be more conserved across species (Ku et al., 2008b). Having assessed the binding profiles of PRC1 and PRC2 components in HC bivalent promoters of mouse ESCs, all HC bivalent promoters were bound by Suz12 and Jarid2 (PRC2), Ring1b and Cbx7 (PRC1), albeit in different levels. We could thus postulate that previously identified PRC2-only promoters are a group of bivalent promoters where low levels of PRC1 were simply considered insignificant due to low sequencing depth of the original experiments. Furthermore, incorporating H2Aub has also validated that all HC bivalent promoters in mouse ESCs are enriched for this histone modification which is a result of PRC1 recruitment.

ChIP-seq profiles of Polycomb components and RNAPII divided bivalent promoters into distinct groups with variable epigenetic signatures, expression levels and functional outputs, with similar patterns emerging at mouse ESCs (Chapter 3) and at the promoter-wise hierarchical clustering across cell types. The TF binding enrichment analysis uncovered very similar binding patterns for the bivalent clusters 3 and 5 (Chapter 5). However, those two clusters were also uniquely bound by some TFs. In addition, their differences in RNAPII occupancy led us to hypothesize that these gene sets might be sensitive to different signalling pathways. Also, a set of HC bivalent promoters was characterized by highly variable expression (using single cell RNA-seq in Chapter 4), and was associated with response to DNA damage and DNA repair. Furthermore, highly variable genes formed tight co-expressed clusters in only one or a few single cells.

The above findings suggest that a fraction of bivalent genes are involved in more general developmental regulatory functions while others are more tissue-specific. The high variance in expression among single ESCs supports the model of ESCs moving transiently between developmental states but not committing. Our results and recent findings in the literature (Illingworth et al., 2016) support that ESCs are free to explore potential differentiation pathways with the assistance of bivalent genes, but not commit until certain developmental and environmental cues reach the appropriate thresholds.

6.4 Transcriptional control is tighter in active than bivalent promoters and sequence motifs are conserved between species

A possible model proposes that the dissimilarity in TF occupancy between active and bivalent loci is responsible for the uninhibited activity of the PRC complexes that leads to the generation of bivalent promoters (Voigt et al., 2013). High abundance of TFs at the active promoters does not allow the Polycomb components to decorate the histone tails with repressive marks and eventually silence the adjacent genes. In

Chapter 3 we show that indeed TF density at the promoter sites decreases as we move from active (H3K4me3-only) to bivalent and last H3K27me3-only promoters, in both human and mouse ESCs. Moreover, active promoters are exclusively bound by pluripotency regulators such as Klf4, Essrb, Pou5f1 (Oct4), Sox2 and Nanog, whereas bivalent promoters are mostly targeted by Polycomb (Ezh1, Suz12, Cbx7, Ring1b) and Polycomb-like (Mtf2, Phf9) components, and Utf1 which is considered to shield away transcription leaking from bivalent loci (Jia et al., 2012b). In Chapter 5, bivalent sub-groupings were less populated by TFs (~ ¼ of the total TFs in the analysis), displaying a largely dissimilar set of TFs from the highly active clusters, which were occupied by more than ¾ of the TFs used in the analysis.

The above findings suggest that active promoters are subject to a tighter transcriptional control than bivalent ones. What is more, bivalent genes highly overlapped (>90%) with differentially expressed genes (up or down regulated) when a set of 91 TFs were over-expressed or knocked-down (each TF separately) in mouse ESCs. It seems that active promoters are regulated by a highly abundant network of TFs with redundant functions, which safeguards their expression equilibrium. This redundancy could dampen the dependency of active genes to only a small set of TFs, in contrast with bivalent genes that exhibit a high degree of volatility in their expression when the function of only one TF is disrupted.

It is intriguing how those two classes of gene promoters (bivalent and active), which are both highly abundant in CpG islands, are distinguishably bound by such a diversified set of TFs and consequently acquire distinct properties. De-novo motif analyses in our studies have successfully detected underlying sequence motifs that could uniquely identify these two promoter classes. Active promoters in ESCs were consistently enriched for GC-rich ETS factor motifs (‘CCGGAA’) and bivalent promoters showed enrichment for a motif conserved across species which contained 4 consecutive cytosine bases (‘TCCCC’). Further research is needed to understand if these motifs are of any biological relevance, especially for the case of bivalent promoters. The similarity of the bivalent promoters’ motif to that of MZF1 (Morris et al., 1994) was not assessed any further, due to poor results after assessing for binding sites for several zinc finger proteins from the HEK293 cell line (Najafabadi et al., 2015). It would be rather interesting to perform the same analysis using ChIP-seq experiments deriving from ESCs, which could potentially uncover a novel role for MZF1, specific to the early stages of development.

As it was previously discussed, bivalent genes seem to be more likely to show discrepancies in their expression than active genes, even when the pluripotency network is minimally distressed. It could be argued that Polycomb is the facilitator of this behavior since it is uniquely enriched at bivalent sites. However, the underlying mechanisms that control and define bivalent genes' sensitivity are still to be determined. In the years to come epigenome editing techniques could target specifically the histone modifications at bivalent promoters and probe the function of the complexes that regulate them (PcG, TrxG) independently of the modifications that they catalyze (Voigt and Reinberg, 2013). More interestingly, knock out of chromatin modifiers specifically at bivalent regions could prevent the out of target genome-wide effects that obscure their real association with bivalency.

6.5 H3K27me3 only promoters have lower CpG density than the active and bivalent promoters and show stronger signatures at cells of lower developmental potential

Over 85% of H3K27me3 marked promoters in ESCs co-incited with H3K4me3, suggesting that bivalency is the default signature for Polycomb silenced promoters in ESCs. It has been shown that high CpG density, un-methylated CpG islands (CGIs) can sufficiently recruit Polycomb components in vertebrates (Farcas et al., 2012; Mendenhall et al., 2010; Riising et al., 2014) and thus define the presence of H3K27me3 in the locus. Moreover, incorporation of synthetic G+C-rich, CpG-rich DNA sequences in the ESC genome has resulted in creation of bivalent domains, reinforcing their status as the default chromatin state of CGIs in ESCs (Wachter et al., 2014). In a cross-species analysis between human and mouse ESCs both in our study and in the literature (Lynch et al., 2012), there was clear CGI erosion correlating with loss of both H3K27me3 and H3K4me3 at the mouse genome, while no CpG-rich bivalent promoters in mouse were eroded in human. Nevertheless, in a small fraction of genes, absence of a CGI in mouse was not followed by the loss of bivalent signature. We could thus argue that bivalency can act as a proxy for the recognition of CGIs in ESCs, when CpG density levels do not pass the arbitrary thresholds necessary for CGI detection.

HC bivalent promoters were most enriched for CGIs, along with the H3K4me3-only promoters. Surprisingly, none or few H3K27me3-only promoters contained a CGI, disagreeing with the model where PcG proteins can be attracted only by CGI promoters. Low numbers of HC H3K27me3-only promoters in ESCs restrained the power for subsequent analysis, so we were very interested to discover that there was a major sub-group marked only by H3K27me3 in our promoter-wise multiple cell type analysis. H3K27me3-only clusters were amongst the clusters with the lowest CpG density and only about half of them overlapped with a known CGI. Higher numbers of identified H3K27me3-only sites have revealed that not all of them lack a CGI, but the findings still suggest that presence of H3K27me3 on CpG-poor promoters could involve an unknown mechanism of Polycomb recruitment, not dependent on CGIs.

B cells, MEFs and BMDMs, were over-represented in the H3K27me3-only cluster 1 and the higher average number of H3K27me3 peaks at those cell types did not agree with the average H3K27me3 signal patterns at the common peaks across cell types. This suggests that there might be a real biological variation where non-pluripotent cell types tend to have H3K27me3 domains (not necessarily accompanied by H3K4me3) which occupy wider regions in the genome. Indeed, studies have shown that the refractory nature of pluripotent chromatin does not allow Polycomb to silence broad domains of the genome and it gets concentrated at the bivalent promoters (Hawkins et al., 2010; Xie et al., 2013; Zhu et al., 2013). In contrast, there is a genome-wide repopulation of the chromatin by H3K27me3 at differentiated cell types, reflecting higher Polycomb efficiency. There is also supporting evidence that some CpG-poor promoters bear only H3K27me3 and are largely found in broad intergenic H3K27me3 domains (Xie et al., 2013), which is in accordance with our results.

6.6 RNAPII pausing levels in poised and active promoters correlate with the developmental potential of the cell type

Our understanding of the diverse aspects of bivalent chromatin can be enlightened by examining its close association with specific variants of phosphorylated RNAPII. More specifically, RNAPII phosphorylated at Serine 5 (S5) was found at the promoters of bivalent genes, indicating that those genes were poised for transcription but the engaged RNAPII could not proceed to productive elongation (Brookes et al. 2012; Brookes & Pombo 2009). RNAPII promoter proximal pausing (Guenther et al. 2007; Muse et al. 2007; Zeitlinger et al. 2007; Krumm et al. 1995) occurs at a wide variety of genes with distinct expression levels and biological functions and it is not solely related with poised chromatin (Guenther et al. 2007; Min et al. 2011; Williams et al. 2015). We have previously shown that RNAPII variants can classify bivalent promoters in distinct groups of variable expression and Polycomb component occupancy (Brookes et al. 2012). Recently, more studies have shed a light on the role of RNAPII pausing in naïve ESCs, where increased levels of RNAPII (Marks et al. 2012) are accompanied by high pausing at cell cycle genes and not in poised developmental genes as previously thought (Williams et al. 2015).

In chapter 5, we incorporated ChIP-seq data from RNAPII (8WG16) in 8 cell types and calculated the RNAPII pausing at the clusters that emerged, but also at a cell type average approach. GRO-seq is the most commonly used type of data for the reliable calculation of RNAPII pausing (Core et al. 2008). Unfortunately, not all the cell types in our study had available GRO-seq experiments. However, the antibody we have used (8WG16) can recognize multiple variants of RNAPII that are not heavily phosphorylated, is considered to highly overlap with GRO-seq data (Williams et al. 2015) and is thus appropriate for the calculation of RNAPII pausing index.

High pausing was more associated with H3K4me3 and high expression signal at the promoter sites, rather than bivalent chromatin and intermediate expression. Surprisingly, when expression was calculated for the whole gene transcript (FPKM), there was a moderate correlation with RNAPII pausing. This could be an indication that highly paused genes do not produce a fully functional transcript (Guenther et al.,

2007) which leads to lower numbers of reads mapping to the exons of the gene body. Moreover, we confirmed that cell cycle genes exhibit higher pausing at their promoter than developmental genes. There are striking similarities in the pausing levels of cell cycle genes with the ones belonging to the H3K4me3 only state and developmental genes with the bivalent state. This result however needs to be validated with more data.

Genes in ESCs, PMNs and MEFs showed higher levels of pausing than in the other cell types. The calculation of the average RNAPII pausing in each cell type demonstrated once more consistently higher levels of pausing for these three cell types. Since ESCs, PMNs and MEFs were the cell types closer to pluripotency in our analysis, we argued that this might provide an explanation for this phenomenon, hence we examined their ES-like characteristics evaluating expression and pausing of genes involved in the pluripotency network of mouse ESCs. There too, ESCs, PMNs and MEFs showed the highest levels of pausing for pro self-renewal, pro-differentiation and ES-signalling genes.

It is quite intriguing that mammalian genes bearing active chromatin signatures and being involved in cell cycle would be subject to this type of transcriptional regulation such a RNAPII pausing which is mainly a control step for developmental genes in *Drosophila* (Muse et al., 2007; Zeitlinger et al., 2007). There is a striking similarity in the pausing patterns of progenitor cells which supports further that RNAPII pausing is deployed by precursor cells as a means to preserve their pluripotent characteristics. MEFs in particular are one of the main cell types used in reprogramming experiments of iPSCs (Takahashi & Yamanaka 2006). Further possible candidate cell types could be inferred if we systematically assess pausing index similarities of important genes involved in the pluripotency network.

6.7 Future research

This study is a meta-analysis of available datasets looking into epigenetic regulation during development with a special focus on the bivalent chromatin. We have so far discussed the challenges we have met in our effort to integrate multiple datasets produced from diverse high throughput experimental techniques and have addressed the major findings. Bivalent promoters in ESCs seem to act as a shelter from an imminent DNA methylation wave that could silence them indefinitely (Voigt et al., 2013) and would not allow them to respond to developmental cues. The major caveat in our analysis is the inconsistencies between the experimental protocols used across the studies we have integrated. The inherent heterogeneity of ESCs (Carter et al., 2008; Graf and Stadtfeld, 2008) should be addressed through comparisons of ESCs grown both in 2i and serum media. As more datasets become available from ESCs grown in 2i media, it could be feasible to detect robust lists of bivalent promoters and compare them with the HC confident promoters in our study.

Another way to address heterogeneity would require the integration of single nucleosome combinatorial chromatin immunoprecipitation techniques (Sadeh et al., 2016; Weiner et al., 2016) with single cell transcriptomics (Tang et al. 2009; Islam et al. 2011; Ramsköld et al. 2012). Cell population variability should be simultaneously assessed with cell-intrinsic variability where allelic differences might influence the resolving of bivalent promoters into monoallelically expressed genes. Combination of prediction of monoallelic expression through molecular signatures and quantification of allele-specific expression of bivalent genes in multiple cell clones (allele specific targeted sequencing – AST-Seq) (Nag et al., 2013) would assist in the discovery of a reliable list of bivalent/random monoallelic expressed genes.

In terms of the functionality and establishment of bivalency, our analysis indicates that expression of the adjacent genes seems to be crucial for the deposition of the respective marks. There is growing evidence suggesting that bivalent genes are regulated with the assistance of miRNAs (Graham et al., 2016) and lncRNAs that bind with high affinity to PRC2 components leading them to the silenced to be loci (Rinn et al., 2007). It is possible that the motif we have discovered here be a common denominator that facilitates the recruitment of regulatory RNAs. Recognition of ncRNA hybridization sites at the transcripts produced by bivalent genes could potentially highlight prospective candidates.

Chromatin organization changes radically during differentiation (Phillips-Cremins et al., 2013) and mapping of epigenetic states of promoters should definitely be combined with enhancer regulatory elements. Chromosome conformation techniques (Dostie et al., 2006) will be instrumental for the discovery of alternating chromatin loop configurations that take place depending on cell fate commitment trajectories. The transcriptional machineries recruited at the interaction points between promoters and enhancers, involve numerous TFs and chromatin modifiers whose function is still mostly undetermined. Orchestration of more large-scale studies encompassing chromatin regulators, histone modifications, transcription initiators and consequently gene expression, could potentially interpret the causal relationship of epigenetic marks and transcription.

Bibliography

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science *252*, 1651–1656.

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat. Rev. Genet. *13*, 720–731.

Adelman, K., Marr, M.T., Werner, J., Saunders, A., Ni, Z., Andrulis, E.D., and Lis, J.T. (2005). Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. Mol. Cell *17*, 103–112.

Adli, M., Zhu, J., and Bernstein, B.E. (2010). Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat. Methods *7*, 615–618.

Akalin, A., Fredman, D., Arner, E., Dong, X., Bryne, J.C., Suzuki, H., Daub, C.O., Hayashizaki, Y., and Lenhard, B. (2009). Transcriptional features of genomic regulatory blocks. Genome Biol. *10*.

Akdemir, K.C., Jain, A.K., Allton, K., Aronow, B., Xu, X., Cooney, A.J., Li, W., and Barton, M.C. (2014). Genome-wide profiling reveals stimulus-specific functions of p53 during differentiation and DNA damage of human embryonic stem cells. Nucleic Acids Res. *42*, 205–223.

Alder, O., Lavial, F., Helness, A., Brookes, E., Pinho, S., Chandrashekran, A., Arnaud, P., Pombo, A., O'Neill, L.P., Azuara, V., et al. (2010). Ring1B and Suv39h1 delineate distinct chromatin states at bivalent genes during early mouse lineage commitment. Development *137*, 2483–2492.

Alexa, A., and Rahnenführer, J. (2016). Gene set enrichment analysis with topGO.

Aloia, L., Di Stefano, B., and Di Croce, L. (2013). Polycomb complexes in stem cells and embryonic development. Dev. {(Cambridge,} England) *140*, 2525–2534.

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., and Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat. Struct. Mol. Biol. *18*, 1435–1440.

Ang, Y.-S., Tsai, S.-Y., Lee, D.-F., Monk, J., Su, J., Ratnakumar, K., Ding, J., Ge, Y., Darr, H., Chang, B., et al. (2011). Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. Cell *145*, 183–197.

Anguita, E., Hughes, J., Heyworth, C., Blobel, G.A., Wood, W.G., and Higgs, D.R. (2004). Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. EMBO J. *23*, 2841–2852.

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

Ballaré, C., Lange, M., Lapinaite, A., Martin, G.M., Morey, L., Pascual, G., Liefke, R., Simon, B., Shi, Y., Gozani, O., et al. (2012). Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. Nat. Struct. Mol. Biol. *19*, 1257–1265.

Bannister, A., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res. *21*, 381–395.

Bannister, A.J., Schneider, R., Myers, F.A., Thorne, A.W., Colyn, C.-R., and Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. J. Biol. Chem. *280*, 17732–17736.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. *41*, D991–D995.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by {DNA} and histone methylation. Cell *143*, 470–484.

Bateman, J.R., Johnson, J.E., and Locke, M.N. (2012). Comparing enhancer action in cis and in trans. Genetics *191*, 1143–1155.

Benaglia, T., Chauveau, D., Hunter, D.R., and Young, D. (2009). {mixtools}: An {R} Package for Analyzing Finite Mixture Models. J. Stat. Softw. *32*, 1–29.

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

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

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *28*, 1045–1048.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. Science *306*, 2242–2246.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Birnbaum, R.Y., Clowney, E.J., Agamy, O., Kim, M.J., Zhao, J., Yamanaka, T., Pappalardo, Z., Clarke, S.L., Wenger, A.M., Nguyen, L., et al. (2012). Coding exons function as tissue-specific enhancers of nearby genes. Genome Res. *22*, 1059–1068.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Stamatoyannopoulos, J.A., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Blackledge, N.P., Rose, N.R., and Klose, R.J. (2015). Targeting Polycomb systems to regulate gene expression: modifications to a complex story. Nat. Rev. Mol. Cell Biol. *16*, 643–649.

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. (2007). Divergence of Transcription Factor Binding Sites Across Related Yeast Species. Science (80-.). *317*.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349–353.

Bracken, A., Dietrich, N., Pasini, D., Hansen, K., and Helin, K. (2006). Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes Dev. *20*, 1123–1136.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

Brien, G.L., Gambero, G., O'Connell, D.J., Jerman, E., Turner, S.A., Egan, C.M., Dunne, E.J., Jurgens, M.C., Wynne, K., Piao, L., et al. (2012). Polycomb PHF19 binds H3K36me3 and recruits PRC2 and demethylase NO66 to embryonic stem cell genes during differentiation.

Nat. Struct. Mol. Biol. *19*, 1273–1281.

Brookes, E., and Pombo, A. (2009). Modifications of RNA polymerase II are pivotal in regulating gene expression states. EMBO Rep. *10*, 1213–1219.

Brookes, E., de Santiago, I., Hebenstreit, D., Morris, K.J., Carroll, T., Xie, S.Q., Stock, J.K., Heidemann, M., Eick, D., Nozaki, N., et al. (2012a). Polycomb associates genome-wide with a specific {RNA} polymerase {II} variant, and regulates metabolic genes in {ESCs.}. Cell Stem Cell *10*, 157–170.

Brookes, E., de Santiago, I., Hebenstreit, D., Morris, K.J., Carroll, T., Xie, S.Q., Stock, J.K., Heidemann, M., Eick, D., Nozaki, N., et al. (2012b). Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. Cell Stem Cell *10*, 157–170.

Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B., and Brom, T.H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. *33*, 155–160.

Cao, R., and Zhang, Y. (2004). {SUZ12} is required for both the histone methyltransferase activity and the silencing function of the {EED-EZH2} complex. Mol. Cell *15*, 57–67.

Cao, K., Lailler, N., Zhang, Y., Kumar, A., Uppal, K., Liu, Z., Lee, E.K., Wu, H., Medrzycki, M., Pan, C., et al. (2013). High-Resolution Mapping of H1 Linker Histone Variants in Embryonic Stem Cells. PLoS Genet. *9*, e1003417.

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., and Lewis, S. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288–289.

Carter, M.G., Stagg, C.A., Falco, G., Yoshikawa, T., Bassey, U.C., Aiba, K., Sharova, L. V, Shaik, N., and Ko, M.S.H. (2008). An in situ hybridization-based screen for heterogeneously expressed genes in mouse ES cells. Gene Expr. Patterns *8*, 181–198.

Chamberlain, S.J., Yee, D., and Magnuson, T. (2008). Polycomb Repressive Complex 2 Is Dispensable for Maintenance of Embryonic Stem Cell Pluripotency. Stem Cells *26*, 1496–1505.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature *450*, 1230–1234.

Chan, Y.-S., Göke, J., Ng, J.-H., Lu, X., Gonzales, K.A.U., Tan, C.-P., Tng, W.-Q., Hong, Z.-Z., Lim, Y.-S., Ng, H.-H., et al. (2013). Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. Cell Stem Cell *13*, 663–675.

Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., and Tyler, J.K. (2015). The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. Mol. Cell. Biol. *36*, 662–667.

Cheng, B., and Price, D.H. (2007). Properties of RNA Polymerase II Elongation Complexes Before and After the P-TEFb-mediated Transition into Productive Elongation. J. Biol. Chem. *282*, 21901–21912.

Choi, J.K., and Kim, Y.-J. (2009). Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat. Genet. *41*, 498–503.

Chopra, V.S., Hendrix, D.A., Core, L.J., Tsui, C., Lis, J.T., Levine, M., Breiling, A., Turner, B.M., Bianchi, M.E., Orlando, V., et al. (2011). The Polycomb Group Mutant esc Leads to Augmented Levels of Paused Pol II in the Drosophila Embryo. Mol. Cell *42*, 837–844.

Clark, T.A., Sugnet, C.W., Ares, M., Staley, J.P., Guthrie, C., Black, D.L., Davis, C.A., Grate, L., Spingola, M., Ares, M., et al. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science *296*, 907–910.

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. *4*, 265–270.

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., Gibson, R., Hoad, G., Hubbard, T., Hunter, C., et al. (2009). Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Res. *37*, D19–D25.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. *38*, 1767–1771.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L.L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biol. *17*, 13.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Cui, K., Zang, C., Roh, T.-Y., Schones, D., Childs, R., Peng, W., and Zhao, K. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell *4*, 80–93.

Czermin, B., Melfi, R., Donna, M., Seitz, V., Imhof, A., and Pirrotta, V. (2002). Drosophila enhancer of {Zeste/ESC} complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. Cell *111*, 185–196.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022.

Deng, Q., Ramsköld, D., Reinius, B., Sandberg, R., Pernis, B., Chiappino, G., Kelus, A.S., Gell, P.G., Malissen, M., Trucy, J., et al. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science *343*, 193–196.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., Pellegrini, M., and Plath, K. (2013). Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. Cell Stem Cell *13*, 602–616.

Dennis, G., Sherman, B.T., Hosack, D. a, Yang, J., Gao, W., Lane, H.C., and Lempicki, R. a (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. *4*, P3.

Devailly, G., Mantsoki, A., Michoel, T., and Joshi, A. (2015). Variable reproducibility in genome-scale public data: A case study using ENCODE ChIP sequencing resource. FEBS Lett. *589*, 3866–3870.

Dillon, S.C., Zhang, X., Trievel, R.C., and Cheng, X. (2005). The SET-domain protein superfamily: protein lysine methyltransferases. Genome Biol. *6*, 227.

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

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Dreos, R., Ambrosini, G., Périer, R.C., and Bucher, P. (2015). The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res. *43*, D92-96.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78–81.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Endoh, M., Endo, T.A., Endoh, T., Isono, K., Sharif, J., Ohara, O., Toyoda, T., Ito, T., Eskeland, R., Bickmore, W.A., et al. (2012). Histone {H2A} mono-ubiquitination is a crucial step to mediate {PRC1-dependent} repression of developmental genes to maintain {ES} cell identity. {PLoS} Genet. *8*.

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

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Ernst, P., Mabon, M., Davidson, A., Zon, L., and Korsmeyer, S. (2004). An Mll-dependent Hox program drives hematopoietic progenitor expansion. Curr. Biol. *14*, 2063–2069.

Farcas, A.M., Blackledge, N.P., Sudbery, I., Long, H.K., McGouran, J.F., Rose, N.R., Lee, S., Sims, D., Cerase, A., Sheahan, T.W., et al. (2012). KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. Elife *1*, e00205.

Feinberg, A.P., and Irizarry, R.A. (2010). Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc. Natl. Acad. Sci. U. S. A. *107 Suppl*, 1757–1764.

Francis, N.J., Kingston, R.E., and Woodcock, C.L. (2004). Chromatin compaction by a polycomb group protein complex. Science *306*, 1574–1577.

Gafni, O., Weinberger, L., Mansour, A.A., Manor, Y.S., Chomsky, E., Ben-Yosef, D., Kalma, Y., Viukov, S., Maza, I., Zviran, A., et al. (2013). Derivation of novel human ground state naive pluripotent stem cells. Nature *504*, 282–286.

Gaidatzis, D., Burger, L., Florescu, M., and Stadler, M.B. (2015). Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation. Nat. Biotechnol. *33*, 722–729.

Galvin-Burgess, K.E., Travis, E.D., Pierson, K.E., and Vivian, J.L. (2013). TGF-β-superfamily signaling regulates embryonic stem cell heterogeneity: self-renewal as a dynamic and regulated equilibrium. Stem Cells *31*, 48–58.

Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., and Reinberg, D. (2012). {PCGF} homologs, {CBX} proteins, and {RYBP} define functionally distinct {PRC1} family complexes. Mol. Cell *45*, 344–356.

García, E., Marcos-Gutiérrez, C., del Mar Lorente, M., Moreno, J.C., and Vidal, M. (1999). RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1. TL - 18. EMBO J. *18 VN-r*, 3404–3418.

Gardiner-Garden, M., and Frommer, M. (1987). CpG Islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Gierliński, M., Cole, C., Schofield, P., Schurch, N.J., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G., Owen-Hughes, T., et al. (2015). Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics *31*, 3625–3630.

De Gobbi, M., Garrick, D., Lynch, M., Vernimmen, D., Hughes, J., Goardon, N., Luc, S., Lower, K., Sloane-Stanley, J., Pina, C., et al. (2011). Generation of bivalent chromatin domains during cell fate decisions. Epigenetics Chromatin *4*, 9.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.

Graf, T., and Stadtfeld, M. (2008). Heterogeneity of Embryonic and Adult Stem Cells. Cell Stem Cell *3*, 480–483.

Graham, B., Marcais, A., Dharmalingam, G., Carroll, T., Kanellopoulou, C., Graumann, J., Nesterova, T.B., Bermange, A., Brazauskas, P., Xella, B., et al. (2016). MicroRNAs of the miR-290–295 Family Maintain Bivalency in Mouse Embryonic Stem Cells.

Grau, D.J., Chapman, B.A., Garlick, J.D., Borowsky, M., Francis, N.J., and Kingston, R.E. (2011). Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. Genes Dev. *25*, 2210–2221.

Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, B.V. (2016). CRAN - Package gplots.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812.

Guberman, J.M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., Chelala, C., Croft, D., Cros, A., Cutts, R.J., et al. (2011). BioMart Central Portal: an open database network for the biological community. Database (Oxford). *2011*, bar041.

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

Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D.H., Sano Marma, M., Meng, Q., Cao, H., Li, X., et al. (2008). Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc. Natl. Acad. Sci. *105*, 9145–9150.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910.

Haider, S., Ballester, B., Smedley, D., and Zhang…, J. (2009). {BioMart} Central Portal—unified access to biological data.

Hanazawa, M., Narushima, H., and Minaka, N. (1995). Generating most parsimonious reconstructions on a tree: A generalization of the Farris-Swofford-Maddison method. Discret. Appl. Math. *56*, 245–265.

Hanna, J., Cheng, A.W., Saha, K., Kim, J., Lengner, C.J., Soldner, F., Cassady, J.P., Muffat, J., Carey, B.W., and Jaenisch, R. (2010). 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.

Harrow, J., Frankish, A., Gonzalez, J.M., and Tapanari…, E. (2012a). {GENCODE:} The reference human genome annotation for The {ENCODE} Project.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012b). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell *6*, 479–491.

Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Sharpe, J.A., Suciu, M.C., et al. (2016). Genetic dissection of the α-globin super-enhancer in vivo. Nat. Genet. *48*, 895–903.

Hayashi, K., Lopes, S.S.M. de S.S.M.C. de S.S.M.C. de S., Tang, F., Surani, M.A.A., Ansel, K.M., Lee, D.U., Rao, A., Avilion, A.A., Nicolis, S.K., Pevny, L.H., et al. (2008). Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. Cell Stem Cell *3*, 391–401.

He, A., Shen, X., Ma, Q., Cao, J., Gise, A. von, Zhou, P., Wang, G., Marquez, V.E., Orkin, S.H., and Pu, W.T. (2012). PRC2 directly methylates GATA4 and represses its transcriptional activity. Genes Dev. *26*, 37.

He, B., Chen, C., Teng, L., and Tan, K. (2014). Global view of enhancer–promoter interactome in human cells. Proc. Natl. Acad. Sci. U. S. A. *111*, E2191.

He, J., Shen, L., Wan, M., Taranova, O., Wu, H., and Zhang, Y. (2013). Kdm2b maintains murine embryonic stem cell status by recruiting {PRC1} complex to {CpG} islands of developmental genes. Nat. Cell Biol. *15*, 373–384.

Heinig, M., Colomé-Tatché, M., Taudt, A., Rintisch, C., Schafer, S., Pravenec, M., Hubner, N., Vingron, M., Johannes, F., Kouzarides, T., et al. (2015). histoneHMM: Differential analysis of histone modifications with broad genomic footprints. BMC Bioinformatics *16*, 60.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

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

Hon, G.C., Hawkins, R.D., and Ren, B. (2009). Predictive chromatin signatures in the mammalian genome. Hum. Mol. Genet. *18*, R195–R201.

Hu, D., Garruss, A., Gao, X., Morgan, M., Cook, M., Smith, E., and Shilatifard, A. (2013). The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1093–1097.

Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C., and Enver, T. (1997). Multilineage gene expression precedes commitment in the hemopoietic system. Genes Dev. *11*, 774–785.

Huang, J., Perez-Burgos, L., Placek, B.J., Sengupta, R., Richter, M., Dorsey, J.A., Kubicek, S., Opravil, S., Jenuwein, T., and Berger, S.L. (2006). Repression of p53 activity by Smyd2-mediated methylation. Nature *444*, 629–632.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods *12*, 115–121.

Huda, A., Mariño-Ramírez, L., Landsman, D., and Jordan, I.K. (2009). Repetitive DNA elements, nucleosome binding and human gene expression. Gene *436*, 12–22.

Illingworth, R.S., Hölzenspies, J.J., Roske, F. V, Bickmore, W.A., Brickman, J.M., Alabert, C., Groth, A., Azuara, V., Perry, P., Sauer, S., et al. (2016). Polycomb enables primitive endoderm lineage priming in embryonic stem cells. Elife *5*, 153–167.

Isagawa, T., Nagae, G., Shiraki, N., Fujita, T., Sato, N., Ishikawa, S., Kume, S., and Aburatani, H. (2011). {DNA} methylation profiling of embryonic stem cell differentiation into the three germ layers. {PloS} One *6*.

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lonnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. *21*, 1160–1167.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2013). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.

Jenuwein, T. (2001). Translating the Histone Code. Science (80-.). *293*, 1074–1080.

Jia, J., Zheng, X., Hu, G., Cui, K., Zhang, J., Zhang, A., Jiang, H., Lu, B., Yates, J., Liu, C., et al. (2012a). Regulation of pluripotency and self- renewal of ESCs through epigenetic-threshold modulation and mRNA pruning. Cell *151*, 576–589.

Jia, J., Zheng, X., Hu, G., Cui, K., Zhang, J., Zhang, A., Jiang, H., Lu, B., Yates, J., Liu, C., et al. (2012b). Regulation of pluripotency and self- renewal of {ESCs} through epigenetic-threshold modulation and {mRNA} pruning. Cell *151*, 576–589.

Jiang, H., Shukla, A., Wang, X., Chen, W., Bernstein, B.E., Roeder, R.G., Andreu-Vieyra, C.V., Chen, R., Agno, J.E., Glaser, S., et al. (2011a). Role for Dpy-30 in ES Cell-Fate Specification by Regulation of H3K4 Methylation within Bivalent Domains. Cell *144*, 513–525.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011b). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. *16*, 167–177.

Ju, J., Kim, D.H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J., et al. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc. Natl. Acad. Sci. *103*, 19635–19640.

Juven-Gershon, T., Hsu, J.-Y.Y., and Kadonaga, J.T. (2006). Perspectives on the RNA polymerase II core promoter. TL - 34. Biochem. Soc. Trans. *34 VN-r*, 1047–1050.

Kalisky, T., and Quake, S.R. (2011). Single-cell genomics. Nat. Methods *8*, 311–314.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. *44*, D457-62.

Kanhere, A., Viiri, K., Araújo, C.C., Rasaiyaah, J., Bouwman, R.D., Whyte, W.A., Pereira, C.F., Brookes, E., Walker, K., Bell, G.W., et al. (2010). Short RNAs Are Transcribed from Repressed Polycomb Target Genes and Interact with Polycomb Repressive Complex-2. Mol. Cell *38*, 675–688.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2014). The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. *42*, D764-70.

Karwacki-Neisius, V., Göke, J., Osorno, R., Halbritter, F., Ng, J.H., Weiße, A.Y., Wong, F.C.K., Gagliardi, A., Mullin, N.P., Festuccia, N., et al. (2013). Reduced Oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by Oct4 and Nanog. Cell Stem Cell *12*, 531–545.

Kennison, J.A. (1995). The Polycomb and Trithorax Group Proteins of *Drosophila* : Trans-Regulators of Homeotic Gene Function. Annu. Rev. Genet. *29*, 289–303.

Kharchenko, P. V, Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. *26*, 1351–1359.

Kikuta, H., Fredman, D., Rinkwitz, S., Lenhard, B., and Becker, T.S. (2007a). Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - a fundamental feature of vertebrate genomes. Genome Biol. *8 Suppl 1*.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A., Engström, P., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., et al. (2007b). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. *17*, 545–555.

Kim, Y.Z. (2014). Altered histone modifications in gliomas. Brain Tumor Res. Treat. *2*, 7–21.

Kim, J.K., Kolodziejczyk, A.A., Illicic, T., Teichmann, S.A., and Marioni, J.C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun. *6*, 8687.

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

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D., and Ren, B. (2005). A high-resolution map of active promoters in the human genome. Nature *436*, 876–880.

Kishore, K., de Pretis, S., Lister, R., Morelli, M.J., Bianchi, V., Amati, B., Ecker, J.R., Pelizzola, M., Bock, C., Lengauer, T., et al. (2015). methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. BMC Bioinformatics *16*, 313.

Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C.H., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Bühler, M., Liu, P., et al. (2015). Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell *17*, 471–485.

Komarnitsky, P., Cho, E.-J.J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev. *14*, 2452–2460.

Kornberg, R.D. (1974). Chromatin structure: a repeating unit of histones and DNA. Science *184*, 868–871.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693–705.

Krogan, N.J., Dover, J., Khorrami, S., Greenblatt, J.F., Schneider, J., Johnston, M., and Shilatifard, A. (2002). COMPASS, a histone H3 (Lysine 4) methyltransferase required for telomeric silencing of gene expression. J. Biol. Chem. *277*, 10753–10755.

Krogan, N.J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D.P., Beattie, B.K., Emili, A., Boone, C., et al. (2003). Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Mol. Cell. Biol. *23*, 4207–4218.

Krumm, A., Hickey, L.B., and Groudine, M. (1995). Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. Genes Dev. *9*, 559–572.

Ku, M., Koche, R., Rheinbay, E., Mendenhall, E., Endoh, M., Mikkelsen, T., Presser, A., Nusbaum, C., Xie, X., Chi, A., et al. (2008a). Genomewide analysis of {PRC1} and {PRC2} occupancy identifies two classes of bivalent domains. {PLoS} Genet. *4*.

Ku, M., Koche, R., Rheinbay, E., Mendenhall, E., Endoh, M., Mikkelsen, T., Presser, A., Nusbaum, C., Xie, X., Chi, A., et al. (2008b). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet. *4*.

Kumar, R.M., Cahan, P., Shalek, A.K., Satija, R., Jay DaleyKeyser, A., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J.J., et al. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature *516*, 56–61.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Lampe, X., Samad, O.A., Guiguen, A., Matis, C., Remacle, S., Picard, J.J., Rijli, F.M., and Rezsohazy, R. (2008). An ultraconserved {Hox-Pbx} responsive element resides in the coding sequence of Hoxa2 and is active in rhombomere 4. Nucleic Acids Res. *36*, 3214–3225.

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

Nat. Methods *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short {DNA} sequences to the human genome.

Lanouette, S., Mongeon, V., Figeys, D., and Couture, J. (2014). The functional diversity of protein lysine methylation. Mol. Syst. Biol. *10*.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., Gentleman, R., Carey, V., et al. (2013). Software for Computing and Annotating Genomic Ranges. PLoS Comput. Biol. *9*, e1003118.

Lehmann, L., Ferrari, R., Vashisht, A.A., Wohlschlegel, J.A., Kurdistani, S.K., and Carey, M. (2012). Polycomb Repressive Complex 1 (PRC1) Disassembles RNA Polymerase II Preinitiation Complexes. J. Biol. Chem. *287*, 35784–35794.

Leisch, F., and Friedrich (2006). A toolbox for -centroids cluster analysis. Comput. Stat. Data Anal. *51*, 526–544.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat. Rev. Genet. *13*, 233–245.

Lercher, M.J., Urrutia, A.O., and Hurst, L.D. (2002). Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. *31*, 180–183.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565–570.

Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nat. Rev. Genet. *3*, 662–673.

Li, B., Dewey, C.N., Wang, Z., Gerstein, M., Snyder, M., Katz, Y., Wang, E., Airoldi, E., Burge, C., Nicolae, M., et al. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., and Ruan…, J. (2009). The sequence alignment/map format and {SAMtools}.

Li, M., He, Y., Dubois, W., Wu, X., Shi, J., and Huang, J. (2012). Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. Mol. Cell *46*, 30–42.

Loh, K.M., Ang, L.T., Zhang, J., Kumar, V., Ang, J., Auyeong, J.Q., Lee, K.L., Choo, S.H., Lim, C.Y.Y., Nichane, M., et al. (2014). Efficient Endoderm Induction from Human Pluripotent Stem Cells by Logically Directing Signals Controlling Lineage Bifurcations. Cell Stem Cell *14*, 237–252.

Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., Young, R.A., Patel, A., et al. (2012). Revisiting global gene expression analysis. Cell *151*, 476–482.

Lu, T., Jackson, M.W., Wang, B., Yang, M., Chance, M.R., Miyagi, M., Gudkov, A. V., and Stark, G.R. (2010). Regulation of NF-κB by NSD1/FBXL11-dependent reversible lysine methylation of p65. Proc. Natl. Acad. Sci. U. S. A. *107*, 46.

Lynch, M.D., Smith, A.J.H., De Gobbi, M., Flenley, M., Hughes, J.R., Vernimmen, D., Ayyub, H., Sharpe, J.A., Sloane-Stanley, J.A., Sutherland, L., et al. (2012). An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. EMBO J. *31*, 317–329.

Mantsoki, A., Devailly, G., and Joshi, A. (2015). CpG island erosion, polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells. Sci. Rep. *5*, 16791.

Mar, J.C., Matigian, N.A., Mackay-Sim, A., Mellick, G.D., Sue, C.M., Silburn, P.A., McGrath, J.J., Quackenbush, J., and Wells, C.A. (2011). Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease. PLoS Genet. *7*, e1002207.

Margueron, R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343–349.

Margueron, R., Trojer, P., and Reinberg, D. (2005). The key to development: interpreting the histone code? Curr. Opin. Genet. Dev. *15*, 163–176.

Margueron, R., Li, G., Sarma, K., Blais, A., Zavadil, J., Woodcock, C.L., Dynlacht, B.D., and Reinberg, D. (2008). Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. Mol. Cell *32*, 503–518.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376.

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. *24*, 496–510.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Stewart, A.F., Smith, A., et al. (2012). The transcriptional and epigenomic foundations of ground state pluripotency. Cell *149*, 590–604.

Mason, E.A., Mar, J.C., Laslett, A.L., Pera, M.F., Quackenbush, J., Wolvetang, E., and Wells, C.A. (2014). Gene Expression Variability as a Unifying Element of the Pluripotency Network. Stem Cell Reports *3*, 365–377.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. *7*, 29–59.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science (80-.). *337*.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. PLoS Genet. *6*, e1001244.

Mercer, E., Lin, Y., Benner, C., Jhunjhunwala, S., Dutkowski, J., Flores, M., Sigvardsson, M., Ideker, T., Glass, C., and Murre, C. (2011). Multilineage priming of enhancer repertoires precedes commitment to the B and myeloid cell lineages in hematopoietic progenitors. Immunity *35*, 413–425.

Meshorer, E., and Misteli, T. (2006). Chromatin in pluripotent embryonic stem cells and differentiation. Nat. Rev. Mol. Cell Biol. *7*, 540–546.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat. Rev. Genet. *11*, 31–46.

Mikkelsen, T., Ku, M., Jaffe, D., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Mikkelsen, T.S., Xu, Z., Zhang, X., Wang, L., Gimble, J.M., Lander, E.S., and Rosen, E.D. (2010). Comparative epigenomic analysis of murine and human adipogenesis. Cell *143*, 156–169.

Miller, T., Krogan, N.J., Dover, J., Erdjument-Bromage, H., Tempst, P., Johnston, M., Greenblatt, J.F., and Shilatifard, A. (2001). COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. Proc. Natl. Acad. Sci. U. S. A. *98*, 12902–12907.

Min, I.M., Waterfall, J.J., Core, L.J., Munroe, R.J., Schimenti, J., and Lis, J.T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. Genes Dev. *25*, 742–754.

Mohn, F., Weber, M., Rebhan, M., Roloff, T., Richter, J., Stadler, M., Bibel, M., Sch"ubeler, D., and Schübeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol. Cell *30*, 755–766.

Morey, L., Aloia, L., Cozzuto, L., Benitah, S.A., and Di Croce, L. (2013). RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells. Cell Rep. *3*, 60–69.

Morgan, H.D., Santos, F., Green, K., Dean, W., and Reik, W. (2005). Epigenetic reprogramming in mammals. Hum. Mol. Genet. *14*, R47–R58.

Morozova, O., and Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. Genomics *92*, 255–264.

Morris, J.F., Hromas, R., and Rauscher, F.J. (1994). Characterization of the DNA-binding properties of the myeloid zinc finger protein MZF1: two independent DNA-binding domains recognize two DNA consensus sequences with a common G-rich core. Mol. Cell. Biol. *14*, 1786–1795.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Müllner, D. (2013). fastcluster : Fast Hierarchical, Agglomerative Clustering Routines for *R* and *Python*. J. Stat. Softw. *53*, 1–18.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat. Genet. *39*, 1507–1511.

Musselman, C.A., Avvakumov, N., Watanabe, R., Abraham, C.G., Lalonde, M.-E., Hong, Z., Allen, C., Roy, S., Nuñez, J.K., Nickoloff, J., et al. (2012). Molecular basis for H3K36me3 recognition by the Tudor domain of PHF1. Nat. Struct. Mol. Biol. *19*, 1266–1272.

Nag, A., Savova, V., Fung, H.-L.L., Miron, A., Yuan, G.-C.C., Zhang, K., and Gimelbrant, A.A. (2013). Chromatin signature of widespread monoallelic expression. Elife *2*.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science *320*, 1344–1349.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., et al. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol.

de Napoles, M., Mermoud, J.E., Wakao, R., Tang, Y.A., Endoh, M., Appanah, R., Nesterova, T.B., Silva, J., Otte, A.P., Vidal, M., et al. (2004). Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. TL - 7. Dev. Cell *7 VN-re*, 663–676.

Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science (80-.). *347*.

Narushima, H., and Hanazawa, M. (1997). A more efficient algorithm for MPR problems in phylogeny (North-Holland).

Nevado, J., Gaudreau, L., Adam, M., and Ptashne, M. (1999). Transcriptional activation by artificial recruitment in mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *96*, 2674–2677.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol. Cell *11*, 709–719.

Ng, S.S., Yue, W.W., Oppermann, U., and Klose, R.J. (2009). Dynamic protein methylation in chromatin biology. Cell. Mol. Life Sci. {CMLS} *66*, 407–422.

Nightingale, K.P., O'Neill, L.P., and Turner, B.M. (2006). Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. Curr. Opin. Genet. Dev. *16*, 125–136.

Niwa, H. (2007). How is pluripotency determined and maintained? Dev. {(Cambridge,} England) *134*, 635–646.

O'Shea, K.S. (2004). Self-renewal vs. differentiation of mouse embryonic stem cells.

Biol. Reprod. *71*, 1755–1765.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., and Wagner, H. (2013). vegan: Community Ecology Package.

Ooi, S.K.T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.-P., Allis, C.D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714–717.

Orlando, D.A., Chen, M.W., Brown, V.E., Solanki, S., Choi, Y.J., Olson, E.R., Fritz, C.C., Bradner, J.E., Guenther, M.G., Paša-Tolić, L., et al. (2014). Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome. Cell Rep. *9*, 1163–1170.

Pan, X. (2014). Single Cell Analysis: From Technology to Biology and Medicine. Single Cell Biol. *3*.

Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G., Stewart, R., and Thomson, J. (2007). Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell *1*, 299–312.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics *20*, 289–290.

Park, P.J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.

Pearson, J., Lemons, D., and McGinnis, W. (2005). Modulating Hox gene functions during animal body patterning. Nat. Rev. Genet. *6*, 893–904.

Pearson, M., Carbone, R., Sebastiani, C., Cioce, M., Fagioli, M., Saito, S., Higashimoto, Y., Appella, E., Minucci, S., Pandolfi, P.P., et al. (2000). PML regulates p53 acetylation and premature senescence induced by oncogenic Ras. Nature *406*, 207–210.

Peng, J.C., Valouev, A., Swigut, T., Zhang, J., Zhao, Y., Sidow, A., and Wysocka, J. (2009). {Jarid2/Jumonji} coordinates control of {PRC2} enzymatic activity and target gene occupancy in pluripotent cells. Cell *139*, 1290–1302.

Percharde, M., Lavial, F., Ng, J.-H., Kumar, V., Tomaz, R.A., Martin, N., Yeo, J.-C., Gil, J., Prabhakar, S., Ng, H.-H., et al. (2012). Ncoa3 functions as an essential Esrrb coactivator to sustain embryonic stem cell self-renewal and reprogramming. Genes Dev. *26*, 2286–2298.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. Cell *153*, 1281–1295.

Polavarapu, N., Mariño-Ramírez, L., Landsman, D., McDonald, J.F., Jordan, I.K., Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., et al. (2008). Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. BMC Genomics *9*, 226.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033–1038.

Pooley, C., Ruau, D., Lombard, P., Gottgens, B., and Joshi, A. (2014). TRES predicts transcription control in embryonic stem cells. *30*, 2983–2985.

Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501–512.

Qin, J., Whyte, W.A., Anderssen, E., Apostolou, E., Chen, H.-H.H., Akbarian, S., Bronson, R.T., Hochedlinger, K., Ramaswamy, S., Young, R.A., et al. (2012). The polycomb group protein L3mbtl2 assembles an atypical {PRC1-family} complex that is essential in pluripotent stem cells and early development. Cell Stem Cell *11*, 319–332.

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

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

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. Cell *141*, 432–445.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA Synthesis in Mammalian Cells. PLoS Biol. *4*, e309.

Raj, A., van Oudenaarden, A., Acar, M., Becskei, A., Oudenaarden, A. van, Acar, M., Mettetal, J.T., Oudenaarden, A. van, Arkin, A., Ross, J., et al. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell *135*, 216–226.

Raj, A., Rifkin, S.A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. Nature *463*, 913–918.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. *30*, 777–782.

Raser, J., and O'Shea, E. (2004). Control of stochasticity in eukaryotic gene expression. Science *304*, 1811–1814.

Rathert, P., Zhang, X., Freund, C., Cheng, X., and Jeltsch, A. (2008). Analysis of the Substrate Specificity of the Dim-5 Histone Lysine Methyltransferase Using Peptide Arrays. Chem. Biol. *15*, 5.

Rea, S., Eisenhaber, F., D, O., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P., Allis, C.D., et al. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature *406*, 593–599.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature *447*, 425–432.

Richly, H., Luciana, R.-V., Ribeiro, J.D., Demajo, S., Gundem, G., Nuria, L.-B., Nakagawa, T., Rospert, S., Ito, T., and Di Croce, L. (2010). Transcriptional activation of polycomb-repressed genes by {ZRF1.}. Nature *468*, 1124–1128.

Riising, E.M., Comet, I., Leblanc, B., Wu, X., Johansen, J.V., and Helin, K. (2014). Gene silencing triggers polycomb repressive complex 2 recruitment to CpG Islands genome wide. Mol. Cell *55*, 347–360.

Ringrose, L., and Paro, R. (2004). Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. Annu. Rev. Genet. *38*, 413–443.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.

Rivera, C.M., and Ren, B. (2013). Mapping Human Epigenomes. Cell *155*, 39–55.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods *4*, 651–657.

Roh, T.-Y.T.-Y., Cuddapah, S., Cui, K., and Zhao, K. (2006). The genomic landscape of histone modifications in human T cells. Proc. Natl. Acad. Sci. U. S. A. *103*, 15782–15787.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352.

Rougvie, A.E., and Lis, J.T. (1988). The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. Cell *54*, 795–804.

Rugg-Gunn, P.J., Cox, B.J., Ralston, A., and Rossant, J. (2010). Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. Proc. Natl. Acad. Sci. U. S. A. *107*, 10783–10790.

Rybtsova, N., Leimgruber, E., Seguin-Estévez, Q., Dunand-Sauthier, I., Krawczyk, M., and Reith, W. (2007). Transcription-coupled deposition of histone modifications during MHC class II gene activation. Nucleic Acids Res. *35*, 3431–3441.

S. Andrews (2010). FastQC A Quality Control tool for High Throughput Sequence Data.

Sadeh, R., Launer-Wachs, R., Wandel, H., Rahat, A., Friedman, N., Baylin, S.B., Jones, P.A., Bernstein, B.E., Mikkelsen, T.S., Xie, X., et al. (2016). Elucidating Combinatorial Chromatin States at Single-Nucleosome Resolution. Mol. Cell *0*, 726–734.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. *4*, 406–425.

Sánchez-Castillo, M., Ruau, D., Wilkinson, A.C., Ng, F.S.L., Hannah, R., Diamanti, E., Lombard, P., Wilson, N.K., and Gottgens, B. (2015). CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. Nucleic Acids Res. *43*, D1117-23.

Sanulli, S., Justin, N., Teissandier, A., Ancelin, K., Portoso, M., Caron, M., Michaud, A., Lombard, B., da Rocha, S.T., Offer, J., et al. (2015). Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation. Mol. Cell *57*, 769–783.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Sasai, N., Saitoh, N., Saitoh, H., and Nakao, M. (2013). The transcriptional cofactor MCAF1/ATF7IP is involved in histone gene expression and cellular senescence. PLoS One *8*, e68478.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. U. S. A. *103*, 1412–1417.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., et al. (2009). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *37*, D5–D15.

Schmid, C.D., and Bucher, P. (2007). ChIP-Seq Data Reveal Nucleosome Architecture of Human Promoters. Cell *131*, 831–832.

Schmitges, F.W., Prusty, A.B., Faty, M., Stützer, A., Lingaraju, G.M., Aiwazian, J., Sack, R., Hess, D., Li, L., Zhou, S., et al. (2011). Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks. Mol. Cell *42*, 330–341.

Schneider, R., Bannister, A.J., Myers, F.A., Thorne, A.W., Colyn, C.-R., and Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nat. Cell Biol. *6*, 73–77.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., and Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. Cell *128*, 735–745.

Schulze, A., and Downward, J. (2001). Navigating gene expression using microarrays--a technology review. Nat. Cell Biol. *3*, E190-195.

Schweikert, G., Cseke, B., Clouaire, T., Bird, A., Sanguinetti, G., Park, P., Wilbanks, E., Facciotti, M., Schmidt, D., Wilson, M., et al. (2013). MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. BMC Genomics *14*, 826.

Seenundun, S., Rampalli, S., Liu, Q.-C., Aziz, A., Palii, C., Hong, S., Blais, A., Brand, M., Ge, K., Dilworth, F.J.F., et al. (2010). UTX mediates demethylation of H3K27me3 at muscle-specific genes during myogenesis. EMBO J. *29*, 1401–1411.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240.

Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S.H., and Waxman, D.J. (2012). MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol. *13*, R16.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. *14*, 618–630.

Sharov, A.A., and Ko, M.S. (2007). Human {ES} cell profiling broadens the reach of bivalent domains. Cell Stem Cell *1*, 237–238.

Shema, E., Jones, D., Shoresh, N., Donohue, L., Ram, O., Bernstein, B.E., Rivera, C.M., Ren, B., Bannister, A.J., Kouzarides, T., et al. (2016). Single-molecule decoding of combinatorially modified nucleosomes. Science *352*, 717–721.

Shilatifard, A. (2012a). The {COMPASS} family of histone {H3K4} methylases: mechanisms of regulation in development and disease pathogenesis. Annu. Rev. Biochem. *81*, 65–95.

Shilatifard, A. (2012b). The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. Annu. Rev. Biochem. *81*, 65–95.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. *15*, 121–132.

Singh, A.M., Hamazaki, T., Hankowski, K.E., and Terada, N. (2007). A heterogeneous expression pattern for Nanog in embryonic stem cells. Stem Cells *25*, 2534–2542.

Singh, A.M., Chappell, J., Trost, R., Lin, L., Wang, T., Tang, J., Matlock, B.K., Weller, K.P., Wu, H., Zhao, S., et al. (2013). Cell-cycle control of developmentally regulated transcription factors accounts for heterogeneity in human pluripotent cells. Stem Cell Reports *1*, 532–544.

Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J.C., Thongjuea, S., Stadhouders, R., Palstra, R.-J., Stevens, M., Kockx, C., van Ijcken, W., et al. (2010). The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. Genes Dev. *24*, 277–289.

Soler, E., Charlotte, A.-S., Boer, E. d, Bryne, J.C., Thongjuea, S., Rijkers, E., Demmers, J., Ijcken, W. v, and Grosveld, F. (2011). A systems approach to analyze transcription factors in mammalian cells. Methods {(San} Diego, Calif.) *53*, 151–162.

Srinivasan, L., and Atchison, M.L. (2004). YY1 DNA binding and PcG recruitment requires CtBP. Genes Dev. *18*, 2596–2601.

Stegle, O., Teichmann, S.A., and Marioni, J.C. (2015). Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. *16*, 133–145.

Stock, J., Giadrossi, S., Casanova, M., Brookes, E., Vidal, M., Koseki, H., Brockdorff, N., Fisher, A., and Pombo, A. (2007a). Ring1-mediated ubiquitination of {H2A} restrains poised {RNA} polymerase {II} at bivalent genes in mouse {ES} cells. Nat. Cell Biol. *9*, 1428–1435.

Stock, J., Giadrossi, S., Casanova, M., Brookes, E., Vidal, M., Koseki, H., Brockdorff, N., Fisher, A., and Pombo, A. (2007b). Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat. Cell Biol. *9*, 1428–1435.

Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F., et al. (2014). Microfluidic single-cell whole-transcriptome sequencing. Proc. Natl. Acad. Sci. U. S. A. *111*, 7048–7053.

Studier, J.A., and Keppler, K.J. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. Mol. Biol. Evol. *5*, 729–731.

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

Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W., et al. (2014). Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. Cell *158*, 1254–1269.

Tan, L., and Shi, Y. (2012). Tet family proteins and 5-hydroxymethylcytosine in development and disease. Development *139*, 1895–1902.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., and Surani, M.A. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell *6*, 468–478.

Taylor, G.C.A., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M.M., and Bickmore, W.A. (2013). H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. Genome Res. *23*, 2053–2065.

Team, R.D.C. (2016). R: A Language and Environment for Statistical Computing.

Tee, W.-W., Shen, S.S., Oksuz, O., Narendra, V., and Reinberg, D. (2014). Erk1/2 activity promotes chromatin features and RNAPII phosphorylation at developmental promoters in mouse ESCs. Cell *156*, 678–690.

Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D.G. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196–199.

Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L., et al. (2014). Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. Cell Stem Cell *15*, 471–487.

Thomson, J., Itskovitz-Eldor, J., Shapiro, S., Waknitz, M., Swiergiel, J., Marshall, V., and Jones, J. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145–1147.

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

Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K., and Niwa, H. (2008). Identification and characterization of subpopulations in undifferentiated ES cell culture. Development *135*, 909–918.

Trapnell, C., and Salzberg, S.L. (2009). How to map billions of short reads onto genomes. Nat. Biotechnol. *27*, 455–457.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Trojer, P., and Reinberg, D. (2007). Facultative heterochromatin: is there a distinctive molecular signature? Mol. Cell *28*, 1–13.

Turner, B.M. (2007). Defining an epigenetic code. Nat. Cell Biol. *9*, 2–6.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. *18*, 1051–1063.

Vastenhouw, N.L., Zhang, Y., Woods, I.G., Imam, F., Regev, A., Liu, X.S., Rinn, J., and Schier, A.F. (2010). Chromatin signature of embryonic pluripotency is established during genome activation. Nature *464*, 922–926.

Vavouri, T., K, M.G., Woolfe, A., Gilks, W.R., and Elgar, G. (2006). Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key.

Trends Genet. {TIG} *22*, 5–10.

Vicente, C., Conchillo, A., García-Sánchez, M.A., and Odero, M.D. (2012). The role of the GATA2 transcription factor in normal and malignant hematopoiesis. Crit. Rev. Oncol. Hematol. *82*, 1–17.

Voigt, P., and Reinberg, D. (2013). Epigenome editing. Nat. Biotechnol. *31*, 1097–1099.

Voigt, P., LeRoy, G., Drury, W.J.J., Zee, B.M.M., Son, J., Beck, D.B.B., Young, N.L.L., Garcia, B.A.A., Reinberg, D., Gary, L., et al. (2012). Asymmetrically modified nucleosomes. Cell *151*, 181–193.

Voigt, P., Tee, W.-W.W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. Genes Dev. *27*, 1318–1338.

Voskoboynik, A., Neff, N.F., Sahoo, D., Newman, A.M., Pushkarev, D., Koh, W., Passarelli, B., Fan, H.C., Mantalas, G.L., Palmeri, K.J., et al. (2013). The genome sequence of the colonial chordate, Botryllus schlosseri. Elife *2*, e00569.

Vousden, K.H., and Prives, C. (2009). Blinded by the Light: The Growing Complexity of p53. Cell *137*, 413–431.

Wachter, E., Quante, T., Merusi, C., Arczewska, A., Stewart, F., Webb, S., and Bird, A. (2014). Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. Elife *3*, e03397.

Walker, E., Chang, W.Y., Hunkapiller, J., Cagney, G., Garcha, K., Torchia, J., Krogan, N.J., Reiter, J.F., and Stanford, W.L. (2010). Polycomb-like 2 associates with {PRC2} and regulates transcriptional networks during mouse embryonic stem cell self-renewal and differentiation. Cell Stem Cell *6*, 153–166.

Wamstad, J.A., Alexander, J.M., Truty, R.M., Shrikumar, A., Li, F., Eilertson, K.E., Ding, H., Wylie, J.N., Pico, A.R., Capra, J.A., et al. (2012). Dynamic and Coordinated Epigenetic Regulation of Developmental Transitions in the Cardiac Lineage. Cell *151*, 206–220.

Wang, H., Wang, L., Hediye, E.-B., Vidal, M., Tempst, P., Jones, R.S., and Zhang, Y. (2004). Role of histone {H2A} ubiquitination in Polycomb silencing. Nature *431*, 873–878.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Ware, C.B., Nelson, A.M., Mecham, B., Hesson, J., Zhou, W., Jonlin, E.C., Jimenez-Caliani, A.J., Deng, X., Cavanaugh, C., Cook, S., et al. (2014). Derivation of naive human embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *111*, 4484–4489.

Weiner, A., Lara-Astiaso, D., Krupalnik, V., Gafni, O., David, E., Winter, D.R., Hanna, J.H., and Amit, I. (2016). Co-ChIP enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution. Nat. Biotechnol.

Wetterstrand K. (2013). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program [online] (2013).

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

Williams, K., Christensen, J., Helin, K., Blackledge, N., Klose, R., Bogdanovic, O., Veenstra, G., Bostick, M., Kim, J., Esteve, P., et al. (2011). DNA methylation: TET proteins—guardians of CpG islands? EMBO Rep. *13*, 28–35.

Williams, L.H., Fromm, G., Gokey, N.G., Henriques, T., Muse, G.W., Burkholder, A., Fargo, D.C., Hu, G., and Adelman, K. (2015). Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. Mol. Cell *58*, 311–322.

Wilson, N., Foster, S., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., Chilarska, P., Kinston, S., Ouwehand, W., Dzierzak, E., et al. (2010). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell *7*, 532–544.

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., et al. (2013). Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods *11*, 41–46.

Wu, Z., Yang, M., Liu, H., Guo, H., Wang, Y., Cheng, H., and Chen, L. (2012). Role of nuclear receptor coactivator 3 (Ncoa3) in pluripotency maintenance. J. Biol. Chem. *287*, 38295–38304.

Xiao, S., Xie, D., Cao, X., Yu, P., Xing, X., Chen, C.-C.C., Musselman, M., Xie, M., West, F.D., Lewin, H.A., et al. (2012). Comparative epigenomic annotation of regulatory {DNA.}. Cell *149*, 1381–1392.

Xie, W., Ling, T., Zhou, Y., Feng, W., Zhu, Q., Stunnenberg, H.G., Grummt, I., and Tao, W. (2012). The chromatin remodeling complex NuRD establishes the poised state of rRNA genes characterized by bivalent histone modifications and altered nucleosome positions. Proc. Natl. Acad. Sci. *109*, 8161–8166.

Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., et al. (2013). Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell *153*, 1134–1148.

Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R., and Ma'ayan, A. (2013). ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database (Oxford). *2013*, bat045.

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

Xu, S., Grullon, S., Ge, K., and Peng, W. (2014). Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. Methods Mol. Biol. *1150*, 97–111.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature *500*, 593–597.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1131–1139.

Ye, T., Krebs, A.R., Choukrallah, M.-A., Keime, C., Plewniak, F., Davidson, I., and Tora, L. (2011). seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res. *39*, e35.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Yip, K.Y., Cheng, C., Bhardwaj, N., Brown, J.B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., et al. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. *13*.

Yu, P., Xiao, S., Xin, X., Song, C.-X., Huang, W., McDee, D., Tanaka, T., Wang, T., He, C., and Zhong, S. (2013). Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation. Genome Res. *23*, 352–364.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355–364.

Yusuf, B., Gopurappilly, R., Dadheech, N., Gupta, S., Bhonde, R., and Pal, R. (2013). Embryonic fibroblasts represent a connecting link between mesenchymal and embryonic stem cells. Dev. Growth Differ. *55*, 330–340.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification {ChIP-Seq} data. Bioinforma. {(Oxford,} England) *25*, 1952–1958.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat. Genet. *39*, 1512–1516.

Zhang, K., Lin, W., Latham, J.A., Riefler, G.M., Schumacher, J.M., Chan, C., Tatchell, K., Hawke, D.H., Kobayashi, R., and Dent, S.Y.R. (2005). The Set1 Methyltransferase Opposes Ipl1 Aurora Kinase Functions in Chromosome Segregation. Cell *122*, 723–734.

Zhang, X., Yang, Z., Khan, S.I., Horton, J.R., Tamaru, H., Selker, E.U., and Cheng, X. (2003). Structural basis for the product specificity of histone lysine methyltransferases. Mol. Cell *12*, 177–185.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, Y., Jurkowska, R., Soeroes, S., Rajavelu, A., Dhayalan, A., Bock, I., Rathert, P., Brandt, O., Reinhardt, R., Fischle, W., et al. (2010). Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. Nucleic Acids Res. *38*, 4246–4253.

Zhang, Y., Xie, S., Zhou, Y., Xie, Y., Liu, P., Sun, M., Xiao, H., Jin, Y., Sun, X., Chen, Z., et al. (2014). H3K36 histone methyltransferase Setd2 is required for murine embryonic stem cell differentiation toward endoderm. Cell Rep. *8*, 1989–2002.

Zhao, X.D., Han, X., Chew, J.L., Liu, J., Chiu, K.P., Choo, A., Orlov, Y.L., Sung, W.-K.K., Shahab, A., Kuznetsov, V.A., et al. (2007). 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.

Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat. Rev. Genet. *12*, 7–18.

Zhu, J., Adli, M., Zou, J.Y.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., et al. (2013). Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell *152*, 642–654.

Appendix I - Additional files (CD-ROM)

Appendix Table 1: Classification of 38,922 mouse promoters in four high confident groups in ESCs. Their CpG ratio, orthologue status (one to one orthologs human-mouse), common chromatin state, expression value (FPKM), expression classification and chromatin status when any one sample was removed or added are also given.

Appendix Table 2: Classification of 57,818 human promoters in four high confident groups in ESCs. Their CpG ratio, orthologue status (one to one orthologs human-mouse), common chromatin state, expression value (FPKM), expression classification and chromatin status when any one sample was removed or added are also given.

Appendix II - Publications

First author publications:

Mantsoki, A., and Joshi, A. (2015). Comparative Analysis of Bivalent Domains in Mammalian Embryonic Stem Cells. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) *9043*, 391–402. (http://link.springer.com/chapter/10.1007/978-3-319-16483-0_39)

Mantsoki, A., Devailly, G., and Joshi, A. (2015). CpG island erosion, polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells. Sci. Rep. *5*, 16791. (http://www.nature.com/articles/srep16791)

Mantsoki, A., Devailly, G., and Joshi, A. (2016). Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data. Comput. Biol. Chem. *63*. (http://www.sciencedirect.com/science/article/pii/S1476927116300330)

Other publications:

Devailly, G., Mantsoki, A., Michoel, T., and Joshi, A. (2015). Variable reproducibility in genome-scale public data: A case study using ENCODE ChIP sequencing resource. FEBS Lett. 589, 3866–3870.

Devailly, G., Mantsoki, A., and Joshi, A. (2016). Heat*seq: an interactive web tool for high-throughput sequencing experiment comparison with public data. Bioinformatics 32, 3354–3356.

Comparative Analysis of Bivalent Domains in Mammalian Embryonic Stem Cells

Anna Mantsoki and Anagha Joshi

Division of Developmental Biology, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, United Kingdom
{anna.mantsoki,anagha.joshi}@roslin.ed.ac.uk

Abstract. Bivalent promoters are defined by the presence of both activating (H3K4me3) and repressive (H3K27me3) chromatin marks. In this paper, we first identified high confidence bivalent promoters in murine ES cells integrating data across eight studies using two methods; peak-based and cutoff-based. We showed that peak-based method is more reliable as promoters are more enriched for developmental regulators than the cutoff-based method. We further identified bivalent promoters in human and pig using the peak-based method to show that the bivalent promoters conserved across species were highly enriched for embryonic developmental processes.

Keywords: ChIP sequencing, embryonic stem cells, chromatin, H3K4me3, H3K27me3, bivalent promoters, developmental genes, comparative genomics.

1 Introduction

The key cellular processes determining the fate of each cell type during development and differentiation are thought to be controlled by gene regulation (Pearson et al. 2005). Genomic regulatory elements such as promoters receive and execute transcriptional signals, dependent on their epigenetic state and chromatin accessibility, controlling the expression of key developmental factors(Wilson et al. 2010). Apart from the transcription control at the promoters and enhancers, gene expression is also controlled epigenetically, by post-translational histone modifications, which transform the chromatin structure and thereby control gene expression (Bannister & Kouzarides 2011).

To unravel key developmental transitions that lead to different types of cell identities, embryonic stem cells (ESCs) offer a valuable model (Thomson et al. 1998) as they have an unlimited potential to self-renew as well as to differentiate in specific lineage when suitable external stimuli are provided. In ESCs, the majority of promoters with high CG content are un-methylated. During differentiation though, some of them become methylated, assisting to the acquisition of their final cell identity (Mohn et al. 2008). Azuara et al., 2006 proposed that particular histone modifications and chromatin structure (Voigt et al. 2013; Thomson et al. 1998) are characteristic of ES cells. Two of the most commonly studied histone modifications related to activation

F. Ortu o and I. Rojas (Eds.): IWBBIO 2015, Part I, LNCS 9043, pp. 391–402, 2015.

and repression of chromatin respectively are H3K4me3 and H3K27me3(Bannister & Kouzarides 2011). Polycomb (PcG) and Trithorax (TrxG) group proteins catalyze H3K27me3 and H3K4me3 respectively, regulating genes involved in development and differentiation (Ringrose & Paro 2004). Bernstein et al., 2006 observed activating (H3K4me3) and repressing (H3K27me3) chromatin signals in promoters of several developmentally regulated genes in murine ES cells. These activating and repressive marks were previously thought to be mutually exclusive and therefore the promoters marked with both modifications were named ‘bivalent’. Mikkelsen et al., 2007 used ChIP sequencing technique to examine the bivalent status and construct chromatin state maps across three cell types: mESCs, neural progenitor cells (NPCs) and mouse embryonic fibroblasts (MEFs). Their study showed for the first time, that bivalent domains also exist in cells of restricted potency and 8-43% of bivalent domains retained their bivalent mark during the differentiation (Mikkelsen et al. 2007). Moreover, Mohn et al., 2008 indicated that bivalent genes that are not present in the pluripotent cells may arise in reduced potency cells.

Bivalent genes were detected also in human ESCs (Pan et al. 2007; Zhao et al. 2007) and the majority of them was shared with bivalent genes in mouse ESCs. Specifically, in two out of three studies there were 2,157 common bivalent genes (Mikkelsen et al. 2007; Pan et al. 2007; Sharov & Ko 2007; Zhao et al. 2007). In agreement with the studies in mice, human ESCs bivalent genes are functionally enriched with developmental transcription factors and genes and most of them lose the repressive H3K27me3 mark during differentiation (Pan et al. 2007; Zhao et al. 2007).

ES cells employ various mechanisms to avoid losing their pluripotency. For example they manage to prevent DNA methylation that would silence important genes indefinitely. Bivalent genes belong to an important category of genes full of developmental factors that need to be poised for activation or repression at the right moment during the differentiation process (Voigt et al. 2013). The bivalent state preserves the plasticity of the developmental genes until certain environmental cues lead to proper differentiation.

Though bivalent genes have been identified across multiple species in ES cells as well as differentiated cells, there is no study so far collecting multiple data sets to build a high-confidence bivalent gene set. We therefore collected genome wide binding patterns of H3K4me3 and H3K27me3 in murine ES cells from eight different studies. We then used two complementary approaches; peak-based and cutoff-based approach to define high confidence bivalent promoters. The high confidence bivalent promoters detected by the peak-based method were more enriched for developmental genes than the cutoff based. Finally, we collected data to identify bivalent promoters in human ES cells and pig induced pluripotent cells to study the evolutionary conservation of bivalency. By performing the comparative analysis of bivalent domains across three species we highlighted the functional relevance of coexistence of these marks on the developmental promoters.

2 Materials and Methods

Data Collection and Processing: Murine ChIP sequencing data for H3K4me3 and H3K27me3 histone marks in ESCs was obtained in fastq format from Gene Expression Omnibus (GEO) database (Barrett et al. 2013). Accession numbers for mouse are: SRX001923, SRX085431, SRX122633, SRX172569, SRX266814, SRX266815, SRX305921, SRX305922, SRX001921, SRX185810, SRX122629, SRX172574, SRX266816, SRX266817, SRX305910, and SRX305911. Human ChIP-Seq data (fastq format) for H3K4me3 and H3K27me3 histone marks in hESCs was obtained from Roadmap to epigenomics (Bernstein et al. 2010) and Gene Expression Omnibus (GEO). Accession numbers for human are: SRX006237, SRX012501, SRX27864, SRX007385, SRX019896, SRX006262, SRX006874, SRX012368, SRX007379, SRX019898, SRX003845, SRX064486, SRX027487, SRX189253, SRX027865, SRX056719, SRX003843, SRX064487, SRX027484, SRX189254, SRX040598, SRX056700. ChIP sequencing data for H3K4me3 and H3K27me3 in pig (Sus Scrofa) induced pluripotent stem cells (iPSCs) was downloaded from a published study with accession number GSE36114 (Xiao et al. 2012). After downloading all the raw sequence files for all the experiments, each technical and biological replicate of the samples was imported in FastQC 0.10.1 (S. Andrews 2010) for quality control. Alignment of reads was done using Bowtie 0.12.9 (Langmead et al. 2009) using reference genomes mm10 for mouse, hg19 for human and susScr3 for pig. For all the species we used single end alignment, seed length=28. We then performed the bowtie execution using custom bash scripts and the samtools (Li et al. 2009) pipeline to convert directly sam format file to a bam format file for each sample. The bam files that belonged to the same experiment (technical replicates) were merged into a common bam file in order to proceed with the further analysis. The biological replicates of each experiment were not merged. We downloaded the Gencode (Harrow et al. 2012) genes for human (Gencode 19) and mouse (Gencode M2). We filtered out and kept only the genes from the initial gtf files. Also, we created bed files for the promoter regions, keeping the areas that were (-1000 bp, +2000 bp) from the Transcription Start Site (TSS). For mouse there were 38,922 promoter regions and for human 57,818. Since there was not a Gencode file available for pig, we downloaded the ensembl gene file available from Biomart (Haider et al. 2009). After doing the same procedure as mentioned above in order to keep only the promoter regions, we got 21,116 regions for pig promoters.

Peak Calling Method: We used SICER (Zang et al. 2009), a tool that is recommended for enrichment analysis of histone modification data, since it outperforms every other tool in its category for peak calling. The input controls were used when they were provided with the samples. When input was available, the SICER parameters were: for H3K4me3, window=200 and gap size=200. For H3K27me3, window=200 and gap size=2x300, since this histone mark is found covering wider chromatin domains. The rest of the parameters (same for both H3K4me3 and H3K27me3) were effective genome fraction =0.7, false discovery rate (FDR) = 0.01, redundancy threshold = 1 and fragment size = 150. When the control library was unavailable, the FDR value parameter was replaced by the E-value parameter equal to

100. We intersected the resulting files after peak calling with the promoter files using the intersect command from BEDtools (Quinlan & Hall 2010).

Cutoff Method: We obtained the read density only at the regions we were interested in, the promoters. Using custom scripts and the coverageBed (BEDtools) (Quinlan & Hall 2010) command, we created bed files for each sample. In the resulting bed files, the column that we kept was the one that contained the number of reads in the promoter regions. We applied logarithmic scale to the read densities of all samples, followed by quantile normalization for H3K4me3 and H3K27me3 samples separately to define a threshold that would reveal the real enrichment for H3K4me3 and H3K27me3 and even out the variability across samples. We generated scatterplots (Figure 1) of the same histone modification samples against each other to examine what type of normalization to choose. To further increase the accuracy of the cutoff method, we created promoter files with sliding windows. Every promoter region was divided in windows of 200 bp, with a sliding step of 50 bp. For all the window regions corresponding to the initial promoter region, the maximum coverage value was chosen as the representative for this region. The distribution pattern of H3K4me3 reads is very close to the bimodal distribution. Following that, we used the mixtools package (Benaglia et al. 2009) in order to fit the bimodal distribution to all of our samples, both for H3K4me3 and H3K27me3. For most of the cases of H3K4me3 bimodal distribution was fitted successfully. In contrast, most H3K27me3 samples were not close to follow the bimodal distribution. For the successfully fitted H3K4me3 samples we kept the mean and standard deviation of the second curve of each distribution. After subtracting each standard deviation value from its respective mean value, we obtained the initial threshold values for each sample. The final threshold value for all the H3K4me3 samples was the average of all the initial values. In the case of H3K27me3 distributions, since we had no successful fitting bimodal distribution, we chose empirically 3 different thresholds and chose the one that would give results best matching to previous studies. The final threshold values used were 4.57 for H3K4me3 and 3.00 for H3K27me3. We used the study of Mikkelsen et al., 2007 to compare peak-based and cut off based method to a published study.

Functional Enrichment Analysis: We conducted gene ontology functional analyses for the bivalent promoters for both approaches, using DAVID (Dennis et al. 2003).

Overlap between Species: To obtain a list of common bivalent, expressed and repressed genes between the species, we used only the orthologous genes that mouse and pig share with human (18,255 genes). We got the common list of genes for all three species between them, but also for each combination by two (human-mouse, human-pig, mouse-pig).

P value Calculation: To calculate if the overlap of two gene lists can happen due to random chance, we used hypergeometric test. Specifically, to compare two lists we used the phyper function in R. When we were comparing more than two lists we used random permutation of the rows and columns of the results table (species in columns, genes in rows) simulated for 1000 times. We used the permatfull function from the vegan package (Oksanen et al. 2013) in R. Then we compared the mean of all the simulations with our result of common genes in order to find if there is significant difference between them.

3 Results and Discussion

3.1 Peak-Based Method to Detect High-Confidence Bivalent Promoters

Bivalent promoters are defined by the presence of both active (H3K4me3) and repressive (H3K27me3) chromatin modifications. In ES cells, they are highly enriched for developmental genes and therefore the identification of high confidence bivalent promoters might lead to discovery of novel developmental regulators. With this rationale, we set to look for high confidence bivalent marked promoters in murine ES cells and collected data for paired (H3K4me3 and H3K27me3) ChIP sequencing samples from eight studies from GEO (methods for details). The samples varied in their or nothing read length, ranging from 27 bp to 115 bp and the total number of mapped reads to the mouse genome assembly mm10, were ranging from 14 million to 200 million reads per sample. We called peaks using SICER (Zang et. al., 2009), the best suited algorithm for peak detection in histone modification data. For eight samples of H3K4me3, between 16 thousand and 66 thousand peaks were identified while for H3K27me3, between 9 thousand and 26 thousand peaks were identified. To check if this variation in peak number can be attributed to the variability in total number of reads across samples, we calculated Pearson's correlation coefficient between number of reads and number of peaks detected across eight samples and found a high correlation. The correlation coefficient for H3K4me3 was 0.84 while for H3K27me3 was 0.75. As the only way to adjust for the sequencing depth for peak based method is to consider the only 7 million reads for peak calling but it suffers a major drawback of not being able to use most of the available data, we defined an approach complementary to peak-based approach - cutoff-based method (described in detail in the following section). We then collected 38,922 transcribed units (genes) in mouse from GENCODE (Harrow et al. 2012) and defined promoters as -1kb and +2kb region around the transcription start site of each transcribed unit. We then intersected these promoters with the H3K4me3 and H3K27me3 peaks. Despite the large variance in the number of H3K4me3 peaks identified in individual samples, the number of peaks within promoters was very consistent across samples ranging from 18 thousand to 20 thousand H3K4me3 marked promoters. This suggests that most promoters have a high peak height of H3K4me3 and therefore H3K4me3 is a distinguishing mark for promoters. In contrast, the number of H3K27me3 promoter peaks showed a large variance ranging from 3 thousand to 9 thousand peaks. The Pearson's correlation coefficient value between the total H3K27me3 peaks and the fraction of these in promoters was 0.5. This suggests that H3K27me3 does not show preference to promoters and therefore is not a distinguishing mark for promoters. The number of bivalent marked promoters varied between 2 thousand and 7 thousand across eight samples. Pearson's correlation coefficient between the number of H3K4me3 promoters and bivalent promoters was 0.58 while between H3K27me3 promoters and bivalent promoters was 0.98. This shows that the classification of a promoter as a bivalent promoter highly depends upon identification of H3K27me3 modification rather than H3K4me3 modification.

In order to identify the high confidence bivalent promoters, we calculated cumulatively the number of promoters identified with the H3K4me3 modification in 'n' or more samples. Over 20 thousand promoters were H3K4me3 marked in at least one sample, while about 15 thousand promoters were H3K4me3 marked in all eight samples. This demonstrates that H3K4me3 modification on promoters across samples is quite stable (Table 1). On the contrary, Over 11 thousand H3K27me3 promoters were detected in at least one sample of which only about 2 thousand were H3K27me3 marked in all samples (Table 1). The rate of decrease in the number of bivalent promoters (ratio of six or more to one or more) was 0.44, in H3K4me3 promoters was 0.81 and in H3K27me3 promoters was 0.37 in 'n' or more samples. This again demonstrates that the number of high confidence bivalent promoters is dependent on the H3K27me3 histone mark. We noted that consistently over 80% of H3K27me3 promoters were marked bivalent. This means that most H3K27me3 marked promoters also have H3K4me3 modification present. This demonstrates that the co-existence of these two chromatin modifications on promoters initially thought as a surprise, is rather a rule than exception. ChIP enrichment signals can be missed during peak calling procedure or by experimental error in an individual sample. Peaks detected in all samples are likely to miss true bivalent promoters. As the ratio of bivalent to H3K27me3 marked promoters was highly consistent when 4, 5 or 6 or more samples are taken into account, we decide to use an arbitrary cut off of six or more to define high confidence bivalent promoters. This resulted into identification of 16,885 high confidence H3K4me3 marked, 4,239 high confidence H3K27me3 marked and 3,740 high confidence bivalent promoters.

We then checked if the high confidence detection was biased towards any individual study or were true representative of all eight studies. About 50% of high confidence peaks were present in individual H3K4me3 samples while the fraction of high confidence H3K27me3 peaks in individual sample varied between 40 and 70%. This again demonstrates that H3K4me3 is consistent while H3K27me3 varies on the promoters.

Table 1. Cumulative count of three categories of promoters in mESCs with the peak based method. The cells with bold font (6 or more) represent the high confidence cut off chosen.

	MOUSE (WITH PEAK CALLING METHOD)			
Samples	**H3K4me3**	**H3K27me3**	**Bivalent**	**Biv./H3K27me3**
1 or more	20761	11610	8515	0.73
2 or more	19980	8931	7252	0.81
3 or more	19358	7413	6343	0.85
4 or more	18523	6198	5458	0.88
5 or more	17848	5175	4679	0.90
6 or more	**16885**	**4239**	**3740**	**0.88**
7 or more	16062	3287	2764	0.84
8 or more	14720	2236	1555	0.69

3.2 Cutoff-Based Method to Detect High-Confidence Bivalent Promoters

As the peak calling method is highly sensitive to the sequencing depth, we defined another independent method to identify enriched genomic regions for a specific histone modification, henceforth called cutoff-based method. We calculated the number of reads mapping to each promoter in each H3K4me3 and H3K27me3 samples by using custom scripts and BEDtools (Quinlan & hall 2010). In order to normalize the reads across multiple samples, the logarithmic scaled promoter read counts across all H3K4me3 and H3K27me3 experiments were quantile normalized separately (see Methods). The H3K4me3 normalized promoter read density followed a clear bimodal distribution separating H3K4me3 unmarked from marked promoters (Figure 1a). We noted further that the H3K4me3 positive and H3K4me3 negative sets were conserved across samples. On the contrary, the normalized promoter read density for H3K27me3 did not show a clear bimodal distribution making it hard to distinguish between the H3K27me3 positive and H3K27me3 negative sets (Figure 1b). Moreover, though H3K27me3 mark was coherent across samples, the distinction of two groups unlike H3K4me3 was not clear (Figure 1). We fitted a bimodal distribution to H3K4me3 log scaled promoter read densities and consistently obtained a cut off of 4.57 to distinguish between H3K4me3 positive and negative promoters (Figure 1a). On the other hand, as bimodal distribution failed to fit, we defined an arbitrary cut off of 3.00 to distinguish between H3K27me3 positive and negative promoters (Figure 1b). The cutoff based method identified consistently about 7 thousand H3K27me3 marked promoters and about 13 thousand H3K4me3 marked promoters.

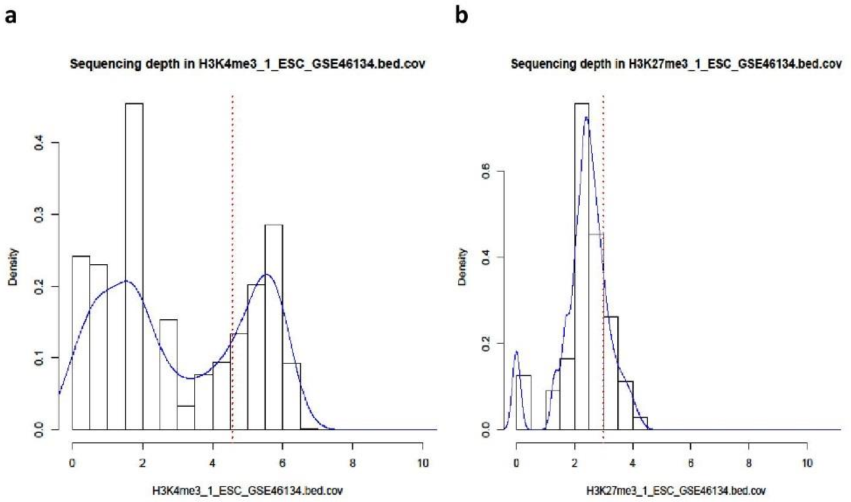


Fig. 1. Representative histograms and density plots for a) H3K4me3 and b) H3K27me3 samples in mouse ESCs. The vertical dotted red line marks the threshold used.

To identify high confidence bivalent promoters using the cutoff method, we calculated cumulatively the number of promoters identified with a given modification in ‘n’ or more samples. Both H3K4me3 and H3K27me3 marks showed a large variability across samples. Over 15 thousand promoters were H3K4me3 marked in at least one sample while only about 11 thousand promoters were H3K4me3 marked in all eight samples. Similarly, Over 16 thousand H3K27me3 promoters were detected in at least one sample from which only about 3 thousand were H3K27me3 marked in all samples (Table 2). The ratio levels were not as consistent as in the case of the peak calling method but for most of the cases (except for the extremes) more than 50% of the bivalent promoters were part of the repressed ones. Similar to the peak calling procedure, we used a threshold of six or more to define high confidence bivalent promoters. This resulted into the identification of 13,034 high confidence H3K4me3 marked, 4,660 high confidence H3K27me3 marked and 2,396 high confidence bivalent promoters.

Table 2. Cumulative count of three categories of promoters in mESCs with the cutoff based method. The cells with bold font (6 or more) represent the high confidence cutoff chosen.

	MOUSE (WITH CUTOFF BASED METHOD)			
Samples	**H3K4me3**	**H3K27me3**	**Bivalent**	**Biv/H3K27me3**
1 or more	15668	16624	7711	0.46
2 or more	14895	10327	5428	0.52
3 or more	14389	7748	4400	0.56
4 or more	13942	6419	3685	0.57
5 or more	13478	5479	3027	0.55
6 or more	**13034**	**4660**	**2396**	**0.51**
7 or more	12378	3846	1708	0.44
8 samples	11190	2829	945	0.33

3.3 Systematic Comparison of Peak-Based and Cutoff-Based Method

We performed a systematic comparison of the peak-based and cutoff-based method. Across individual samples, the variability in the total number of peaks identified by cutoff-based method was much lower compared to the peak-based method for both H3K4me3 and H3K27me3 data sets. Though cutoff-based method showed high consistency across samples for both modifications, there was more variability for the cutoff-based method when the cumulative analysis was performed (Table 1 & 2). We then compared the high confidence bivalent promoters obtained by both methods by defining the same threshold of six or more samples. The cutoff-based method concluded that only about 50% of H3K27me3 marked promoters were bivalent whereas peak-based method predicted this fraction to be over 80%. The peak-based method results are thus in agreement with literature. This was expected since the peak-based method is the most widely used approach in the literature. Over 80% of bivalent peaks detected by the cutoff method were also found by the peak method. The peak-based method is therefore able to identify high confidence bivalent promoters missed by the cutoff method (Figure 2a). Finally we calculated functional enrichment for bivalent promoters using both methods. Though both promoter sets were enriched for developmental categories such as anatomic structure development, and developmental process the enrichment was higher with the peak method than for the cutoff method (Figure 2b). Taken together, the peak-based method was more reliable in detecting high confidence bivalent promoters.

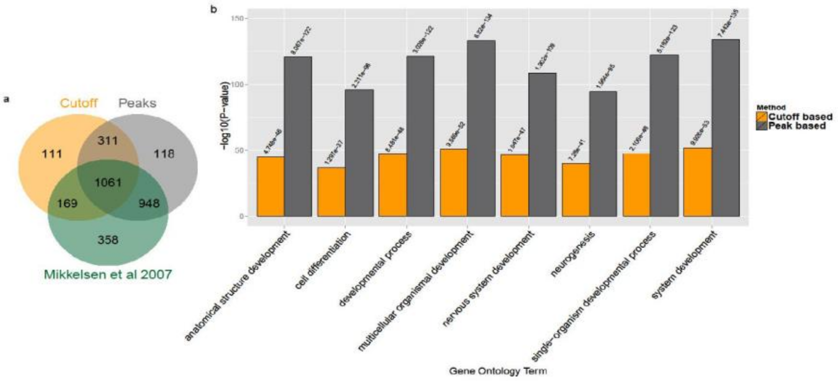


Fig. 2. a) Common bivalent promoters between the cutoff based method, the peak based method and from Mikkelsen et al., 2007 b) Functional enrichment values (-log10Pvalue) for the most enriched gene ontology terms for the two methods (P-value indicated on top of each bar)

3.4 Comparison of Serum-Grown High-Confidence Bivalent Promoters with 2i

Having established that peak detection method predicts reliable high confidence bivalent promoters, we used the bivalent promoters detected by the peak-based method for further analysis. Murine ES cells can be maintained in two distinct culture conditions in vitro, 2i (with inhibitors of two kinases Mek and GSK3) and serum. All eight samples used for high confidence bivalent promoter detection were grown in serum culture condition. Marks et al., 2012 identified 1,014 bivalent genes in murine ES cells grown under 2i media and 2,936 bivalent genes grown in serum and stated that the identification of fewer bivalent genes in ‘2i’ was in agreement with the postulated na ve ground state of ES cells grown in ‘2i’ and not in serum. If this were the case, the high confidence bivalent promoters should have higher overlap with 2i grown bivalent genes than bivalent genes detected in a serum grown sample. 76% of 2i-grown bivalent genes and 68% of serum-grown bivalent genes overlapped with our high confidence bivalent promoters respectively. 2i grown show higher overlap than serum-grown suggesting 2i might be more similar to na ve ground state. A fraction of 2i-grown bivalent genes were not identified bivalent in any of the ten samples grown in serum. This suggests that there are genes specifically bivalent marked in 2i and not in serum culture condition.

3.5 Identification of Bivalent Regions in Other Mammalian Species

In order to check if the high confidence bivalent regions are more conserved across species, we collected genome wide binding profiles for H3K4me3 and H3K27me3 in human ES cells and pig induced pluripotent cells. We collected 11 paired samples in humans from six studies with reads ranging from 13 million to 60 million in individual samples. We used the peak based method to call peaks in individual samples. These peaks were then mapped to promoters of 57,818 transcribed units defined by GENCODE (Harrow et al. 2012). Similar to mouse, the number of H3K4me3 promoter peaks were highly consistent across samples (mean 19,219.73, SD 462.88) while the number of H3K27me3 promoter peaks was variable (mean 8,035.73, SD 2,626.27). In order to identify high confidence human bivalent promoters we calculated bivalent promoters identified in 'n' or more samples. The rate of decrease in the number of bivalent promoters (ratio of eight or more to one or more) was 0.39, H3K4me3 promoters was 0.89 and H3K27me3 promoters was 0.31 in 'n' or more samples. The fraction of bivalent to H3K27me3 promoters was consistently higher than 80%. We used an arbitrary threshold of eight or more samples to define high confidence bivalent promoters. This resulted into the identification of 18,744 high confidence H3K4me3 marked, 5,841 high confidence H3K27me3 marked and 5,116 high confidence human bivalent promoters.

In pig (Sus Scrofa), only one study was available in the public domain hindering detection of high confidence bivalent promoters. Using 21,116 promoter regions we detected 8,383 H3K4me3 marked, 2,816 H3K27me3 marked and 1,561 bivalent marked promoters again demonstrating that over half of H3K27me3 marked promoters also contain an H3K4me3 modification.

Table 3. Cumulative count of three categories of promoters in mESCs with the peak based method. The cells with bold font (8 or more) represent the high confidence cut off chosen.

HUMAN (WITH PEAK CALLING METHOD)				
Samples	**H3K4me3**	**H3K27me3**	**Bivalent**	**Biv./H3K27me3**
1 or more	21167	18701	13206	0.70
2 or more	20275	12066	9778	0.81
3 or more	19865	9825	8236	0.83
4 or more	19602	8560	7308	0.85
5 or more	19341	7789	6713	0.86
6 or more	19123	7102	6177	0.86
7 or more	18944	6480	5660	0.87
8 or more	**18744**	**5841**	**5116**	**0.87**
9 or more	18489	5171	4505	0.87
10 or more	18189	4087	3495	0.85
11 samples	17678	2771	2202	0.79

3.6 Comparative Analysis of Bivalent and Promoters Across Three Species

Finally, we computed the overlap of bivalent promoters across three species by considering only one-to-one mapping orthologs. The bivalent promoters were less conserved across three species compared to the active promoters (Figure 3a and b). Specifically less than 10% of human bivalent promoters were conserved across three species while over 25% of H3K4me3 marked promoters were conserved across three species. The functional enrichment of common bivalent genes resulted in development processes more specific to embryogenesis, such as pattern specification process, embryonic morphogenesis and embryonic organ development, suggesting that the three species have more commonalities during embryonic development.

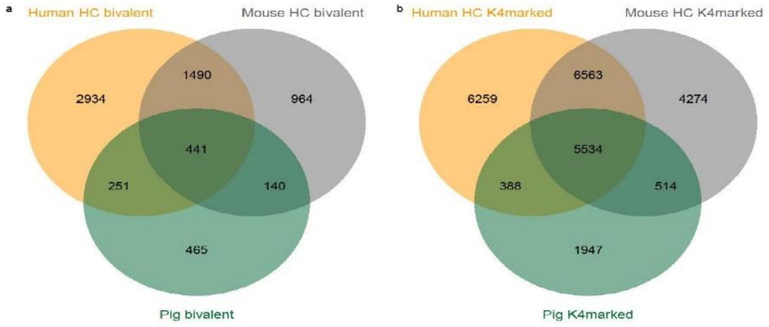


Fig. 3. Venn diagram of a) bivalent and b) K4marked promoters between human, mouse and pig using the peak calling method

4 Conclusion

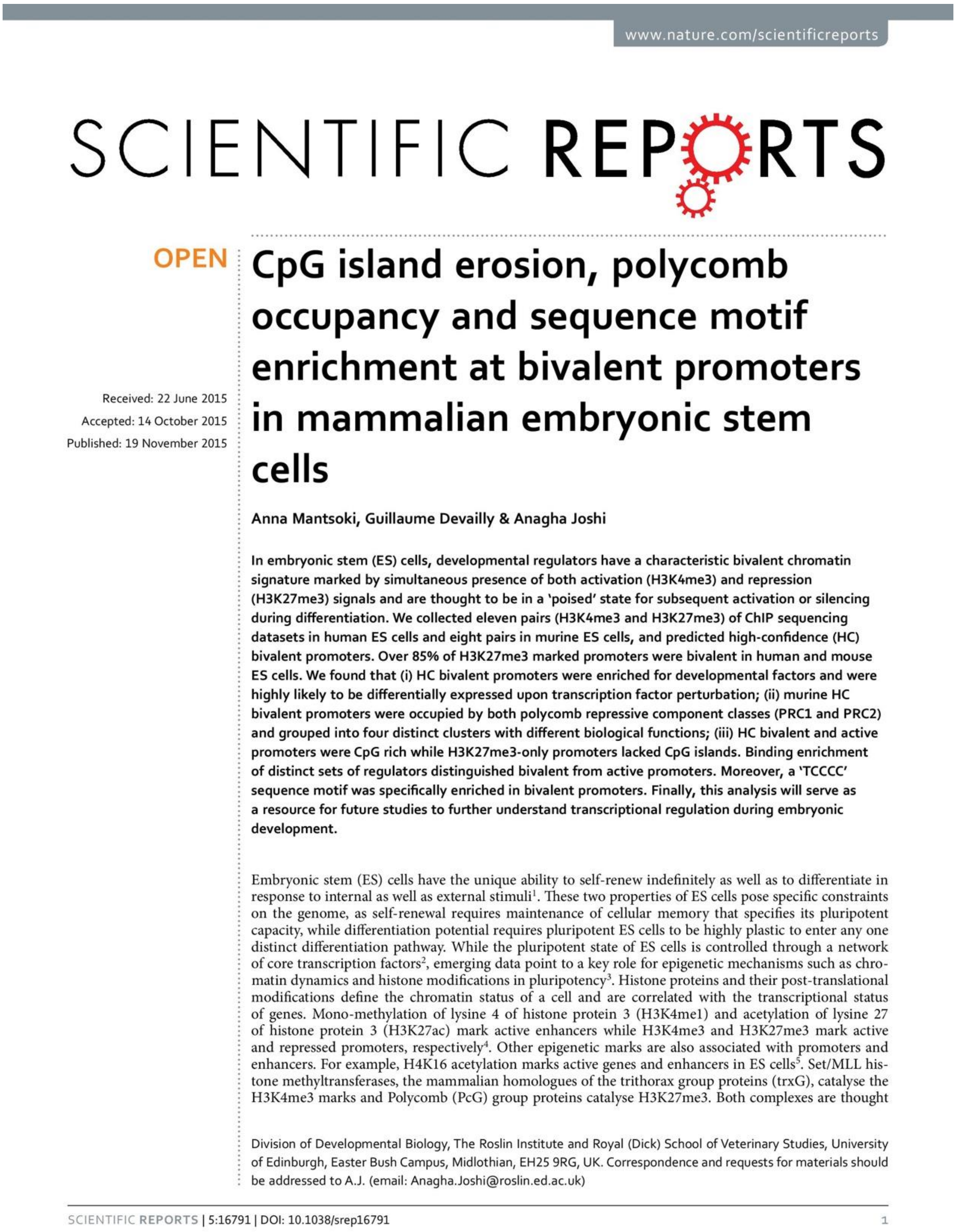
In summary, we identified high confidence bivalent domains in murine ES cells by integrating data across eight studies using two methods; peak-based and cutoff-based and demonstrated that the peak-based method is more reliable. We then identified bivalent promoters in human and pig and performed a multi-species comparative analysis of bivalent promoters to show that the conserved bivalent promoters were highly enriched for embryonic developmental processes.

Acknowledgements. A.J. is a Chancellors Fellow at the University of Edinburgh. This work was supported by the Roslin Institute Strategic Grant funding from the BBSRC.

References

1. Azuara, V., et al.: Chromatin signatures of pluripotent cell lines. Nature Cell Biology 8(5), 532–538 (2006)
2. Bannister, A., Kouzarides, T.: Regulation of chromatin by histone modifications. Cell Research 21(3), 381–395 (2011)

3. Barrett, T., Wilhite, S.E., Ledoux, P.: NCBI GEO: archive for functional genomics data sets—update (2013)
4. Benaglia, T., et al.: mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software 32(6), 1–29 (2009)
5. Bernstein, B., et al.: A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125(2), 315–326 (2006)
6. Bernstein, B.E., et al.: The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotech. 28(10), 1045–1048 (2010)
7. Dennis, G., et al.: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology, 4(5), P3 (2003)
8. Haider, S., et al.: BioMart Central Portal—unified access to biological data (2009)
9. Harrow, J., et al.: GENCODE: The reference human genome annotation for The ENCODE Project (2012)
10. Langmead, B., et al.: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome (2009)
11. Li, H., et al.: The sequence alignment/map format and SAMtools (2009)
12. Marks, H., et al.: The transcriptional and epigenomic foundations of ground state pluripotency. Cell 149(3), 590–604 (2012)
13. Mikkelsen, T., et al.: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448(7153), 553–560 (2007)
14. Mohn, F., et al.: Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Molecular Cell 30(6), 755–766 (2008)
15. Oksanen, J., et al.: vegan: Community Ecology Package (2013)
16. Pan, G., et al.: Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. Cell Stem Cell 1(3), 299–312 (2007)
17. Pearson, J., Lemons, D., McGinnis, W.: Modulating Hox gene functions during animal body patterning. Nature reviews. Genetics, 6(12), 893–904 (2005)
18. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6), 841–842 (2010)
19. Ringrose, L., Paro, R.: Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. Annual Review of Genetics 38, 413–443 (2004)
20. Andrews, S.: FastQC A Quality Control tool for High Throughput Sequence Data (2010)
21. Sharov, A.A., Ko, M.S.: Human {ES} cell profiling broadens the reach of bivalent domains. Cell Stem Cell 1(3), 237–238 (2007)
22. Thomson, J., et al.: Embryonic stem cell lines derived from human blastocysts. Science 282(5391), 1145–1147 (1998)
23. Voigt, P., Tee, W.-W., Reinberg, D.: A double take on bivalent promoters. Genes & Development 27(12), 1318–1338 (2013)
24. Wilson, N., et al.: Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell 7(4), 532–544 (2010)
25. Xiao, S., et al.: Comparative epigenomic annotation of regulatory DNA. Cell 149(6), 1381–1392 (2012)
26. Zang, C., et al.: A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25(15), 1952–1958 (2009)
27. 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(3), 286–298 (2007)



www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

CpG island erosion, polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells

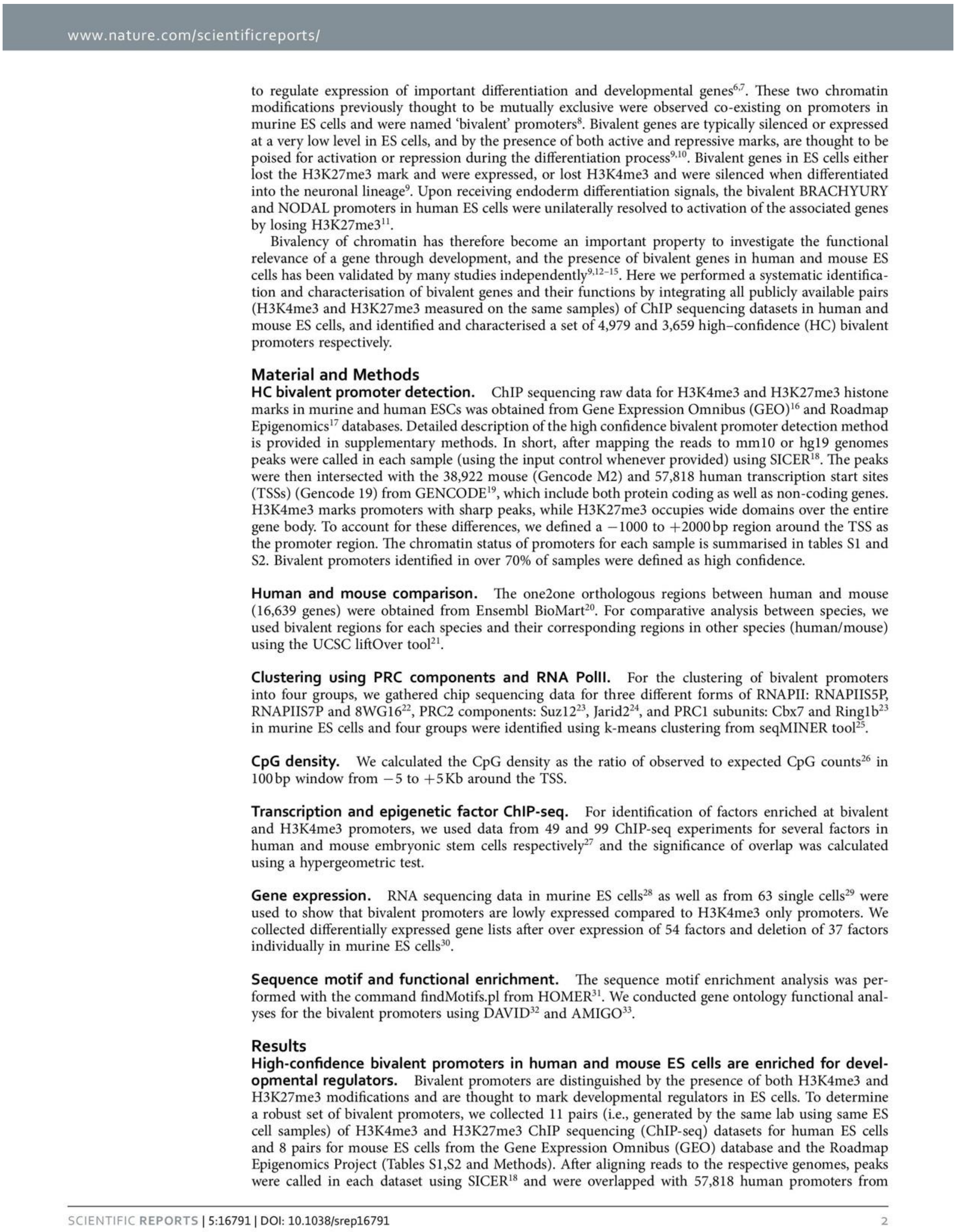
Received: 22 June 2015
Accepted: 14 October 2015
Published: 19 November 2015

Anna Mantsoki, Guillaume Devailly & Anagha Joshi

In embryonic stem (ES) cells, developmental regulators have a characteristic bivalent chromatin signature marked by simultaneous presence of both activation (H3K4me3) and repression (H3K27me3) signals and are thought to be in a 'poised' state for subsequent activation or silencing during differentiation. We collected eleven pairs (H3K4me3 and H3K27me3) of ChIP sequencing datasets in human ES cells and eight pairs in murine ES cells, and predicted high-confidence (HC) bivalent promoters. Over 85% of H3K27me3 marked promoters were bivalent in human and mouse ES cells. We found that (i) HC bivalent promoters were enriched for developmental factors and were highly likely to be differentially expressed upon transcription factor perturbation; (ii) murine HC bivalent promoters were occupied by both polycomb repressive component classes (PRC1 and PRC2) and grouped into four distinct clusters with different biological functions; (iii) HC bivalent and active promoters were CpG rich while H3K27me3-only promoters lacked CpG islands. Binding enrichment of distinct sets of regulators distinguished bivalent from active promoters. Moreover, a 'TCCCC' sequence motif was specifically enriched in bivalent promoters. Finally, this analysis will serve as a resource for future studies to further understand transcriptional regulation during embryonic development.

Embryonic stem (ES) cells have the unique ability to self-renew indefinitely as well as to differentiate in response to internal as well as external stimuli[1]. These two properties of ES cells pose specific constraints on the genome, as self-renewal requires maintenance of cellular memory that specifies its pluripotent capacity, while differentiation potential requires pluripotent ES cells to be highly plastic to enter any one distinct differentiation pathway. While the pluripotent state of ES cells is controlled through a network of core transcription factors[2], emerging data point to a key role for epigenetic mechanisms such as chromatin dynamics and histone modifications in pluripotency[3]. Histone proteins and their post-translational modifications define the chromatin status of a cell and are correlated with the transcriptional status of genes. Mono-methylation of lysine 4 of histone protein 3 (H3K4me1) and acetylation of lysine 27 of histone protein 3 (H3K27ac) mark active enhancers while H3K4me3 and H3K27me3 mark active and repressed promoters, respectively[4]. Other epigenetic marks are also associated with promoters and enhancers. For example, H4K16 acetylation marks active genes and enhancers in ES cells[5]. Set/MLL histone methyltransferases, the mammalian homologues of the trithorax group proteins (trxG), catalyse the H3K4me3 marks and Polycomb (PcG) group proteins catalyse H3K27me3. Both complexes are thought

Division of Developmental Biology, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, UK. Correspondence and requests for materials should be addressed to A.J. (email: Anagha.Joshi@roslin.ed.ac.uk)



to regulate expression of important differentiation and developmental genes[6,7]. These two chromatin modifications previously thought to be mutually exclusive were observed co-existing on promoters in murine ES cells and were named 'bivalent' promoters[8]. Bivalent genes are typically silenced or expressed at a very low level in ES cells, and by the presence of both active and repressive marks, are thought to be poised for activation or repression during the differentiation process[9,10]. Bivalent genes in ES cells either lost the H3K27me3 mark and were expressed, or lost H3K4me3 and were silenced when differentiated into the neuronal lineage[9]. Upon receiving endoderm differentiation signals, the bivalent BRACHYURY and NODAL promoters in human ES cells were unilaterally resolved to activation of the associated genes by losing H3K27me3[11].

Bivalency of chromatin has therefore become an important property to investigate the functional relevance of a gene through development, and the presence of bivalent genes in human and mouse ES cells has been validated by many studies independently[9,12–15]. Here we performed a systematic identification and characterisation of bivalent genes and their functions by integrating all publicly available pairs (H3K4me3 and H3K27me3 measured on the same samples) of ChIP sequencing datasets in human and mouse ES cells, and identified and characterised a set of 4,979 and 3,659 high–confidence (HC) bivalent promoters respectively.

Material and Methods

HC bivalent promoter detection. ChIP sequencing raw data for H3K4me3 and H3K27me3 histone marks in murine and human ESCs was obtained from Gene Expression Omnibus (GEO)[16] and Roadmap Epigenomics[17] databases. Detailed description of the high confidence bivalent promoter detection method is provided in supplementary methods. In short, after mapping the reads to mm10 or hg19 genomes peaks were called in each sample (using the input control whenever provided) using SICER[18]. The peaks were then intersected with the 38,922 mouse (Gencode M2) and 57,818 human transcription start sites (TSSs) (Gencode 19) from GENCODE[19], which include both protein coding as well as non-coding genes. H3K4me3 marks promoters with sharp peaks, while H3K27me3 occupies wide domains over the entire gene body. To account for these differences, we defined a −1000 to +2000 bp region around the TSS as the promoter region. The chromatin status of promoters for each sample is summarised in tables S1 and S2. Bivalent promoters identified in over 70% of samples were defined as high confidence.

Human and mouse comparison. The one2one orthologous regions between human and mouse (16,639 genes) were obtained from Ensembl BioMart[20]. For comparative analysis between species, we used bivalent regions for each species and their corresponding regions in other species (human/mouse) using the UCSC liftOver tool[21].

Clustering using PRC components and RNA PolII. For the clustering of bivalent promoters into four groups, we gathered chip sequencing data for three different forms of RNAPII: RNAPIIS5P, RNAPIIS7P and 8WG16[22], PRC2 components: Suz12[23], Jarid2[24], and PRC1 subunits: Cbx7 and Ring1b[23] in murine ES cells and four groups were identified using k-means clustering from seqMINER tool[25].

CpG density. We calculated the CpG density as the ratio of observed to expected CpG counts[26] in 100 bp window from −5 to +5 Kb around the TSS.

Transcription and epigenetic factor ChIP-seq. For identification of factors enriched at bivalent and H3K4me3 promoters, we used data from 49 and 99 ChIP-seq experiments for several factors in human and mouse embryonic stem cells respectively[27] and the significance of overlap was calculated using a hypergeometric test.

Gene expression. RNA sequencing data in murine ES cells[28] as well as from 63 single cells[29] were used to show that bivalent promoters are lowly expressed compared to H3K4me3 only promoters. We collected differentially expressed gene lists after over expression of 54 factors and deletion of 37 factors individually in murine ES cells[30].

Sequence motif and functional enrichment. The sequence motif enrichment analysis was performed with the command findMotifs.pl from HOMER[31]. We conducted gene ontology functional analyses for the bivalent promoters using DAVID[32] and AMIGO[33].

Results

High-confidence bivalent promoters in human and mouse ES cells are enriched for developmental regulators. Bivalent promoters are distinguished by the presence of both H3K4me3 and H3K27me3 modifications and are thought to mark developmental regulators in ES cells. To determine a robust set of bivalent promoters, we collected 11 pairs (i.e., generated by the same lab using same ES cell samples) of H3K4me3 and H3K27me3 ChIP sequencing (ChIP-seq) datasets for human ES cells and 8 pairs for mouse ES cells from the Gene Expression Omnibus (GEO) database and the Roadmap Epigenomics Project (Tables S1,S2 and Methods). After aligning reads to the respective genomes, peaks were called in each dataset using SICER[18] and were overlapped with 57,818 human promoters from

GENCODE 19[19] and 38,922 murine promoters from GENCODE M2[19]. The number of H3K4me3 marked promoters across data sets was highly consistent (human: mean 18,632.55 relative SD 2.8%, mouse: mean 17,554.25 relative SD 11%), in contrast to the number of H3K27me3 marked promoters (human: mean 7,523.45 relative SD 37%, mouse: mean 6,128.75 relative SD 35%) (Tables S3 and S4). Moreover, the same promoters were consistently identified as H3K4me3 marked across samples, as demonstrated by incrementally intersecting the peaks from multiple datasets (Fig. 1A, green curve). In contrast, the H3K27me3 marked promoters (Fig. 1A, purple curve) varied across datasets, strongly influencing the number of bivalent promoters detected (Fig. 1A, yellow curve). Assigning a bivalent status to a promoter is therefore largely subject to H3K27me3 peak identification on the promoter. Over 85% of H3K27me3 marked promoters in both human and mouse were bivalent promoters (Fig. 1A, Tables S5 and S6). Thus, we reconfirm that bivalency at the H3K27me3 marked promoters is rather a rule than an exception[15]. The sequencing depth across samples varied from 14 million to over 100 million which might contribute to the variation of bivalent promoter detection in individual datasets. Indeed there was a high correlation between the number of reads and number of peaks across murine datasets (for H3K27me3 Pearson's correlation coefficient (***r***) = 0.75, for H3K4me3 ***r*** = 0.84), but not across human datasets (for H3K27me3 ***r*** = −0.20, for H3K4me3 ***r*** = 0.14). There are other factors contributing to the variation between samples, for example ES cells were grown in diverse culture conditions, and using different cell lines as well as various antibodies across datasets (Tables S1 and S2). We therefore defined bivalent promoters identified in more than 70% of the datasets (eight or more human datasets and six or more murine datasets) as high confidence (HC), resulting in 4,979 human and 3,659 murine HC bivalent promoters (Fig. 1A). Eight HC bivalent regions were validated by ChIP qPCR for the presence of H3K27me3 modification[9] (Table S11). Adding or removing a sample in defining HC promoters did not change the key findings of the downstream analysis (see supplementary methods and Figure S1). There was no strong correlation between the fraction of HC bivalent promoters detected in a sample and the sequencing depth of that sample for both histone modifications (Pearson's Correlation: Human: r = −0.34 H3K27me3, r = −0.38 H3K4me3, Mouse: r = 0.35 H3K27me3, r = −0.112 H3K4me3) (Figure S2).

HC bivalent promoters had higher H3K27me3 read density than H3K27me3-only promoters in any individual dataset (Student's t-test, P-value < 0.0001) (Fig. 1B and S7), while H3K4me3 read density at HC bivalent promoters was lower than at H3K4me3-only promoters (Student's t-test, P-value < 0.0001) (Figures S3 and S8). To test whether integration of multiple samples simply resulted in selecting the peaks with the strongest signal (peak height) from individual H3K27me3 samples, we selected the top (highest H3K27me3 signal) 4,979 human and 3,659 murine bivalent promoter peaks in each dataset and calculated the overlap with HC bivalent promoters. Less than 2/3rd of H3K27me3 top promoters in any individual dataset overlapped with HC bivalent promoters (Figure S4).

We also checked whether the peaks of H3K27me3 and H3K4me3 modifications were present at the same genomic location within a promoter region and found that over 95% of H3K27me3 and H3K4me3 peaks overlapped in each pair of samples at HC bivalent promoters. Both chromatin modifications were indeed present at the same genomic location (Figure S5). We compared the functional enrichment between high-confidence and non-high-confidence (detected as bivalent in less than 70% of datasets) bivalent promoters and found that only the high-confidence promoters were strongly enriched for processes such as 'cell differentiation' and 'system development' (Fig. 1C). Interestingly, metabolic processes were enriched in murine but not human HC bivalent promoters.

In summary, by integrating data from multiple studies we identified HC human and murine bivalent promoters, which could not be identified by simply selecting the top peaks from individual samples. The HC bivalent promoters were highly enriched for developmental regulators compared to non-HC bivalent promoters.

High-confidence bivalent promoters are marked by PRC1, PRC2 and RNA polymerase II. Bivalent promoters are known to show variation in their levels of occupancy by RNA polymerase II[22] and PRC complexes[12] . To further characterize HC bivalent promoters, we gathered ChIP-seq data in murine ES cells for various forms of RNAPII phosphorylated in different residues (RNAPIIS5P and RNAPIIS7P) as well as RNAPII8WG16 (an antibody that recognizes mostly unphosphorylated PolII)[22], together with ChIP-seq data for the SUZ12, a subunit of PRC2, responsible for catalysing the histone modification H3K27me3, the RING1B and CBX7 subunits of PRC1[23], responsible for catalysing H2Aub1 and for compacting chromatin, and Jarid2[24]. Jarid2 is a co-factor of PRC2 and is methylated by PRC2 which in turn promotes PRC2 activity[34]. All HC bivalent promoters were marked by both PRC1 and PRC2 components albeit at different levels (Fig. 2A).

HC bivalent promoters could be classified in four distinct clusters based on the presence of PRC1 components and forms of RNAPII (Fig. 2A). The first two clusters had low PRC1 (Ring1b) levels and high RNAPII (8WG16) levels compared to clusters 3 and 4. The second cluster distinguished from the first cluster by the presence of RNAPII (8WG16 and S5P) modifications as a sharp peak on the promoter. The second cluster consisted of the only group of bivalent promoters marked with RNAPII (S7P). This cluster was enriched for genes involved in metabolic processes. The third and fourth clusters were marked by strong PRC1 (Ring1b), PRC2 (Suz12) and RNAPII (S5P) modifications. Cluster 3 and 4 were distinguished based on the fact that PRC components formed wide domains on cluster 3 and narrow peaks on cluster 4 promoters. Cluster 3 promoters were enriched for regulation of transcription (P value $< 10^{-51}$)

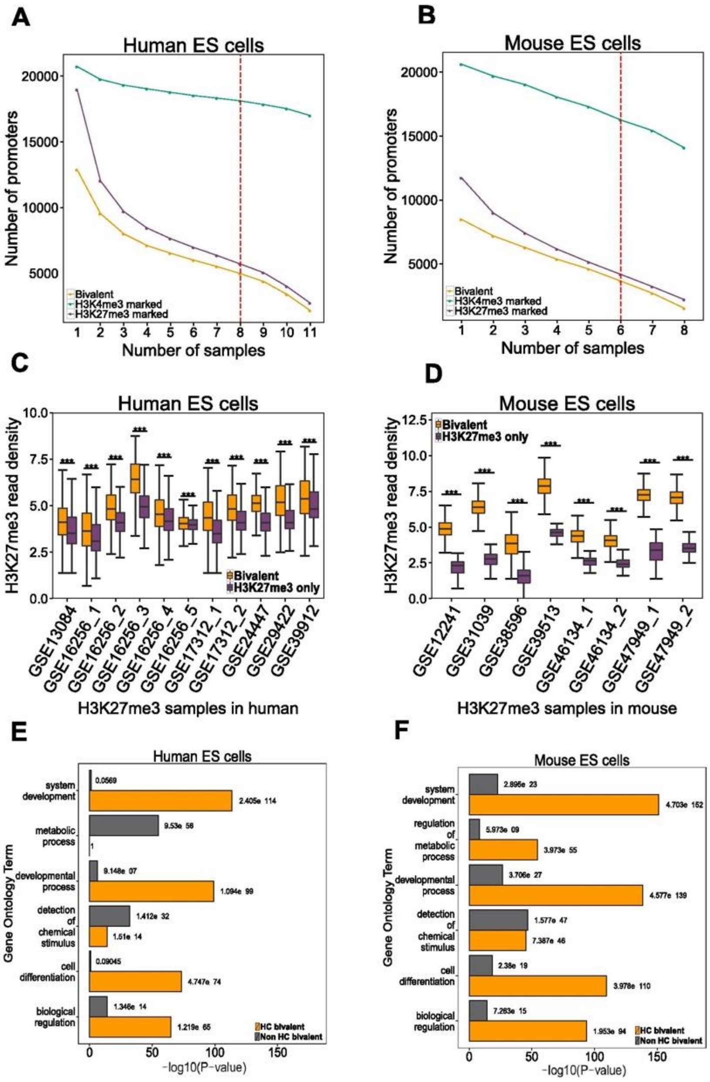


Figure 1. Identification of high confidence bivalent promoters in human and mouse ES cells. (**A**) The number of H3K4me3 (green), H3K27me3 (purple) and bivalent (yellow) promoters detected in 'n' or more samples (x axis) in human (left) and mouse ES cells (right). The red dotted line represents the cut off used to define high-confidence bivalent promoters. (**B**) H3K27me3 read density (in log scale) at bivalent and H3K27me3 only promoters in each sample designated by their GEO accession number (x axis) in human (left) and mouse (right) ES cells (***P-value $< 10^{-4}$). (**C**) Gene Ontology terms enriched in HC bivalent promoter list (yellow) or non HC bivalent promoter list (grey) in human (left) and mouse (right) ES cells with their corresponding P-value.

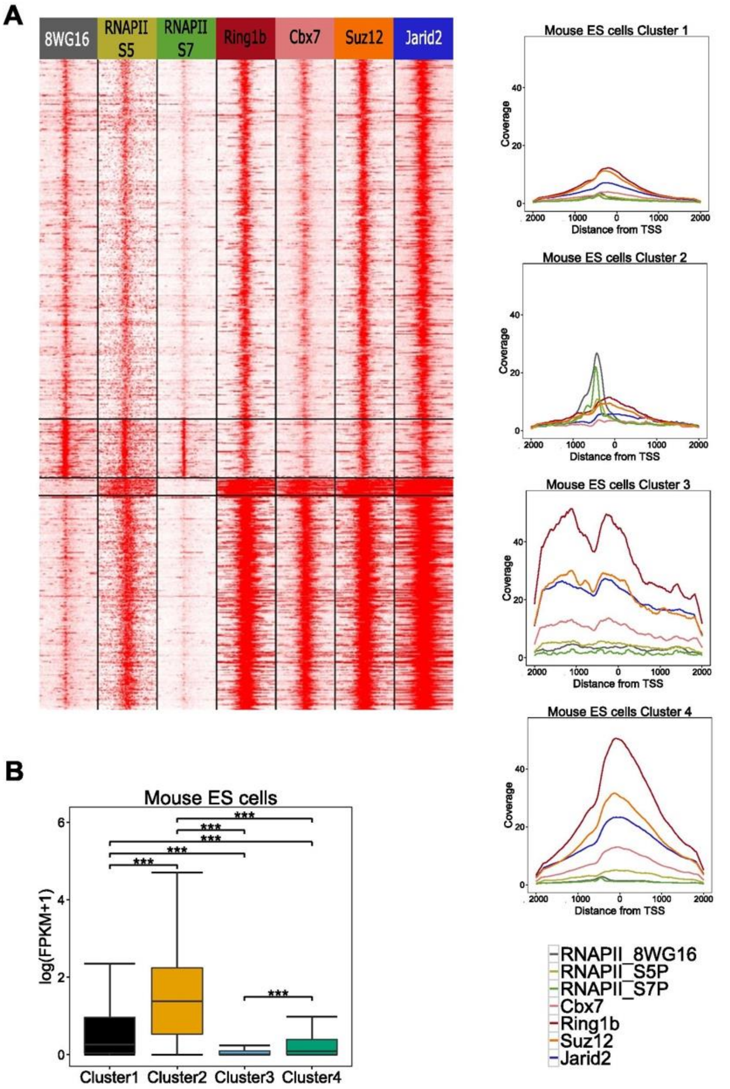


Figure 2. Four groups of HC bivalent promoters with distinct biological features. (**A**) HC bivalent promoters in murine ES cells classified in four subgroups based on occupancy of PRC1 components (Ring1b, Cbx7), PRC2 (Suz12), Jarid2 and RNA polymerase II (Ser7P, Ser5P, 8WG16). Each line represents one single promoter while color code summarizes ChIP-seq read densities, from −5kb to +5Kb around TSS. For each cluster, mean read coverage around TSS is shown on the right. (**B**) Expression levels in mouse ES cells using RNA sequencing data for each of the four clusters. FPKM: Fragment per kilo-base per million (***P-value $< 10^{-4}$).

while cluster 4 promoters were enriched for developmental functions such as organ morphogenesis (P value $< 10^{-27}$). Cluster 3 promoters contained transcription factors important for specific lineages like haematopoiesis factors Gfi1 and Meis1, whereas cluster 4 contained multiple members of transcription factor families controlling development such as winged helix/forkhead box (Fox) and Hox families.

We noted that bivalent promoters could be distinguished into two groups based on PRC1 occupancy: PRC1 low (cluster 1 & 2) and PRC1 high (cluster 3 & 4). Ku *et al.* (12) suggested that PRC1 was absent in our PRC1 low bivalent promoter (Figure S11). Ring1b ChIP sequencing at higher sequencing depth confirms that all bivalent promoter are bound by PRC1 albeit at different levels. The PRC1 high group separated into two distinct groups each enriched for a distinct functional category, namely cluster 3 for transcription factors and cluster 4 for developmental controllers. Based on RNA PolII occupancy, PRC1 low consisted of two distinct gene sets: PolII-low (S7P) and PolII-high (S7P). The difference in chromatin signature of these two clusters was also reflected in the expression level namely PolII-high (cluster 2) promoters were expressed at higher levels than PolII-low or cluster 1 promoters (Kruskal-Wallis test P-value < 0.0001) (Fig. 2B).

In summary, all HC bivalent promoters are occupied by components of both PRC1 and PRC2. There exists a distinct set of metabolic genes (cluster 2) which though bivalently marked has RNAPII (S7P) and is expressed at a higher level than other bivalent genes.

Bivalent promoters are lowly expressed and highly sensitive to perturbations in ES cells. RNA polymerase II (PolII) may be present but stalled at the promoters of bivalent genes and short (abortive) transcripts may be detected at their promoters[35]. To check whether bivalent genes indeed show a low or leaky expression, we collected RNA sequencing data for murine[28] and human[36] ES cells and calculated the mean expression level for the following categories of promoters: We classified promoters into four HC groups (Table S9 and S10) depending on the presence or absence of one or both chromatin modifications in over 70% of samples as bivalent promoters, promoters marked only with H3K27me3 (H3K27me3-only), promoters marked only with H3K4me3 (henceforth called 'active') and latent promoters (unmarked for H3K27me3 and H3K4me3). Promoters that belonged to any of the previous four categories in less than 70% of the samples, and thus were not considered in that category were marked as unclassified. In human and mouse ES cells, most active promoters were expressed at higher levels than bivalent promoters, and latent promoters were mostly not expressed (FPKM $= 0$) (Kruskal-Wallis test, P-value < 0.0001) (Fig. 3A). Low expression can result from two scenarios: either a gene is expressed at low levels in most cells or few cells express a gene while others do not. To determine whether lowly expressed genes in the four groups can be classified into one of the two scenarios, we downloaded single cell RNA sequencing data for 63 mouse ES cells[29]. Lowly expressed (i.e. FPKM < 4, or log(FPKM) < 1.4) active promoters were expressed in a similar number of single cells as lowly expressed bivalent promoters (Kruskal-Wallis test, P-value > 0.05) (Fig. 3B) demonstrating that single cell gene expression data cannot distinguish between bivalent and active lowly expressed genes.

As bivalent genes are thought to be poised for activation or repression, we hypothesised that these genes might be more likely to be differentially expressed upon perturbation of ES cells. We therefore used a collection of differentially expressed genes upon deletion or over-expression of 91 transcription and epigenetic factors in mouse ES cells, and found that 98% of differentially expressed gene sets by the overexpression of at least one TF significantly overlapped (Hypergeometric test, P value $<$ 1e-3) with bivalent genes, and 89% differentially expressed gene sets by the down-regulation of at least one TF (Fig. 3C). To check whether this is a property of bivalent genes or lowly expressed genes in general, we also calculated the overlap of active and latent lowly expressed genes with the differentially expressed gene sets upon transcription and epigenetic factor perturbation. We confirmed that bivalent genes are highly susceptible to perturbations compared to active or latent lowly expressed genes (Kruskal-Wallis test, P-value < 0.001) (Figure S6).

Over 50% of bivalent promoters maintain their chromatin status as well as gene expression profile across species. To perform a systematic comparison of chromatin status between human and mouse promoters in ES cells, we used 16,639 one-to-one orthologous genes between the two species[20]. We classified orthologous promoters into four HC groups – active (H3K4me3-only), H3K27me3-only, bivalent and latent. Promoters that did not belong to any of the previously mentioned groups were designated as 'unclassified'. We confirmed that HC H3K27me3-only and active promoters indeed had low or no other chromatin modification (Figures S7 and S8). We then calculated the overlap of the five groups across species (Fig. 4A). Over 40% of murine orthologous promoters (n $= 6964$) contain an activating mark (H3K4me3-only), in contrast to only 24% of human orthologous promoters (n $= 3961$). There was a 47% overlap of murine active promoters with human active promoters; while 84% of human active promoters overlapped with murine active promoters i.e. most active promoters in human are also active in mouse but not vice versa. Bivalent promoters constitute 17% (n $= 2854$) and 20% (n $= 3342$) of mouse and human orthologous genes respectively. 66% of murine bivalent promoters are also bivalent in human and 56% of human bivalent promoters are bivalent in mouse. The promoters with the H3K27me3-only modification form a very small fraction of orthologous promoters reaching merely 0.2% (n $= 45$) and 0.3% (n $= 66$) in mouse and human respectively. About 20% of H3K27me3-only promoters in one species are bivalent in the other species. Conserved bivalent promoters were enriched

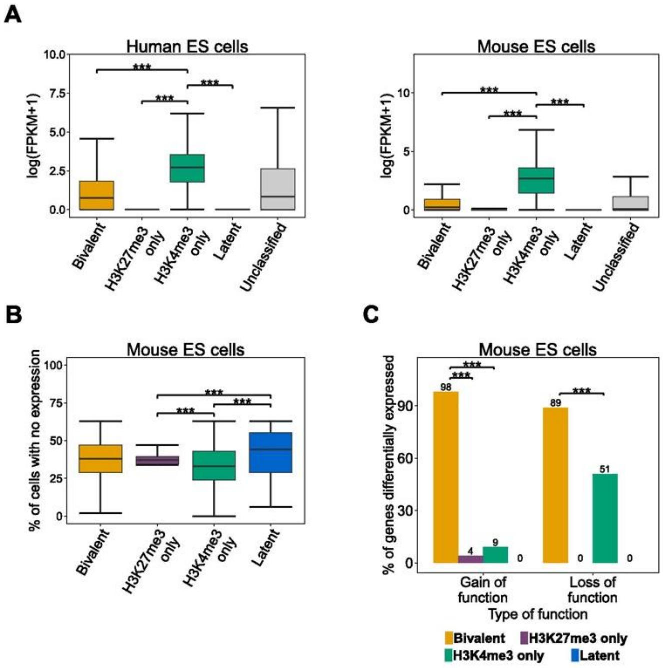


Figure 3. Bivalent promoters are lowly expressed in ES cells and are more likely to be differentially expressed upon perturbation. (**A**) Expression levels according to human (left) or mouse (right) ES cells RNA-seq for HC bivalent promoters (yellow), promoters marked with H3K27me3 only (purple), promoters marked with H3K4me3 only (green), latent promoter (blue), or unclassified promoters (grey, see text) (***P-value $< 10^{-4}$). (**B**) From single cell RNA-seq data of mouse ES cells, percentage of cells non expressing the lowly expressed genes (i.e. FPKM < 4) was computed for different classes of promoters (bivalent, H3K27me3 only, H3K4me3 only and latent) (***P-value $< 10^{-4}$). (**C**) HC bivalent promoters are hypersensitive to changes in the transcription network perturbation. Differentially expressed gene lists were collected from studies overexpressing one of 54 factors (gain of function) or down-regulation of one of 37 factors in ES cells. Percentage of significantly overlapping (P value $<$ 1e-3) bivalent, H3K27me3 only, H3K4me3 only and latent genes with differentially expressed in at least one of the experiments is represented (***P-value $< 10^{-4}$).

for functional categories developmental protein (P value $< 10^{-71}$) and transcription factor activity (P value $< 10^{-65}$); whereas species-specific promoters were not enriched for the two above terms (Table S7). Specifically, the mouse-specific bivalent promoters were enriched for membrane (P value $< 10^{-16}$) and glycoprotein (P value $< 10^{-13}$) and the human-specific for plasma membrane part (P value $< 10^{-5}$) and alternative splicing (P value $< 10^{-3}$).

To check whether the chromatin status across species is reflected in the gene expression status, we focused on five groups of promoters (Fig. 4B): three groups (I, II and III) with conserved chromatin status and two groups with divergent chromatin status (IV and V) across species. The gene expression profiles of conserved chromatin groups across species were also conserved. Specifically, active promoters (II) were expressed at higher level than bivalent promoters (I) which in turn were expressed at higher level than latent promoters (III) in both human and mouse ES cells (Kruskal-Wallis test, P-value < 0.0001) (Fig. 4B). The divergence of chromatin status promoters across species was not reflected in the gene expression level. For example the orthologous promoters with bivalent status in human and active status in mouse (IV) were expressed at intermediate levels between active (II) and bivalent promoters (I) in both species (Fig. 4B).

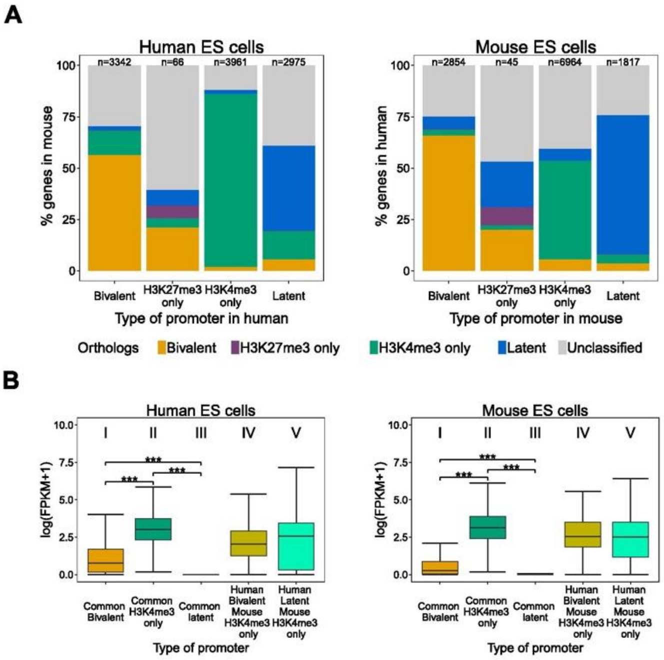


Figure 4. Over 50% of bivalent promoters maintain their chromatin status as well as gene expression profile across species. (**A**) Overlap of high confidence (HC) H3K4me3 only (green), H3K7me3 only (purple), bivalent (yellow) and latent (blue, absence of both H3K4me3 and H3K7me3 modifications) in human ES cells with the corresponding categories in mouse ES cells (left), and vice versa (right). Grey: Unclassified promoters (see text). (**B**) Expression levels in human (right) and mouse (left) ES cells using RNA sequencing data for each of the five groups of orthologous genes identified in (**A**) (***P-value $< 10^{-4}$).

Bivalent promoters are CpG rich while H3K27me3-only promoters are CpG poor. As shown in the first section, the bivalent status of promoters is primarily determined by the detection of an H3K27me3 modification (Fig. 1A). CpG islands (CGIs) have been implicated in polycomb recruitment and therefore H3K27me3 modification[37–39]. CGIs are CpG-rich genomic regions and are sites of transcription initiation[40]. CGI promoters are silenced by either DNA methylation or polycomb group proteins with approximately a fifth of CGI promoters accounting for bivalent promoters in ES cells[12]. About 35% of all GENCODE genes in both human and mouse overlapped with at least one CGI. When only protein coding genes were considered, this overlap increased to 67% for human and 54% for mouse (Fig. 5A). Mouse promoters in most categories showed lower overlap with CGIs than human promoters (Fig. 5A). 89% of human active (H3K4me3-only) as well as 82% of murine active promoters contained at least one CGI (Fig. 5A).

Over 90% of our HC bivalent promoters in ES cells in both species overlap with at least one CGI region, whereas only 8% (37 of 397) of human H3K27me3 only promoters contained a CGI and no mouse H3K27me3 only promoters (none of 152) contained a CGI (Fig. 5A). Previously CGIs have been associated with H3K27me3 modification in mammalian ES cells[41,42], but our results show that this is the case for bivalent promoters but not for H3K27me3 only promoters. We confirmed that the lack of CGIs on active promoters is not due to the CGI detection threshold and that the CpG density at repressed promoters is indeed significantly lower than at CGIs (Kruskal-Wallis test, P-value < 0.0001)(Fig. 5B). It has been proposed that a high density of un-methylated CpG is sufficient for vertebrate polycomb recruitment[42]. The fact that H3K27me3-only promoters are specifically CpG-poor (Fig. 5A,B), suggests that, although highly unmethylated CpG islands might be sufficient for polycomb recruitment, they might not be necessary.

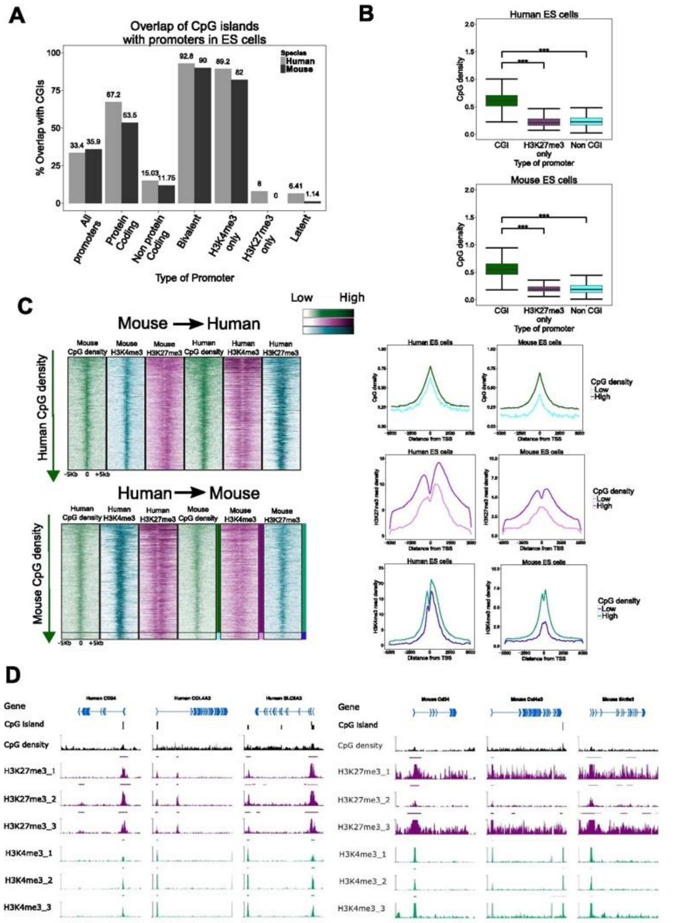


Figure 5. Bivalent promoters are CpG island rich while H3K27me3 only are CGI poor. (**A**) Percentage of promoters overlapping with one or more CpG island in human (grey) or mouse (black). (**B**) CpG ratio at H3K27me3 only promoters is similar to non-CGI promoters in human (top) and mouse (bottom) ES cells (***P-value $< 10^{-4}$). (**C**) Relationship between CpG density, H3K27me3 modification and H3K4me3 modification in human and mouse ES cells. There is a loss of human CGI promoters in mouse (bottom, below marked black line) but no loss of mouse CGI promoters in human (top). This loss is linked with decreasing H3K4me3 and H3K27me3 in mouse as compare with human. Left panels indicate mean CpG densities, mean H3K27me3 read densities and mean H3K4me3 densities in human and mouse. (**D**) Exemplar murine promoters where CGI loss on promoters does not correspond to the loss H3K27me3 modification. These promoters despite losing CGI keep bivalent promoter status in murine ES cells.

The loss of H3K27me3 in rodents (mouse and rat) compared to human ES cells at many developmental genes has been associated with depletion of CGIs; mouse CGI erosion has been characterised at MYO1G, CLEC4G and MYF6 gene loci with corresponding H3K27me3 loss[41]. We performed a cross-species comparison of CpG density, H3K4me3 and H3K27me3 profiles of bivalent promoters

(Fig. 5C). Indeed, about 5% of bivalent human promoters lost CGIs in mouse but not vice versa (indicated by black horizontal line). There was a high correlation between CpG density and H3K4me3 as well as H3K27me3 profiles within each species as well as across species (Fig. 5C), but the concordance between loss/gain of CGIs and H3K4me3 and/or H3K27me3 mark does not always hold true. Of 70 orthologous CpG-rich bivalent promoters in human where CGI was lost in mouse and analysed their chromatin status, only 18% of these promoters had clearly lost their H3K27me3 mark in mouse ES cells, of which half were classified as H3K4me3-only and the rest as latent in murine ES cells (Figure S9). Despite losing CGI on murine promoters, 20% of these orthologous promoters maintained a bivalent chromatin status including Col4a3, Cd34 and Slc6a3 (Fig. 5D).

In summary, the H3K27me3-only CpG-poor promoters demonstrate that polycomb recruitment does not only depend on CpG density. Although the CpG density largely correlates with H3K4me3 and H3K27me3 profiles across promoters, the loss of CGI on a promoter does not always imply a corresponding loss of the H3K4me3 and/or H3K27me3 modification on that promoter.

Bivalent promoters are occupied by fewer transcription factors than active promoters and are specifically enriched in a 'TCCCC' sequence motif. As both active (H3K4me3-only) and bivalent promoters are CpG-rich, we investigated possible modes of distinction between the two in ES cells. Voigt, Tee, and Reinberg[43] proposed a model where the density of transcription factors at the promoters determines establishment of bivalent domains. Specifically, the model suggests that PcG proteins are inhibited from binding at active promoters by an abundance of transcription factors, while at promoter sites with a low occupancy of transcription factors, PcG proteins can easily be recruited at CpG islands to establish the H3K27me3 modification. To test this model, we used publicly available genome-wide TF and epigenetic modifier binding profiles (ChIP-seq data) in murine and human ES cells[44] and calculated the number of transcription factors bound (TF density) at the four classes of promoters. Indeed the TF density decreases from active to bivalent to H3K27me3-only promoters in both human and mouse ES cells (Kruskal-Wallis test, P-value < 0.0001) (Fig. 6A).

To identify factors preferentially binding to bivalent promoters, we calculated the overlap between transcription and epigenetic factor binding sites (peaks) and bivalent promoters. Four out of 49 and eleven out of 99 factors characterised by ChIP-seq preferred bivalent promoters in human and mouse respectively (Fig. 6B). As expected, members of the PcG family were enriched at both human and mouse bivalent promoters (P value $< 10^{-256}$). Moreover, the co-repressor c-terminal binding protein 2 (CTBP2), required for PcG recruitment in Drosophila[45], and the RBBP5 (MLL subunit) were enriched at human bivalent promoters (P value < 0.005). The components of both PRC2 (Ezh2, Suz12) and PRC1 (Cbx7, Ring1b) together with two polycomb-like proteins (Mtf2, Phf9) were enriched at mouse bivalent promoters. Mtf2 and Phf19 recruit the PRC2 complex and are thought to silence transcriptionally active loci (H3K36me3) by recruiting H3K36me3 histone demethylases such as Kdm2b to further recruit PRC2 components for H3K27me3[46–48]. Accordingly, Kdm2b was also enriched at mouse bivalent promoters (P value $< 10^{-3}$). Four other epigenetic regulators, Utf1, Tet1, Rest and Setdb1 were highly enriched at mouse bivalent regions. Utf1 (P value $< 10^{-256}$) was recently identified as a component of bivalent chromatin by acting as a buffer against full activation of bivalent genes[13].

As expected, many TFs (33 out of 49 factors in human and 39 out of 99 factors in mouse) were enriched at active (H3K4me3-only) promoters. This included known regulators of pluripotency in ES cells such as Klf4, Esrrb, Oct4, Sox2, and Nanog (Table S8). Only two factors enriched in bivalent promoters, Kdm2b and Tet1, were also enriched at active promoters. All other factors showed preference to either bivalent promoters or active but not both. For example, C-Myc can stimulate Pol II elongation[48] and was enriched in active promoters in both human and mouse ES cells but not in bivalent promoters.

The observation that some factors are enriched specifically at bivalent promoters suggests that sequence motifs specific to bivalent promoters may determine their binding. We performed *de novo* motif identification on bivalent promoters by providing active promoter sequences as background in HOMER software[31] and found several AG-rich and GC-rich motifs specific to bivalent promoters (Figure S10). These resemble the sequence motifs of Jarid2[49] and Utf1[13] identified from ChIP-seq data. Interestingly, a 'TCCCC' sequence motif was enriched and found in about 50% of bivalent promoters in both human and mouse (Fig. 6C). This motif was not enriched in active promoters in either of the species (the number of repressed promoters was not large enough to perform a reliable *de novo* motif discovery). The 'TCCCC' motif was most similar to the known binding sequence of the Mzf1 transcription factor[50]. The Mzf1 promoter both in mouse and human ES cells is characterized as HC H3K4me3 only and belonged to the low expressed genes in our analysis. However, in recent Mzf1 ChIP-seq experiment performed in HEK293 cell line[51], the "TCCCC" motif was not enriched in Mzf1 peak list (Table S9). When *de novo* motif enrichment was performed on active human and mouse promoters using bivalent promoter sequences as background, they were enriched for a 'CGGAA' motif found in 40% of the active promoter sequences, which was not enriched in bivalent promoters. This motif is the most similar to the known motif for Elk1 transcription factor (Fig. 6C).

In summary, bivalent promoters are bound by fewer transcription factors than active (H3K4me3-only) promoters, but more than H3K27me3 only and latent promoters. Active promoters were preferentially occupied by pluripotency factors. On the other hand, bivalent promoters were enriched for Polycomb factors as well as other chromatin modifiers. The factors enriched at bivalent promoters show very little

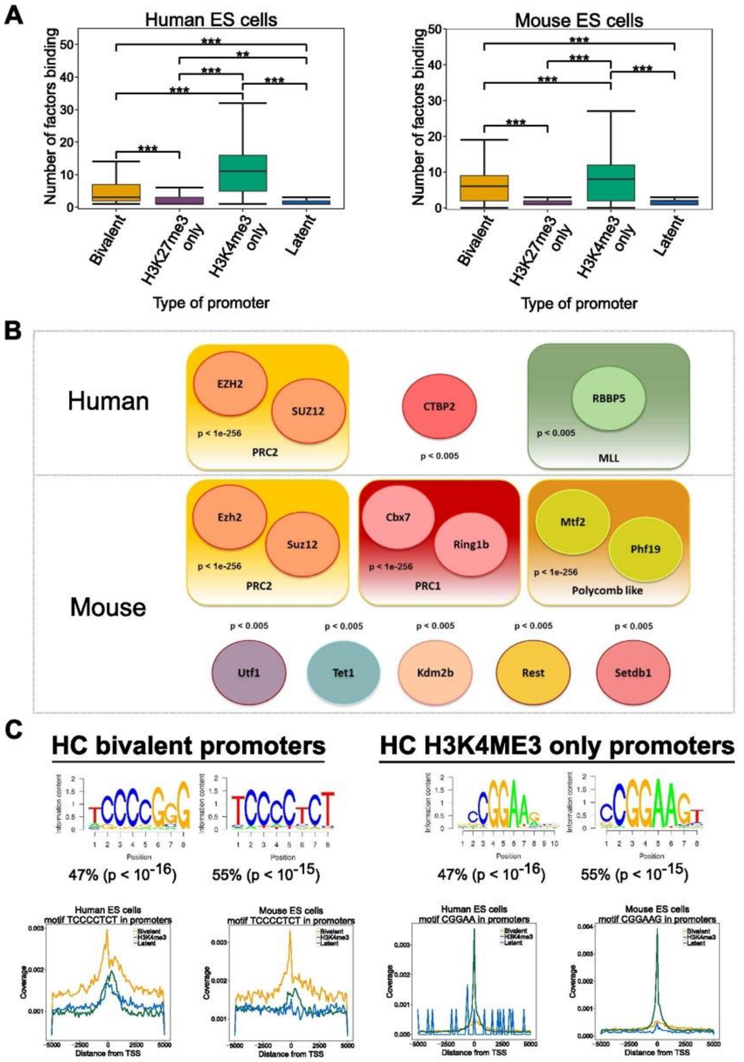


Figure 6. (**A**) 'TCCCC' sequence motif is specifically enriched in bivalent promoters. (**A**) The average occupancy of factors at HC H3K4me3-only promoters (green) is higher than at HC bivalent promoters (yellow) which is higher than at HC H3K27me3 only promoters (purple) and latent promoters (blue) in human (left) and mouse (right) ES cells (***equals to $<10^{-4}$). (**B**) Transcription and epigenetic factors with statistically significant overlap with HC bivalent promoters from ChIP sequencing data for 49 in human (up) and 99 factors in mouse (down) ES cells. (**C**) A 'TCCCC' sequence motif is specifically enriched in HC bivalent promoters in both human and mouse ES cells. Similarly a 'CGGAA' motif is enriched HC H3K4me3 promoters in both human and mouse ES cells. Each motif was then mapped to the genome, and motif densities around TSSs of bivalent (black), H3K4me3-only (yellow) and latent (blue) promoters are shown in the left (human) and right (mouse) panels.

overlap with the ones enriched at active promoters. These findings are consistent with the observed spatial segregation of transcriptional networks in ES cells where Nanog and Polycomb proteins were shown to occupy distinct nuclear spaces[52]. Finally, we identified a 'TCCCC' sequence motif specifically at bivalent promoters and a 'CGGAA' sequence motif at active promoters.

Discussion

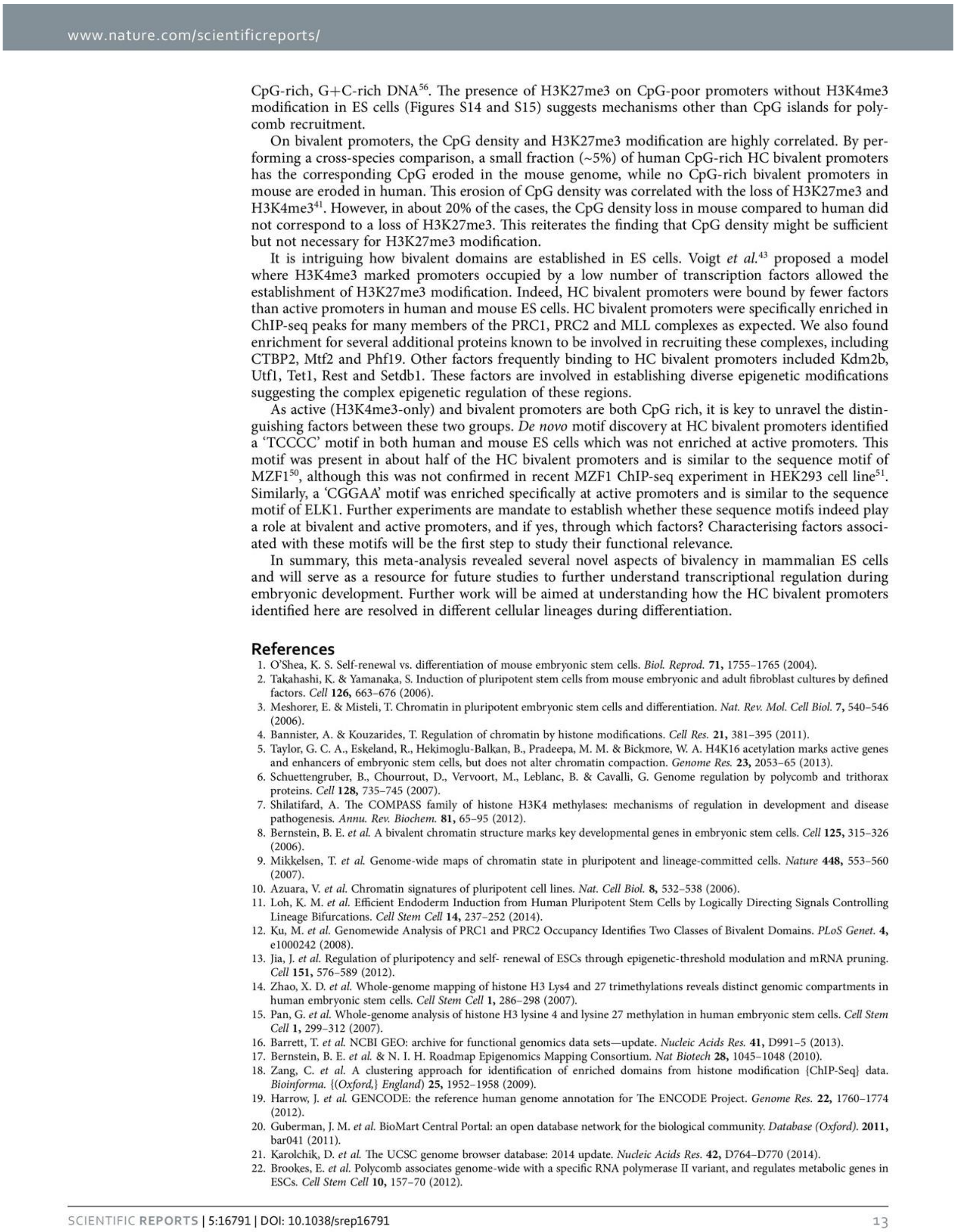
Bivalent chromatin domains bearing both H3K4me3 and H3K27me3 modifications have been shown to be a key feature of developmentally regulated genes in ES cells[8,9,13–15]. These domains are thought to be 'poised', with an ability to quickly become active (losing H3K27me3) or inactive (losing H3K4me3) during differentiation[9,53]. While many studies have produced ChIP-seq data for both H3K4me3 and H3K27me3 in ES cells in both humans[14,15] and mice[9,13], differences in species, ES growth conditions, ChIP protocols (shearing, cross link, antibodies used) and high throughput sequencing setup (with or without replicate, with or without input) have rendered a comparison across studies challenging. By systematic integration of available data, we identified robust lists of 4,979 and 3,659 high confidence bivalent promoters in human and mouse respectively. Since our work is using the data of previous studies using H3K4me3 and H3K27me3 ChIP-seq to define bivalency in ES cells, we are biased toward a confirmation of the original studies, as their data is integrated in our dataset. However our integrative approach (see methods) renders this analysis resistant to any outlier experiments. By cumulatively integrating the samples, it became evident that the detection of bivalency on promoters is dependent on the reliable detection of the H3K27me3 modification. Over 85% of H3K27me3 promoters were bivalent, i.e. they also had the H3K4me3 mark. This confirms that bivalency in ES cells is rather the rule than the exception. The three main chromatin states on promoters in ES cells are thus active, bivalent and latent (no mark). Correspondingly, active promoters were expressed, bivalent were lowly expressed and latent were mostly not expressed.

Bivalent promoters are thought to be poised for rapid activation or inactivation during differentiation[13,43]. To tease out whether the low expression at bivalent promoters is a result of some cells expressing the genes while others not, or the genes are expressed at low levels in most cells, we used single cell gene expression data. Bivalent genes were expressed in a similar number of single cells as lowly expressed active genes. It is therefore unlikely that bivalency is a result of mixture of cell populations in ES cells. Similarly, H3K27me3 read density was higher at HC bivalent promoters than at H3K27me3-only promoters, again arguing in disfavour of a mix-population model. The low transcription level can be interpreted as a "leaking" transcription rate, in the absence of a strong repressive chromatin environment. During development, these poised domains have been shown to resolve as either active (by losing the H3K27me3 mark) or inactive (by losing the H3K4me3 mark), and in some cases gaining DNA methylation[37], depending on the cellular lineage. In agreement with this model, we have found that >90% of differentially expressed (either up-regulated or down-regulated) gene sets when any one of a set of 91 transcription factors was either overexpressed or knocked down in mouse ES cells were enriched for bivalent genes. This finding suggests that bivalent genes are hypersensitive to most perturbations of the regulatory network in ES cells.

We computed binding profiles of PRC components (PRC1 and PRC2) and various forms of RNA polymerase II at bivalent promoters in murine ES cells. All HC bivalent promoters were marked by Suz12, Jarid2, Ring1b and Cbx7. To note, the PRC2-only group defined by[12] overlapped with PRC1-low clusters, the PRC1 signal detected due to higher sequencing depth in latter case (Figure S11). Thus all bivalent promoters were occupied by both PRC1 and PRC2. Accordingly, H2Aub showed enrichment at HC bivalent promoters (Figure S12). Recent studies have suggested that true bivalency is better associated with H2Aub than H3K27me3[22]. We note that H2Aub predominantly but not exclusively marks bivalent promoters (Table S10) as it also marks a fraction of H3K4me3-only expressed gene promoters (Figure S13). Based on PRC1 and RNA PolII occupancy, bivalent promoters grouped into four clusters. Clusters 1 and 2 had low PRC1 occupancy and high RNA PolII (8WG16) levels while clusters 3 and 4 were PRC-rich with low RNA PolII (8WG16) levels. Cluster 2 was enriched for metabolic genes and marked with RNA PolII (S7P) and cluster 2 genes were expressed at higher levels than the other three clusters. The bivalent promoters therefore consist of sub-groups of genes which at functional, epigenetic and transcriptional level are quite different from each other.

More than half of high-confidence bivalent promoters were conserved between human and mouse, suggesting the existence of a set of genes bivalently marked across most mammalian ES cells (Table S9 and S10). These genes were very highly enriched for transcription regulators and developmental factors, compared to the species specific bivalent promoters. On the other hand, divergence of epigenetic status across species did not imply divergence of gene expression i.e. promoters with bivalent chromatin status in human and active chromatin status in mouse did not have gene expression profiles similar to bivalent genes in human and active genes in mouse. Further analysis is necessary to understand whether the differences between mouse and human ES cells are indeed species-specific or developmental stage specific as human ES cells do not share the same developmental state as mouse ES cells[54,55].

Since a high density of un-methylated CpG is sufficient for vertebrate polycomb recruitment[38,39,42], it is assumed that the presence of CpG islands determines H3K27me3 modification. Over 90% of bivalent promoters contained a CpG island while few to none of the H3K27me3-only promoters had a CpG island. Wachter *et al.* (2014) recently suggested that bivalency is the default chromatin structure for



CpG-rich, G+C-rich DNA[56]. The presence of H3K27me3 on CpG-poor promoters without H3K4me3 modification in ES cells (Figures S14 and S15) suggests mechanisms other than CpG islands for polycomb recruitment.

On bivalent promoters, the CpG density and H3K27me3 modification are highly correlated. By performing a cross-species comparison, a small fraction (~5%) of human CpG-rich HC bivalent promoters has the corresponding CpG eroded in the mouse genome, while no CpG-rich bivalent promoters in mouse are eroded in human. This erosion of CpG density was correlated with the loss of H3K27me3 and H3K4me3[41]. However, in about 20% of the cases, the CpG density loss in mouse compared to human did not correspond to a loss of H3K27me3. This reiterates the finding that CpG density might be sufficient but not necessary for H3K27me3 modification.

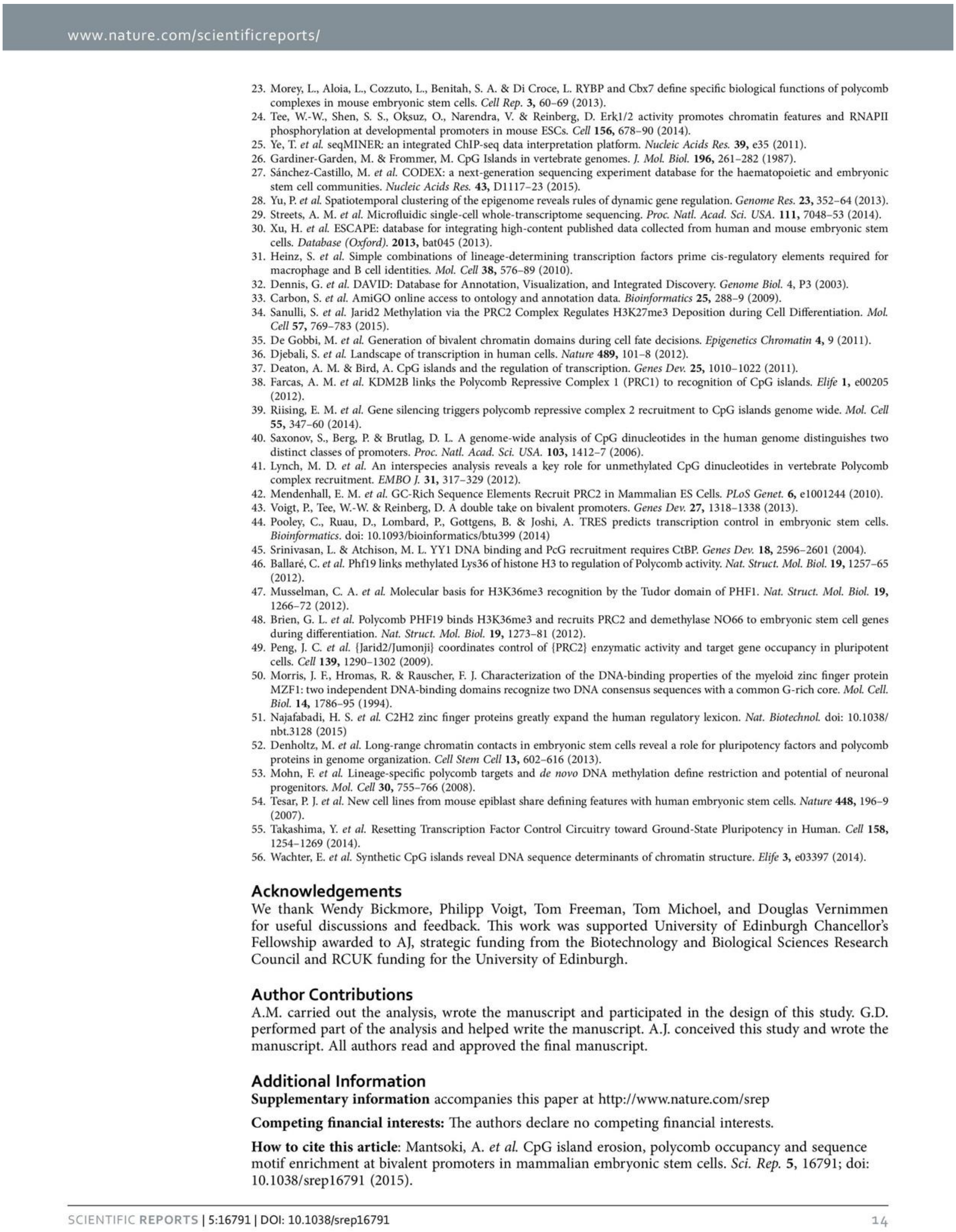
It is intriguing how bivalent domains are established in ES cells. Voigt *et al.*[43] proposed a model where H3K4me3 marked promoters occupied by a low number of transcription factors allowed the establishment of H3K27me3 modification. Indeed, HC bivalent promoters were bound by fewer factors than active promoters in human and mouse ES cells. HC bivalent promoters were specifically enriched in ChIP-seq peaks for many members of the PRC1, PRC2 and MLL complexes as expected. We also found enrichment for several additional proteins known to be involved in recruiting these complexes, including CTBP2, Mtf2 and Phf19. Other factors frequently binding to HC bivalent promoters included Kdm2b, Utf1, Tet1, Rest and Setdb1. These factors are involved in establishing diverse epigenetic modifications suggesting the complex epigenetic regulation of these regions.

As active (H3K4me3-only) and bivalent promoters are both CpG rich, it is key to unravel the distinguishing factors between these two groups. *De novo* motif discovery at HC bivalent promoters identified a 'TCCCC' motif in both human and mouse ES cells which was not enriched at active promoters. This motif was present in about half of the HC bivalent promoters and is similar to the sequence motif of MZF1[50], although this was not confirmed in recent MZF1 ChIP-seq experiment in HEK293 cell line[51]. Similarly, a 'CGGAA' motif was enriched specifically at active promoters and is similar to the sequence motif of ELK1. Further experiments are mandate to establish whether these sequence motifs indeed play a role at bivalent and active promoters, and if yes, through which factors? Characterising factors associated with these motifs will be the first step to study their functional relevance.

In summary, this meta-analysis revealed several novel aspects of bivalency in mammalian ES cells and will serve as a resource for future studies to further understand transcriptional regulation during embryonic development. Further work will be aimed at understanding how the HC bivalent promoters identified here are resolved in different cellular lineages during differentiation.

References

1. O'Shea, K. S. Self-renewal vs. differentiation of mouse embryonic stem cells. *Biol. Reprod.* **71,** 1755–1765 (2004).
2. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
3. Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat. Rev. Mol. Cell Biol.* **7,** 540–546 (2006).
4. Bannister, A. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21,** 381–395 (2011).
5. Taylor, G. C. A., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M. M. & Bickmore, W. A. H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Res.* **23,** 2053–65 (2013).
6. Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. & Cavalli, G. Genome regulation by polycomb and trithorax proteins. *Cell* **128,** 735–745 (2007).
7. Shilatifard, A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu. Rev. Biochem.* **81,** 65–95 (2012).
8. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
9. Mikkelsen, T. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
10. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8,** 532–538 (2006).
11. Loh, K. M. *et al.* Efficient Endoderm Induction from Human Pluripotent Stem Cells by Logically Directing Signals Controlling Lineage Bifurcations. *Cell Stem Cell* **14,** 237–252 (2014).
12. Ku, M. *et al.* Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains. *PLoS Genet.* **4,** e1000242 (2008).
13. Jia, J. *et al.* Regulation of pluripotency and self- renewal of ESCs through epigenetic-threshold modulation and mRNA pruning. *Cell* **151,** 576–589 (2012).
14. 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 (2007).
15. Pan, G. *et al.* Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* **1,** 299–312 (2007).
16. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41,** D991–5 (2013).
17. Bernstein, B. E. *et al.* & N. I. H. Roadmap Epigenomics Mapping Consortium. *Nat Biotech* **28,** 1045–1048 (2010).
18. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification {ChIP-Seq} data. *Bioinforma. {(Oxford,} England)* **25,** 1952–1958 (2009).
19. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
20. Guberman, J. M. *et al.* BioMart Central Portal: an open database network for the biological community. *Database (Oxford).* **2011,** bar041 (2011).
21. Karolchik, D. *et al.* The UCSC genome browser database: 2014 update. *Nucleic Acids Res.* **42,** D764–D770 (2014).
22. Brookes, E. *et al.* Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. *Cell Stem Cell* **10,** 157–70 (2012).



23. Morey, L., Aloia, L., Cozzuto, L., Benitah, S. A. & Di Croce, L. RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells. *Cell Rep.* **3,** 60–69 (2013).
24. Tee, W.-W., Shen, S. S., Oksuz, O., Narendra, V. & Reinberg, D. Erk1/2 activity promotes chromatin features and RNAPII phosphorylation at developmental promoters in mouse ESCs. *Cell* **156,** 678–90 (2014).
25. Ye, T. *et al.* seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.* **39,** e35 (2011).
26. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196,** 261–282 (1987).
27. Sánchez-Castillo, M. *et al.* CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.* **43,** D1117–23 (2015).
28. Yu, P. *et al.* Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation. *Genome Res.* **23,** 352–64 (2013).
29. Streets, A. M. *et al.* Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. USA.* **111,** 7048–53 (2014).
30. Xu, H. *et al.* ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database (Oxford).* **2013,** bat045 (2013).
31. 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,** 576–89 (2010).
32. Dennis, G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4, P3 (2003).
33. Carbon, S. *et al.* AmiGO online access to ontology and annotation data. *Bioinformatics* **25,** 288–9 (2009).
34. Sanulli, S. *et al.* Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation. *Mol. Cell* **57,** 769–783 (2015).
35. De Gobbi, M. *et al.* Generation of bivalent chromatin domains during cell fate decisions. *Epigenetics Chromatin* **4,** 9 (2011).
36. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8 (2012).
37. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
38. Farcas, A. M. *et al.* KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. *Elife* **1,** e00205 (2012).
39. Riising, E. M. *et al.* Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide. *Mol. Cell* **55,** 347–60 (2014).
40. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. USA.* **103,** 1412–7 (2006).
41. Lynch, M. D. *et al.* An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *EMBO J.* **31,** 317–329 (2012).
42. Mendenhall, E. M. *et al.* GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. *PLoS Genet.* **6,** e1001244 (2010).
43. Voigt, P., Tee, W.-W. & Reinberg, D. A double take on bivalent promoters. *Genes Dev.* **27,** 1318–1338 (2013).
44. Pooley, C., Ruau, D., Lombard, P., Gottgens, B. & Joshi, A. TRES predicts transcription control in embryonic stem cells. *Bioinformatics*. doi: 10.1093/bioinformatics/btu399 (2014)
45. Srinivasan, L. & Atchison, M. L. YY1 DNA binding and PcG recruitment requires CtBP. *Genes Dev.* **18,** 2596–2601 (2004).
46. Ballaré, C. *et al.* Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. *Nat. Struct. Mol. Biol.* **19,** 1257–65 (2012).
47. Musselman, C. A. *et al.* Molecular basis for H3K36me3 recognition by the Tudor domain of PHF1. *Nat. Struct. Mol. Biol.* **19,** 1266–72 (2012).
48. Brien, G. L. *et al.* Polycomb PHF19 binds H3K36me3 and recruits PRC2 and demethylase NO66 to embryonic stem cell genes during differentiation. *Nat. Struct. Mol. Biol.* **19,** 1273–81 (2012).
49. Peng, J. C. *et al.* {Jarid2/Jumonji} coordinates control of {PRC2} enzymatic activity and target gene occupancy in pluripotent cells. *Cell* **139,** 1290–1302 (2009).
50. Morris, J. F., Hromas, R. & Rauscher, F. J. Characterization of the DNA-binding properties of the myeloid zinc finger protein MZF1: two independent DNA-binding domains recognize two DNA consensus sequences with a common G-rich core. *Mol. Cell. Biol.* **14,** 1786–95 (1994).
51. Najafabadi, H. S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat. Biotechnol.* doi: 10.1038/nbt.3128 (2015)
52. Denholtz, M. *et al.* Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell* **13,** 602–616 (2013).
53. Mohn, F. *et al.* Lineage-specific polycomb targets and *de novo* DNA methylation define restriction and potential of neuronal progenitors. *Mol. Cell* **30,** 755–766 (2008).
54. Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448,** 196–9 (2007).
55. Takashima, Y. *et al.* Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. *Cell* **158,** 1254–1269 (2014).
56. Wachter, E. *et al.* Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *Elife* **3,** e03397 (2014).

Acknowledgements

We thank Wendy Bickmore, Philipp Voigt, Tom Freeman, Tom Michoel, and Douglas Vernimmen for useful discussions and feedback. This work was supported University of Edinburgh Chancellor's Fellowship awarded to AJ, strategic funding from the Biotechnology and Biological Sciences Research Council and RCUK funding for the University of Edinburgh.

Author Contributions

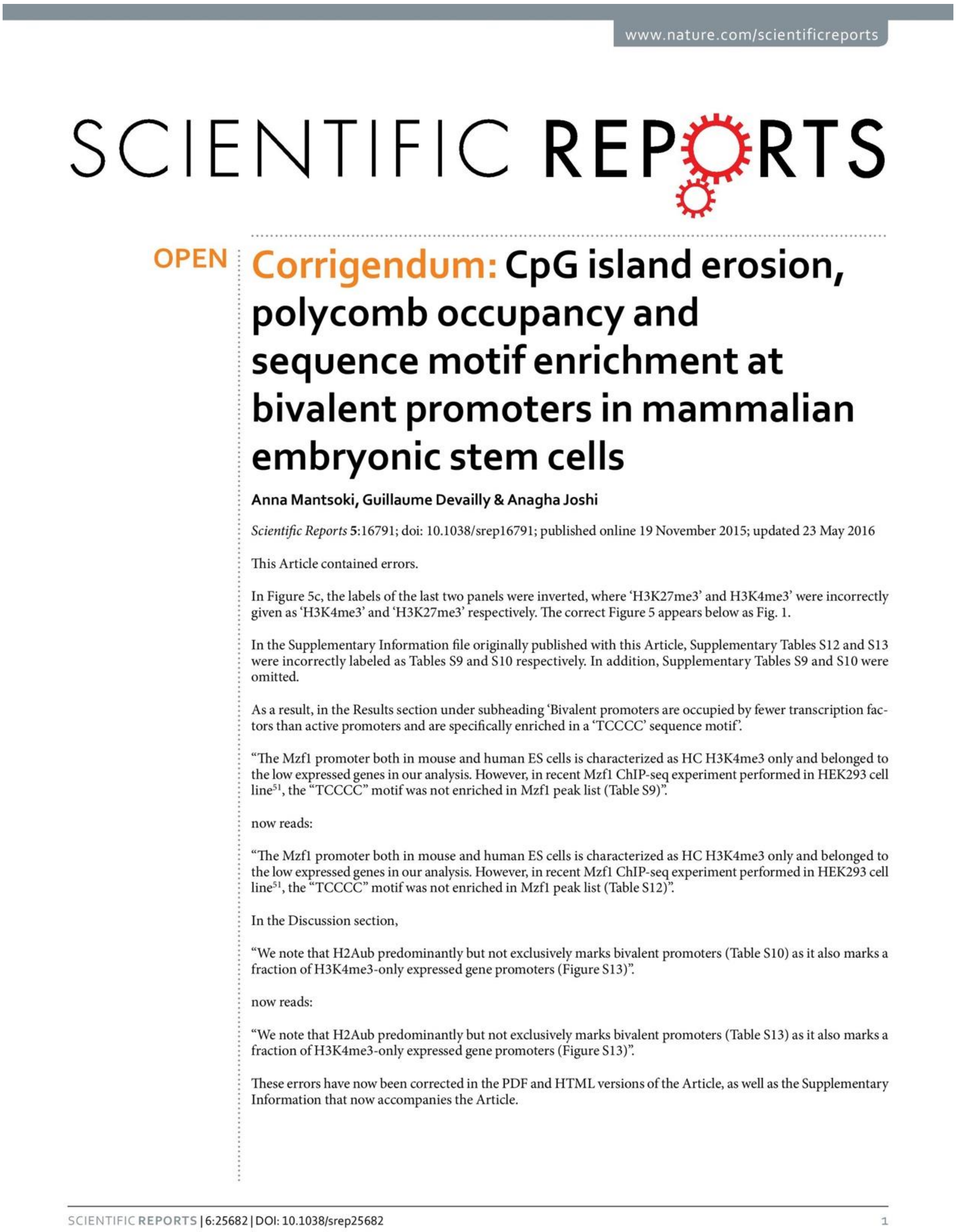
A.M. carried out the analysis, wrote the manuscript and participated in the design of this study. G.D. performed part of the analysis and helped write the manuscript. A.J. conceived this study and wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Mantsoki, A. *et al.* CpG island erosion, polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells. *Sci. Rep.* **5**, 16791; doi: 10.1038/srep16791 (2015).



www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Corrigendum: CpG island erosion, polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells

Anna Mantsoki, Guillaume Devailly & Anagha Joshi

Scientific Reports **5**:16791; doi: 10.1038/srep16791; published online 19 November 2015; updated 23 May 2016

This Article contained errors.

In Figure 5c, the labels of the last two panels were inverted, where 'H3K27me3' and H3K4me3' were incorrectly given as 'H3K4me3' and 'H3K27me3' respectively. The correct Figure 5 appears below as Fig. 1.

In the Supplementary Information file originally published with this Article, Supplementary Tables S12 and S13 were incorrectly labeled as Tables S9 and S10 respectively. In addition, Supplementary Tables S9 and S10 were omitted.

As a result, in the Results section under subheading 'Bivalent promoters are occupied by fewer transcription factors than active promoters and are specifically enriched in a 'TCCCC' sequence motif'.

"The Mzf1 promoter both in mouse and human ES cells is characterized as HC H3K4me3 only and belonged to the low expressed genes in our analysis. However, in recent Mzf1 ChIP-seq experiment performed in HEK293 cell line[51], the "TCCCC" motif was not enriched in Mzf1 peak list (Table S9)".

now reads:

"The Mzf1 promoter both in mouse and human ES cells is characterized as HC H3K4me3 only and belonged to the low expressed genes in our analysis. However, in recent Mzf1 ChIP-seq experiment performed in HEK293 cell line[51], the "TCCCC" motif was not enriched in Mzf1 peak list (Table S12)".

In the Discussion section,

"We note that H2Aub predominantly but not exclusively marks bivalent promoters (Table S10) as it also marks a fraction of H3K4me3-only expressed gene promoters (Figure S13)".

now reads:

"We note that H2Aub predominantly but not exclusively marks bivalent promoters (Table S13) as it also marks a fraction of H3K4me3-only expressed gene promoters (Figure S13)".

These errors have now been corrected in the PDF and HTML versions of the Article, as well as the Supplementary Information that now accompanies the Article.

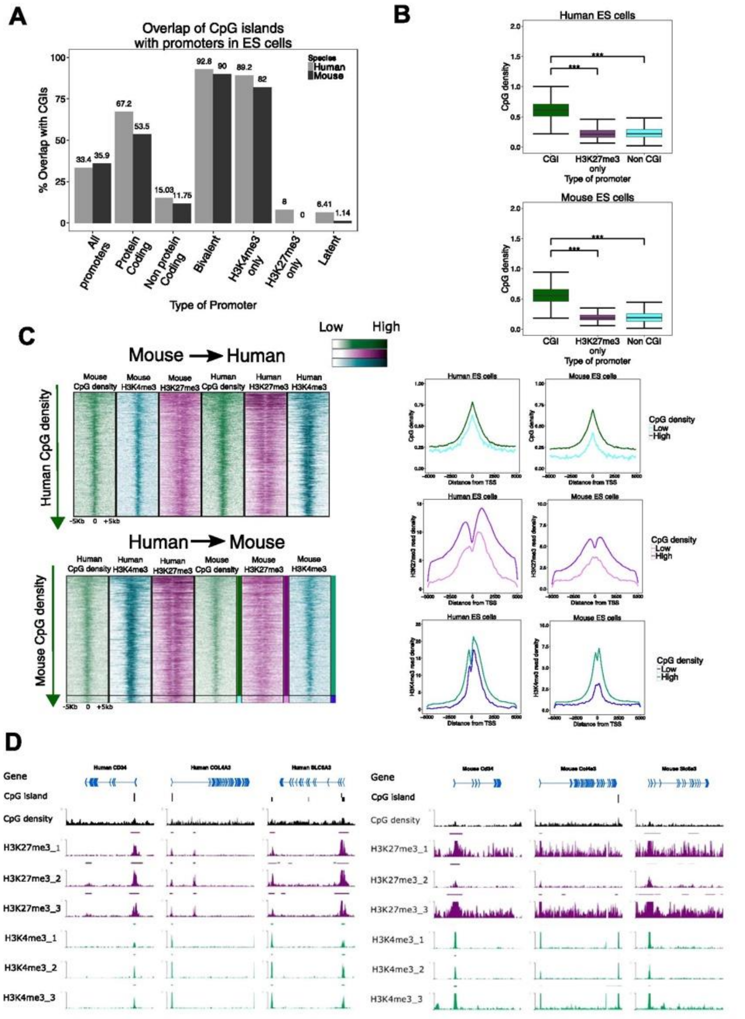
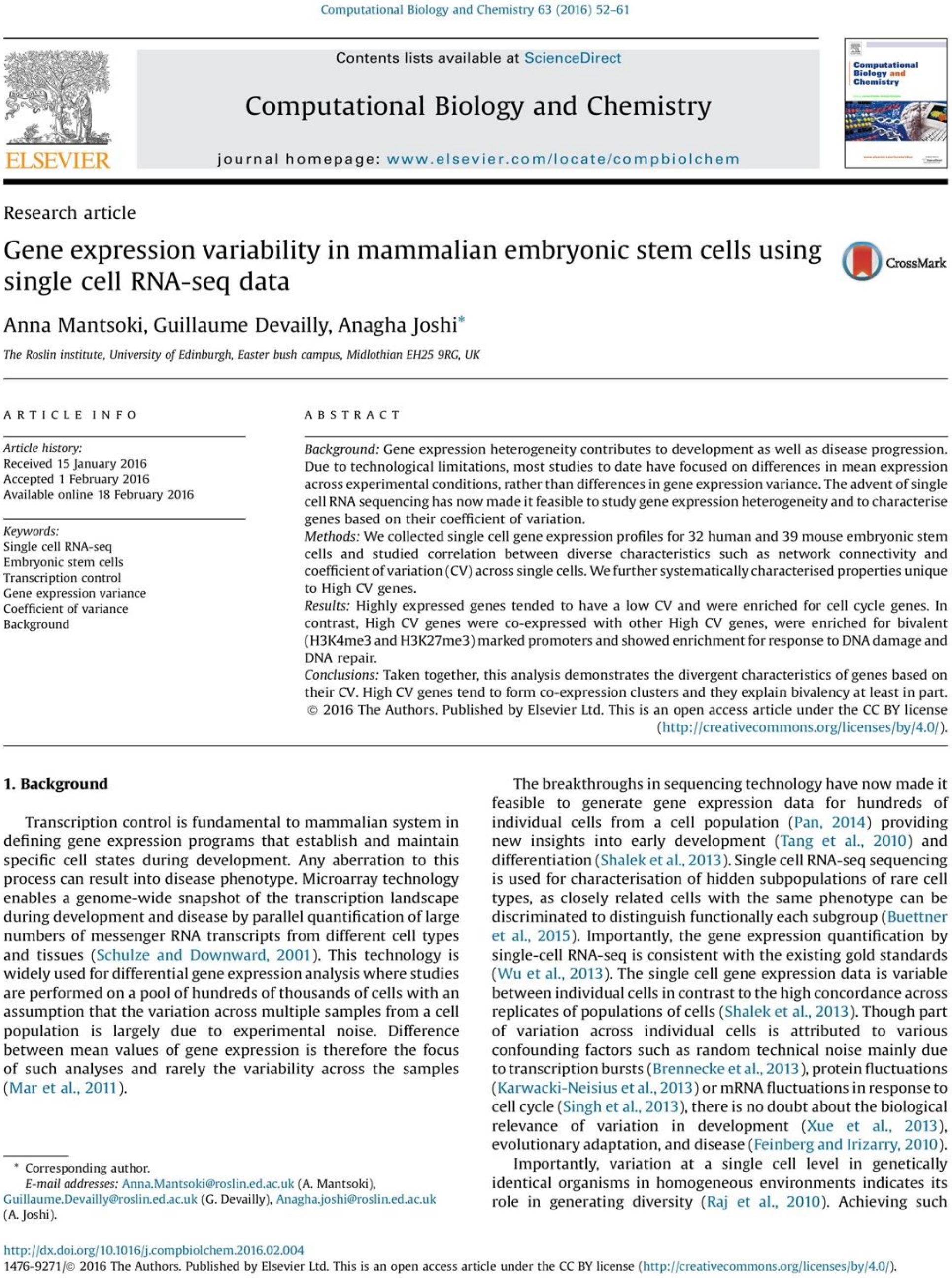

A
Overlap of CpG islands with promoters in ES cells
% Overlap with CGIs
100
75
50
25
0
Species
Human
Mouse
33.4
35.9
67.2
53.5
15.03
11.75
92.8
90
89.2
82
8
0
6.41
1.14
All promoters
Protein Coding
Non protein Coding
Bivalent
H3K4me3 only
H3K27me3 only
Latent
Type of Promoter
B
Human ES cells
Mouse ES cells
CpG density
2.0
1.5
1.0
0.5
0.0
***
***
CGI
H3K27me3 only
Non CGI
Type of promoter
C
Low
High
Mouse → Human
Mouse CpG density
Mouse H3K4me3
Mouse H3K27me3
Human CpG density
Human H3K27me3
Human H3K4me3
Human CpG density
-5Kb 0 +5kb
Human → Mouse
Human CpG density
Human H3K4me3
Human H3K27me3
Mouse CpG density
Mouse H3K27me3
Mouse H3K4me3
Mouse CpG density
-5Kb 0 +5kb
Human ES cells
Mouse ES cells
CpG density
H3K27me3 read density
H3K4me3 read density
Distance from TSS
CpG density
Low
High
D
Human CD34
Human COL4A3
Human SLC6A3
Mouse Cd34
Mouse Col4a3
Mouse Slc6a3
Gene
CpG Island
CpG density
H3K27me3_1
H3K27me3_2
H3K27me3_3
H3K4me3_1
H3K4me3_2
H3K4me3_3

Figure 1.



Computational Biology and Chemistry 63 (2016) 52–61

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research article

Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data

Anna Mantsoki, Guillaume Devailly, Anagha Joshi*

The Roslin institute, University of Edinburgh, Easter bush campus, Midlothian EH25 9RG, UK

ARTICLE INFO

Article history:
Received 15 January 2016
Accepted 1 February 2016
Available online 18 February 2016

Keywords:
Single cell RNA-seq
Embryonic stem cells
Transcription control
Gene expression variance
Coefficient of variance
Background

ABSTRACT

Background: Gene expression heterogeneity contributes to development as well as disease progression. Due to technological limitations, most studies to date have focused on differences in mean expression across experimental conditions, rather than differences in gene expression variance. The advent of single cell RNA sequencing has now made it feasible to study gene expression heterogeneity and to characterise genes based on their coefficient of variation.
Methods: We collected single cell gene expression profiles for 32 human and 39 mouse embryonic stem cells and studied correlation between diverse characteristics such as network connectivity and coefficient of variation (CV) across single cells. We further systematically characterised properties unique to High CV genes.
Results: Highly expressed genes tended to have a low CV and were enriched for cell cycle genes. In contrast, High CV genes were co-expressed with other High CV genes, were enriched for bivalent (H3K4me3 and H3K27me3) marked promoters and showed enrichment for response to DNA damage and DNA repair.
Conclusions: Taken together, this analysis demonstrates the divergent characteristics of genes based on their CV. High CV genes tend to form co-expression clusters and they explain bivalency at least in part.

1. Background

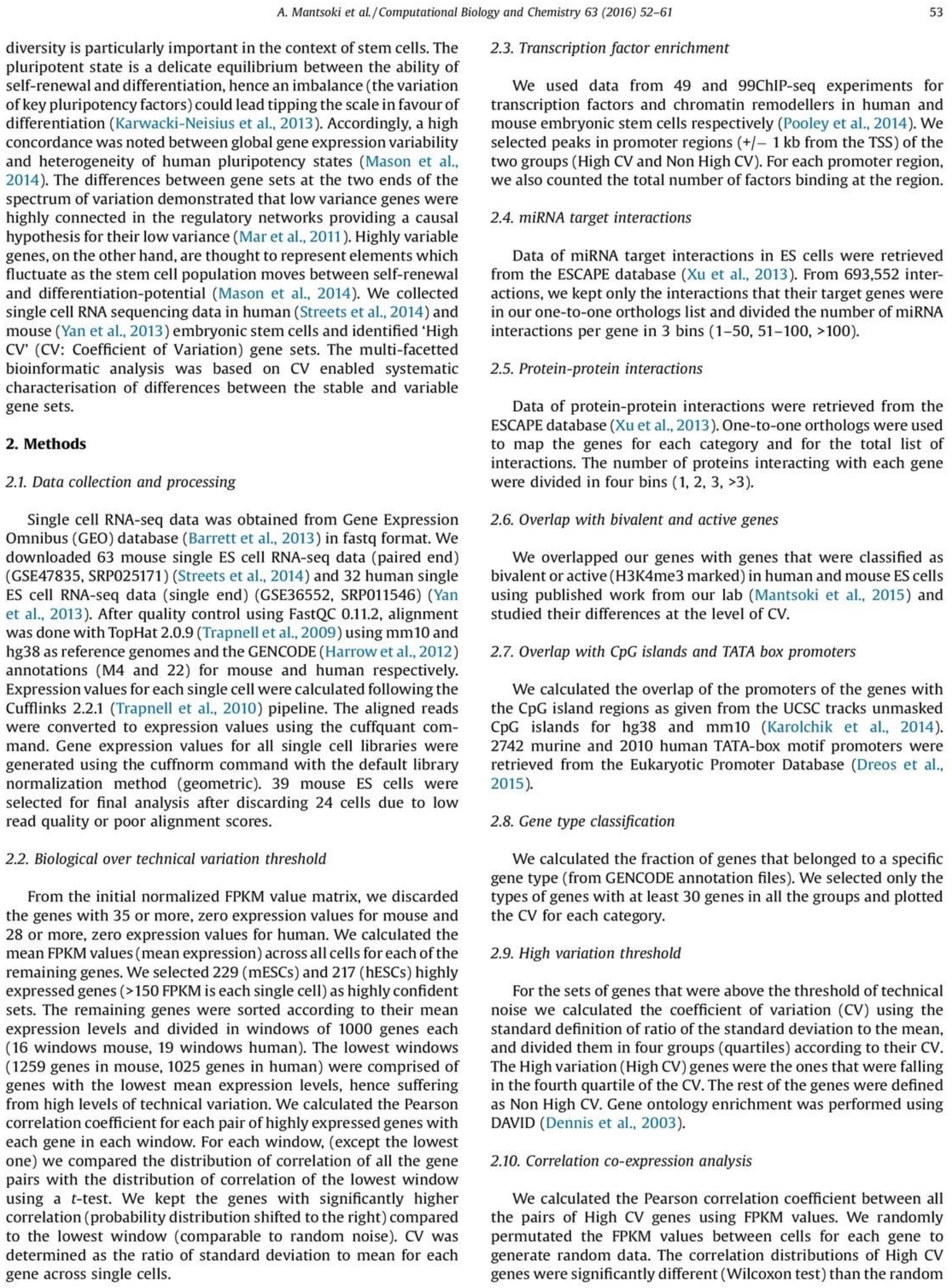
Transcription control is fundamental to mammalian system in defining gene expression programs that establish and maintain specific cell states during development. Any aberration to this process can result into disease phenotype. Microarray technology enables a genome-wide snapshot of the transcription landscape during development and disease by parallel quantification of large numbers of messenger RNA transcripts from different cell types and tissues (Schulze and Downward, 2001). This technology is widely used for differential gene expression analysis where studies are performed on a pool of hundreds of thousands of cells with an assumption that the variation across multiple samples from a cell population is largely due to experimental noise. Difference between mean values of gene expression is therefore the focus of such analyses and rarely the variability across the samples (Mar et al., 2011).

The breakthroughs in sequencing technology have now made it feasible to generate gene expression data for hundreds of individual cells from a cell population (Pan, 2014) providing new insights into early development (Tang et al., 2010) and differentiation (Shalek et al., 2013). Single cell RNA-seq sequencing is used for characterisation of hidden subpopulations of rare cell types, as closely related cells with the same phenotype can be discriminated to distinguish functionally each subgroup (Buettner et al., 2015). Importantly, the gene expression quantification by single-cell RNA-seq is consistent with the existing gold standards (Wu et al., 2013). The single cell gene expression data is variable between individual cells in contrast to the high concordance across replicates of populations of cells (Shalek et al., 2013). Though part of variation across individual cells is attributed to various confounding factors such as random technical noise mainly due to transcription bursts (Brennecke et al., 2013), protein fluctuations (Karwacki-Neisius et al., 2013) or mRNA fluctuations in response to cell cycle (Singh et al., 2013), there is no doubt about the biological relevance of variation in development (Xue et al., 2013), evolutionary adaptation, and disease (Feinberg and Irizarry, 2010).

Importantly, variation at a single cell level in genetically identical organisms in homogeneous environments indicates its role in generating diversity (Raj et al., 2010). Achieving such

* Corresponding author.
E-mail addresses: Anna.Mantsoki@roslin.ed.ac.uk (A. Mantsoki), Guillaume.Devailly@roslin.ed.ac.uk (G. Devailly), Anagha.joshi@roslin.ed.ac.uk (A. Joshi).

http://dx.doi.org/10.1016/j.compbiolchem.2016.02.004



diversity is particularly important in the context of stem cells. The pluripotent state is a delicate equilibrium between the ability of self-renewal and differentiation, hence an imbalance (the variation of key pluripotency factors) could lead tipping the scale in favour of differentiation (Karwacki-Neisius et al., 2013). Accordingly, a high concordance was noted between global gene expression variability and heterogeneity of human pluripotency states (Mason et al., 2014). The differences between gene sets at the two ends of the spectrum of variation demonstrated that low variance genes were highly connected in the regulatory networks providing a causal hypothesis for their low variance (Mar et al., 2011). Highly variable genes, on the other hand, are thought to represent elements which fluctuate as the stem cell population moves between self-renewal and differentiation-potential (Mason et al., 2014). We collected single cell RNA sequencing data in human (Streets et al., 2014) and mouse (Yan et al., 2013) embryonic stem cells and identified 'High CV' (CV: Coefficient of Variation) gene sets. The multi-facetted bioinformatic analysis was based on CV enabled systematic characterisation of differences between the stable and variable gene sets.

2. Methods

2.1. Data collection and processing

Single cell RNA-seq data was obtained from Gene Expression Omnibus (GEO) database (Barrett et al., 2013) in fastq format. We downloaded 63 mouse single ES cell RNA-seq data (paired end) (GSE47835, SRP025171) (Streets et al., 2014) and 32 human single ES cell RNA-seq data (single end) (GSE36552, SRP011546) (Yan et al., 2013). After quality control using FastQC 0.11.2, alignment was done with TopHat 2.0.9 (Trapnell et al., 2009) using mm10 and hg38 as reference genomes and the GENCODE (Harrow et al., 2012) annotations (M4 and 22) for mouse and human respectively. Expression values for each single cell were calculated following the Cufflinks 2.2.1 (Trapnell et al., 2010) pipeline. The aligned reads were converted to expression values using the cuffquant command. Gene expression values for all single cell libraries were generated using the cuffnorm command with the default library normalization method (geometric). 39 mouse ES cells were selected for final analysis after discarding 24 cells due to low read quality or poor alignment scores.

2.2. Biological over technical variation threshold

From the initial normalized FPKM value matrix, we discarded the genes with 35 or more, zero expression values for mouse and 28 or more, zero expression values for human. We calculated the mean FPKM values (mean expression) across all cells for each of the remaining genes. We selected 229 (mESCs) and 217 (hESCs) highly expressed genes (>150 FPKM is each single cell) as highly confident sets. The remaining genes were sorted according to their mean expression levels and divided in windows of 1000 genes each (16 windows mouse, 19 windows human). The lowest windows (1259 genes in mouse, 1025 genes in human) were comprised of genes with the lowest mean expression levels, hence suffering from high levels of technical variation. We calculated the Pearson correlation coefficient for each pair of highly expressed genes with each gene in each window. For each window, (except the lowest one) we compared the distribution of correlation of all the gene pairs with the distribution of correlation of the lowest window using a *t*-test. We kept the genes with significantly higher correlation (probability distribution shifted to the right) compared to the lowest window (comparable to random noise). CV was determined as the ratio of standard deviation to mean for each gene across single cells.

2.3. Transcription factor enrichment

We used data from 49 and 99ChIP-seq experiments for transcription factors and chromatin remodellers in human and mouse embryonic stem cells respectively (Pooley et al., 2014). We selected peaks in promoter regions (+/− 1 kb from the TSS) of the two groups (High CV and Non High CV). For each promoter region, we also counted the total number of factors binding at the region.

2.4. miRNA target interactions

Data of miRNA target interactions in ES cells were retrieved from the ESCAPE database (Xu et al., 2013). From 693,552 interactions, we kept only the interactions that their target genes were in our one-to-one orthologs list and divided the number of miRNA interactions per gene in 3 bins (1–50, 51–100, >100).

2.5. Protein-protein interactions

Data of protein-protein interactions were retrieved from the ESCAPE database (Xu et al., 2013). One-to-one orthologs were used to map the genes for each category and for the total list of interactions. The number of proteins interacting with each gene were divided in four bins (1, 2, 3, >3).

2.6. Overlap with bivalent and active genes

We overlapped our genes with genes that were classified as bivalent or active (H3K4me3 marked) in human and mouse ES cells using published work from our lab (Mantsoki et al., 2015) and studied their differences at the level of CV.

2.7. Overlap with CpG islands and TATA box promoters

We calculated the overlap of the promoters of the genes with the CpG island regions as given from the UCSC tracks unmasked CpG islands for hg38 and mm10 (Karolchik et al., 2014). 2742 murine and 2010 human TATA-box motif promoters were retrieved from the Eukaryotic Promoter Database (Dreos et al., 2015).

2.8. Gene type classification

We calculated the fraction of genes that belonged to a specific gene type (from GENCODE annotation files). We selected only the types of genes with at least 30 genes in all the groups and plotted the CV for each category.

2.9. High variation threshold

For the sets of genes that were above the threshold of technical noise we calculated the coefficient of variation (CV) using the standard definition of ratio of the standard deviation to the mean, and divided them in four groups (quartiles) according to their CV. The High variation (High CV) genes were the ones that were falling in the fourth quartile of the CV. The rest of the genes were defined as Non High CV. Gene ontology enrichment was performed using DAVID (Dennis et al., 2003).

2.10. Correlation co-expression analysis

We calculated the Pearson correlation coefficient between all the pairs of High CV genes using FPKM values. We randomly permutated the FPKM values between cells for each gene to generate random data. The correlation distributions of High CV genes were significantly different (Wilcoxon test) than the random

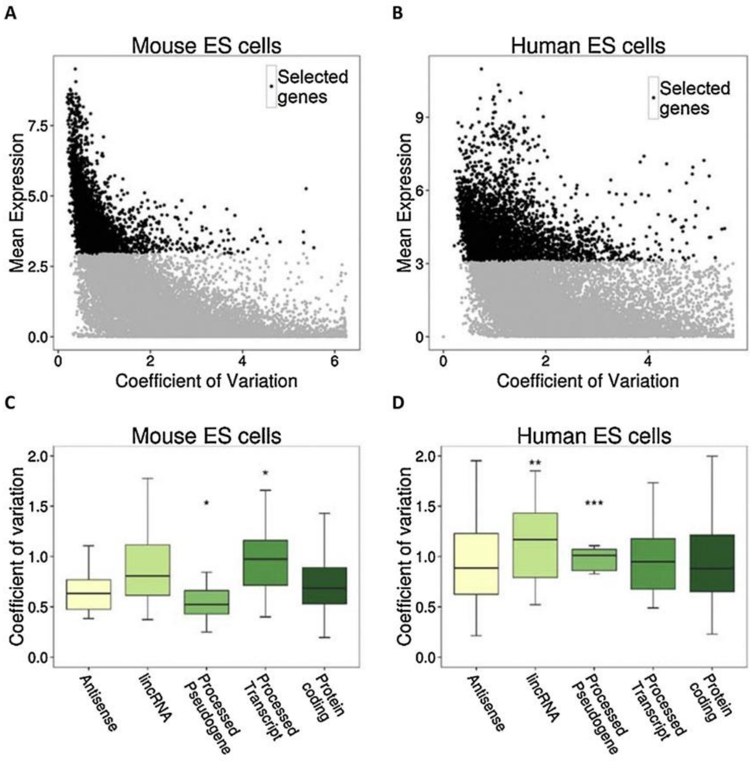


Fig. 1. Correlation based approach for the identification of genes above the threshold of technical variation (A, B) Scatterplots showing genes according to their mean expression (log (mean FPKM + 1)) and coefficient of variation in mouse and human ES cells. The genes highlighted in black were chosen for the analysis, since they were more correlated with the highly expressed genes. (C, D) Gene types in mouse and human ES cells and their respective CV levels (shown only the genes types that were found in 30 genes or more).

ones and we investigated their co-expression patterns by hierarchical clustering (flashClust package in R) visualised with heatmaps (heatmap.2 in R).

2.11. Conservation analysis

17,009 one-to-one orthologs from ensembl BioMart (Guberman et al., 2011) were used to calculate CV values in each species. After intersecting the orthologs with the 4000 genes (for both mouse and human) we end up with a gene set containing 2363 orthologous genes.

2.12. Topological associated domains

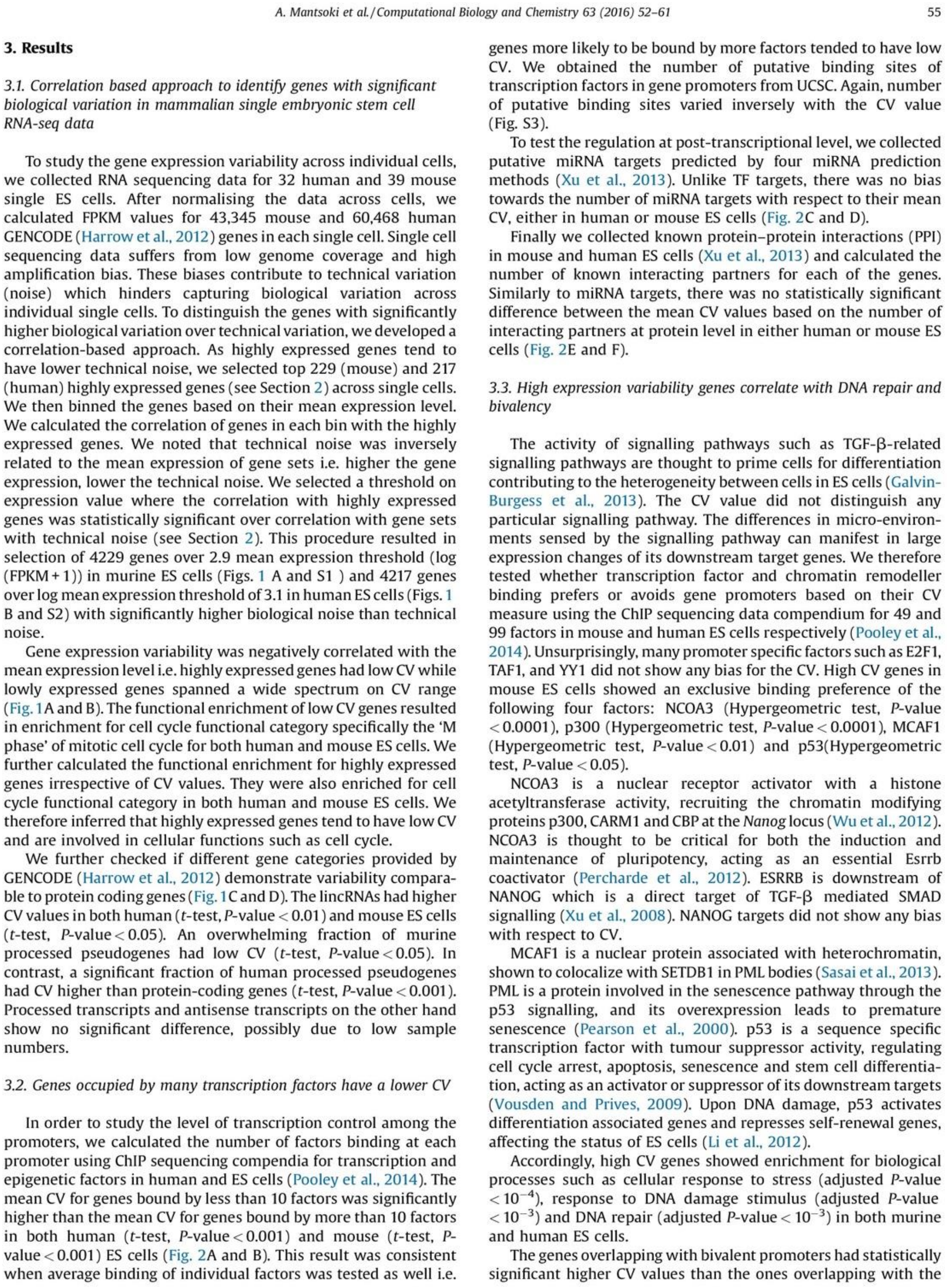
A lists of topological associated domains (TADs) for mouse and human ES cells (Dixon et al., 2012) was used to calculate the number of genes per TAD for the High CV and Non High CV genes in our analysis.

2.13. Bulk expression data

For the bulk RNA analysis we used 3 biological replicates of Microarray data from mouse ES cells (GSM1326660-2) (Zhang et al., 2014) and 4 biological replicates of RNA-seq data from hESCs (GSE33480) (Djebali et al., 2012).

2.14. Sequence conservation

The sequence conservation scores where obtained from PhyloP100way (Human) and PhyloP60way (Mouse) tracks available at UCSC.



3. Results

3.1. Correlation based approach to identify genes with significant biological variation in mammalian single embryonic stem cell RNA-seq data

To study the gene expression variability across individual cells, we collected RNA sequencing data for 32 human and 39 mouse single ES cells. After normalising the data across cells, we calculated FPKM values for 43,345 mouse and 60,468 human GENCODE (Harrow et al., 2012) genes in each single cell. Single cell sequencing data suffers from low genome coverage and high amplification bias. These biases contribute to technical variation (noise) which hinders capturing biological variation across individual single cells. To distinguish the genes with significantly higher biological variation over technical variation, we developed a correlation-based approach. As highly expressed genes tend to have lower technical noise, we selected top 229 (mouse) and 217 (human) highly expressed genes (see Section 2) across single cells. We then binned the genes based on their mean expression level. We calculated the correlation of genes in each bin with the highly expressed genes. We noted that technical noise was inversely related to the mean expression of gene sets i.e. higher the gene expression, lower the technical noise. We selected a threshold on expression value where the correlation with highly expressed genes was statistically significant over correlation with gene sets with technical noise (see Section 2). This procedure resulted in selection of 4229 genes over 2.9 mean expression threshold (log (FPKM + 1)) in murine ES cells (Figs. 1 A and S1) and 4217 genes over log mean expression threshold of 3.1 in human ES cells (Figs. 1 B and S2) with significantly higher biological noise than technical noise.

Gene expression variability was negatively correlated with the mean expression level i.e. highly expressed genes had low CV while lowly expressed genes spanned a wide spectrum on CV range (Fig. 1A and B). The functional enrichment of low CV genes resulted in enrichment for cell cycle functional category specifically the ‘M phase’ of mitotic cell cycle for both human and mouse ES cells. We further calculated the functional enrichment for highly expressed genes irrespective of CV values. They were also enriched for cell cycle functional category in both human and mouse ES cells. We therefore inferred that highly expressed genes tend to have low CV and are involved in cellular functions such as cell cycle.

We further checked if different gene categories provided by GENCODE (Harrow et al., 2012) demonstrate variability comparable to protein coding genes (Fig. 1C and D). The lincRNAs had higher CV values in both human (*t*-test, P-value < 0.01) and mouse ES cells (*t*-test, P-value < 0.05). An overwhelming fraction of murine processed pseudogenes had low CV (*t*-test, P-value < 0.05). In contrast, a significant fraction of human processed pseudogenes had CV higher than protein-coding genes (*t*-test, P-value < 0.001). Processed transcripts and antisense transcripts on the other hand show no significant difference, possibly due to low sample numbers.

3.2. Genes occupied by many transcription factors have a lower CV

In order to study the level of transcription control among the promoters, we calculated the number of factors binding at each promoter using ChIP sequencing compendia for transcription and epigenetic factors in human and ES cells (Pooley et al., 2014). The mean CV for genes bound by less than 10 factors was significantly higher than the mean CV for genes bound by more than 10 factors in both human (*t*-test, P-value < 0.001) and mouse (*t*-test, P-value < 0.001) ES cells (Fig. 2A and B). This result was consistent when average binding of individual factors was tested as well i.e. genes more likely to be bound by more factors tended to have low CV. We obtained the number of putative binding sites of transcription factors in gene promoters from UCSC. Again, number of putative binding sites varied inversely with the CV value (Fig. S3).

To test the regulation at post-transcriptional level, we collected putative miRNA targets predicted by four miRNA prediction methods (Xu et al., 2013). Unlike TF targets, there was no bias towards the number of miRNA targets with respect to their mean CV, either in human or mouse ES cells (Fig. 2C and D).

Finally we collected known protein–protein interactions (PPI) in mouse and human ES cells (Xu et al., 2013) and calculated the number of known interacting partners for each of the genes. Similarly to miRNA targets, there was no statistically significant difference between the mean CV values based on the number of interacting partners at protein level in either human or mouse ES cells (Fig. 2E and F).

3.3. High expression variability genes correlate with DNA repair and bivalency

The activity of signalling pathways such as TGF-β-related signalling pathways are thought to prime cells for differentiation contributing to the heterogeneity between cells in ES cells (Galvin-Burgess et al., 2013). The CV value did not distinguish any particular signalling pathway. The differences in micro-environments sensed by the signalling pathway can manifest in large expression changes of its downstream target genes. We therefore tested whether transcription factor and chromatin remodeller binding prefers or avoids gene promoters based on their CV measure using the ChIP sequencing data compendium for 49 and 99 factors in mouse and human ES cells respectively (Pooley et al., 2014). Unsurprisingly, many promoter specific factors such as E2F1, TAF1, and YY1 did not show any bias for the CV. High CV genes in mouse ES cells showed an exclusive binding preference of the following four factors: NCOA3 (Hypergeometric test, P-value < 0.0001), p300 (Hypergeometric test, P-value < 0.0001), MCAF1 (Hypergeometric test, P-value < 0.01) and p53(Hypergeometric test, P-value < 0.05).

NCOA3 is a nuclear receptor activator with a histone acetyltransferase activity, recruiting the chromatin modifying proteins p300, CARM1 and CBP at the *Nanog* locus (Wu et al., 2012). NCOA3 is thought to be critical for both the induction and maintenance of pluripotency, acting as an essential Esrrb coactivator (Percharde et al., 2012). ESRRB is downstream of NANOG which is a direct target of TGF-β mediated SMAD signalling (Xu et al., 2008). NANOG targets did not show any bias with respect to CV.

MCAF1 is a nuclear protein associated with heterochromatin, shown to colocalize with SETDB1 in PML bodies (Sasai et al., 2013). PML is a protein involved in the senescence pathway through the p53 signalling, and its overexpression leads to premature senescence (Pearson et al., 2000). p53 is a sequence specific transcription factor with tumour suppressor activity, regulating cell cycle arrest, apoptosis, senescence and stem cell differentiation, acting as an activator or suppressor of its downstream targets (Vousden and Prives, 2009). Upon DNA damage, p53 activates differentiation associated genes and represses self-renewal genes, affecting the status of ES cells (Li et al., 2012).

Accordingly, high CV genes showed enrichment for biological processes such as cellular response to stress (adjusted P-value $< 10^{-4}$), response to DNA damage stimulus (adjusted P-value $< 10^{-3}$) and DNA repair (adjusted P-value $< 10^{-3}$) in both murine and human ES cells.

The genes overlapping with bivalent promoters had statistically significant higher CV values than the ones overlapping with the

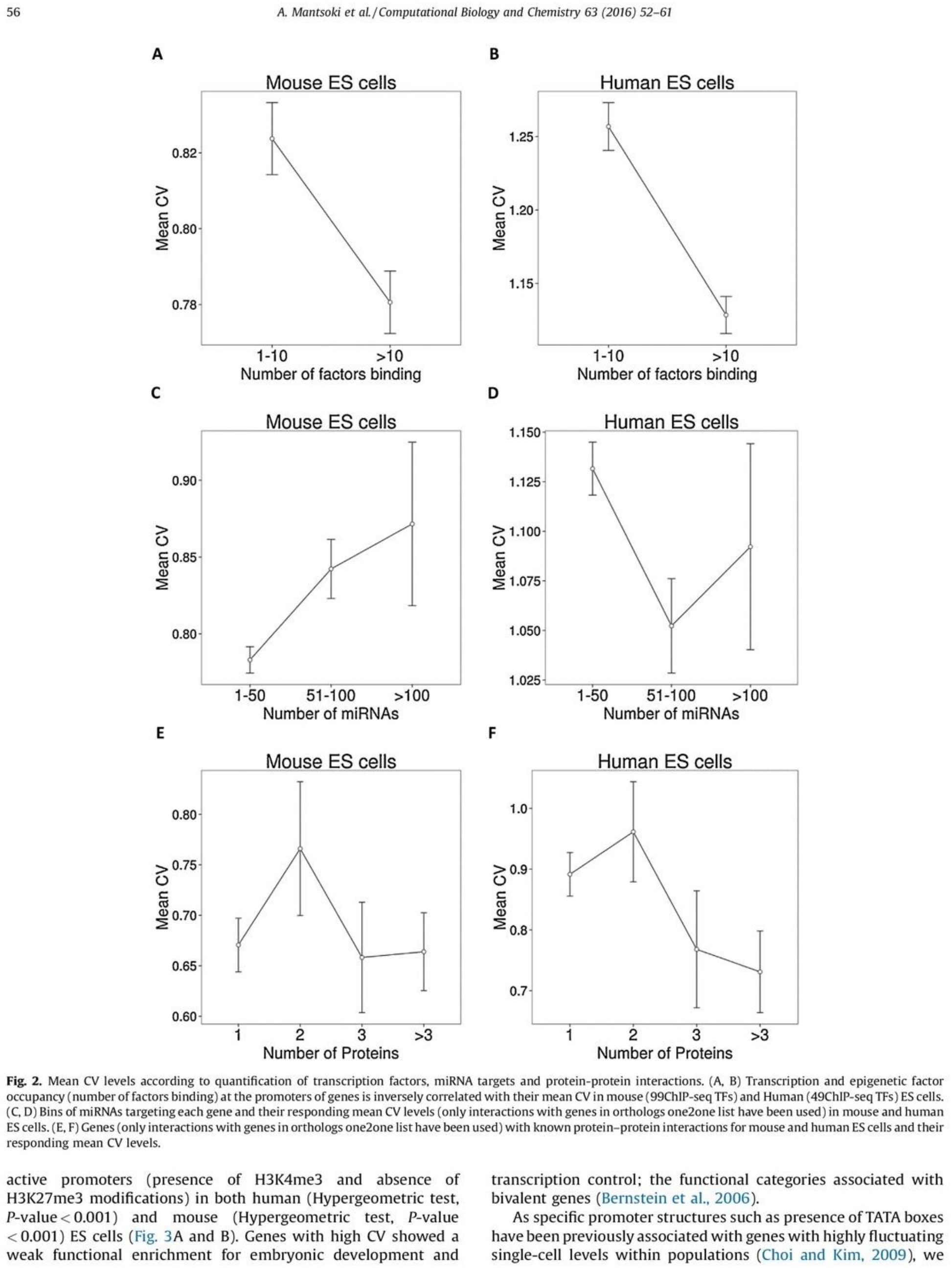


Fig. 2. Mean CV levels according to quantification of transcription factors, miRNA targets and protein-protein interactions. (A, B) Transcription and epigenetic factor occupancy (number of factors binding) at the promoters of genes is inversely correlated with their mean CV in mouse (99ChIP-seq TFs) and Human (49ChIP-seq TFs) ES cells. (C, D) Bins of miRNAs targeting each gene and their responding mean CV levels (only interactions with genes in orthologs one2one list have been used) in mouse and human ES cells. (E, F) Genes (only interactions with genes in orthologs one2one list have been used) with known protein–protein interactions for mouse and human ES cells and their responding mean CV levels.

active promoters (presence of H3K4me3 and absence of H3K27me3 modifications) in both human (Hypergeometric test, *P*-value < 0.001) and mouse (Hypergeometric test, *P*-value < 0.001) ES cells (Fig. 3A and B). Genes with high CV showed a weak functional enrichment for embryonic development and transcription control; the functional categories associated with bivalent genes (Bernstein et al., 2006).

As specific promoter structures such as presence of TATA boxes have been previously associated with genes with highly fluctuating single-cell levels within populations (Choi and Kim, 2009), we calculated TATA and CpG island fraction for all human and mouse promoters (−/+ 1Kb from TSS). The CpG-rich promoters showed lower CV values than the CpG-poor promoters and the difference was statistically significant in both human and mouse ES cells (*t*-test *P*-value < 0.001) (Fig. 3C and D). Unlike CpG promoters, TATA box promoters could not be distinguished based on the CV value (Fig. 3E and F).

3.4. High CV genes form dense highly co-expressed clusters

In order to study the characteristics of genes with high variability, we defined genes with CV value greater than 0.92 (3rd quartile value) as High CV in mouse (Fig. 4A) and genes with CV value greater than 1.45 (3rd quartile value) in human ES cells (Fig. 4B). We then checked whether the expression of High CV

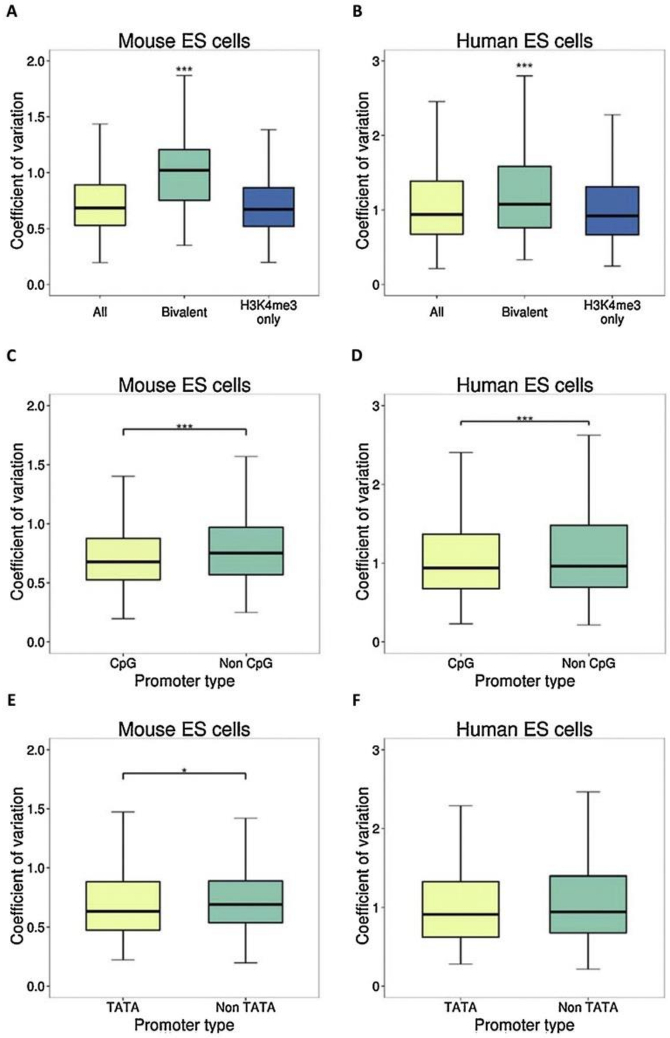


Fig. 3. Chromatin modifications and sequence features of genes and their corresponding coefficient of variation. (A, B) Overlapping genes with bivalent and active (H3K4me3 marked) gene promoters in response to their CV, in mouse and human ES cells. Bivalent genes show significantly higher CV levels than all the promoters (irrespective of overlap) and the active promoters (pairwise *t*-test, *P*-value < 0.001) (C) CV levels of genes having a CpG island and a non- CpG island promoter. (D) CV levels of genes having a TATA box and a non-TATA box promoter.

genes varies concordantly across single cells by calculating Pearson's correlation coefficient between all pairs of High CV genes. A subset of High CV genes were significantly more correlated with each other compared to expected from a random permutation (Fig. 4C (mouse) and D (human)).

The highly correlated network (Pearson's correlation coefficient >0.95) of High CV genes grouped them mainly into only few tightly co-expressed clusters in both human and mouse ES cells (Figs. S4 and S5). Interestingly, the genes in each cluster were highly expressed only in one individual cell (Fig. 4E (mouse) and F (human)). We firstly confirmed that these single cells (e.g. single cell 24 and 26 in humans) did not suffer from poor technical quality of samples (Fig. S6). We also removed these two cells and redefined the High CV gene set (Fig. S7) to find a similar result. This assured that the significant co-expression among High CV genes is not an artefact of few aberrant single cells.

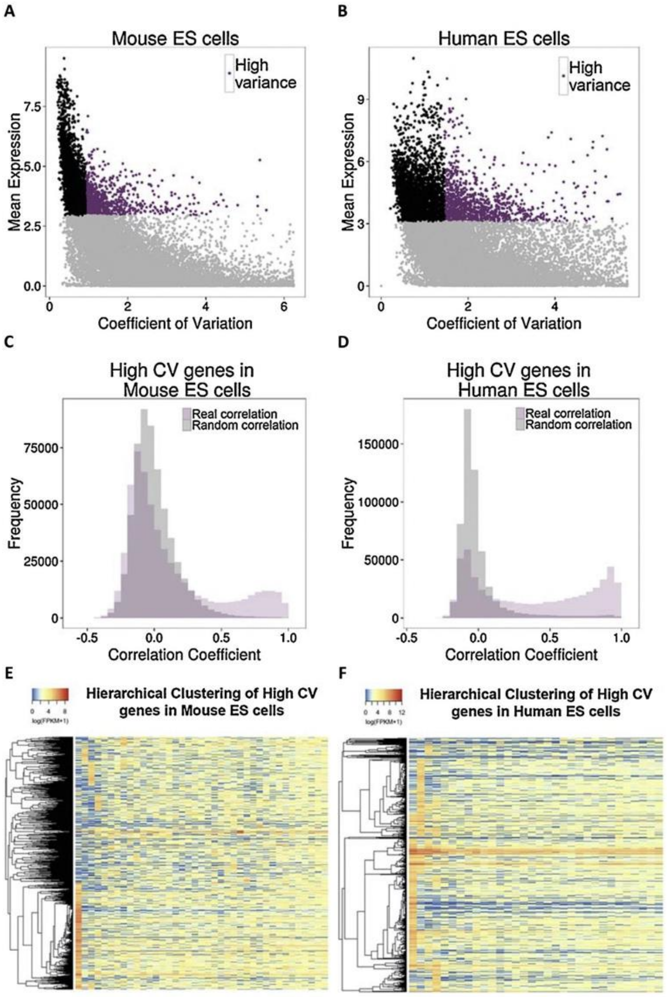


Fig. 4. High variance genes are more correlated than expected by chance (A, B) Scatterplot of genes in response to their CV and mean expression. Highlighted in purple are the High variance genes, selected based on their CV (CV value greater than the third quartile of the distribution). (C, D) Correlation coefficient distributions for the High variance (High CV) genes in mouse and human ES cells (statistically significant difference ($p < 0.001$, Wilcoxon test) between the real and random distributions). (E, F) Heatmaps of gene expression (in log(FPKM + 1) values) for the High variance genes (High CV) in mouse and human ES cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The co-expressed genes derived from large-scale analyses of mammalian expression data have demonstrated that neighbouring genes tend to have similar expression profiles (Lercher et al., 2002). As high CV genes formed tight co-expression clusters, we checked whether they tend to be in gene neighbourhoods with each other compared to other genes. We did not observe any tendency of genes clustering based on CV value. We also checked whether there was any bias towards similar CV genes co-existing in

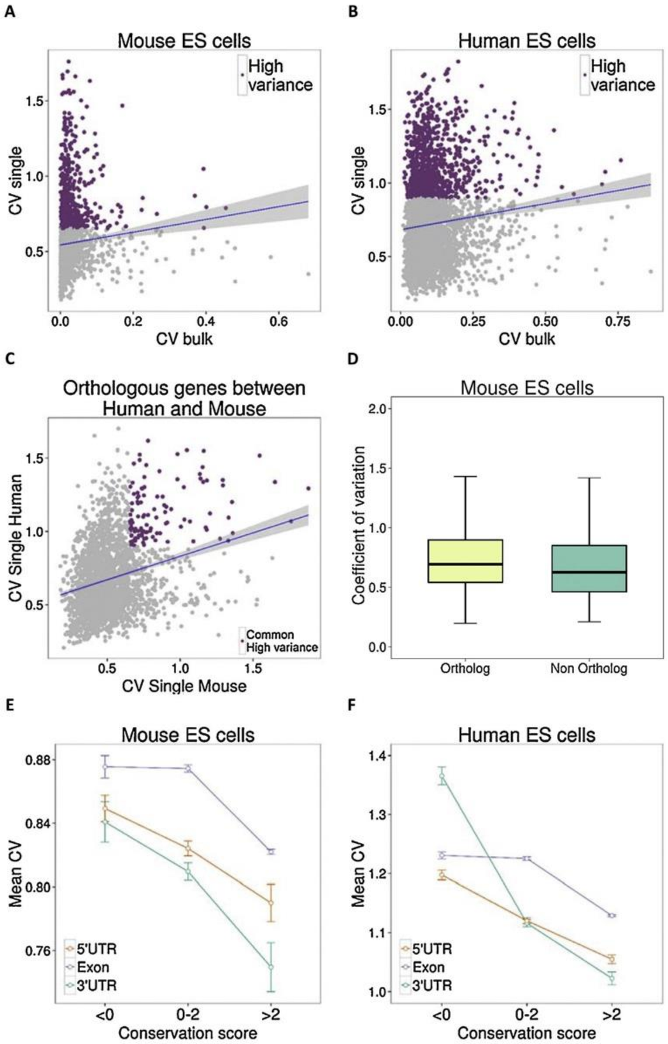
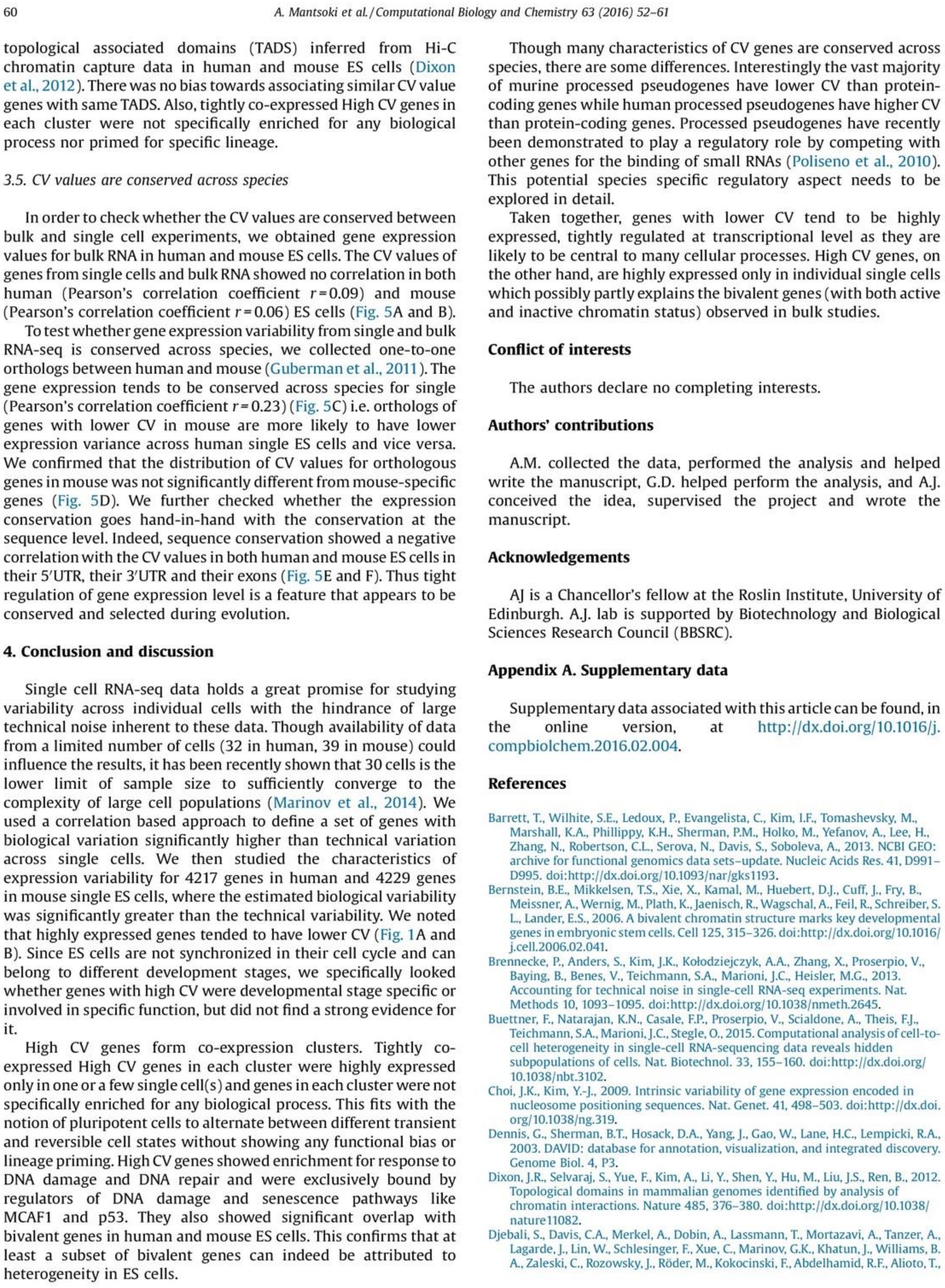


Fig. 5. Conservation of expression variability across technologies and species. (A, B) Scatterplot of CV values in a bulk expression study against CV values in a single cell RNA–seq study in mouse and human ES cells. There is a positive correlation between the CV values of the two technologies (Pearson's $r = 0.06$ for mouse, $r = 0.09$ for human). (C) Scatterplot of CV values of orthologous genes between human and mouse from single RNA-seq studies in ESCs. There is a positive correlation of CV values between species (Pearson's $r = 0.23$) and 10% of High CV genes (highlighted in purple) are conserved as highly variant between species (D) Boxplot of CV values of orthologous and non-orthologous genes between human and mouse in ESCs (3675 orthologs and 554 non-orthologs out of 4229 genes in our analysis). (E, F) Sequence conservation scores and their corresponding mean CV values for 5′UTR, Exons and 3′UTRs in mouse and human ES cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



topological associated domains (TADS) inferred from Hi-C chromatin capture data in human and mouse ES cells (Dixon et al., 2012). There was no bias towards associating similar CV value genes with same TADS. Also, tightly co-expressed High CV genes in each cluster were not specifically enriched for any biological process nor primed for specific lineage.

3.5. CV values are conserved across species

In order to check whether the CV values are conserved between bulk and single cell experiments, we obtained gene expression values for bulk RNA in human and mouse ES cells. The CV values of genes from single cells and bulk RNA showed no correlation in both human (Pearson's correlation coefficient $r = 0.09$) and mouse (Pearson's correlation coefficient $r = 0.06$) ES cells (Fig. 5A and B).

To test whether gene expression variability from single and bulk RNA-seq is conserved across species, we collected one-to-one orthologs between human and mouse (Guberman et al., 2011). The gene expression tends to be conserved across species for single (Pearson's correlation coefficient $r = 0.23$) (Fig. 5C) i.e. orthologs of genes with lower CV in mouse are more likely to have lower expression variance across human single ES cells and vice versa. We confirmed that the distribution of CV values for orthologous genes in mouse was not significantly different from mouse-specific genes (Fig. 5D). We further checked whether the expression conservation goes hand-in-hand with the conservation at the sequence level. Indeed, sequence conservation showed a negative correlation with the CV values in both human and mouse ES cells in their 5′UTR, their 3′UTR and their exons (Fig. 5E and F). Thus tight regulation of gene expression level is a feature that appears to be conserved and selected during evolution.

4. Conclusion and discussion

Single cell RNA-seq data holds a great promise for studying variability across individual cells with the hindrance of large technical noise inherent to these data. Though availability of data from a limited number of cells (32 in human, 39 in mouse) could influence the results, it has been recently shown that 30 cells is the lower limit of sample size to sufficiently converge to the complexity of large cell populations (Marinov et al., 2014). We used a correlation based approach to define a set of genes with biological variation significantly higher than technical variation across single cells. We then studied the characteristics of expression variability for 4217 genes in human and 4229 genes in mouse single ES cells, where the estimated biological variability was significantly greater than the technical variability. We noted that highly expressed genes tended to have lower CV (Fig. 1A and B). Since ES cells are not synchronized in their cell cycle and can belong to different development stages, we specifically looked whether genes with high CV were developmental stage specific or involved in specific function, but did not find a strong evidence for it.

High CV genes form co-expression clusters. Tightly co-expressed High CV genes in each cluster were highly expressed only in one or a few single cell(s) and genes in each cluster were not specifically enriched for any biological process. This fits with the notion of pluripotent cells to alternate between different transient and reversible cell states without showing any functional bias or lineage priming. High CV genes showed enrichment for response to DNA damage and DNA repair and were exclusively bound by regulators of DNA damage and senescence pathways like MCAF1 and p53. They also showed significant overlap with bivalent genes in human and mouse ES cells. This confirms that at least a subset of bivalent genes can indeed be attributed to heterogeneity in ES cells.

Though many characteristics of CV genes are conserved across species, there are some differences. Interestingly the vast majority of murine processed pseudogenes have lower CV than protein-coding genes while human processed pseudogenes have higher CV than protein-coding genes. Processed pseudogenes have recently been demonstrated to play a regulatory role by competing with other genes for the binding of small RNAs (Poliseno et al., 2010). This potential species specific regulatory aspect needs to be explored in detail.

Taken together, genes with lower CV tend to be highly expressed, tightly regulated at transcriptional level as they are likely to be central to many cellular processes. High CV genes, on the other hand, are highly expressed only in individual single cells which possibly partly explains the bivalent genes (with both active and inactive chromatin status) observed in bulk studies.

Conflict of interests

The authors declare no completing interests.

Authors' contributions

A.M. collected the data, performed the analysis and helped write the manuscript, G.D. helped perform the analysis, and A.J. conceived the idea, supervised the project and wrote the manuscript.

Acknowledgements

AJ is a Chancellor's fellow at the Roslin Institute, University of Edinburgh. A.J. lab is supported by Biotechnology and Biological Sciences Research Council (BBSRC).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2016.02.004.

References

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A., 2013. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 41, D991–D995. doi:http://dx.doi.org/10.1093/nar/gks1193.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S. L., Lander, E.S., 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell 125, 315–326. doi:http://dx.doi.org/10.1016/j.cell.2006.02.041.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., Heisler, M.G., 2013. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10, 1093–1095. doi:http://dx.doi.org/10.1038/nmeth.2645.

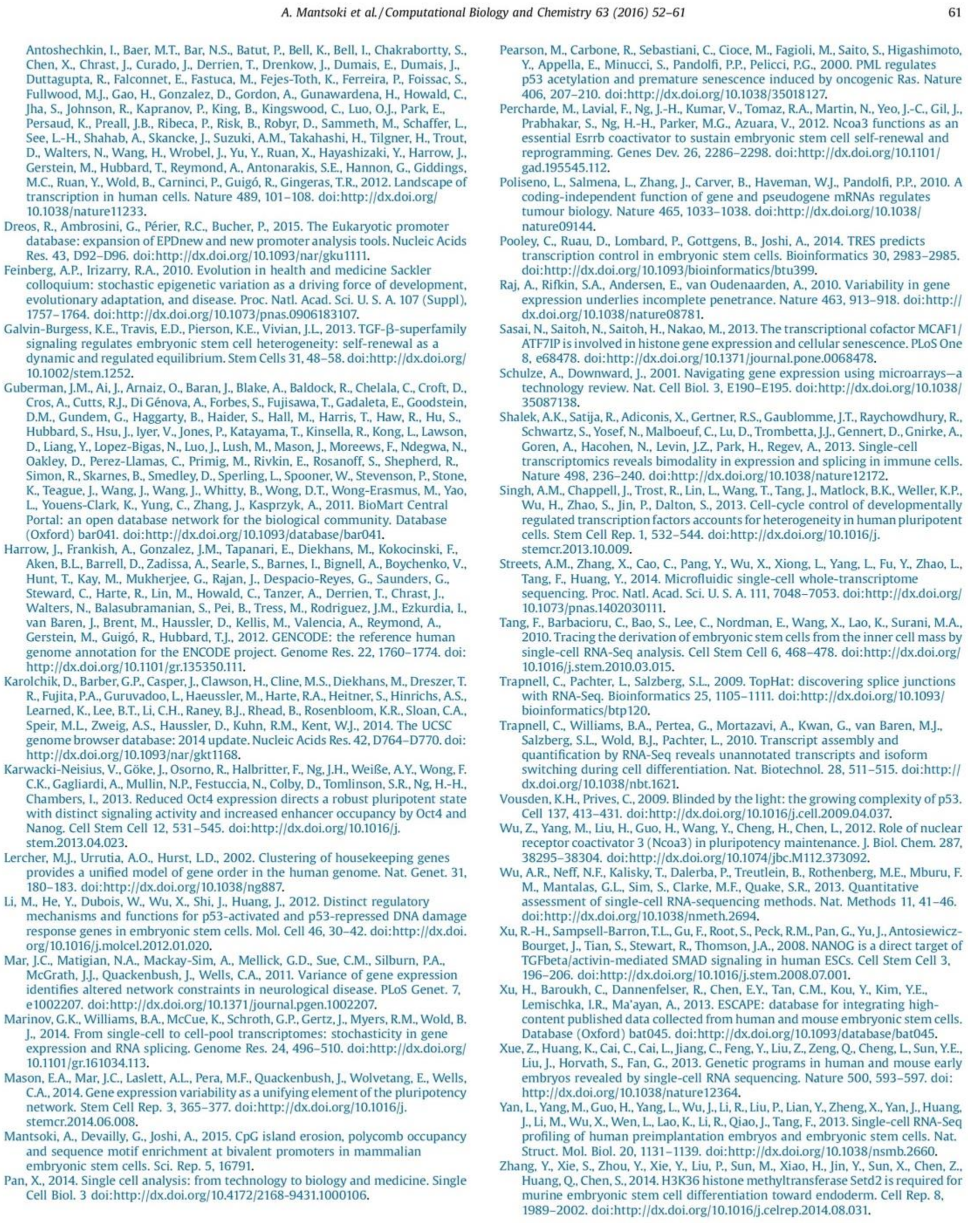
Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., Stegle, O., 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. 33, 155–160. doi:http://dx.doi.org/10.1038/nbt.3102.

Choi, J.K., Kim, Y.-J., 2009. Intrinsic variability of gene expression encoded in nucleosome positioning sequences. Nat. Genet. 41, 498–503. doi:http://dx.doi.org/10.1038/ng.319.

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A., 2003. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 4, P3.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380. doi:http://dx.doi.org/10.1038/nature11082.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T.,



Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T.R., 2012. Landscape of transcription in human cells. Nature 489, 101–108. doi:http://dx.doi.org/10.1038/nature11233.

Dreos, R., Ambrosini, G., Périer, R.C., Bucher, P., 2015. The Eukaryotic promoter database: expansion of EPDnew and new promoter analysis tools. Nucleic Acids Res. 43, D92–D96. doi:http://dx.doi.org/10.1093/nar/gku1111.

Feinberg, A.P., Irizarry, R.A., 2010. Evolution in health and medicine Sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc. Natl. Acad. Sci. U. S. A. 107 (Suppl), 1757–1764. doi:http://dx.doi.org/10.1073/pnas.0906183107.

Galvin-Burgess, K.E., Travis, E.D., Pierson, K.E., Vivian, J.L., 2013. TGF-β-superfamily signaling regulates embryonic stem cell heterogeneity: self-renewal as a dynamic and regulated equilibrium. Stem Cells 31, 48–58. doi:http://dx.doi.org/10.1002/stem.1252.

Guberman, J.M., Ai, J., Arnaiz, O., Baran, J., Blake, A., Baldock, R., Chelala, C., Croft, D., Cros, A., Cutts, R.J., Di Génova, A., Forbes, S., Fujisawa, T., Gadaleta, E., Goodstein, D.M., Gundem, G., Haggarty, B., Haider, S., Hall, M., Harris, T., Haw, R., Hu, S., Hubbard, S., Hsu, J., Iyer, V., Jones, P., Katayama, T., Kinsella, R., Kong, L., Lawson, D., Liang, Y., Lopez-Bigas, N., Luo, J., Lush, M., Mason, J., Moreews, F., Ndegwa, N., Oakley, D., Perez-Llamas, C., Primig, M., Rivkin, E., Rosanoff, S., Shepherd, R., Simon, R., Skarnes, B., Smedley, D., Sperling, L., Spooner, W., Stevenson, P., Stone, K., Teague, J., Wang, J., Wang, J., Whitty, B., Wong, D.T., Wong-Erasmus, M., Yao, L., Youens-Clark, K., Yung, C., Zhang, J., Kasprzyk, A., 2011. BioMart Central Portal: an open database network for the biological community. Database (Oxford) bar041. doi:http://dx.doi.org/10.1093/database/bar041.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., Hubbard, T.J., 2012. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 22, 1760–1774. doi:http://dx.doi.org/10.1101/gr.135350.111.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T. R., Fujita, P.A., Guruvadoo, L., Haeussler, M., Harte, R.A., Heitner, S., Hinrichs, A.S., Learned, K., Lee, B.T., Li, C.H., Raney, B.J., Rhead, B., Rosenbloom, K.R., Sloan, C.A., Speir, M.L., Zweig, A.S., Haussler, D., Kuhn, R.M., Kent, W.J., 2014. The UCSC genome browser database: 2014 update. Nucleic Acids Res. 42, D764–D770. doi:http://dx.doi.org/10.1093/nar/gkt1168.

Karwacki-Neisius, V., Göke, J., Osorno, R., Halbritter, F., Ng, J.H., Weiße, A.Y., Wong, F. C.K., Gagliardi, A., Mullin, N.P., Festuccia, N., Colby, D., Tomlinson, S.R., Ng, H.-H., Chambers, I., 2013. Reduced Oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by Oct4 and Nanog. Cell Stem Cell 12, 531–545. doi:http://dx.doi.org/10.1016/j.stem.2013.04.023.

Lercher, M.J., Urrutia, A.O., Hurst, L.D., 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet. 31, 180–183. doi:http://dx.doi.org/10.1038/ng887.

Li, M., He, Y., Dubois, W., Wu, X., Shi, J., Huang, J., 2012. Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. Mol. Cell 46, 30–42. doi:http://dx.doi.org/10.1016/j.molcel.2012.01.020.

Mar, J.C., Matigian, N.A., Mackay-Sim, A., Mellick, G.D., Sue, C.M., Silburn, P.A., McGrath, J.J., Quackenbush, J., Wells, C.A., 2011. Variance of gene expression identifies altered network constraints in neurological disease. PLoS Genet. 7, e1002207. doi:http://dx.doi.org/10.1371/journal.pgen.1002207.

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., Wold, B. J., 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 24, 496–510. doi:http://dx.doi.org/10.1101/gr.161034.113.

Mason, E.A., Mar, J.C., Laslett, A.L., Pera, M.F., Quackenbush, J., Wolvetang, E., Wells, C.A., 2014. Gene expression variability as a unifying element of the pluripotency network. Stem Cell Rep. 3, 365–377. doi:http://dx.doi.org/10.1016/j.stemcr.2014.06.008.

Mantsoki, A., Devailly, G., Joshi, A., 2015. CpG island erosion, polycomb occupancy and sequence motif enrichment at bivalent promoters in mammalian embryonic stem cells. Sci. Rep. 5, 16791.

Pan, X., 2014. Single cell analysis: from technology to biology and medicine. Single Cell Biol. 3 doi:http://dx.doi.org/10.4172/2168-9431.1000106.

Pearson, M., Carbone, R., Sebastiani, C., Cioce, M., Fagioli, M., Saito, S., Higashimoto, Y., Appella, E., Minucci, S., Pandolfi, P.P., Pelicci, P.G., 2000. PML regulates p53 acetylation and premature senescence induced by oncogenic Ras. Nature 406, 207–210. doi:http://dx.doi.org/10.1038/35018127.

Percharde, M., Lavial, F., Ng, J.-H., Kumar, V., Tomaz, R.A., Martin, N., Yeo, J.-C., Gil, J., Prabhakar, S., Ng, H.-H., Parker, M.G., Azuara, V., 2012. Ncoa3 functions as an essential Esrrb coactivator to sustain embryonic stem cell self-renewal and reprogramming. Genes Dev. 26, 2286–2298. doi:http://dx.doi.org/10.1101/gad.195545.112.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., Pandolfi, P.P., 2010. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 465, 1033–1038. doi:http://dx.doi.org/10.1038/nature09144.

Pooley, C., Ruau, D., Lombard, P., Gottgens, B., Joshi, A., 2014. TRES predicts transcription control in embryonic stem cells. Bioinformatics 30, 2983–2985. doi:http://dx.doi.org/10.1093/bioinformatics/btu399.

Raj, A., Rifkin, S.A., Andersen, E., van Oudenaarden, A., 2010. Variability in gene expression underlies incomplete penetrance. Nature 463, 913–918. doi:http://dx.doi.org/10.1038/nature08781.

Sasai, N., Saitoh, N., Saitoh, H., Nakao, M., 2013. The transcriptional cofactor MCAF1/ATF7IP is involved in histone gene expression and cellular senescence. PLoS One 8, e68478. doi:http://dx.doi.org/10.1371/journal.pone.0068478.

Schulze, A., Downward, J., 2001. Navigating gene expression using microarrays—a technology review. Nat. Cell Biol. 3, E190–E195. doi:http://dx.doi.org/10.1038/35087138.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., Trombetta, J.J., Gennert, D., Gnirke, A., Goren, A., Hacohen, N., Levin, J.Z., Park, H., Regev, A., 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498, 236–240. doi:http://dx.doi.org/10.1038/nature12172.

Singh, A.M., Chappell, J., Trost, R., Lin, L., Wang, T., Tang, J., Matlock, B.K., Weller, K.P., Wu, H., Zhao, S., Jin, P., Dalton, S., 2013. Cell-cycle control of developmentally regulated transcription factors accounts for heterogeneity in human pluripotent cells. Stem Cell Rep. 1, 532–544. doi:http://dx.doi.org/10.1016/j.stemcr.2013.10.009.

Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F., Huang, Y., 2014. Microfluidic single-cell whole-transcriptome sequencing. Proc. Natl. Acad. Sci. U. S. A. 111, 7048–7053. doi:http://dx.doi.org/10.1073/pnas.1402030111.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., Surani, M.A., 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6, 468–478. doi:http://dx.doi.org/10.1016/j.stem.2010.03.015.

Trapnell, C., Pachter, L., Salzberg, S.L., 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25, 1105–1111. doi:http://dx.doi.org/10.1093/bioinformatics/btp120.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515. doi:http://dx.doi.org/10.1038/nbt.1621.

Vousden, K.H., Prives, C., 2009. Blinded by the light: the growing complexity of p53. Cell 137, 413–431. doi:http://dx.doi.org/10.1016/j.cell.2009.04.037.

Wu, Z., Yang, M., Liu, H., Guo, H., Wang, Y., Cheng, H., Chen, L., 2012. Role of nuclear receptor coactivator 3 (Ncoa3) in pluripotency maintenance. J. Biol. Chem. 287, 38295–38304. doi:http://dx.doi.org/10.1074/jbc.M112.373092.

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F. M., Mantalas, G.L., Sim, S., Clarke, M.F., Quake, S.R., 2013. Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods 11, 41–46. doi:http://dx.doi.org/10.1038/nmeth.2694.

Xu, R.-H., Sampsell-Barron, T.L., Gu, F., Root, S., Peck, R.M., Pan, G., Yu, J., Antosiewicz-Bourget, J., Tian, S., Stewart, R., Thomson, J.A., 2008. NANOG is a direct target of TGFbeta/activin-mediated SMAD signaling in human ESCs. Cell Stem Cell 3, 196–206. doi:http://dx.doi.org/10.1016/j.stem.2008.07.001.

Xu, H., Baroukh, C., Dannenfelser, R., Chen, E.Y., Tan, C.M., Kou, Y., Kim, Y.E., Lemischka, I.R., Ma'ayan, A., 2013. ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. Database (Oxford) bat045. doi:http://dx.doi.org/10.1093/database/bat045.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E., Liu, J., Horvath, S., Fan, G., 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature 500, 593–597. doi:http://dx.doi.org/10.1038/nature12364.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J., Tang, F., 2013. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. 20, 1131–1139. doi:http://dx.doi.org/10.1038/nsmb.2660.

Zhang, Y., Xie, S., Zhou, Y., Xie, Y., Liu, P., Sun, M., Xiao, H., Jin, Y., Sun, X., Chen, Z., Huang, Q., Chen, S., 2014. H3K36 histone methyltransferase Setd2 is required for murine embryonic stem cell differentiation toward endoderm. Cell Rep. 8, 1989–2002. doi:http://dx.doi.org/10.1016/j.celrep.2014.08.031.

Appendix III – Springer License