The Role of DNA Methylation on Transcription Factor Occupancy and Transcriptional Activity

Martin Cusack

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Clinical Medicine

St Catherine's College

Ludwig Institute for Cancer Research

Nuffield Department of Clinical Medicine

University of Oxford

Michaelmas 2017

This thesis is the result of my own work unless otherwise stated and does not constitute part of any other thesis. The work herein described was carried out at the Ludwig Institute for Cancer Research, under the supervision of Dr. Skirmantas Kriaucionis.

Word count: approx. 48350 (including citations and tables)

Martin Cusack

The Role of DNA Methylation on Transcription Factor Occupancy and Transcriptional Activity

Martin Cusack Michaelmas 2017

Doctor of Philosophy in Clinical Medicine

St Catherine's College

Abstract

DNA methylation is an epigenetic mark that is deposited throughout the genome of mammals and plays an important role in the maintenance of transcriptionally repressive states across cell divisions. There are two major mechanisms by which DNA methylation has been proposed to act: one involves the recognition of the mark by protein complexes containing histone deacetylases (HDACs) that can remodel the local chromatin. Alternatively, methylation has been suggested to directly affect the interaction between transcription factors and their cognate binding sequence. The aim of this research was to determine the contributions of these two mechanisms in cells.

The importance of HDAC activity in mediating DNA methylation-dependent transcriptional repression was assessed by comparing the genes and retrotransposons that are upregulated in response to DNA methylation loss or the disruption of HDAC activity. To this purpose, we performed whole-genome transcriptional analysis in wild type and DNA methylation-deficient mouse embryonic stem cells (DNMT.TKO mESCs) in the presence and absence of the HDAC inhibitor trichostatin A. Our data suggests that there are few genes whose repression is solely dependent on the recruitment of HDACs by DNA methylation in mESCs. Rather it appears that DNA methylation and HDAC-mediated silencing represent two independent layers of repression that converge at certain transcriptional elements.

To investigate the contribution of DNA methylation on the genome-wide occupancy of transcription factors, we compared the global chromatin accessibility landscape and the binding profile of candidate transcription factors in the absence or presence of DNA methylation. We found that loss of DNA methylation associates with localised gains in accessibility, some of which can be linked to the novel binding of transcription factors such as GABPA, MAX, NRF1 and YY1.

Altogether, our results present new insights into the interplay between DNA methylation and histone deacetylation and their impact on the localisation of transcription factors from different families.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Dr. Skirmantas Kriaucionis for hosting me in his research group and guiding me throughout my PhD studies. I particularly appreciate his enthusiasm for science, depth of knowledge and optimism that have allowed me to make the most of these last four years and helped me through the more challenging periods.

I thank the Nuffield Department for Medicine and the Medical Research Council for granting me a PhD studentship and the Ludwig Institute for Cancer Research in providing me with a supportive environment in which to pursue my doctoral degree. Much of what I have learnt, I owe to my wonderful colleagues. Within my lab group, I am grateful to Anandhakumar Chandran, Hiromi Tagoh, Marie Gerus-Durand, Marketa Tomkova, Melania Zauri, Michael McClellan, Paolo Spingardi, Pijus Brazauskas, Sophie Kirschner and Wenjun Huang, not only for accompanying me to lunch; more importantly, for sharing their scientific knowledge and suggestions and for their unwavering friendship. It was my privilege to work alongside such caring and fun people as Federica Lari, Kira Fischer, Yurena Vivas of whom I will have fond memories. A special mention to Emma Fenech, who worked opposite me (albeit separated by a solid partition) since the beginning is imperative. She always had ears for my complaints and grievances without losing her patience or her cool.

I am especially thankful to Hamish King who has helped me carry out and analyse the ATAC-seq experiments and with whom I have had many interesting discussions.

During my second year, I had the opportunity to supervise an undergraduate student, an experience that I greatly valued. It was a pleasure to work with Barbara Dzino and we were close friends outside of the lab.

I had the opportunity to meet many fantastic people during my time in Oxford. My flatmates have frequently provided my with joy: Emily Simmonds through her kindness, Lara Veylit through her compassion, Josie Mahony by always being full of happiness and good humour and Erik Sandvig by providing me with home-made beers. I am grateful to Mattia Zucca, Derek Leske and Bertrand Gallet for their easy friendship and for joining me on memorable cycling adventures.

Last but not least, I would like to thank my parents, my brother Jeremy and my loving girlfriend Angela for their support and encouragement throughout this period.

Table of Contents

List of Abbreviations

2C	2-cell
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
A	Adenine
ATAC-seq	Assay for Transposase-Accessible Chromatin coupled with high-throughput sequencing
BAM	Binary sequence Alignment/Map (file type)
BED	Browser Extensible Data (file type)
bHLH	basic Helix-Loop-Helix
bp	Base pair
C	Cytosine
CEBPB	CCAAT/Enhancer Binding Protein Beta
CGI	CpG Island
ChIP	Chromatin Immunoprecipitation
Chr	Chromosome
CIMP	CpG Island Methylator Phenotype
CpG	Genomic sequence with C followed by G in a 5' to 3' direction
CRE	cAMP Response Element
CREB	CRE Binding protein
CTCF	CCCTC-binding Factor
DHS	DNaseI Hypersensitive Site
DMR	Differentially Methylated Region
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic Acid
DNMT	DNA Methyltransferase
DNMT.TKO	Cell line carrying *Dnmt1, Dnmt3a* and *Dnmt3b* inactivating mutations
DSG	Disuccinimidyl glutarate
EGS	Ethylene glycol bis(succinimidyl succinate)
EMSA	Electrophoretic Mobility Shift Assay
ERV	Endogenous retrovirus
ESC	Embryonic Stem Cell
ETS	E-twenty-six transcription factor family
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase per Million
G	Guanine
GABPA	GA Binding Protein Alpha
h	hour(s)
H3ac	Acetylated Histone protein 3
HAT	Histone Acetyltransferase
HDAC	Histone Deacetylase
HSC	Haematopoietic Stem Cell
HT-SELEX	High Throughput-Systematic Evolution of Ligands by Exponential enrichment
IAP	Intracisternal A-particle

ICR	Imprinting Control Region
IP	Immunoprecipitation
KAP1	KRAB-Associated Protein 1
KLF4	Kruppel-Like Factor 4
KRAB	Kruppel associated box
LIF	Leukaemia Inhibitory Factor
LINE	Long Interspersed Nuclear Element
LMR	Lowly Methylated Region
LTR	Long Terminal Repeat
MBD	Methyl-Binding Domain
mESC	mouse Embryonic Stem Cell
min	minute
ml	milliliter
mM	milliMolar
NFY	Nuclear Factor Y
NRF1	Nuclear Respiratory Factor 1
°C	degrees centigrade
PAGE	Polyacrylamide Gel Electrophoresis
PBS	Phosphate Buffered Saline solution
PcG	Polycomb Group protein
PE	Paired-end (sequencing reads)
PGC	Primordial Germ Cells
piRNA	piwi-interacting RNA
Pol2	RNA Polymerase 2
PWM	Position Weight Matrix
qPCR	quantitative-Polymerase Chain Reaction (interchangeable with real-time PCR)
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
RPKM	Reads Per Kilobase per Million
RPM	Reads Per Million
s	seconds
SETDB1	SET Domain Bifurcated 1
SINE	Short Interspersed Nuclear Element
SP1	Specificity Protein 1
T	Thymine
TBST	Tris-Buffered Saline supplemented with Tween-20
TET	Ten-eleven Translocation methylcytosine dioxygenase
TF	Transcription Factor
TKO	Triple Knockout Out (used interchangeably with DNMT.TKO)
TSA	Trichostatin A
TSS	Transcription Start Site
TTS	Transcription Termination Site
µl	microliter
WGBS	Whole Genome Bisulfite Sequencing
WT	Wild Type
XCI	X Chromosome Inactivation
YY1	Ying-Yang 1

Chapter 1:

Introduction

1.1 The biological significance of DNA methylation in vertebrates

In all known cellular life, genetic information is coded for by the sequence of nucleotides, making up the deoxyribonucleic acid (DNA) polymer. In addition to the four prevalent nucleobases guanine (G), cytosine (C), adenine (A) and thymine (T); a number of chemically modified versions of these bases have been identified to occur naturally in the DNA of all organisms (Figure 1.1). These modifications are formed through the action of specific enzymes as opposed to modified bases that occur as a result of DNA damage.

The most frequently modified bases are cytosine and adenine and these have a variety of fundamental biological roles in different domains of life. The genomes of prokaryotes contain substantial amounts of N6-methyladenine (6mA) (Dunn and Smith, 1958) and 5-methylcytosine (5mC) (Doskočil and Šormová, 1965). An additional methylated base, N4-methylcytosine has only been identified in certain bacterial strains (Ehrlich et al., 1985). The best-known function of these modifications in bacteria is their involvement in defence mechanisms against phage DNA based on restriction-modification systems (Arber, 1974). In these systems, bacterial cells express pairs of site-specific methylases and restriction enzymes that target the same DNA sequences. Methylation of the host DNA differentiates it from foreign unmodified DNA and protects it from cleavage by the restriction enzyme.

In eukaryotes, there is no compelling evidence that DNA modifications are involved in such immune mechanisms to restrict viruses in a manner equivalent to that seen in bacteria. Instead, covalent modifications to the DNA along with those to the chromatin-associated proteins have been implicated in providing local information that controls or preserves transcriptional regulatory states without involving changes to the underlying sequence (Suzuki and Bird, 2008). The most abundant DNA modification in eukaryotes is 5-methylcytosine (5mC), first identified in the mammalian genome in 1948 by paper

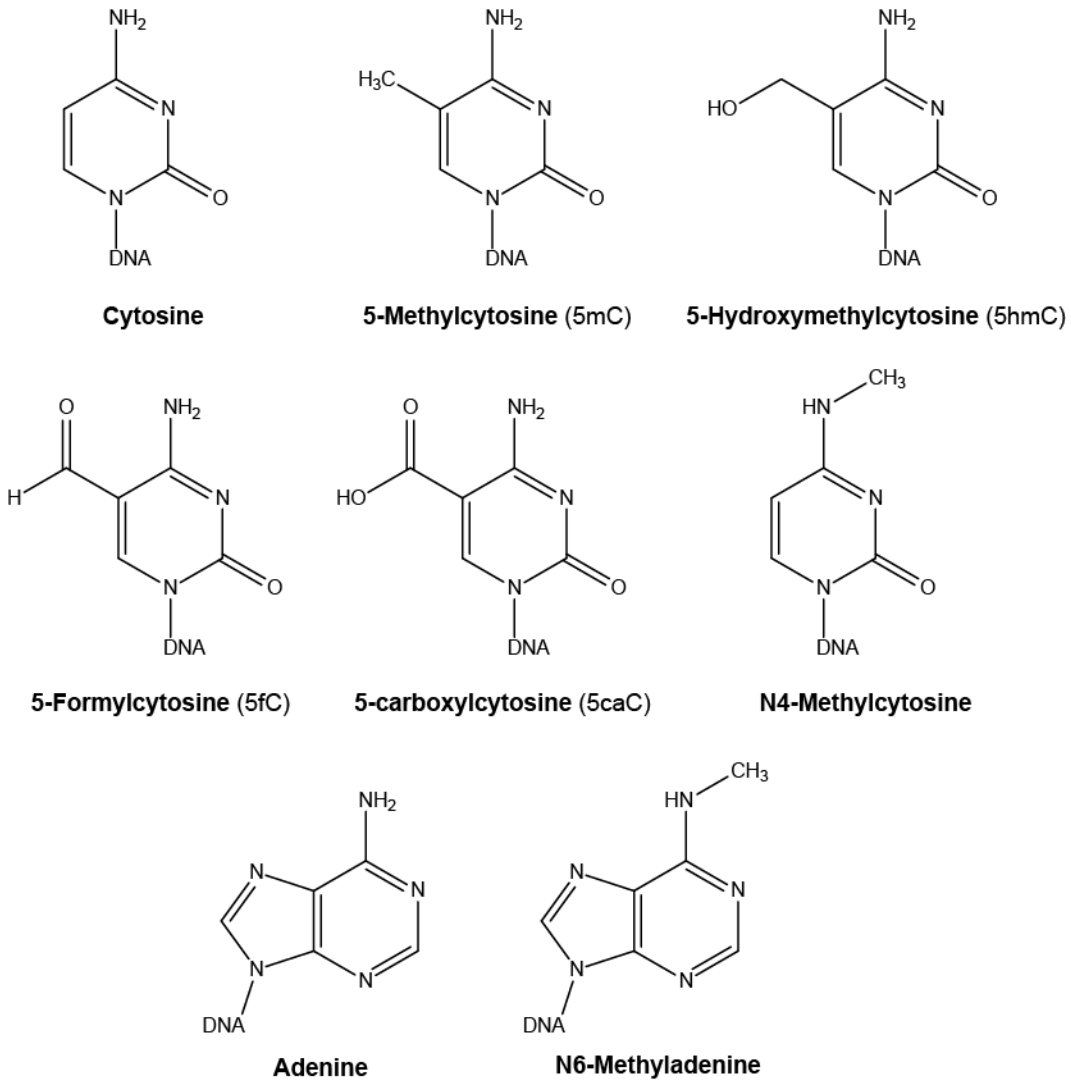


Figure 1.1: Chemical structures of modified bases found in genomic DNA. All these modified bases apart from N4-methylcytosine have been detected in mammalian genomes.

chromatography (Hotchkiss, 1948). Cytosine methylation is deposited by a conserved family of DNA methyltransferases (DNMT) members of which are expressed in metazoan, fungi, flowering plants and some protist taxa (Goll and Bestor, 2005; Huff and Zilberman, 2014). A variety of roles have been attributed to DNA methylation in these different species perhaps most widespread are functions in nucleosome positioning (Huff and Zilberman, 2014) or in the transcriptional repression of parasitic transposable elements (Huff and Zilberman, 2014). Other modified cytosines have been detected in vertebrate genomes including 5-

hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Figure 1.1), although in relatively low levels compared to 5mC and the latter two being particularly rare (Kriaucionis and Heintz, 2009; Ito et al., 2011). Certain unicellular eukaryotes contain large amounts of N6-methyladenine (Ratel et al., 2006). However the overall abundance of this modified base in metazoans, particularly mammalians, is very low (Koziol et al., 2016; Liu et al., 2016a; Wu et al., 2016) and its importance towards biological processes in mammals remains elusive.

In vertebrates, the vast majority of cytosine methylation occurs at the palindromic CpG dinucleotide (Lister et al., 2009; Stadler et al., 2011). Strikingly, the CpG dinucleotide is globally depleted from vertebrate genomes, present on average at ~20% of the expected frequency (Josse et al., 1961; Swartz et al., 1962; Lander et al., 2001). This is attributed to the fact that cytosine and its modified derivatives are prone to spontaneous deamination (Duncan and Miller, 1980). While cytosine deaminates to uracil, which can be recognised by the DNA repair machinery and efficiently removed from DNA, 5-methylcytosine deaminates to thymine, which normally occurs in the genome and is therefore repaired less efficiently (Shen et al., 1994). Notably, the accumulation of CpG to TpG transitions caused by the spontaneous deamination of 5mC during the evolution of vertebrates is likely to be responsible for the genome-wide depletion of CpG dinucleotides. In contrast, the genome of other eukaryotic organisms such as *Caenorhabditis elegans* (nematodes) or *Drosophila melanogaster* (fruit fly) possess very little cytosine methylation and do not display CpG depletion (Bird, 1980). This indicates that DNA methylation has a fundamental biological function that justifies the conservation and activity of DNA methyltransferases despite their adverse mutagenic effects.

This introduction and our research focus on the function and biological relevance of 5mC in mammals. In this context, cytosine methylation has been most widely studied in two species: humans (*Homo sapiens)* and mice (*Mus musculus)*.

1.1.1 DNA methyltransferases establish and maintain DNA methylation

Three conserved DNA methyltransferase enzymes were originally identified as catalysing the methylation of cytosine in mammals, namely DNMT1, DNMT3A and DNMT3B (Bestor and Ingram, 1983; Bestor et al., 1988; Okano et al., 1998). All DNMT proteins share a similar catalytic domain, located at the C-terminus (Goll and Bestor, 2005). The methyltransferase domains of DNMTs contain ten sequence motifs highly conserved with bacterial 5mC methyltransferase proteins and operate through the same catalytic mechanism which involves *S*-adenosyl-L-methionine (SAM) as a methyl-donor (Goll and Bestor, 2005). Last year, DNMT3C was identified as an additional catalytically active member of the DNMT family present in mice but not humans. The novel *Dnmt3c* gene evolved from a duplication of the *Dnmt3b* gene in the last common ancestor of the muroid rodents (Barau et al., 2016).

Two additional DNMT genes are present in all mammalian genomes. DNMT2 has a catalytic domain that bears remarkable structural and sequence similarity to other 5mC DNA methyltransferases however it was shown not to methylate DNA but small RNA molecules instead (Goll et al., 2006). DNMT3L is related to both DNMT3A and DNMT3B although has lost its catalytic activity; nevertheless it plays an important role in the establishment and maintenance of DNA methylation in germ cells (Bourc'his et al., 2001). Whilst the major DNA methyltransferases DNMT1, DNMT3A and DNMT3B are expressed ubiquitously, DNMT3L is restricted to germ cells, the thymus, extra-embryonic tissues and pluripotent embryonic stem cells (Aapola et al., 2000; Hata et al., 2002) while DNMT3C is exclusive to male germ cells (Barau et al., 2016).

Despite the similarity in their C-terminal catalytic region, the N-terminal regions of DNMT1 and DNMT3 proteins differ substantially and play a role in specifying the activities of these

enzymes in the maintenance and establishment of DNA methylation respectively (Goll and Bestor, 2005; Ambrosi et al., 2017).

DNMT1 has the exceptional property of having preferential activity towards hemi-methylated DNA (Yoder et al., 1997). Intrinsic preference for hemi-methylated DNA *in vitro* is explained by the binding of DNMT1's zinc finger-CxxC domain to the unmethylated CpG dinucleotide which interposes and auto-inhibitory stretch of highly acidic amino acids between the DNA and the active site (Song et al., 2011). Hemi-methylated CpG dinucleotides are typically formed immediately after DNA replication at which point only the parental strand is methylated but not the daughter strand. A role for DNMT1 in copying CpG methylation onto the newly synthesised strand has further been supported by its association with the replication fork machinery. Notably DNMT1 interacts with the Proliferating Cell Nuclear Antigen (PCNA) (Suetake et al., 2006) and UHRF1 (also named Np95). The latter binds hemi-methylated CpG dinucleotides through its SRA domain and is essential for recruiting DNMT1 to replication forks and ensuring the maintenance of methylation during replication (Bostick et al., 2007; Sharif et al., 2007).

The capacity for DNMT1 to faithfully maintain CpG methylation patterns across cell divisions is significant as it supports the hypothesis that DNA methylation is an epigenetic mark (Holliday and Pugh, 1975; Riggs, 1975). The term "epigenetic" is used to describe the mechanisms involved in the inheritance of gene expression patterns without altering the underlying DNA sequence (Allis and Jenuwein, 2016). The maintenance of gene expression patterns through epigenetic mechanisms is thought to be involved in the stabilisation of cell fate decisions and the durable silencing of deleterious transcriptional elements. To date, CpG methylation is the only chromatin mark in vertebrates for which a mechanism of stable propagation through mitosis has been characterised. Through the action of DNMT1, DNA

methylation patterns can be maintained unchanged across eighty or more cell divisions and these can confer long-term transcriptional repression (Schübeler et al., 2000).

In line with DNA methylation as a critical epigenetic mark, the generation and characterisation of knockout mouse models has revealed that DNA methylation is essential for normal mammalian development. *Dnmt1*$^{-/-}$ mice fail to develop and die from developmental defects at mid-gestation (Li et al., 1992). Uhrf1$^{-/-}$ embryos show early gestational lethality in a similar manner to *Dnmt1*$^{-/-}$ embryos and both show global reduction in DNA methylation (Sharif et al., 2007). Together, these observations suggest that a failure to maintain DNA methylation patterns is incompatible with mammalian life.

DNMT3A and DNMT3B catalyse the symmetrical methylation of unmethylated DNA and are referred to as *de novo* DNA methyltransferases (Goll and Bestor, 2005). These enzymes can compensate for incomplete methylation maintenance and are responsible for novel methylation of CpG dinucleotides after transient exposure to the DNA methyltransferase inhibitor 5-aza-2'-deoxycytidine in the absence of DNMT1 (Liang et al., 2002). These two enzymes have mostly overlapping but not completely redundant functions. In particular, DNMT3A is essential for the establishment of methylation patterns in mouse gametes along with DNMT3L (Bourc'his et al., 2001; Kaneda et al., 2004). DNMT3A is also responsible for the formation of non-CpG methylation, reported to be present at low levels in embryonic stem cells and neurons, and the biological role of which has not been elucidated (Ramsahoye et al., 2000; Stadler et al., 2011; Lister et al., 2009, 2013). On the other hand, DNMT3B has been reported to be important for the establishment of DNA methylation during pre-implantation embryogenesis at which point it is highly expressed (Borgel et al., 2010; Auclair et al., 2014).

The targeting of DNMT3 enzymes to certain genomic regions is thought to rely on interactions in their N-terminal regions. For example the ATRX-DNMT3-DMT3L (ADD) domain mediates the association between the DNMT3L/DNMT3A complex and histone tails unmethylated on

lysine 4 (H3K4) (Ooi et al., 2007; Otani et al., 2009). Additionally the PWWP domain of DNMT3A and DNMT3B recognises H3K36me3 and is believed to contribute to the preferential localisation of DNMT3B at gene bodies (Baubec et al., 2015; Neri et al., 2017). Nevertheless the principles that guide the targeting of DNA methyltransferases to certain genomic loci have not been fully uncovered.

Knockout of the *Dnmt3a* or *Dnmt3b* genes result in severe phenotypes in mice: *Dnmt3b*$^{-/-}$ embryos are not viable, dying from developmental defects at mid-gestation while *Dnmt3a*$^{-/-}$ mice develop normally until birth but pups are runted and die postnately within 4 weeks (Okano et al., 1999). This indicates that these two enzymes possess non-redundant roles and that their ability to deposit novel DNA methylation marks is indispensible for normal development and post-natal biological processes.

1.1.2 The importance of DNA methylation in development and mammalian physiological processes.

Decades of study have provided evidence that DNA methylation plays major roles in parental imprinting, chromosome X inactivation, stabilising cell fate decisions and repressing repetitive elements; failure to contribute to these processes leads to embryonic lethality. A brief description of these processes is given below although the involvement of DNA methylation in the repression of repetitive elements will be expanded on in section 1.4.2.

Parental imprinting is an epigenetic heritable mechanism which results in mono-allelic expression of a subset of genes based on the parent of origin. Imprinted genes are commonly found in clusters where their expression is controlled by the epigenetic status of imprinted control loci (ICR). DNA methylation is the primary epigenetic mark that ensures the inter-generational transmission of imprints: it is deposited in gametes in a sex-specific manner and mono-allelic methylation is maintained after fertilisation (Barlow and Bartolomei, 2014).

An example of a differentially methylated ICR is the one regulating the expression of the *Igf2* and *H19* genes. Methylation of a CCCTC-binding Factor (CTCF) recognition site on the paternal allele, established during spermatogenesis, impedes CTCF binding and its function as an insulator. This results in the expression of *Igf2* through *cis*-acting mechanisms from the paternal allele while *H19* is transcribed from the maternal allele uniquely (Figure 1.2) (Bell and Felsenfeld, 2000; Hark et al., 2000).

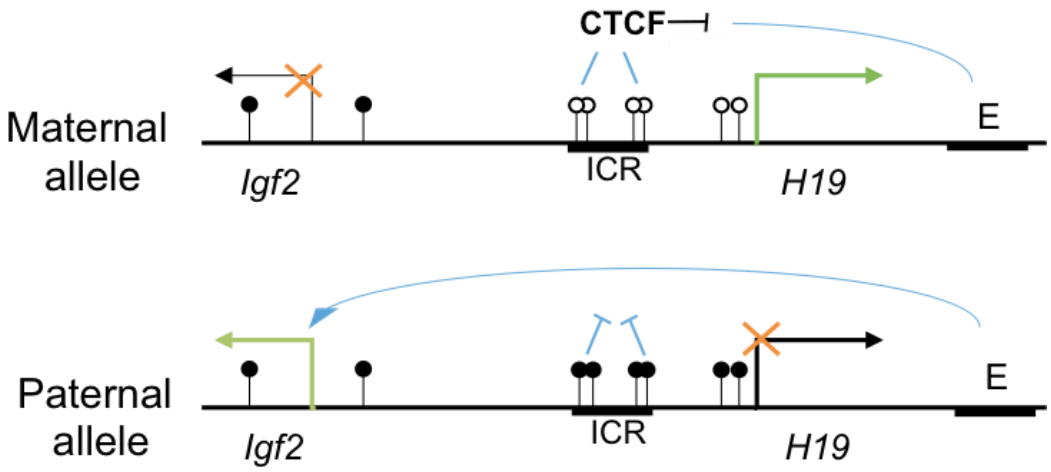


Figure 1.2: Allele-specific expression of the *Igf2* and *H19* genes controlled by methylation of the Imprinted Control Region (ICR). On the maternally-inherited allele, the ICR is unmethylated (clear circles on sticks). This allows the binding of CTCF to sequences within the ICR which blocks the activation of the *Igf2* gene by the *H19* enhancer (E). On the paternally-inherited allele, the ICR is methylated (black circles on sticks), impeding the binding of CTCF and suppressing its insulator function. This results in activation of the *Igf2* gene by the *H19* enhancer (E) while the *H19* gene is silenced. This model was drawn based on mechanisms determined by Bell and Felsenfeld and Hark *et al.* (Bell and Felsenfeld, 2000; Hark et al., 2000).

DNMT3A along with its cofactor DNMT3L are essential for establishing the methylation imprints during male or female gametogenesis (Kaneda et al., 2004). While oogenesis is normal in *Dnmt3l*$^{-/-}$ mothers, their heterozygous progeny do not develop past 9.5 days post-coitum (Bourc'his et al., 2001). Similar developmental defects were observed in progeny from females where *Dnmt3a* had been conditionally disrupted in germ cells (Kaneda et al., 2004) highlighting the crucial role for DNA methylation and imprinting in development. In humans,

loss of imprinting at particular loci contributes to the development of a number of diseases including, Angelman Syndrome, Beckwith-Wiedemann Syndrome and Prader-Willi Syndrome (Lalande, 1996).

X-chromosome inactivation (XCI) is a process that occurs in female mammalian development to transcriptionally silence one of the pair of X chromosomes, thereby achieving dosage equivalency with males that have a single X chromosome (Lyon, 1961). Chromosome X inactivation relies on the expression of the non-coding Xist RNA and the action of Polycomb Group (PcG) protein complexes that mediate the spreading of heterochromatin across the inactive chromosome (Brockdorff, 2011). DNA methylation is not required for the initiating phase of XCI (Sado et al., 2004) however DNMT1 is essential for the maintenance of methylation patterns at the *Xist* promoter and the preservation of XCI in somatic cells (Beard et al., 1995; Panning and Jaenisch, 1996). Furthermore, gene promoters are frequently methylated on the inactive but not the active X chromosome with implications on allele specific transcription (Cotton et al., 2015; Sharp et al., 2011).

In itself, DNA methylation is not critical for cell proliferation as proven by *Dnmt1*$^{-/-}$ *Dnmt3a*$^{-/-}$ *Dnmt3b*$^{-/-}$ triple knockout (DNMT.TKO) mouse embryonic stem cells (ESCs) that are viable despite a complete deficiency in DNA methyltransferase activity (Tsumura et al., 2006). ESC cell lines are derived directly from the inner cell mass of blastocyst stage embryos. ESCs retain their pluripotent attributes without differentiating when cultured in the presence of leukaemia inhibitory factor (LIF, supplied as a recombinant protein in the medium or expressed by co-cultured feeder cells). Importantly wild type ESCs can fully participate to fetal development when reintroduced into the embryo during implantation and, *in vitro,* can be induced to differentiate into the three primary germ layers and further into a wide-range of defined cell types (Smith, 2001). ESC cells are widely used as a model system in molecular biology and epigenetics as they display a normal karyotype, are easy to culture to generate

vast amounts of material for biochemical studies, and genetic manipulation is relatively straightforward.

DNMT.TKO mouse ES cells maintain stem cell properties but are unable to differentiate (Tsumura et al., 2006). Early passage *Dnmt3a*$^{-/-}$ *Dnmt3b*$^{-/-}$ mouse ES cells which have > 10% methylated CpGs are able to form teratomas when injected into nude mice (Chen et al., 2003) however this was not the case for cells lacking DNMT1 and carrying similar global levels of DNA methylation (Tucker et al., 1996; Jackson et al., 2004). Therefore, the maintenance of DNA methylation patterns, rather than the lack of *de novo* methylation is believed to limit the differentiation potential of pluripotent stem cells.

However DNMT.TKO mouse ES cells are the only mammalian cell line that are known to grow in the complete absence of DNA methylation, an ability that does not extend to human embryonic stem cells (Liao et al., 2015). Hypomethylated mouse embryonic fibroblasts display chromosomal instability and DNMT1 abrogation leads to p53-dependent apoptosis (Jackson-Grusby et al., 2001; Dodge et al., 2005) suggesting that lack of DNA methylation maintenance is incompatible with the growth of non-embryonic cell lines. In contrast, human colorectal cancer HCT116 cells deficient for *DNMT1* and *DNMT3B* are viable despite a 95% reduction in genomic DNA methylation (Rhee et al., 2002). That 5-aza-deoxycytosine treatment is toxic in these cells has been taken to indicate that the low levels of DNA methylation deposited by DNMT3A are sufficient but essential to maintain cell viability, however, an alternative explanation is that DNA damage caused as a side-effect of 5-aza-deoxycytosine treatment is responsible for cell toxicity.

Alterations in DNA methylation have also been shown to disrupt normal haematopoietic stem cell (HSC) function and differentiation. DNMT1 is essential for the self-renewal and differentiation capability of HSCs *in vivo,* and decision to differentiate along myeloerythroid rather than lymphoid lineages is sensitive to DNA methylation levels (Bröske et al., 2009;

Trowbridge et al., 2009). Furthermore, genetic ablation of *Dnmt3a* in HSCs causes an increase in HSC self-renewal whilst limiting the differential potential. This is associated with the inability to repress genes normally restricted to multipotent cells during cell type specification (Challen et al., 2012). These findings are relevant to human biology as 15-25% cases of acute myeloid leukemia and roughly 10% patients with myelodysplastic syndrome present with somatic mutations in *DNMT3A* (Ley et al., 2010; Walter et al., 2011). Recent data indicates that focal regions of abnormal hypomethylation appear in pre-leukaemic cells carrying inactivating $DNMT3A^{R882}$ mutations although their influence on leukemogenesis remains unknown (Spencer et al., 2017).

In humans, a variety of disorders have been associated with pathogenic germline mutations in DNMTs. Homozygous loss-of-function mutations in *DNMT3B* cause Immunodeficiency, Centromere instability and Facial anomalies (ICF) syndrome type 1, a disorder that is usually fatal prior to adulthood (Xu et al., 1999). Satellite DNA is almost completely unmethylated in patients with ICF syndrome type 1, however the methylation abnormalities responsible for the pathogenesis of this disease have not been confidently specified (Edwards et al., 2017). A heterogeneous group of adult-onset neurological disorders, collectively known as Autosomal Dominant DNMT1 Complex disorder are caused by mutations in a small domain of *DNMT1* which mediates its interaction with UHRF1 (Baets et al., 2015). The mechanism of pathogenesis has been attributed to the formation of DNMT1 cytoplasmic aggregates rather than aberrant DNA methylation patterning (Baets et al., 2015). Patients with Tatton-Brown-Rahman syndrome have heterozygous mutations in *DNMT3A* and present with overgrowth and early onset intellectual disabilities. Although no DNA methylation data has been obtained at this point, there are similarities to the imprinting disorder Beckwith-Wiedemann syndrome (Tatton-Brown et al., 2014). Although these disorders have non-overlapping phenotypes they are indicative that DNA methylation plays a role in many biological processes, some of which are not fully understood.

1.2 The genomic distribution of CpG methylation

In the previous section I have described how DNA methylation patterns are established and maintained by the cooperative activity of DNA methyltransferases and that these play crucial roles in mammalian development and biological processes. To understand the role of DNA methylation in mammalian biology it is important to know how CpG methylation is distributed across the genome. The following sections will introduce general aspects of DNA methylation patterns in mammalian genomes.

1.2.1 CpG Islands: a key feature of the DNA methylation landscape

Early studies made use of methylation-sensitive restriction enzymes such as HpaII (recognition pattern CCGG) to study the genome-wide distribution of DNA methylation (Cooper et al., 1983; Bird et al., 1985). Analysis of the size distribution of DNA fragments obtained after HpaII digestion of genomic DNA revealed the existence of a discrete fraction of small DNA fragments (Cooper et al., 1983). These were indicative of the presence of regions in the genome that contain a high density of non-methylated CpG sites (therefore being frequently cut by HpaII). Such regions were found to cover 1-2% of the mouse genome and were identified consistently across different tissues types (Bird et al., 1985). These CpG clusters were found to contain a significantly higher occurrence of CpG dinucleotides, close to the expected frequency, as opposed to the bulk genome and were therefore named CpG islands (CGI) (Bird, 1986). CGIs were defined as regions with a GC content of at least 50% and a CpG frequency (observed/expected) of over 0.6. The higher retention of CpGs in CGIs is attributed to the fact that these are rarely methylated, including in germ cells and therefore less prone to undergo spontaneous deamination (Cooper et al., 1983; Bird et al., 1985; Gardiner-Garden and Frommer, 1987).

The mapping of CGIs in the genome has revealed that these are on average 1kb in length and associate with 60-70% transcription start sites suggesting an important regulatory role for these regions (Gardiner-Garden and Frommer, 1987; Lander et al., 2001). In fact, CGIs localise to the promoter of almost all constitutively active genes (housekeeping genes) and 40% of genes with a tissue-restricted expression pattern (Illingworth and Bird, 2009). In line with this, they are enriched for the presence of cognate sequences bound by transcription factors such as SP1, NRF1 or members of the ETS or E2F families (Rozenberg et al., 2008; Landolin et al., 2010). CpG islands are a distinctive feature in the landscape of mammalian genomes that demarcate a large fraction of gene promoters. They represent sites of highly accessible DNA and are generally characterised by the presence of histone H3 methylated on lysine 4 (H3K4me3), a mark associated with active transcription (Guenther et al., 2007; Mikkelsen et al., 2007). CpG islands at the promoters of non-transcribed genes typically remain unmethylated but are marked by repressive chromatin modifications particularly H3K27me3, deposited through the action of the Polycomb-repressive complexes PRC1 and PRC2 (Mikkelsen et al., 2007; Ku et al., 2008).

CpG islands methylation is infrequent: in somatic cells, it occurs at the promoters of a subset of genes whose expression is restricted to the germ-line as well as on the inactive X chromosome in females (Maatouk et al., 2006; Mohn et al., 2008; Borgel et al., 2010; Sharp et al., 2011). Additionally, some CpG islands associated with tumour suppressor genes are aberrantly methylated in cancer cells (Toyota et al., 1999; Rhee et al., 2002; Robert et al., 2003). Transcription initiation from CpG islands is regulated instead by the binding of activating transcription factors. The opposing actions of PRC complexes on one hand and the SET1 complex (which methylates H3K4 residues) or related proteins of the Trithorax group on the other, act to reinforce repressive and active transcriptional states respectively (Deaton and Bird, 2011; Klose et al., 2013).

1.2.2 Genome-wide mapping of DNA methylation

The current gold standard for the mapping of DNA methylation is bisulfite sequencing. Based on the different kinetics with which sodium bisulfite reacts with methylated and unmethylated cytosines, this method allows the accurate measurement of methylation frequency at a given cytosine (Figure 1.3) (Frommer et al., 1992). Bisulfite sequencing coupled with high-throughput sequencing technologies now enables the genome-wide mapping of DNA methylation (often referred to as the methylome) at single base resolution (Lister et al., 2009). The first use of Whole Genome Bisulfite Sequencing (WGBS) in a mammalian model was to interrogate the methylome of human and mouse embryonic stem cells (Lister et al., 2009; Stadler et al., 2011). WGBS has now been applied to determine DNA methylation patterns for a wide range of cell types and tissues.

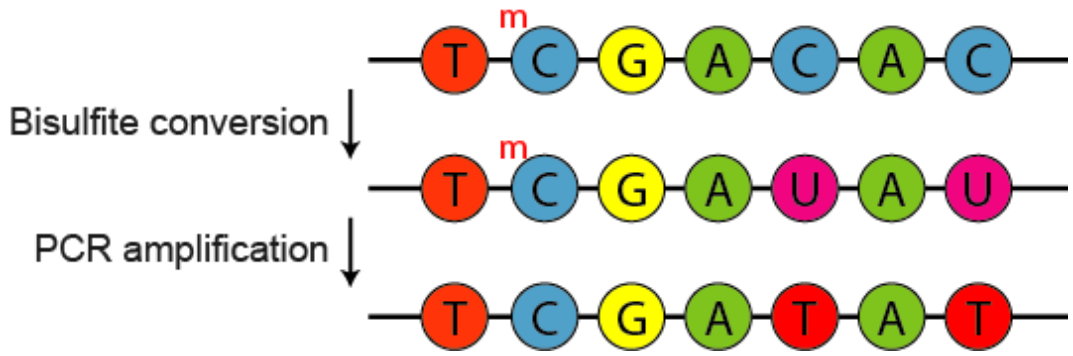


Figure 1.3: Schematic of bisulfite sequencing, developed as a method to allow positive identification of methylated cytosines residues within DNA sequences (Frommer et al., 1992). Bisulfite sequencing relies on the treatment of DNA with sodium bisulfite which converts cytosine (C) but not 5-methylcytosine (mC) into uracil. During the subsequent PCR amplification of the resulting DNA, thymine residues are incorporated instead of uracil while unconverted 5-methylcytosines are amplified as cytosines. Only 5-methylcytosines positions, which were not converted, will be read as cytosines by sequencing.

DNA methylation mapping in ESCs and other cell lines confirmed previous conceptions regarding the distribution of DNA methylation. The methylation frequency of CpG displays a bimodal distribution where 60-80% of genomic CpGs are highly methylated while a smaller

group of cytosines remain unmethylated particularly those residing in CpG islands and near transcription start sites (Lister et al., 2009; Stadler et al., 2011). In addition, regions with lower levels of DNA methylation (LMRs) were identified and these overlapped to a large extent with distal regulatory elements. Such regions are CpG poor, sensitive to DNaseI digestion and contain binding sites for numerous transcription factors such as CTCF, REST, KLF4 or GABPA (Stadler et al., 2011). The relationship between transcription factor binding and DNA methylation will be discussed in more detail in a following section.

WGBS is typically carried out on a bulk population of cells in which each allele carries a methylated or unmethylated cytosine at a particular position. Bisulfite sequencing provides a value between 0 and 1 (or 0-100) that represents the fraction (or percentage) of cytidines that are methylated in the population (Frommer et al., 1992; Cokus et al., 2008). Recent efforts are providing the first methylomes at the level of single cells although the practical and bioinformatic technologies involved are still in their infancy and not widely used (Smallwood et al., 2014; Clark et al., 2016).

1.3 The dynamics of DNA methylation

While most cell types and cell lines display an overall similar distribution of DNA methylation (i.e. elevated methylated CpG levels throughout the genome-wide with the exception of CpG islands) (Hon et al., 2013; Ziller et al., 2013), DNA methylation patterns are not static during development. The most pronounced changes in DNA methylation occur during primordial germ cell development and during the pre-implantation stage after fertilisation at which points levels of DNA methylation vary across the entire genome (Iurlaro et al., 2017). In addition, locus-specific changes in DNA methylation have been observed to accompany a plethora of biological processes such as those involved in differentiation, reprogramming,

phenotype switching and disease progression. In this section we will first introduce the mechanisms that lead to the loss of CpG methylation and then discuss the dynamic changes in DNA methylation patterns.

1.3.1 Active and passive DNA demethylation mechanisms

Removal of DNA methylation can be achieved through DNA replication in the absence of DNMT1 activity. As the daughter strand is synthesised in an unmethylated state during DNA replication, CpG methylation is diluted and can be completely absent from a locus after two cycles if DNMT1 is lacking, mis-targeted or inhibited. This process is typically called “passive demethylation” (Chen and Riggs, 2011). Mechanisms through which the removal of DNA methylation is facilitated by enzymatic activity have been characterised during the last decade. A breakthrough discovery was the detection of 5-hydroxymethylcytosine (5hmC) in mammalian genomes (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). Three mammalian Ten-Eleven Translocation (TET1-3) enzymes were identified as being responsible for the oxidation of 5mC to 5hmC and further to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Iyer et al., 2009; Tahiliani et al., 2009; Ito et al., 2011). Based on the *in vitro* preferential activity of DNMT1 on hemi-methylated rather than hemi-hydroxymethylated CpG substrates, 5mC oxidation is believed to facilitate site-specific passive demethylation in the presence of the maintenance machinery (Hashimoto et al., 2012). Additionally, replication-independent demethylation can be achieved through the efficient excision of 5fC and 5caC by Thymidine DNA Glycosylase (TDG) and the reintroduction of an unmodified cytosine by base-excision repair (termed as “active demethylation”) (Maiti and Drohat, 2011; Zhang et al., 2012).

The contribution of active or passive DNA demethylation pathways in shaping DNA methylation patterns is an area of great interest and is extensively studied (Iurlaro et al., 2017). In mouse ES cells, 5hmC and 5fC are preferentially found at CpG islands and active gene

promoters contrasting with the distribution of 5mC (Ficz et al., 2011; Raiber et al., 2012). Suggestive for a role for active demethylation in preserving the unmethylated state of CGIs, TET1 was found to be enriched at these regions, likely guided by the interaction of its CXXC domain with unmethylated CpGs (Williams et al., 2011). Furthermore 5hmC is enriched at distal regulatory elements and LMRs (Stadler et al., 2011). A functional role for TET proteins in 5hmC deposition and the active demethylation of enhancers has been demonstrated: deletion of the three *Tet* genes in mouse ESCs leads to the hypermethylation of enhancer elements and mis-regulation of their associated genes (Lu et al., 2014). These findings highlight an interesting relationship between DNA demethylation, transcription factor binding and enhancer activity.

Similarly to *DNMT3A*, *TET2* mutations are frequent in myeloid malignancies (Delhommeau et al., 2009). Although these two enzymes have opposite biochemical effects on 5-methylcytosine, the fact that disruption of either enzyme results in a similar clinical outcome is intriguing. A functional study of DNMT3A and TET2 in mouse haematopoetic stem cells has painted a complicated picture of cooperation and competition between these two enzymes. Both proteins have opposing roles in the maintenance of large unmethylated regions termed "canyons" but cooperate for the correct expression of lineage-specific genes (Zhang et al., 2016). Nevertheless these studies highlight the importance of precise DNA methylation dynamics in cell fate decisions.

1.3.2 DNA methylation dynamics in early embryogenesis and germ cell development

During mammalian development the genome undergoes two waves of near global erasure of DNA demethylation followed by remethylation (Monk et al., 1987). The first wave of demethylation occurs in primordial germ cells (PGCs) during their proliferation and migration to the genital ridge. PGCs are the precursors to male and female gametes and genome-wide

epigenetic reprogramming is important for the erasure of parental imprints, X-chromosome inactivation and the activation of meiotic transcriptional programs. This has led to the model that global hypomethylation is associated with the loss of cellular memory and the resetting of developmental potency (Iurlaro et al., 2017).

Demethylation in PGCs is uneven both in terms of timing and genomic context. CpG rich sequences with regulatory functions, such as those found at imprinted differentially methylated regions (DMRs), at CpG islands on the inactive X chromosome or near genes associated with meiosis or gamete generation are demethylated at a later stage of PGC development (Seisenberger et al., 2012). Interestingly, several sequences escape this wave of demethylation in mice including a small number of non-repetitive CpG islands and a specific family of retrotransposons: the intracisternal A particles (IAPs) (Guibert et al., 2012; Seisenberger et al., 2012). In humans, another class of evolutionarily young and active transposons, L1Pa and SINE-VNTR-Alus (SVAs) are known to resist global hypomethylation in PGC development (Tang et al., 2015). Sex-specific de novo methylation occurs during the subsequent development of male and female gametes. Oocytes and sperm cells display remarkably different methylation patterns reflective of the difference in the timing and mechanisms of de novo methylation (Messerschmidt et al., 2014). Despite their differences, these mechanisms ensure the establishment of sex-specific imprints (Bourc'his et al., 2001; Hata et al., 2002; Kaneda et al., 2004; Kato et al., 2007).

After fertilisation, the differential methylation patterns established in gametes are largely removed during a second wave of genome-wide demethylation. This occurs with different timing in the female and male pro-nuclei but ultimately results in the global hypomethylation of naïve pluripotent cells that make up the inner cell mass of the blastocyst (Leitch et al., 2013; Smith et al., 2012). Compared to the bulk of the genome, methylation at imprinting control regions and IAP sequences is maintained post-fertilisation by the action of DNMT1 (Gaudet et

al., 2003; Lane et al., 2003; Hirasawa et al., 2008) and subsequently preserved through embryogenesis and into adulthood. After implantation, the action of DNA methyltransferases restores high levels of DNA methylation throughout the genome and at a small number of CpG islands (Borgel et al., 2010; Auclair et al., 2014).

Pluripotent mouse embryonic stem cells (mESCs) derived from the inner cell mass are traditionally grown in medium containing fetal bovine serum and LIF. Under these culture conditions, they display high levels of DNA methylation similar to that seen in the post-implantation epiblast and somatic tissues (Stadler et al., 2011; Leitch et al., 2013). However when propagated in serum-free conditions and in the presence of two small molecule inhibitors of the FGF/ERK and GSK3/WNT signalling pathways (2i conditions), mESCs are globally hypomethylated (Ying et al., 2008; Leitch et al., 2013). Whilst 4% of cytosines (80% of CpGs) are methylated in serum-grown mESCs under 1% remain methylated after transition to 2i growth condition with most methylated CpGs clustering at IAP elements and ICRs (Habibi et al., 2013). Under 2i conditions, mESCs are believed to better recapitulate the hypometylated “naïve” state of the pre-implantation ICM in terms of gene expression and developmental potency (Habibi et al., 2013). The transition between serum and 2i-grown mESCs has been widely used in recent years to study the epigenetic reprogramming that accompanies global DNA methylation dynamics at the molecular level (Iurlaro et al., 2017).

1.3.3 DNA methylation dynamics in somatic tissues

The DNA methylation landscape established in early embryogenesis is relatively stable in somatic tissues and CpG methylation is largely maintained through cell divisions. Increasingly comprehensive and high-resolution mapping of DNA methylation in a variety of tissues has shown that only a small fraction of CpG sites switch their methylation status (Hon et al., 2013; Ziller et al., 2013). In mice, approximately 6.7% of the genome displays tissue variable DNA

methylation. Tissue-specific differential methylation regions (ts-DMR) are preferentially located distal from transcription start sites, do not include CpG-rich CGIs and overlap with distal regulatory elements (Hon et al., 2013). Similarly to the low methylated regions identified in mES cells and neural progenitor cells (Stadler et al., 2011), ts-DMRs overlap with transcription factors binding sites (Hon et al., 2013). Importantly, tissue-specific hypomethylation at enhancers is associated with their lineage-specific activation (Hon et al., 2013; Ziller et al., 2013). For example, hematopoietic transcription factors SPI1/PU.1 and RUNX1 motifs were specifically enriched in ts-DMRs found in blood-producing organs while the endodermal FOXA1 was found to be enriched at regions hypomethylated specifically in endodermal tissues (Hon et al., 2013).

Only a small fraction of promoter-associated autosomal CpG islands are methylated in somatic cells (Shen et al., 2007). The promoters of pluripotency genes such as *Oct4/Pou5f1* and *Nanog* are methylated in somatic cells compared to mESCs, suggesting that methylation, along with other chromatin modifications, could serve as a barrier to dedifferentiation (Farthing et al., 2008). In agreement with this hypothesis, successful reprogramming of somatic cells into induced pluripotent cells is characterised by the hypomethylation of pluripotency promoters and is facilitated by genetic or pharmacological interference with the DNA methylation machinery (Hill et al., 2014). On the other hand, the ability of *Dnmt3a*$^{-/-}$ *Dnmt3b*$^{-/-}$ mES cells to differentiate terminally puts into question the necessity for de novo methylation in the exit from pluripotency and preservation of differentiated states (Jackson et al., 2004).

A growing number of studies are highlighting the importance of DNA methylation in a vast range of processes associated with disease and aging. Most significantly, a great effort has been invested in describing and understanding the DNA methylation changes that accompany cancer development. Over 8000 cancer methylomes have been generated largely through the

use of microarray based technologies to capture CpG methylation levels throughout the genome (Stirzaker et al., 2014). Common features that have emerged through the study of DNA methylation abnormalities in cancer include the genome-wide hypomethylation of cancer cells compared to normal cells (Portela and Esteller, 2010). Loss of methylation at centromeric sequences and intergenic regions containing repetitive elements has been linked with chromosomal instability and activation of endogenous mobile DNA elements (Eden et al., 2003; Gaudet et al., 2003; Daskalos et al., 2009). Furthermore a variety of cancers from different origins have been described as having a CpG island methylator phenotype (CIMP) (Hughes et al., 2013), first identified in colorectal cancer (Toyota et al., 1999) and defined by frequent and aberrant hypermethylation of promoter-associated CpG islands. This phenotype can be associated with specific clinical or genetic features such as somatic mutations in *IDH1/2* in glioblastoma and leukaemias (Figueroa et al., 2010; Turcan et al., 2012) or *BRAF* mutations in colorectal cancer (Hinoue et al., 2012). Nevertheless a causative role of aberrant CpG island DNA methylation in transformation is often hard to establish. A recent study has suggested that hypermethylation is likely to be a consequence of hyperdysplasia in haematopoietic stem cells (Spencer et al., 2017). In colorectal cancer cells that carry activating *BRAF* mutations, increased CpG island methylation at gene promoters has been demonstrated to result from the increased binding of the transcriptional repressor MAFG and DNMT3B recruitment following aberrant activation of the *BRAF/MEK/ERK* signalling pathway (Fang et al., 2014).

1.4 The role of DNA methylation in transcriptional regulation

1.4.1 DNA methylation as a repressive mark

DNA methylation has widely been invoked to play a role in transcriptional repression. Tissue-specific changes in DNA methylation were first proposed to participate in controlling gene expression (Waalwijk and Flavell, 1978) although the evidence was lacking at the time. In the following decades correlations between site-specific methylation and expression were reported while several lines of evidence suggested a causative role for DNA methylation in gene silencing at individual loci. For example, *in vitro* methylation of constructs known to be sensitive to DNaseI (a readout for chromatin accessibility) when inserted into cells, impeded the formation of accessible chromatin conformations (Keshet et al., 1986). Several studies determined that transcriptional inhibition of human γ-globin or β-globin transgenes could be achieved by the methylation of specific CpG elements prior to their transfection into cells (Busslinger et al., 1983; Yisraeli et al., 1988). *In vivo,* artificial modulation of the methylation status at CpG elements upstream of a reporter gene could lead to its differential expression (Siegfried et al., 1999).

Genes with high levels of DNA methylation at their promoter can be re-activated by the DNMT inhibitor 5-aza-deoxycytidine (5-azadC) or by mutation of the DNMT genes suggesting that promoter methylation is necessary for the stable repression of genes in certain contexts. This is the case for tumour suppressor genes aberrantly methylated in cancer cells such as *CDKN2A* (Rhee et al., 2002; Robert et al., 2003) or at the CpG island and pathogenic CGG repeat expansion at the *FMR1* gene locus (Chiurazzi et al., 1998). Despite these observations, emerging evidence indicates that a global role for promoter DNA methylation on gene expression is unlikely. The promoters of most genes are associated with CpG islands and remain largely unmethylated in somatic and embryonic tissues whether they are transcribed

or not (Weber et al., 2007). Along these lines, complete loss of DNA methylation in mESCs does not lead to global deregulation in gene transcription (Fouse et al., 2008; Karimi et al., 2011).

Nevertheless a subset of genes has been described whose expression is directly influenced by developmental DNA methylation changes occurring *in cis.* These exceptions include genes on the X chromosome in female somatic cells and imprinted genes controlled by differentially methylated regions. DNA methylation is essential for the maintenance of XCI in female somatic cells and *Dnmt1* disruption or 5-azadC treatment leads to defective inactivation and the re-expression of X-linked genes or the *Xist* transcript (Beard et al., 1995; Hansen et al., 1996). Genes escaping inactivation show low levels of DNA methylation compared to repressed ones (Cotton et al., 2015; Weber et al., 2007) however the relationship between gene activation and DNA methylation of *cis*-regulatory elements on the X chromosome remains mostly correlative. More compelling evidence exists for the direct role of DNA methylation in the control of imprinted genes. These are upregulated when DNMT genes are mutated (Jackson-Grusby et al., 2001; Rhee et al., 2002) and mechanisms have been determined for the direct readout of DNA methylation and its effect on gene expression (Bell and Felsenfeld, 2000; Quenneville et al., 2011). Expression of the imprinted gene *Snrpn* can even be used as an indicator of the methylation state of its promoter (Stelzer et al., 2015). This paints the picture of a role for DNA methylation in the heritable transcriptional repression of genes that require mono-allelic expression to achieve appropriate protein levels.

Another group of genes has been identified as requiring DNA methylation for appropriate tissue-specific expression. Post-migratory germ-cell specific genes such as *Dazl*, *Mvh*, and *Scp3* are methylated in somatic cells and are expressed after *Dnmt1* knockout (Maatouk et al., 2006). The DNMT3B-dependent methylation at the CpG island promoters of germline genes is established at the implantation stage and is the primary mechanism for their silencing

in adult tissues (Auclair et al., 2014; Borgel et al., 2010). Additionally the methylation of a cis-regulatory region of the X-linked *Rhox* cluster is established in epiblast cells and limits their expression to extra-embryonic trophectoderm cells. The cells giving rise to trophectoderm are specified in the blastocyst before de novo methylation of the genome (Borgel et al., 2010; Oda et al., 2006).

A role for intragenic methylation in gene regulation has also been studied with a special interest in the role of DNMT3B, which is directed to these regions through its interaction with H3K36 methylated histone tails (Baubec et al., 2015). Such intragenic DNA methylation appears to enhance genic transcription (Yang et al., 2014) potentially through the repression of spurious transcription within the gene body. Indeed, novel transcription start sites have been identified in intragenic regions in the absence of DNMT3B mouse ES cells (Neri et al., 2017).

1.4.2 The role of DNA methylation in the repression of repetitive elements

Yet another established role for DNA methylation is its importance in the heritable silencing of repetitive elements. Approximately 40% of the mouse genome is comprised of transposable elements. These are mostly retrotransposons, derived from ancient germline viral integrations that duplicate through the formation of an RNA intermediate (Stocking and Kozak, 2008). Transposable elements (TE) fall into two broad categories: Long Terminal Repeat (LTR) sequences which include the Endogenous Retrovirus (ERV) classes ERV1, ERVK and ERVL; and the non-LTR elements comprising Long and Short Interspersed Elements (LINE and SINEs). Although the majority of TEs have accumulated inactivating mutations, a few of the evolutionarily young LINEs and ERVs are capable of encoding proteins necessary for their mobilisation (Stocking and Kozak, 2008). Activation of transposition-capable ERVs is associated with germline insertional mutagenesis (Maksakova et al., 2006) and genomic

instability. To counteract the mutational capacity of transposable elements, both epigenetic and cytoplasm-based mechanisms have evolved to limit their mobility (Goodier, 2016).

The importance of DNA methylation in the repression of transposable elements is dependent on the developmental stage and the type of mobile sequences. DNA methylation is the main transcriptional repressor for young retrotransposons such as LTR-ERVKs (that include IAPs) and LINE-1A elements in somatic cells and in the germline. IAPs are highly transcribed in embryos treated with 5-azadC or carrying *Dnmt1* inactivating mutations (Davis et al., 1989; Walsh et al., 1998). Both IAP elements as well as the 5'UTR sequences of certain LINE-1 subtypes are heavily methylated in all tissues and resist the waves of demethylation that occur during germ cell development and post-fertilisation (Gaudet et al., 2004; Seisenberger et al., 2012; Smith et al., 2012; Hill et al., 2018). Furthermore, unique pathways have evolved to ensure high levels of methylation at IAPs and LINE-1A elements in male germ cells. Specifically DNMT3A, DNMT3L guided by piwi-interacting RNAs as well as DNMT3C cooperate to maintain silencing (Bourc'his and Bestor, 2004; Aravin et al., 2008; Kuramochi-Miyagawa et al., 2008; Barau et al., 2016).

In mouse embryonic stem cells, the expression of IAP elements is significantly elevated after the removal of DNA methylation (Karimi et al., 2011) however other mechanisms are known to be involved in keeping them repressed. Particularly, methylation at the K9 position of histone H3 tails (H3K9me3), deposited by SETDB1, directly participates in repressing ERVK transposons (Karimi et al., 2011; Matsui et al., 2010; Rowe et al., 2013) and has been implicated in the maintenance of DNA methylation at these loci (Leung et al., 2014). In fact, IAP reactivation is more pronounced in mES cells following *Setdb1* inactivation compared to that seen after DNA methylation loss (Karimi et al., 2011). Additionally SETDB1 has a repressive role on a wider range of ERVs compared to DNA methylation. For example, the ERVL class of TEs appear to be regulated independently of DNA methylation through

mechanisms that involve H3K9 methylation (Karimi et al., 2011; Maksakova et al., 2013). At transposable elements, SETDB1 functions in cooperation with Krüppel-associated box domain (KRAB)-Associated Protein 1 (KAP1, also known as TRIM28) (Matsui et al., 2010; Rowe et al., 2013). KAP1 mediates the recruitment of chromatin remodelling complexes to particular DNA sequences through its binding to KRAB-containing Zinc Finger DNA binding Proteins (KRAB-ZFPs). This large family of proteins have expanded and co-evolved to match and bind new LTR retroelement insertions in mammalian genomes (Friedli and Trono, 2015; Thomas and Schneider, 2011). Unlike ERVK elements substantial demethylation of CpG-rich L1 LINEs is not always mirrored by transcriptional activation (Seisenberger et al., 2012) although these are heavily methylated in most somatic tissues (Hon et al., 2013).

In summary, the emerging view is that DNA methylation together with H3K9 methylation participates in the stable repression of a selection of transcriptional elements regulated by common transcription factors.

1.5 Mechanisms of transcriptional regulation by DNA methylation

Multiple models have been put forward to explain how DNA methylation impacts the transcription of genes or transposable elements. DNA methylation has been suggested to impact the positioning of nucleosomes: using *in vitro* nucleosomal reconstitution of bacterial artificial chromosomes (BACs) containing sequences derived from mammalian chromosomal DNA, methylation of CpG rich regions was shown to enhanced their tight association with nucleosomes (Collings et al., 2013). This observation has potential in explaining a repressive role for DNA methylation at CpG islands. However, *in vivo* DNA methylation also modulates interactions between non-histone proteins and DNA.

A wide range of proteins whose binding is affected by cytosine modifications has been identified (Spruijt et al., 2013; Yin et al., 2017). Through these interactions, DNA methylation

is believed to affect the local chromatin structure, histone tail modifications and transcriptional potential (Bogdanović and Veenstra, 2009). In this introduction, we will first describe the proteins which bind differentially to DNA containing methylated or unmethylated CpGs with little influence from the sequence context. We will then turn our attention to the impact of DNA methylation on the binding of sequence-specific DNA binding proteins.

1.5.1 MBD proteins and their associated complexes

The first proteins reported to bind specifically to methylated CpG dinucleotides were members of a family of Methyl-binding domain (MBD)-containing proteins including MBD1, MBD2, MBD4 and MECP2 (Hendrich and Bird, 1998). Another member of the family, MBD3 is highly similar in amino acid sequence to MBD2 but does not preferentially bind 5mC over unmethylated cytosine (Hendrich and Bird, 1998; Zhang et al., 1999). Although binding of MBD-containing proteins is mostly mediated through specific recognition of symmetrical CpG methylation, MeCP2 has a preference for mCpG sites adjacent to A/T-rich regions (Klose et al., 2005). The domain organisation of the MBD-containing proteins is depicted in Figure 1.4. The main point of homology between the five proteins is their MBD domain. Other domains present in individual MBD-containing proteins may confer specialised functions. For example, MBD1 contains another DNA-binding module, the CXXC domain through which it has been shown to bind unmethylated DNA (Jørgensen et al., 2004).

In mouse embryonic stem cells, MeCP2, MBD1, MBD2 and MBD4 proteins have affinity for chromatin independently of the underlying sequence although this association is stronger at regions with higher methylated CpG content. Reports suggest that MBD3 localisation also

Figure 1.4: Domain organisation of mammalian Methyl-Binding Domain (MBD)-containing proteins. All five proteins contain homologous MBD domains. Transcription Repression Domains (TRD) have been described for MeCP2, MBD1 and MBD2. MBD1 contains three CxxC motifs, the third of which binds unmethylated DNA. A glycine and arginine rich domain in MBD2 is indicated. MBD4 possesses a C-terminal thymine glycosylase domain.

coincides with high levels of 5mC although this is the subject of conflicting reports (Baubec et al., 2013; Hainer et al., 2016). As specific amino acid changes have abrogated the specificity of MBD3 for binding to methylated cytosines, localisation of this protein is likely to depend more strongly on protein-protein interactions or its reported association with 5hmC (Saito and Ishikawa, 2002; Yildirim et al., 2011). Nevertheless, different studies agree that reduction or complete removal of DNA methylation in mouse ES cells leads to the loss of enrichment of MECP2, MBD1, MBD2 and MBD4 at methylated regions with high CpG density (Baubec et al., 2013; Hainer et al., 2016; Hendrich and Bird, 1998).

MBD1, MBD2 and MECP2 have been extensively characterised and are known for their association with repressor complexes that contain histone deacetylases (HDAC) or other chromatin modifying activities (Bogdanović and Veenstra, 2009). This positions them as major players in the crosstalk between DNA methylation, histone modifications and transcriptional regulation.

MeCP2 was first reported to interact with mSin3A (or homologues in *Xenopus laevis*) in mammalian nuclear extracts and Xenopus oocytes through its repressor domain (Jones et al., 1998; Nan et al., 1998). Sin3a belongs to a transcriptional repressor complex that contains class I HDACs. HDAC inhibition by Trichostatin A (TSA) relieved transcriptional silencing mediated by MeCP2's repressor domain (Jones et al., 1998; Nan et al., 1998). Subsequently,

MeCP2 was shown to interact with two other repressor complexes Ski and N-CoR (Kokura et al., 2001). Notably, N-CoR/SMRT is a Sin3a-independent transcriptional repressor complex that contains class II HDACs necessary for its function (Huang et al., 2000). Mutations in MeCP2 are the cause of Rett syndrome, an X-linked neurodevelopmental disorder (Amir et al., 1999). Some mutations that have been found in patients with Rett syndrome abolish MeCP2's binding to mCpG while others disrupt its interaction with N-CoR/SMRT (Lyst et al., 2013). Both types of mutations were demonstrated to cause abnormalities in mice that mimicked some aspects of the human disease highlighting the importance of the MeCP2 protein in bridging DNA methylation and co-repressive complexes (Lyst et al., 2013).

MBD3 and MBD2 are components of mutually exclusive versions of the nucleosome remodelling and deacetylase (NuRD) complex (Wade et al., 1999; Zhang et al., 1999; Saito and Ishikawa, 2002). Similarly to MeCP2, MBD2 preferentially binds methylated DNA and can repress transcription in reporter assays, a function relieved by HDAC inhibition (Ng et al., 1999). MBD1 has been found to interact with histone methyltransferase proteins, particularly Suv39h1 and SETDB1 both of which methylate histone H3K9 residues and associate with the heterochromatic protein HP1 (Fujita et al., 2003; Sarraf and Stancheva, 2004). The interaction with SETDB1 was shown in human HeLa cells and reported to lead to the establishment of H3K9 methylation specifically during S-phase (Sarraf and Stancheva, 2004). However, the genomic distribution of H3K9me3 is largely unaltered in DNMT.TKO mouse ES that are devoid of DNA methylation. Furthermore, SETDB1 and DNMTs were found to target distinct sets of genes in mESCs suggesting that additional factors modulate the genomic localisation of SETDB1 (Karimi et al., 2011).

MBD4 has mostly been studied in relation to its role in DNA repair rather than transcriptional regulation. This protein is a thymine glycosylase and *in vitro* has preferential binding to TpG-mCpG mismatches. These mismatches are likely to arise from the spontaneous deamination

of 5mC nucleobases and can result in C-T transition mutations if unrepaired. Therefore, through its binding specificity and enzymatic activity, MBD4 has been proposed to minimise mutations occurring at methylated CpGs (Hendrich et al., 1999; Millar et al., 2002).

1.5.2 Histone Deacetylases and their relationship to DNA methylation

MBDs share the common feature of interacting with complexes that contain HDAC activity. In higher eukaryotes, HDAC enzymes are divided into four classes based on phylogenetic analysis. Class I, II and IV proteins are evolutionarily related and share a common enzymatic mechanism that is zinc-dependent. Class III enzymes (also called sirtuins) are phylogenetically distinct and operate through a different catalytic process (de Ruijter et al., 2003). MBDs are known to associate with HDACs belonging to classes I and II (Huang et al., 2000) whose activity is inhibited by TSA and other clinically used HDAC inhibitors such as SAHA (Gallinari et al., 2007).

The activity of HDACs counteracts that of histone acetylases (HATs) and together these two groups of enzymes tightly control the acetylation state of lysine residues present on histone tails. In turn, the acetylation state of histone tails impacts chromatin conformation and transcription in several ways. Histone acetylation reduces the positive charge of histone tails thereby weakening histone tail-DNA interactions contributing to chromatin decompaction (Mutskov et al., 1998; Tse et al., 1998). Additionally, acetylated or deacetylated histone tails can be specifically recognised by transcriptional regulators or chromatin modifiers that possess bromodomain (Dhalluin et al., 1999) or SANT domains respectively (Yu et al., 2003). Finally, HDAC-mediated deacetylation of lysine residues (such as H3K9 or H3K27) allows for their subsequent methylation by histone methyltransferases and recruitment of methyl lysine-binding proteins such as HP1 leading to a compact chromatin configuration.

Newly synthesised histones are acetylated and rapidly deacetylated after deposition onto DNA (Ruiz-Carrillo et al., 1975; Jackson et al., 1976). Class I and II HDACs are thought to be responsible for this global deacetylation process (Vogelauer et al., 2000) however they can also be targeted to specific loci through their interaction with sequence-specific repressors for example unliganded nuclear hormone receptors (Franco et al., 2001; Hartman et al., 2005) or Mad:Max heterodimers (Hassig et al., 1997).

Despite the association between MBD proteins and class I/II HDAC repressor complexes, it is unclear to what extent DNA methylation relies on histone deacetylation for its repressive effect across the genome.

Mouse models in which *Mecp2*, *Mbd1-4* or *Kaiso* were knocked-out did not show major developmental defects (relatively mild neurological phenotypes) which could point to a certain level of redundancy (Bogdanović and Veenstra, 2009). However when *Mbd2*, *Mecp2* and *Kaiso* were deleted, no phenotype similar to that seen in *Dnmt* knockout mice was observed suggesting that DNA methylation could be playing its developmental role through mechanisms independent of the recruitment of HDACs by MBD proteins (Martín Caballero et al., 2009). In addition, HDAC inhibition with TSA was insufficient to reactivate heavily methylated tumour suppressor genes in cancer cells or the *FMR1* promoter that contained a methylated CGG tract. However these genes were reactivated by 5-aza-deoxycytidine treatment and further potentiated by addition of TSA (Cameron et al., 1999; Coffee et al., 1999). Furthermore, in mouse ES cells, *Hdac1* and *Dnmt* knockout led to the activation of different classes of ERV repeats suggesting that both HDACs and methylation can act through independent mechanisms depending on the sequence context (Reichmann et al., 2012).

The relationship between DNA methylation and histone deacetylation is not as straightforward as originally thought. It remains to be examined to what extent DNA

methylation relies on HDAC activity for its role at CpG islands, transcription factor binding sites and other genomic elements.

1.5.3 Other DNA methylation-sensitive proteins

In addition to MBD-proteins several proteins containing zinc finger domains have been reported to bind methylated CpGs including Kaiso (ZBTB33), ZBTB4 and ZBTB38 (Prokhortchouk et al., 2001; Filion et al., 2006). Kaiso can bind two adjacent methylated CpG dinucleotides site and recruit the NCoR complex for repression (Yoon et al., 2003). ZBTB4 and ZBTB38 only require a single mCpG site and were also able to repress transcription in transfection assays however these proteins remain poorly characterised (Filion et al., 2006). Despite its ability to bind methylated CGCG sequences *in vitro*, Kaiso ChIP-seq data showed that most of its target sequences in human cell lines are unmethylated (Blattler et al., 2013) putting into question its role as a mediator of DNA methylation dependent repression.

Another important class of DNA binding proteins that are sensitive to cytosine methylation are those containing CXXC domains. The CXXC domain, which is found in a dozen nuclear proteins binds specifically to unmethylated CpG dinucleotides (Voo et al., 2000). Proteins that have a CXXC domain include KDM2A and B (H3K36me3 demethylases), FBXL19, CFP1, DNMT1, TET1 and 3, IDAX, MBD1, MLL1 and 2 (H3K4 methyltransferases) and CXX5 many of which are epigenetic modifiers. These proteins are mostly targeted to unmethylated CpG islands and have a variety and often opposing roles in dictating epigenetic states at gene promoters (Long et al., 2013). Although CXXC domain proteins have important functions in gene regulation, I will not discuss them further.

1.5.4 The impact of cytosine methylation to sequence-specific transcription factor binding

Another major mechanism by which methylated CpGs are proposed to contribute to transcriptional repression is by direct interference with the binding of sequence specific transcription factors (TFs). Several transcription factors whose consensus recognition sequences contain a CpG dinucleotide have been shown to preferentially or specifically bind unmethylated DNA probes *in vitro*. These include E2F1 (Campanero et al., 2000), SP1 (Clark et al., 1997), YY1 (Kim et al., 2003), ETS-2 (Gaston and Fried, 1995), CREB and c-Jun (Iguchi-Ariga and Schaffner, 1989; Jones and Chen, 2006; Rishi et al., 2010; Mann et al., 2013; Spruijt et al., 2013), HIF1a (Wenger et al., 1998), GR (Wiench et al., 2011) and MYC (Prendergast and Ziff, 1991).

The repertoire of DNA-binding proteins described as being sensitive to CpG methylation has increased dramatically in recent years with the methodological advances in systematic *in vitro* binding assays. These techniques have measured the binding of thousands of vertebrate transcription factors to arrays of methylated or unmethyated oligonucleotides (Hu et al., 2013; Spruijt et al., 2013) or determined the sequence preference of DNA-binding proteins using High-Throughput Systematic Evolution of unmethylated or methylated Ligands by Exponential Enrichment (methylation sensitive HT-SELEX) (Kribelbauer et al., 2017; Yin et al., 2017).

Complicating the original view that DNA methylation was antagonistic to transcription factor binding, these experiments have identified many DNA binding proteins that displayed preferential binding to certain methylated sequences (Kribelbauer et al., 2017; Mann et al., 2013; Rishi et al., 2010; Yin et al., 2017). In fact, preferential binding to unmethylated or methylated DNA varies between transcription factor families. Specifically binding of basic Helix Loop Helix (bHLH)-, Leucine Zipper (bZIP)- and E-Twenty-Six (ETS)-family factors that are

known to recognise sequences containing CpGs was generally inhibited by DNA methylation while proteins containing POU, NFAT and homeo-domain displayed preferential binding to the methylated versions of certain CpG-containing oligonucleotides (Yin et al., 2017). Other examples of proteins that can bind methylated CpGs include p53 (Kribelbauer et al., 2017), KLF4 (Hu et al., 2013), RBP-J (Bartels et al., 2011) or CEBPB (Rishi et al. 2010; Mann et al. 2013). Interestingly, many of these factors could also bind their canonical recognition sequence in an unmethylated state and the requirement for DNA methylation in the localisation of these TFs awaits confirmation *in vivo*.

The above findings mostly relate to experiments performed *in vitro* with sparse and locus specific *in vivo* confirmation. The extent to which methylation of a CpG site within a motif interferes with transcription factor binding on a genome-wide level *in vivo* has only been explored for a select few (Mann et al., 2013; Spruijt et al., 2013; Maurano et al., 2015; Domcke et al., 2015). As an example, *in vitro* evidence indicated that CTCF binding to DNA is abrogated by CpG methylation (Bell and Felsenfeld, 2000). Using ChIP coupled with bisulfite sequencing it was shown that CTCF bound irrespective of DNA methylation status genome-wide in mESCs (Feldmann et al. 2013) although a more recent report suggests that there are in fact a subset of proven CTCF binding sites that are bound uniquely in the absence of DNA methylation (Maurano et al., 2015). It is therefore likely that the influence of DNA methylation varies from TF to TF and that these effects are sequence and context dependent *in vivo*.

The idea of an antagonistic relationship between DNA methylation and transcription factor binding partly emerged from the observation that gene promoter regions are almost all unmethylated during normal development and are frequently methylated in cancer cells leading to aberrant silencing (Hon et al., 2013; Hughes et al., 2013; Ziller et al., 2013). Conversely, genome-wide mapping efforts have identified that transcription factors most often occupy regions with no or low levels of DNA methylation at CpG islands or promoter-

distal regulatory elements respectively (Stadler et al., 2011; Hon et al., 2013; Ziller et al., 2013). However these correlations are not sufficient to prove a causative role for DNA methylation in specifying transcription factor occupancy *in vivo* and this has indeed been put into question.

Notably, certain DNA sequences containing intact binding sites for SP1, CTCF, REST or other transcription factors have been shown to induce demethylation when incorporated into the genome (Brandeis et al., 1994; Siegfried et al., 1999; Stadler et al., 2011; Feldmann et al., 2013). In fact, genetic abrogation of the REST DNA binding protein leads to a decrease in 5-hydroxymethylcytosine at its known binding sites implying that the binding of some "pioneering" transcription factors may dictate DNA methylation patterns by recruiting TET proteins (Feldmann et al., 2013). In addition, integration of transcription factor expression and bisulfite sequencing data from a variety of tissues has indicated that DNA methylation of TF binding sites may follow protein evacuation (Thurman et al., 2012).

The causal relationship between differential transcription factor binding and DNA methylation is likely to depend from protein to protein; research is active in this area. In a recent study, Domcke *et al.* searched for transcription factors binding sites enriched within regions that were differentially sensitive to DNaseI in wild-type or DNA methylation-deficient mES cells and identified NRF1 as a factor whose localisation is dictated by DNA methylation (Domcke et al., 2015).

1.6 Aims of the study

The aim of this study is to determine the *in vivo* contribution of two mechanisms by which DNA methylation is proposed to mediate transcriptional regulation: the recruitment of HDAC-containing repressor complexes and the differential binding of transcription factors (see models in Figure 1.5). Our first goal was to identify the transcriptional units throughout the genome whose activity is under the influence of DNA methylation or HDACs. We planned to assess the overlap between the elements that respond to disruption of the DNA methylation or HDAC machinery hence allowing us to appreciate the contribution of HDAC-containing complexes in linking DNA methylation and transcription. Our second goal consisted in identifying genomic regulatory elements whose function is modulated by DNA methylation and/or HDAC activity. We hoped this data would provide some mechanistic insight into the different transcriptional effects that were observed after HDAC inhibition and DNA methylation loss. In parallel, an attempt was made to examine the extent to which DNA methylation affects transcription factor occupancy genome wide. Our third goal was therefore to identify transcription factors whose genomic distribution differed in the presence and absence of DNA methylation.

To study the role of DNA methylation we made use of mouse ES cells in which DNA methyltransferase activity has been abolished by mutation of the *Dnmt1*, *Dnmt3a* and *Dnmt3b* genes (Tsumura et al., 2006) (subsequently referred to as DNMT.TKO or TKO in figures) and compared these to a corresponding wild-type ES cell line. High-throughput sequencing methods were used to address the objectives described above. The relative effects of DNA methylation removal and HDAC inhibition on gene expression were investigated through whole-genome transcriptional analysis in wild type and DNMT-TKO cells in the presence and absence of the HDAC inhibitor TSA (**Chapter 3**). Genome-wide chromatin accessibility mapping was performed under the same conditions to determine regulatory

elements that respond to DNA methylation loss or HDAC inhibition (**Chapter 4**). Finally, to investigate the differential binding of transcription factors in the absence or presence of DNA methylation, we chose to profile the occupancy of candidate transcription factors in WT or DNMT-TKO cells through chromatin immunoprecipitation coupled with high-throughput sequencing (**Chapter 5**) as well as make use of the chromatin accessibility data.

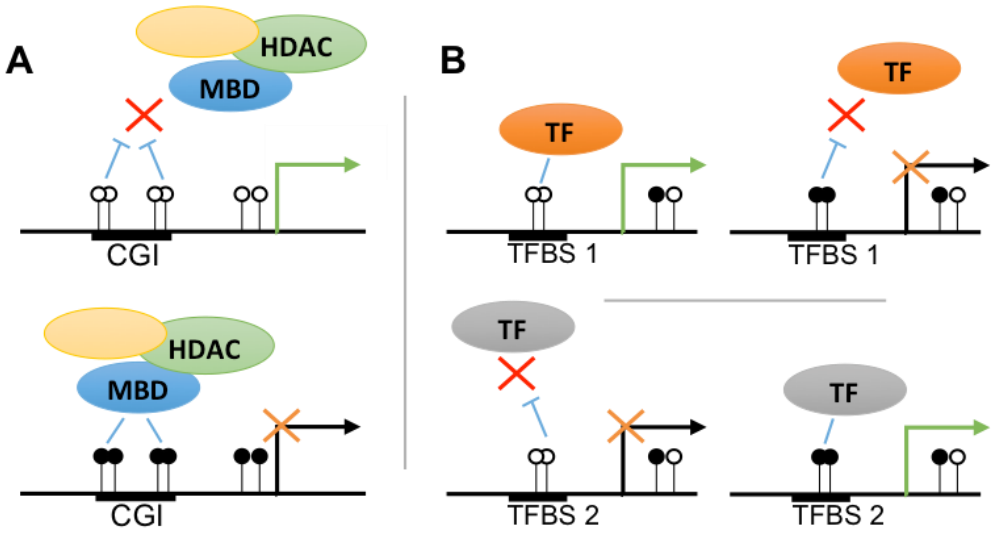


Figure 1.5: Schemes of mechanisms by which DNA methylation is proposed to affect transcriptional regulation. A. One suggested mechanism is that regions that harbour methylated CpGs (black circles on sticks) are preferentially occupied by repressive protein complexes that contain HDAC enzymes. Methyl-Binding Domain (MBD)-containing proteins (such as MeCP2, MBD1 or MBD2) within these complexes are responsible for the specific interaction with methylated CpGs. **B.** A second proposed mechanism is that CpG methylation directly affects the binding of methylation-sensitive transcription factors (TF) to their cognate transcription factor binding sequences (TFBS). Differential binding of TFs to promoters or distal regulatory elements is likely to affect transcription *in cis*. Best characterised are transcription factors whose interaction with DNA is disrupted by CpG methylation (top, eg. NRF1 or CTCF). Other transcription factors preferentially have been found to bind certain sequences when these are methylated (bottom, eg. KLF4).

Chapter 2:

Materials and Methods

2.1 General notes

Unless stated otherwise, all centrifugation steps were carried out in a 5424R microcentrifuge (Eppendorf) for 1.5ml or 2ml Eppendorf tubes or a Heraeus Multifuge 3SR+ centrifuge (Thermo Scientific, 75004371) for 15ml or 50ml Falcon tubes. All chemicals were purchased from Sigma-Aldrich unless specified otherwise.

2.2 Cell culture techniques

All cell culture procedures were carried out in a sterile environment in a dedicated room inside a vertical flow hood (Envair EcoSafe Comfort Plus, SCS Evo 2-4). Before beginning work, all surfaces and materials were wiped down with 70% ethanol or industrial methylated spirit (Fischer Chemicals, M/4450/PB17). All cells were incubated in a humidified incubator (Binder) set at 37% and 5% atmospheric CO_2. Cells were routinely checked for Mycoplasma contamination using the MycoAlert™ mycoplasma detection kit (Lonza).

2.2.1 Cell lines and growth media

Dnmt1$^{-/-}$, *Dnmt3a*$^{-/-}$ and *Dnmt3b*$^{-/-}$ TKO cells (DNMT.TKO, (Tsumura et al., 2006)) and their parental wild-type J1 mouse embryonic stem cells were obtained as a gift from the Rob Klose lab (Department of Biochemistry, University of Oxford). These cells were grown on 0.1% gelatin-coated dishes in DMEM (Lonza, BE12-604F) complemented with 10% Fetal Bovine Serum (Biosera, product FB-1001G/500, lot #11484), 2mM L-glutamine (Life Technologies, 25030081), 1x non-essential amino acids (Life Technologies, 11140050), 1x penicillin-streptomycin (Life Technologies, 15140122), leukaemia inhibitory factor (LIF, produced in-house following protocol by Tomala et al., 2010) and 0.1mM beta-mercaptoethanol (Sigma, M6250-100ML). 0.1% gelatin was prepared by first dissolving gelatin (Sigma, G1890-100G) in water, autoclaving and then adding a 10x sterile PBS stock (Life Technologies, 14200067).

Chicken DF-1 (Doug Foster strain 1) fibroblast cells were generously provided by Alison Turner at the Viral Vector Core Facility in the Jenner Institute (Oxford). DF-1 cells were grown in DMEM (Lonza, BE12-604F) supplemented with 10% Fetal Bovine Serum (Biosera, product FB-1001/500 lot #014BS386), 1x penicillin-streptomycin (Life Technologies, 15140122), 2mM L-glutamine (Life Technologies, 25030081) and 0.15% Sodium Bicarbonate (Life Technologies, 25080060).

2.2.2 Cell passaging

When cells were at least 70% confluent, they were harvested and split at a 1:4 or 1:6 ratio into a new flask. First, the old medium was aspirated and the cells washed with PBS. A trypsin-EDTA solution (Life Technologies, 25300054) was added to the flask and the cells incubated at 37°C for approximately 1 minute until they were detached. Trypsin was quenched by adding growth medium at a 1:1 volume ratio following which the resulting cell suspension was spun down in a 15 ml or 50 ml Falcon tube for 3 minutes at 220xg at room temperature. The resulting cell pellet was resuspended in growth medium and the appropriate fraction of cells seeded down in a new dish.

2.2.3 Cell counting

When a specific number of cells were to be seeded or used for an experiment, cell suspensions were prepared in medium or PBS and mixed by pipetting to break up clusters. Two 10 µl samples were removed, each mixed with 10 µl Trypan Blue (Life Technologies, 15250061) in a 1.5 ml Eppendorf tube after which 10 µl samples of the resulting mixtures were loaded onto a counting slide. A TC10 automated cell counter (Biorad) was used to obtain a cell count for each duplicate sample and the average count was used to calculate the number and concentration of cells in the original suspension. Only live cells were taken under consideration.

2.2.4 Freezing and defrosting cells

Mouse ES cells were frozen down using the following procedure. Cells were grown to 70-80% confluency in 75 cm^2 (T75 flasks) and harvested by trypsinisation. After pelleting, cells were resuspended in growth medium supplemented with 10% DMSO (Sigma, D2650) and 1ml of the resulting suspension was transferred to a 1.8ml CryoTube vial (Thermo Scientific, 377267). To avoid instant freezing, the tubes were first placed in a Mr. Frosty™ Freezing Container (Thermo Scientific, 5100-0001) and kept at -80°C. The next day, frozen cells were transferred to liquid-nitrogen for long-term storage.

Frozen cells were thawed in a 37°C bead bath or by holding in the palm of a hand following which 5ml growth medium was added. Next, cells were pelleted for 3 minutes at 220 x g at room temperature following which they were resuspended in 10ml growth medium and seeded in a 25 cm^2 flask.

2.2.5 Trichostatin A treatment of mouse embryonic stem cells

Trichostatin A was purchased as a ready-made 5mM solution in DMSO (Sigma, T1952). Upon the first thawing, 5 µl aliquots were made and stored at -20°C. At each use, 45 µl DMSO (Sigma, D2650) was added to a new aliquot and this was further diluted 300x in mESC growth medium (500ng/µl TSA solution).

Prior to TSA treatment, 5 x 10^6 mES cells were seeded onto a T75 flask. After 4-6h, once cells had attached, growth medium was replaced with medium containing 5ng/µl TSA, prepared by diluting freshly made up 500ng/µl TSA solution. The same TSA supplemented medium solution was used for all cells treated in the same experiment. TSA-containing media was changed after 24h and cells were harvested for downstream applications after 36h. For larger

experiments, cell numbers were scaled proportionally to the total surface area so as to keep the same cell density.

Control DMSO-treated cells were seeded and treated in a similar manner however medium containing 0.33% DMSO was used instead of the 500ng/µl TSA solution.

2.3 Histone isolation and immunoblot analysis of histone acetylation levels

In order to monitor global increases in histone acetylation after TSA treatment, histones were extracted from cell pellets and immunoblotting was carried out using antibodies against pan-acetylated histone. For each experiment involving TSA treated cells, a small proportion of treated cells was set aside and used to monitor global increases in histone acetylation. Typically, 1.25 x 10^6 cells were pelleted (211 x g, 3min 4°C), flash frozen and stored at -80°C for this purpose. Histones were extracted from cell pellets and immunoblotting was carried out using antibodies against pan-acetylated histone according to the following procedures.

2.3.1 Histone isolation from cell pellets

Unless stated otherwise, all steps were performed on ice, spins performed in a centrifuge chilled to 4°C and rotations set up in a cold room. Cell pellets containing 1 to 2 x 10^6 cells were resuspended in 300 µl Lysis Buffer (PBS supplemented with 0.5% Triton-X100, 1x cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche, 05056489001) and 10mM sodium butyrate (Sigma , 303410-100G)) then rotated for 10 min a 4°C. A 10min spin at 10,000g was used to pellet the nuclei and the supernatant was discarded. The pelleted nuclei were washed by resuspension in 300 µl lysis buffer. The nuclei were spun again at 10,000 x g for 5 min after which the supernatant was discarded. Nuclei were resuspended in 125 µl 0.2M H_2SO_4 and rotated overnight at 4°C to allow for solubilisation of the histone proteins. The next day,

samples were spun at 16,000 x g for 10min and the supernatant containing the solubilised histones were transferred to a new 1.5ml Eppendorf tube.

In order to obtain a histone preparation in neutral pH solution, histone proteins were precipitated by adding trichloroacetic acid (TCA, Sigma, T0699-100ML) to each sample to a final concentration of 25% TCA. After incubation on ice for at least 2 hours, the samples were spun at 16,000 x g for 10min and the supernatant was discarded. This was followed by two washes with 1ml ice-cold acetone. After the final wash, the precipitated histones were air-dried and solubilised in 100 µl 1x SDS-PAGE Loading Buffer (prepared by diluting 5x SDS-PAGE loading buffer (200mM Tris-HCl pH 6.8, 400mM DTT, 8% SDS, 0.4% bromophenol blue, 40% glycerol)) and boiled at 95°C for 5min. These samples were directly loaded onto 15% SDS-PAGE gels to estimate histone concentration and immunoblotting.

2.3.2 SDS-PAGE electrophoresis and immunoblotting of acetylated histones

Handcast 15% SDS-PAGE gels were run in a Mini Protean Tetra cell (Biorad, 1658005EDU) in 1x SDS-PAGE Running Buffer (25mM Tris-base, 192mM glycine, 0.1% SDS) for 1.5h at 100V. For every gel, 5 µl of a PageRuler Plus prestained protein molecular weight marker (Thermo Scientific, 26619) was run alongside the samples.

In order to achieve equal loading for immunoblotting, the relative amounts of histones present in each sample were estimated by running 5 µl of each sample on a 15% SDS-PAGE gel and staining for total protein using the GelCode Blue safe protein stain (Thermo Scientific, 24594). Gels were imaged using a ChemiDoc instrument (Biorad) and histone band intensity was measured using the associated software. Based on the measured intensities, adjusted volumes to be loaded on a gel were calculated so as to achieve equal loading.

Once equal loading could be achieved, the samples were run on a 15% PAGE in order to proceed with immunoblotting. After separation through electrophoresis, proteins were

transferred to a nitrocellulose membrane (10600002, GE Healthcare) in a 2-Gel Tetra Blotting Module (1660827EDU, Biorad) filled with chilled 1x Transfer Buffer (25mM Tris-HCl pH 7.6, 192mM glycine, 20% methanol) for 1 hour at 50 V in the presence of an ice pack. To confirm the loading equality and assess the success of the transfer, the membrane was stained with Ponceau S (P7170, Sigma). After imaging, membranes were de-stained with 1x TBST (50mM Tris-HCl pH 7.5, 150mM NaCl, 0.1% Tween-20 (Fisher Chemicals BP337-500)). Next membranes were blocked in 5% Milk in TBST (Marvel Original dried skinned milk) for 1h at room temperature. Following this, individual membranes were placed within a 50ml Falcon tube and incubated with a primary antibody overnight at 4°C with rotation. The antibodies used to probe for acetylated histones, along with the dilutions used are presented in Table 2.1.

Antibody	Source (lot)	Dilution	Species	Usage
Anti-acetylated Histones H3 and H4 (pan-acetyl)	Upstate, 06-866 (lot #2664262)	1:5000	Rabbit	Primary
Anti-acetyl H3 (pan-acetyl)	Millipore, 06-599 (lot #2768436)	1:2000	Rabbit	Primary
Anti-Rabbit	Santa Cruz, sc-2004 (lot #F0515)	1:10000	Goat	Secondary, HRP-conjugated
Anti-Histone H3	Abcam, ab1791 (lot #GR242835-1)	1:40000	Rabbit	Primary

Table 2.1: Antibodies used for immunoblotting experiments

The following day, membranes were washed 3 x 5 min in TBST and were then incubated with an HRP-conjugated secondary antibody (Table 2.1) for 45-60min at room temperature. Once again, the membranes were washed three times in TBST for 5min per wash. HRP activity was detected using Amersham ECL Select Western Blotting Detection Reagents (GE Healthcare, RPN2235) and imaged using a ChemiDoc instrument (BioRad).

To stain for total unmodified histone H3 protein, membranes were stripped by incubation in Restore™ WB Stripping Buffer (Thermo Scientific, 21059) for 15min at room temperature then blocked again in 5% Milk TBST for 1h at room temperature. Membranes were incubated for 30min with the primary antibody towards total H3 protein (Table 2.1) after which washes, incubation with secondary antibody and detection were carried out as described above.

2.4 RNA-seq sample preparation and sequencing

Total RNA was isolated from DMSO- or TSA- treated mES cells grown in 75cm^2 (T75) flasks using the TRI Reagent (Sigma, T9424) following the protocol provided by the manufacturer. Briefly, 5ml TRI Reagent was directly added to the flasks, nucleic acid-protein complexes were dissociated by pipetting, and the lysates were transferred to 15ml Falcon tubes. Phase separation was achieved by adding 0.5ml (0.1 volume of original TRI reagent) 1-bromo-3-chloropropane (Sigma, B9673-200ML), vortexing for 15 seconds and leaving the samples to stand for 5 min at room temperature after which they were spun at 4538 x g for 15min at 4°C. The top aqueous phase was transferred to a separate 15ml Falcon tube to which 2.5ml (0.5 volume of original TRI Reagent) isopropanol was added. Samples were mixed and left at room temperature for 10min then spun for 1h at 4538 x g at 4°C. The supernatant was discarded and the resulting RNA pellet was washed with 2ml freshly prepared 75% ethanol. After spinning for 5 min at 4538 x g, ethanol was thoroughly aspirated and the pellet air-dried for 15min at room temperature. The RNA pellets were dissolved in 100 µl RNase-free water. Concentration was measured using a NanoDrop 2000 spectrophotometer (Thermo Scientific) and samples were stored at -80°C.

Once total RNA was isolated from all biological replicates, DNaseI treatment was carried out on 10ug RNA from each sample to digest contaminating genomic DNA. Reactions were set up in a total volume of 100 µl with 10U of DNaseI (Thermo Scientific, EN0525) and 1x provided

Reaction Buffer then incubated for 30min at 37°C. The reaction was stopped by adding 10 µl of 50mM EDTA. RNA was extracted by adding 110 µl Phenol-Choloroform-isoamyl alcohol mix (Sigma 77617-500ML) to each sample then vortexing and spinning at 15,871 x g for 5min at 4°C. The aqueous upper phase was collected and transferred to a new 1.5ml Eppendorf tube. Next, RNA was precipitated by adding 40 µl (~ 0.5 volume) 8M ammonium acetate and 550 µl 100% ethanol (~ 7 volumes) and incubating for 2h at -80°C. Samples were then spun at 15,871 x g for 20min at 4°C and the supernatant was removed. The RNA pellet was washed with 75% ethanol, air-dried and dissolved in 50 µl RNase-free water. RNA was quantified using a Nanodrop 2000 spectrophotometer (Thermo Scientific) and RNA integrity verified on an agarose gel. 1-6ug RNA was provided to the Oxford Genomics Centre where ribosomal RNA depletion and library preparation for single-stranded RNA-seq was carried out. Libraries were sequenced on an Illumina HiSeq2500 instrument to generate 50bp paired-end reads.

RNA-seq sample preparation was carried by Dr Skirmantas Kriaucionis.

2.5 ATAC-seq sample preparation and sequencing

Assay for Transposase Accessible Chromatin (ATAC)-seq experiments were carried out as described below using mES cells treated with TSA or DMSO. The protocol was based on that published by (Buenrostro et al., 2013).

2.5.1 Isolation of nuclei

Cells were first washed with PBS, detached by trypsinisation, pelleted, resuspended anew in PBS then counted. For each sample, 5 x 10^6 cells were transferred to fresh container and washed once more time in ice-cold PBS and pelleted at 1000 x g for 5min at 4°C. The pellets were gently resuspended in 1ml Lysis Buffer (50mM KCl, 10mM MgSO4.7H20, 5mM HEPES, 0.05% NP40 (IGEPAL CA630), freshly supplemented with 1mM PMSF and 3mM DTT),

transferred to 2ml Eppendorf tubes and incubated at room-temperature for 1 min. Nuclei were centrifuged at 1000 x g for 5min at 4°C and washed twice in 1ml ice-cold RSB Buffer (10mM NaCl, 10mM Tris (pH 7.4), 3mM $MgCl_2$) so as to remove mitochondrial and cytoplasmic material. All the supernatant was carefully removed and nuclei resuspended once more in RSB Buffer. Nuclei were then counted using a haemocytometer, whilst high nuclei integrity and Trypan Blue (Life Technologies, 15250061) staining (indicative of nuclear permeability) was confirmed visually. Nuclei were kept on ice until ready to proceed.

2.5.2 Assays for transposase accessible chromatin

5 x 10^5 nuclei were aliquoted into a fresh 1.5ml Eppendorf tube and spun at 1000xg for 5min at 4°C. These were resuspended in 100 µl water and used immediately in an ATAC reaction. ATAC reactions were performed in 1.5ml DNA LoBind tubes (Eppendorf, 022431021) in 50 µl reactions in the presence of Tn5 reaction buffer (10mM TAPS, 5mM $MgCl_2$ and 10% dimethylformamide), 5 x 10^4 nuclei (10 µl from aliquots) and 2 µl 12uM Tn5 transposase. The latter was provided as a generous gift by the Rob Klose lab where it was produced in house according to a previously published protocol (Picelli et al., 2014). ATAC reactions were incubated for 30min at 37°C. Reactions were carried out in technical duplicates although only one was used for sequencing while the other was stored as a backup.

During the incubation step, the nuclei suspension in water were visualised using a haemocytometer and the relative numbers of nuclei was compared between samples. If the relative number of cells between samples was noticeably different, the reactions were repeated with re-adjusted cell numbers.

Once the reactions were finished, tagmented DNA was purified using the QiaQuick MinElute columns (Qiagen, 28004) following manufacturer's instructions. 10 µl of the provided Elution Buffer was used to elute the DNA from the columns.

To control for sequence bias of the Tn5 transposase, naked genomic DNA from mES cells was digested with Tn5 for 7min at 55°C. In this study, this Tn5 control DNA was provided by the Klose lab and had been isolated from a different batch of mES cells than that used for ATAC-seq experiments.

2.5.3 Library preparation and sequencing

ATAC-seq libraries were prepared by PCR amplification using custom made Illumina barcodes and the NEBNext High-Fidelity 2x PCR Master Mix (NEB, M0541L) with 8-10 cycles. Libraries were purified with two rounds of Agencourt XP Bead Cleanup (1.5x beads to sample ratio, Beckman Coulter) as instructed by manufacturer and then quantified by real time PCR using SensiMix SYBR (Bioline) and KAPA Library Quantification DNA standards (KAPA Biosystems). ATAC-seq libraries were sequenced on an Illumina HiSeq4000 instrument at the Oxford Genomics Centre to generate 75bp paired-end reads.

ATAC-seq sample preparation was jointly performed by myself and Hamish King, a former member of the Rob Klose lab. Library preparations were carried out by Hamish King.

2.6 Native chromatin immunoprecipitation with exogenous calibration

Native Chromatin Immunoprecipitation (ChIP) followed by high-throughput sequencing was carried out in order to map the position of acetylated histone H3 proteins genome-wide. Native ChIP samples were prepared from wild-type J1 or DNMT.TKO mES cells, treated with TSA or DMSO, as detailed in the steps below. To allow for the accurate comparison of ChIP sequencing signal between different samples, Chicken DF-1 cells were spiked-in on a per-cell basis in a manner adapted from (Orlando et al., 2014).

2.6.1 Isolation of nuclei

Both mouse and chicken cells were detached by trypsinisation, pelleted and counted after being resuspended in PBS. 5 x 10^7 mES cells were transferred to a fresh 50ml falcon tube for each condition and 2.5 x 10^7 chicken DF-1 cells were transferred to a separate tube. All cells were spun down at 100 x g for 3min at room temperature and the supernatants discarded. Both mouse and chicken cells were resuspended in 10ml PBS following which 2ml of the chicken cell suspension (containing 5 x 10^6 cells) was added to the mouse cells from each condition (containing 5 x 10^7 cells each). Cell mixtures were once again spun down and resuspended in 1ml ice-cold RSB Buffer (10mM NaCl, 10mM Tris (pH 8.0), 3mM $MgCl_2$56) then kept on ice.

For the isolation of nuclei, 28ml RSB Buffer supplemented with 0.1% NP-40 (IGEPAL CA630), 1x cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche, 05056489001), and 10mM sodium butyrate (used to inhibit histone deacetylases) was added to each sample and the tubes were gently inverted ten times. Nuclei were spun down at 1500 x g for 5min at 4°C and washed once in 5ml RSB buffer supplemented with 0.25M sucrose, 3mM $CaCl_2$, 1x cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche, 05056489001) and 10mM sodium butyrate, before being resuspended in 1ml of this same buffer and transferred to a 1.5ml tube.

2.6.2 Chromatin fragmentation with micrococcal nuclease

Chromatin was fragmented by adding 200 units of Micrococcal Nuclease (MNase, Fermentas EN0181) and incubating at 37°C for 5min. The tubes were gently inverted every minute to stop nuclei from settling at the bottom. The reaction was stopped by the addition of 8 µl 0.5M EDTA and mixing. Samples were spun at 2348 x g for 5min at 4°C and the supernatant (approximately 1ml) was transferred to a fresh 1.5ml Protein Lobind tube (S1 fraction). 1.5ml

Protein LoBind tubes (Eppendorf, 022431081) were used for all subsequent steps unless indicated otherwise.

To extract nucleosomes from the remaining insoluble fraction, the pellet was resuspended in 300 µl Nucleosome Release Buffer (10mM Tris pH 7.5, 10mM NaCl, 0.2mM EDTA, freshly supplemented with 1x cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche, 05056489001) and 10mM sodium butyrate) and rotated for 1h at 4°C. Following this, the samples were passed through a 25G needle five times using a 1ml syringe and then centrifuged at 2348 x g for 5min at 4°C. The supernatant was removed and combined to the S1 chromatin fraction. At this point, a 40 µl aliquot was removed to check the DNA fragment size within the samples. DNA was purified using a QiaQuick PCR Purification kit (Qiagen, 28104) following by the manufacturer's instructions. DNA was isolated from the columns in 30 µl of the provided EB buffer, the concentration was determined using a NanoDrop 2000 spectrophotometer and approximately 1ug DNA was run on a 1% agarose gel. The presence of an intense band at the correct size for mononucleosomal DNA (150-200bp) and relatively faint bands at di- and tri- nucleosomal sizes was indicative of successful chromatin digestion.

2.6.3 Chromatin immunoprecipitation for acetylated histone H3

For chromatin immunoprecipitation of acetylated histone H3, two duplicate IPs were performed in parallel for every condition, each using 100 µl of the solubilised chromatin. Material from the two repeats was combined at a later step. In addition, 100 µl of the solubilised chromatin was processed for use as an input sample. Sufficient chromatin for all IP and input samples was transferred to a 15ml Falcon tube and diluted ten times in Native ChIP Incubation Buffer (10mM Tris pH 7.5, 70mM NaCl, 2mM $MgCl_2$, 2mM EDTA and 0.1% Triton X-100 freshly supplemented with 1x cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche, 05056489001) and 10mM sodium butyrate). 100 µl aliquots of the remaining

chromatin were flash frozen on dry ice and stored at -20°C. Diluted chromatin was centrifuged at 845 x g for 5 min at 4°C to pellet any precipitate and then aliquoted into the appropriate number of 1.5ml tubes (one for each IP and one for the input sample). Immunoprecipitation was carried out at 4°C, overnight with rotation using 3 µg of an anti-acetylated histone H3 antibody (Millipore, 06-599). The input samples were kept at -20°C overnight.

Immune complexes were captured using Protein A agarose beads (Roche, 11134515001). The total required amount of beads was blocked and washed prior to their addition to the IP samples. Agarose beads were first removed from their storage buffer by collecting them using a 30s spin at 1000 x g and discarding the supernatant. Beads were then taken back to their original slurry volume (1:1 bead bed volume to buffer volume) in blocking buffer (Native ChIP Incubation buffer containing 1mg/ml BSA (Sigma, A7906-100G) and 1mg/ml yeast tRNA (Roche, 10109517001) then rotated at 4°C overnight; alongside the IPs. The next morning, the beads were washed thrice in Native ChIP Incubation buffer and finally resupended in this same buffer. IP samples were transferred to a fresh tube containing 20 µl of packed, blocked beads and rotated at 4°C for 1h. Next, beads were collected by spinning at 1000 x g for 30s and washed four times in ice-cold Native ChIP wash buffer (20mM Tris pH 7.5, 2mM EDTA, 125mM NaCl, 0.1% Triton-X100) and once in TE buffer (10mM Tris HCl pH 8.0, 1mM EDTA).

2.6.4 DNA isolation, library preparation and sequencing

Beads were resuspended in 100 µl elution buffer (1% SDS, 0.1M $NaHCO_3$, prepared freshly before each use) and shaken vigorously for 30 min at room temperature to elute the ChIP material. Beads were collected to the bottom of the tube by centrifugation and the eluate was transferred to a fresh 1.5ml tube. At this point input samples were thawed and subsequently processed in an identical way to the ChIP samples. DNA was purified using a ZymoResearch ChIP DNA clean and concentrator kit (Zymo Research, D5205) following the

manufacturer's instructions. Technical duplicate IPs from each condition were combined at this step: the ChIP material from both samples was loaded onto a single column. 12 µl of the provided EB buffer was used to elute the DNA into 1.5ml DNA Lobind tubes. DNA concentration in 1 µl aliquots was measured using a Qubit dsDNA Assay quantitation kit (Thermo Scientific, Q32851) and a Qubit 3.0 fluorometer (Thermo Scientific) following the manufacturer's instructions.

10ng ChIP material was submitted to the Oxford Genomics Centre where libraries were prepared and sequenced on an Illumina HiSeq4000 instrument to generate 75bp paired-end reads.

2.7 Chromatin immunoprecipitation sample preparation and sequencing

Chromatin immunoprecipitation (ChIP) for the real-time PCR analysis or high throughput sequencing analysis of transcription factor binding was carried out following the steps detailed below.

2.7.1 Chromatin crosslinking and sonication

Chromatin was prepared from mouse ES cells grown to 70-90% confluency in 15 cm cell culture dishes. Cells were washed with PBS, detached by trypsinisation, pelleted and counted after being resuspended in PBS. 1 x 10^8 cells were transferred to a new 15ml tube, washed again and finally resuspended in 10ml PBS. 625 µl 16% methanol-free formaldehyde (Thermo Scientific, 28908) was added for a final concentration of 1% and cells were crosslinking for 10min at room temperature on a rocker.

Optionally, a first crosslinking step was carried out prior to the addition of formaldehyde (see Chapter 5 - Section 5.2 and Table 2.2) using disuccinimidyl glutarate (DSG, Thermo Scientific, 20593) or ethylene glycol bis(succinimidyl succinate) (EGS, Thermo Scientific, 21565). In these

experiments DSG or EGS in powder form was dissolved in DMSO shortly before use and added to the 10ml cell suspensions to a final concentration of 2mM. Cells were incubated with DSG or EGS for 1h at room temperature with gentle rotation before the addition of formaldehyde.

The crosslinking reaction was quenched by the addition of 1ml 2M glycine and the samples incubated for a further 3min at room temperature. Crosslinked cells were pelleted by centrifugation at 211 x g for 4min at 4°C and the supernatant discarded. Subsequently, samples were kept on ice and all centrifugation, rotation or incubation steps were carried out at 4°C. Chromatin was prepared by rotating the cells for 10min first in 10ml buffer LB1 (50mM HEPES KOH pH 7.9, 140mM NaCl, 1mM EDTA, 10% Glycerol, 0.5% NP40, 0.25% Triton x-100, freshly complemented with 1x cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche, 05056489001)) and then 10ml buffer LB2 (10mM Tris-HCl pH 8.0, 200mM NaCl, 1mM EDTA, 0.5mM EGTA, freshly complemented with 1x cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche, 05056489001)). After either rotation step, cells were pelleted by centrifugation at 211 x g for 4min and the supernatant was discarded. Next buffer LB3 (10mM Tris-HCl pH 8.0, 100mM NaCl, 1mM EDTA, 0.5mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine, freshly complemented with 1x cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche, 05056489001)) was added to a total volume of 1ml and the chromatin was transferred to a 1ml milliTUBE with AFA Fiber (Covaris, 520130) for sonication. Chromatin was sheared using a Covaris S220 AFA ultrasonicator (set at 150W Peak Incident Power, 8% Duty Factor, 200 Cycles/Burst) for 30min. Following sonication, the sheared chromatin was spun at 15,000 x g for 20min and the soluble fraction collected and aliquoted for immediate use or storage at -80°C. 1.5ml Protein LoBind microcentrifuge tubes (Eppendorf, 022431081) were used for all steps following sonication unless indicated otherwise.

2.7.2 Chromatin immunoprecipitation

For immunoprecipitation (IP), the sheared chromatin was diluted 11-fold in Chromatin Dilution Buffer (16.7mM Tris pH 8.1, 167mM NaCl, 0.01% SDS, 1.1% Triton X-100, 1.2mM EDTA, freshly complemented with 1x cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche, 05056489001)) so that the rough equivalent of 10^7 cells was used for each IP. Prior to immunoprecipitation, diluted chromatin in 1.1ml was pre-cleared using 25 µl bed volume of blocked Protein A or G agarose beads (Roche, 11134515001 and 11719416001 respectively) for 30min with rotation in the absence of antibodies.

Protein A or G agarose beads were prepared prior to this step by first washing the total required amount of beads twice in 0.5% BSA made up in PBS. To isolate beads from suspensions, these were collected at the bottom of tubes by a gentle spin at 1000 x g for 30s. Beads were blocked by incubation for 1h with rotation in at least 1ml 0.5% BSA/PBS.

Pre-cleared chromatin was removed from the beads and 1ml was incubated with 5µg antibody overnight at 4°C. 50 µl (5%) samples of pre-cleared chromatin were taken to serve as an input control. The antibodies that were used for ChIP in this study and the type of agarose beads used for their capture are indicated in Table 2.2.

Immune complexes were captured with 50 µl Protein A or G agarose beads by incubation for 4h with rotation after which beads were washed with the following buffers: Low Salt Buffer (20mM Tris pH8.1, 150mM NaCl, 0.2% SDS, 1% Triton X-100, 5mM EDTA, and 5% w/v sucrose), High Salt Buffer (50mM HEPES pH7.5, 500mM NaCl, 0.1% deoxycholic acid, 1% Triton X-100 and 1mM EDTA), Lithium Chloride Buffer (10mM Tris pH 8.0, 250mM LiCl, 0.5% deoxycholic acid, 0.5% NP-40 and 1mM EDTA) and twice with TE Buffer (10mM Tris-HCl pH 8.0, 1mM EDTA). For each wash, the beads were resuspended in the appropriate wash buffer, rotated for 4 min and collected by gentle centrifugation (1000 x g for 30s). Wash steps were

performed in a cold room. Crosslinked material was eluted from the beads in Elution Buffer (1% SDS and 0.1M $NaHCO_3$) by vortexing and then shaking vigorously for 30min at room temperature. The eluate was transferred to a new 1.5ml Eppendorf tube. Input samples were made up to 120 µl with Elution Buffer and subsequently processed like the IP samples. 4 µl 5M NaCl was added to both IP and input samples and crosslinks were reversed overnight at 65°C. Samples were treated with 0.5ug RNaseA (Thermo Scientific, EN0531) (1h at 37°C) and 20ug Proteinase K (Qiagen, 1019499) (1h at 45°C) following which DNA was purified using the QIAquick PCR Purification kit (Qiagen, 28104) following the manufacturer's instructions (pH was adjusted before binding to the column by addition of 5 µl 4M NaAc pH5.2). DNA was eluted in 55 µl nuclease-free water and collected in 1.5 ml DNA Lobind Eppendorf tubes.

If ChIP material was to be used solely for real-time PCR analysis, the protocol was stopped at this stage and purified DNA was diluted and directly used in PCR reactions (see section 2.8 below).

2.7.3 ChIP-seq library preparation and sequencing

For sequencing, further sonication of the ChIP DNA was carried out in order to obtain DNA fragments in the 100-300bp range. Without this step, the significant amount of larger molecular weight DNA that remained after chromatin shearing would interfere with quantification and downstream amplification. Sonication was carried out in a Bioruptor sonication system (Diagenode, UCD-200) set at full power for thirty {30s ON/30s OFF} cycles in a water-bath chilled to 4°C. Following this, DNA was lyophilised using a SpeedVac concentrator (Savant, ISS110) set on medium warmth for 30-40min and dissolved in 26 µl nuclease-free water. DNA concentration in 1.5 µl aliquots was measured using a Qubit dsDNA Assay quantitation kit (Thermo Scientific, Q32851) and a Qubit 3.0 fluorometer (Thermo Scientific) following manufacturer's instructions. Repeats of the same immunoprecipitation

using the same batch of crosslinked chromatin were pooled in order to obtain the 10ng necessary for submission to the Oxford Genomics Centre. Prior to submission, a final sample was removed to validate the enrichment of DNA fragments mapping to known transcription factor binding sites in a real-time PCR assay.

Library preparation was performed at the Oxford Genomics Centre and the libraries were sequenced on an Illumina HiSeq4000 instrument to generate 75bp paired-end reads.

Antibody	Source (lot #)	Agarose beads used to capture immune complexes	Species and isotype	DSG double crosslinking step?
Anti-CEBPB	Santa Cruz, sc-150 (lot #E2814)	Protein A	Rabbit polyclonal IgG	No
Anti-CREB-1	Santa Cruz, sc240 (lot #B2509)	Either	Mouse monoclonal IgG1	No
Anti-GABPA	Santa Cruz, sc-22810 (lot #D0715)	Protein A	Rabbit polyclonal IgG	No
Anti-E2F1	Millipore, 05-379 (lot #2750328)	Either	Mouse monoclonal IgGs	No
Anti-KLF4	Santa Cruz, sc-20691 (lot #H2714)	Protein A	Rabbit polyclonal IgG	Yes
Anti-MAX	Santa Cruz, sc-197 (lot #B2414)	Protein A	Rabbit polyclonal IgG	No
Anti-c-Myc	Santa Cruz, sc-764 (lot #I2613)	Either	Rabbit polyclonal IgG	No
Anti-NRF1	Abcam, ab55744 (lot #GR280462-1)	Protein G	Mouse monoclonal IgG1	No
Anti-SP1	Santa Cruz, sc-17824 (lot #G0114)	Protein A	Mouse monoclonal IgG2a	No
Anti-YY1	Santa Cruz, sc-1703 (lot #F0415)	Protein A	Rabbit polyclonal IgG	Yes
Mouse G3A1 IgG1 isotype control	Cell Signalling Technologies, #5415 (lot #9)	Protein G	Mouse monoclonal IgG1	No
Normal Rabbit IgG	Cell Signalling Technologies, #2729 (lot #7)	Protein A	Rabbit polyclonal IgG	No

Table 2.2: Antibodies used in ChIP-qPCR or ChIP-seq experiments

2.8 Real time PCR analyses

Real-time PCR was carried out in 15 µl reactions containing 1µM forward and 1µM reverse primer, 7.5 µl 2x QuantiFast SYBR Green PCR Master Mix (Qiagen, 204056), and 4 µl DNA in 0.1ml Capped Strip Tubes (Qiagen, 981106). Thermal cycling and fluorescence detection were performed in a Qiagen Rotor-Gene Q platform. Samples were heated at 95°C for 5min, followed by 40 cycles at 95°C for 10s and 60°C for 30s after which a melt curve analysis was performed by ramping the temperature from 50°C to 95°C (5s per °C). Crossing threshold (Ct) values were determined for a threshold at which the fluorescence increased linearly. Reactions were carried out in technical duplicates and the Ct value averaged between the two.

2.8.1 Real time PCR analysis of ATAC-seq libraries

For the analysis of ATAC-seq libraries by real time PCR, samples were analysed using an absolute quantitation method. For each primer pair, PCR efficiency was determined by running real-time PCR on a 1:10 dilution series prepared from a known amount of naked genomic DNA. ATAC-sec libraries were diluted to an equal molar concentration and 4 µl was used for each reaction. DNA fragment levels were determined using locus specific primers (Table 2.3) by measuring Ct values in relation to the standard. Values obtained at each locus were made relative to the signal from a background locus.

Locus Name	Forward and reverse primer sequences (5'-3')	Oligo name	Mm10 coordinates
Sepsecs_1	CTCTAGAAGCAGCCGGATGAGG	SK2439	chr5:52669578-52669655
	CGGGTGTCTCCGGCTTAC	SK2440	
Ube2d1_1	GGGTTCCTGGGGACATCTTG	SK2441	chr10:71285048-71285117
	TGAAGCGGATCCAGAAAGTGAG	SK2442	
Tbp_promoter	CGTCATTTTCTCCGCAGTG	SK2300	chr17:15500001-15500084
	GCTGCAGAAGTCGCCTTTAT	SK2301	
Mllt6	CTTGGGTCGCTTTCCATCAAAC	SK2443	chr11:97662221-97662299
	ACAGTAGCGTACAAAAACGCAC	SK2444	
Utf1_1	CTGGGTCCCTAAGGAAAGGACT	SK2445	chr7:139943637-139943716
	CTCGCCTACCTAGTTCCTCTTG	SK2446	
Utf1_2	GCATAACAATGAGGGCTCATCC	SK2447	chr7:139945685-139945762
	TCTTTGGGTTAGCAGTAAGGGC	SK2448	
Sepsecs_2	TAGGTAACTGCTTCAGTGGGAG	SK2449	chr5:52644498-52644585
	GAAGCCTTGGAAGTCAGCCTTA	SK2450	
Mir6927	AGAACTCCAGCTGCTGAAAAGA	SK2451	chr11:97671298-97671386
	GAAGTCTGAGCCACCCCATTAC	SK2452	
Ube2d1_2	AATCCGCAATTTCCTGGTCACT	SK2453	chr10:71310476-71310550
	ACAACAAAAGATGAGCAAAGGC	SK2454	
negative control Chr1	CATAGATGAAGCTGCCACATAGGT	SK2374	chr1:157832873-157832953
	GTGGGCAAGGACAAAGCATTA	SK2375	
negative control Chr19	CTCCTGGGTCATCCTGCTATTC	SK2455	chr19:59,638,068-59,638,712
	CCCATGAGCCCATAAGGTTTGA	SK2456	
negative control Chr2	AAGTGAGTATTGTCAACTGGCT	SK2457	chr2:107,656,953-107,657,537
	TACATGGTGTATCCCAGAAGGG	SK2458	

Table 2.3: Primers used in ATAC-qPCR experiments

2.8.2 Real-time PCR analysis of ChIP-seq DNA

ChIP and input material obtained after DNA purification were diluted equally in all samples and 4 µl DNA was used for analyses using locus specific primers (Table 2.4). For each primer pair, PCR efficiency was determined in each experiment by running real-time PCR on a 1:10 dilution series prepared from a ChIP input sample. To ensure that measured Ct values fall within the range covered by the standard, all analysed input samples were diluted a further 5x compared to ChIP samples and run in reactions independent to the standards. The difference in Ct values between the immunoprecipitated and input samples was used to calculate the percentage enrichment over input (% input value) for each sample and primer

pair relative to the PCR efficiency. For each condition, the % input values obtained for one primer pair were normalised to that obtained at a reference locus.

Locus Name	Forward and reverse primer sequences (5'-3')	Oligo name	Mm10 coordinates
YY1-1	GGAGAAGAGGCTACGCTTGA	SK1296	chr1:93478798-93478902
	GACGCGATCCTGCCGTC	SK1297	
YY1-2	CCTGAGTGCCTCGCACC	SK1298	chr2:30416123-30416192
	TAGGACCGGAAGCCGAGAA	SK1299	
YY1-3	AGAGATGGGGAAGCAGACCA	SK1724	chr12:85620976-85621073
	GTGCTTGGAGAGCGGAGATT	SK1725	
YY1-4	CCAGCGCCGTTTCTATGGTT	SK1722	chr4:137048624-137048697
	TCCTCAATGGCGCTGTCAAT	SK1723	
KLF4-1	GTTCCCACCGCCTATTTTGA	SK1779	chr8:105852032-105852102
	TGTTGGACTACAGTTCCCGT	SK1780	
KLF4-2	TTGGCCGATTCCTTACTGGG	SK1781	chr7:118656822-118656895
	GCTCTGGTTACAGCCCACTC	SK1782	
KLF4-3	TCTAATTTTGGAGGCCACACCT	SK1783	chr6:125433830-125433926
	GAGCATTACGCATACTCACGC	SK1784	
KLF4-4	CCATTTATCCCCAGCCGACA	SK1785	chr5:139484423-139484498
	TTGGTGTGGCGTCATCAGAG	SK1786	
Bg-3	GAGAGGTCAGAAGTCCCCCT	SK1677	chr7:101431761-101431856
	TCAGGTCAGTGAGTCAGGGT	SK1678	
Bg-2	CCTCTGGTGCTGAATGAGGG	SK1063	chr2:173106089-173106149
	GAAGTGCCGATACTGTCCCG	SK1064	
Bg-1	GGGGAAGTCAGTGAGCCTTA	SK1726	chr4:137051229-137051300
	TTAGGGATGCTCAGTGTCACG	SK1727	

Table 2.4: Primers used in the ChIP-qPCR experiments presented in this thesis

2.9 Data analysis of high-throughput sequencing experiments

2.9.1 General notes about genome builds and data analysis software

The methods described below refer to the analysis of sequencing data obtained from RNA-seq, ATAC-seq, ChIP-seq and Calibrated Native H3ac ChIP-seq experiments. The sample preparation and sequencing steps for these experiments were described above. The mouse mm10 genome build was used throughout. Frequently used programs or tool suites include:

Bedtools (v2.25.0) (http://bedtools.readthedocs.io/en/latest/index.html),
Deeptools (v2.4.2) (http://deeptools.readthedocs.io/en/latest/, Ramírez et al., 2016),
Samtools (v1.3) (http://www.htslib.org/doc/samtools-1.3.html),
HOMER (v4.7) (http://homer.ucsd.edu , Heinz et al., 2010), and
R version 3.4.0 (https://www.R-project.org/).

2.9.2 Quality control, mapping and filtering

The quality of sequencing data was assessed using FASTQC (v0.10.1) in default mode.

RNA-seq 50bp paired-end reads were aligned to the mouse mm10 genome to generate .BAM files using TopHat (v2.0.13) (Trapnell et al., 2009) with the --library-type option fr-firststrand, --mate-inner-dist set to 100 and coverage based searching for junctions disabled. TopHat was run in conjunction with bowtie2 version 2.2.5.0.

ATAC-seq 75bp paired-end reads were also aligned using bowtie2 but using the options --no-mixed and --no-discordant to keep only concordant paired-end alignments against the forward reference strand. ATAC-seq fragments mapping to a custom “blacklist” of artificially high regions of the genome, obtained from Rob Klose’s lab, were discarded.

ChIP-seq 75bp paired-end reads were aligned to the mouse mm10 genome using bowtie2 (v2.2.5.0) using default parameters.

Samtools was used to compress .SAM alignment files to .BAM format, sort alignments according to genomic coordinates, and index .BAM files. Duplicate reads were removed using the Picard Tools (v1.105) function 'MarkDuplicates' with the REMOVE_DUPLICATES option set to "true". Multiply mapped reads were excluded by applying a mapping quality (MAPQ) filter of 10 using Samtools. The Bedtools 'bamtobed' tool was used to generate .BED files that contain the genomic coordinates for every fragment in each library for which both mate pairs aligned concordantly.

For the Native H3ac ChIP-seq experiment using exogenous calibration, sequencing libraries contained DNA fragments originating from both the mouse and spiked-in chicken genomes. Paired-end reads were aligned to a concatenated mm10 + galGal4 genome, containing sequences from both genomes, using bowtie2 with the options --no-mixed and --no-discordant. Reads that had multiple alignments (SAM tag "XS:" present) were discarded and duplicate reads were removed. The number of reads mapping to the mm10 or galGal4 genome were counted and used to calculate the ratio of galGal4 to mm10 reads for both IP and inputs samples. To adjust for the variation in mouse cells used for the four experimental conditions from a single biological replicate, the number of galGal4 reads in the input samples were randomly downsampled so that the resulting mm10:galGal4 ratio in the four input samples was equal (see Table 4.9 in Chapter 4). galGal4 reads from each IP sample were then downsampled applying the same downsampling factor used for their respective input. This adjusted number of galGal4 was used when normalising the ChIP-seq signal to the exogenous spike-in. The Picard Tool 'DownSampleSam' was used to downsample .BAM alignment files.

For all experiments, alignments obtained from biological replicate samples were merged using the Samtools 'merge' command. These were used to create coverage files for visualisation and in metaplot analyses.

2.9.3 Generation of bigWig genome coverage files

In order to visualise ChIP-seq, ATAC-seq or gene expression data in the UCSC Genome Browser, we generated and used bigWig files. These files contain read coverage information for each genomic position. BigWig files were created for data derived from single biological replicates as well as data obtained after merging alignments from biological replicate samples. The Deeptools function 'bamCoverage' was used to generate these files taking .BAM alignment files as input.

For RNA-seq data, the length of each read was not extended (default mode), bin size was set to 1bp (-bs option) and RPKM was used to normalise the number of reads per bin to the total number of aligned reads (--normalizeUsingRPKM option). The formula to calculate Reads Per Kilobase per Million (RPKM) was:

RPKM = number of reads per bin / (number of mapped reads (in millions) * bin length (kb))

In addition to creating bigWig files for all alignments irrespective of strandedness, the --filterRNAstrand option was used to generate two distinct bigWig files for each sample, each containing data for reads split by their strand of origin.

For ATAC-seq data: read length was not extended (default mode), bin size was set to 5bp (-bs option) and RPKM was used to normalise the number of reads per bin to the total number of aligned reads (--normalizeUsingRPKM option).

For ChIP-seq data: paired-end reads were extended to match the fragment size defined by the two mate-reads using the -e option; bin size was set to 2bp (-bs option), and RPKM was used to normalise the number of reads per bin to the total number of aligned reads (--normalizeUsingRPKM option).

For H3ac Native ChIP-seq data, RPKM normalised bigWig files were generated from mm10 reads in the same manner as other ChIP-seq data. However, in order to compare quantitative ChIP-seq signal between different conditions, independent bigWig files were created where the number of mm10 reads per bin was normalised to the total number of galGal4 reads. To achieve this, the Bedtools function genomeCoverageBed (v2.26.0) was used to generate a bedGraph file. The option –pc was used for extending reads to match the fragment size. Normalisation to the exogenous genome was achieved by multiplying the number of reads at each position by a scaling factor calculated as:

$1 / (\text{number of galGal4 after downsampling} / 10^{-6})$

The UCSCtools (v1.0) command 'wigToBigWig' was subsequently used to convert the bedGraph file to bigWig.

Note that the use of scaling factors in the generation of bigwig files is not the optimal approach to allow for comparison between samples, as it does not correct for subsampling issues. A more appropriate method is to first randomly downsample BAM files containing the mm10 alignments in each sample based on the number of reads mapping to the galGal4 genome for that sample (in a manner which results in an equal mm10:galGal4 reads ratio in all samples). BigWig files can subsequently be generated from the downsampled BAM files without scaling.

2.9.4 Peak calling

Tn5 hyper-accessible sites and regions occupied by transcription factors were identified from paired-end ATAC-seq and ChIP-seq data respectively using the dpeak peak-calling algorithm of DANPOS2 version 2.2.2 (Chen et al., 2013). The options -kd (minimal distance between two adjacent peaks) and –kw (minimal peak pidth) were set to 250bp and 150bp respectively. The dpeak tool accepts alignment data from multiple biological replicates and makes use of

experimental input samples to measure background signal. The peak calling threshold was set at a significance level of $p < 10^{-40}$ and $p < 10^{-100}$ for ATAC-seq and ChIP-seq peaks respectively. These values were chosen after comparing the peak regions obtained using a range of significance thresholds to the coverage data as visualised in the UCSC genome browser.

To identify regions enriched in H3ac from aligned Native ChIP-seq data, we used the MACS2 (v2.0.10, https://github.com/taoliu/MACS) (Zhang et al., 2008) function 'callpeak' with the genome size set to 1.87×10^9 as recommended for the mouse genome. Peak calling was performed independently for each biological replicate using all uniquely-mapped mm10 reads. Regions identified as enriched in more than one sample from any condition were kept for downstream analyses.

2.9.5 Annotation of genomic regions

Genomic regions, for example ChIP-seq or ATAC-seq peaks, were annotated using the HOMER script annotatePeaks.pl (Heinz et al., 2010) with the option -CpG and the genome argument set to 'mm10'. This script was used to associate genomic intervals to their nearest gene transcription start site, assign regions to genomic features (such as Exon (Coding), 5' UTR Exon, 3' UTR Exon, Intronic, or Intergenic), identify overlaps with repetitive elements and calculate percent CpG and GC contents. The gene annotation files for mm10 or galGal4 used in this study were downloaded from UCSC. The positions of interspersed repeats and low complexity DNA sequences in the mm10 genome were obtained from UCSC (created using RepeatMasker, from the RepBase library of repetitive elements (Jurka et al., 2005)).

2.9.6 Read quantitation at defined genomic intervals and differential analyses

2.9.6.1 Differential gene expression analysis (RNA-seq data)

FeatureCounts version v1.4.5-p1 (Liao et al., 2014) was used to count aligned RNA-seq reads that mapped to the exons of genes or transcripts annotated in a .GTF file (downloaded from UCSC). Read counting was strand specific (option -s set to 2). If read counting was performed for individual genes (option -g set to "gene_id"), reads which mapped to multiple genes were discarded (default mode). If read counting was performed for individual transcripts (option -g set to "transcript_id"), reads mapping to multiple features were kept (option -O): using the default mode, all reads mapping to exons shared by multiple transcripts of the same gene would be discarded.

Differential analysis was performed using the R package edgeR (Robinson et al., 2010; McCarthy et al., 2012) with raw read counts supplied as input. Genes with a low level of expression in all samples were filtered out using a counts per million (cpm) threshold of 1. The generalized linear model (GLM) likelihood test was used to determine differential expression between multiple groups of biological replicate samples. Genes or transcripts were considered as significantly differentially expressed if they had a p-value < 10^{-5} after adjusting for multiple comparisons using the Benjamini and Hochberg procedure.

2.9.6.2 Differential analysis of transcription factor occupancy (ChIP-seq data) or Tn5 accessibility (ATAC-seq data)

The R package DiffBind (Stark and Brown, 2011) was used to identify ChIP-seq or ATAC-seq peak regions with significant changes in enrichment between experimental conditions. These are indicative of differential transcription factor occupancy or Tn5 sensitivity respectively. In both cases, significant changes were determined using the DBA_DESEQ2 method.

2.9.6.3 Quantitation of ATAC-seq reads at other genomic intervals

Samtools was used to count the total number of ATAC-seq alignments that mapped to genomic intervals defined in .BED format (eg. .BED file containing the coordinates of all ATAC-seq peaks or all RepMask elements). Bedtools 'multicov' was used to count the number of ATAC-seq alignments that mapped to each genomic interval defined in .BED format (eg. .BED file containing the coordinates of each chromosome).

The number of ATAC-seq fragments fully overlapping with PCR amplicons (defined as the interval between the forward and reverse primer (see Table 2.3)) was determined using the Bedtools command "intersect" with the options -f 1 and -c. This counted the number of entries in a .BED file, which contain the genomic coordinates for every fragment in each library for which both mate pairs aligned concordantly, that fully overlap genomic intervals specified in a separate .BED file (containing the coordinates of PCR amplicons).

2.9.6.4 Differential analysis of H3ac enriched regions with exogenous calibration

The featureCounts tool (version v1.4.5-p1) (Liao et al., 2014) was used to count the number of mouse or chicken reads, derived from the H3ac ChIP-seq experiment, that mapped to each genomic intervals defined in a .BED file. Genomic regions with differential H3ac enrichment were determined using the R package DeSeq2 (Love et al., 2014). First, normalisation factors for each sample were calculated by providing DeSeq2 with read count data obtained by quantifying the number of galGal4 reads that map to gene bodies in the chicken genome. These normalisation factors were then used to scale the read count data obtained by quantifying the number of mm10 reads that map to the genomic regions of interest. In this way the H3ac ChIP-seq signal in each region is normalised to the number of reads originating from the calibration genome. Subsequently the standard DeSeq2 differential analysis pipeline was followed. A p-value threshold of < 0.05 was used to identify significant differences after

correction for multiple testing using the Benjamini & Hochberg procedure and disabling independent filtering. The R script used for this analysis is provided in the Appendix.

2.9.6.5 Quantification of reads mapping to repetitive genomic features and differential analysis.

This section describes the analysis workflow used to first count the number of sequencing reads that mapped to interspersed repeats or low complexity DNA sequences and then perform differential analysis for each repeat type. The same workflow was used for RNA-seq, ATAC-seq and ChIP-seq data. For each sample, total read counts for every type of repetitive element were generated using a custom script (provided in Appendix). All alignments, including those ascribed to multimapping reads, were taken into account. The position and annotation of interspersed repeats and low complexity DNA sequences in the mm10 genome was obtained from UCSC (created using RepeatMasker). The custom script makes use of the Samtools "view" command to count reads at genomic intervals, the Bedtools "nuc" command to determine the CpG content of genomic intervals and the 'st' tool (https://github.com/nferraz/st) to calculate CpG content and nucleotide substitution statistics across all copies of a repeat element type. The R package DeSeq2 (Love et al., 2014) was used to identify repeat element types that show significant differences in coverage between conditions. A p-value threshold of < 0.05 was used to identify significant differences after correction for multiple testing using the Benjamini & Hochberg procedure.

2.9.7 Analysis of sequence motifs bound by transcription factors.

The analysis of transcription factor binding motifs within sets of genomic intervals was performed using multiple scripts within the HOMER tool suite HOMER (v4.7) (Heinz et al., 2010). HOMER works with specifically formatted sequence motifs. These consist of a probability weight matrix (PWM) which, for each position within a sequence of a given length, specifies the probability of finding each nucleotide A, C, G or T. Any given sequence can be

scored against a particular PWM by summing up the log-odds probability for each nucleotide to be present at a specific position in the sequence. In HOMER, motifs are associated with a detection threshold; a sequence is said to match a given motif if the sum of its log-odds probabilities is greater than this threshold.

HOMER motifs were obtained from three main sources:

1) a database of 264 known vertebrate transcription factor binding motifs included in HOMER and mostly derived from pre-existing ChIP-seq datasets (known motifs);
2) motifs for specific sequences were created using the seq2profile.pl script (custom motifs);
3) motifs identified as enriched in particular genomic regions through *de novo* discovery.

In this study, *de novo* motif discovery was carried out using the findMotifsGenome.pl script to define motifs enriched in a set of chosen target regions against a background of regions automatically generated by HOMER or specified using the -bg option. The same script was used to screen for the enrichment of known or custom motifs in target versus background regions. Default parameters were used although the size of the target regions was specified either as the given lengths of the provided genomic regions (-size option set to “given”) or as a constant (-size option set to the desired number of bp surrounding the centre of a genomic interval). The latter was used when screening for the enrichment of known motifs in the region surrounding small sequences identified as potential transcription factor binding sites (see Chapter 5 - Section 5.5.3). The hypergeometric distribution was used to calculate the significance of motif enrichment (default mode in HOMER script).

Sequences that matched a particular PWM motif were identified genome-wide or within a subset of genomic intervals using the scanMotifGenomeWide.pl and annotatePeaks.pl scripts respectively. The number of sequences that matched a given PWM motif within each interval in a .BED file was determined using the annotatePeaks.pl script with the options -nmotifs (and -rmrevopp to avoid double counting reverse opposites). The maximum log-odds score for a

given PWM motif within each interval in a .BED file was determined using the annotatePeaks.pl script with the -mscore option. Peak centering onto a specific motif was used to examine the spatial relationship between the motif and other surrounding sequence features and was carried out using the annotatePeaks.pl script with the -center option and the -size argument set to "given". Finally the distribution of sequences that match a given PWM relative to the centre of a genomic interval was determined using the -hist option of the annotatePeaks.pl script.

2.9.8 Plotting and statistics

Heatmaps and metaplots used to display and summarise ATAC-seq, ChIP-seq or Native H3ac ChIP-seq signals were generated using Deeptools functions (Ramírez et al., 2016). As input, these tools were provided with genome coverage files (bigWigs) created after merging alignments from biological replicate samples.

R release 3.4.0 was used for data processing, formatting, statistical analyses and generating figures. The graphical functions of the ggplot2 package were used to create many figures.

The Pacific Northwest National Laboratory Venn Diagram Plotter tool (https://omics.pnl.gov/software/venn-diagram-plotter) was used to draw Venn diagrams.

2.9.9 Accession codes for publicly available datasets

Previously published datasets used for analysis include RNA-seq data from HA36CB1/159-2 wild-type mES cells and corresponding DNMT.TKO mESCs (GSE67867) (Domcke et al., 2015); RNA-seq data for wild-type J1 and DNMT.TKO mESCs (the same cell lines as used in this study) (GSE29413 , Karimi et al., 2011) and whole genome bisulfite sequencing data from E14 mouse ESCs (GSM1027571, Habibi et al., 2013).

The whole genome bisulfite sequencing data was downloaded and analysed using custom scripts written by a fellow student, Marketa Tomkova. BedGraph files were obtained that contained methylation values for each cytosine in CpG context and the read coverage. Only cytosine with a read coverage of five or more were taken into account in this study.

Chapter 3:

Investigating the relationship between the roles of DNA methylation and HDAC activity in gene expression

3.1 Introduction

Functional experiments that have disrupted DNA methyltransferase through genetic manipulation or the use of small inhibitors have shed a light on the biological role of DNA methylation. A consensus view is that DNA methylation plays an important role in maintaining the transcriptional repressed state of imprinted genes (Kaneda et al., 2004), a selection of tissue-specific genes (De Smet et al., 1999; Maatouk et al., 2006; Borgel et al., 2010), evolutionarily young repetitive elements (Davis et al., 1989; Walsh et al., 1998) and genes on the inactive X chromosome in female somatic cells (Beard et al., 1995; Hansen et al., 1996) (see also Section 1.4 in Chapter 1). A number of mechanisms have been put forward to explain how local DNA methylation influences transcription. Locus specific studies and *in vitro* experiments implicate an effect on histone modifications (Jones et al., 1998; Fujita et al., 2003; Sarraf and Stancheva, 2004) and/or direct affects on the binding of site-specific transcription factors to regulatory elements (Hu et al., 2013; Jones and Chen, 2006) (see Section 1.5).

One of the key mechanisms that have been put forward involves the recruitment of HDAC-containing protein complexes to methylated loci. At these regions, HDAC activity likely promotes the formation of a transcriptionally repressive chromatin environment. This model was developed following the identification of multiple proteins that bind with high affinity and specificity towards methylated CpGs (such as MBD-domain proteins or Kaiso) while stably interacting with HDAC enzymes (Nan et al., 1998; Zhang et al., 1999; Kokura et al., 2001; Yoon et al., 2003). Importantly, HDAC inhibition was shown to alleviate the transcriptional silencing mediated by DNA methylation (Jones et al., 1998; Nan et al., 1998; Ng et al., 1999). Nevertheless the contribution of HDAC activity to the repressive function of DNA methylation *in vivo* has not been fully resolved (see Section 1.5.2). To address this, we aimed to compare the transcriptional alterations that result from either the loss of DNA methylation or histone

deacetylase activity. We expected that HDAC inhibition would mirror the transcriptional effects of DNA methylation loss for transcriptional elements at which HDAC activity is required for DNA methylation mediated silencing.

In order to assess the impact of DNA methylation deficiency we chose to compare the transcriptional landscape of *Dnmt3a*, *Dnmt3b* and *Dnmt1* triple knockout (DNMT.TKO) mouse embryonic stem cells (Tsumura et al., 2006) to their corresponding wild-type cells. DNMT.TKO mES cells were generated through the stepwise genetic disruption of the DNMT genes: first *Dnmt3a* and *Dnmt3b* in 1999 (Okano et al., 1999) and subsequently *Dnmt1* (Tsumura et al., 2006). The original genetic deletion of *Dnmt3a* was performed using the J1 mES cell line and these have been used as the reference wild-type cells against which DNMT.TKO cells are compared (Fouse et al., 2008; Karimi et al., 2011). DNA methylation levels in DNMT.TKO cells are below detection level (Tsumura et al., 2006). The fourth enzymatically active DNA methyltransferase DNMT3C, recently identified in rodents, was found not to be expressed in this cell line (Barau et al., 2016) and our analysis confirms that the longer, enzymatically active isoform is not transcribed (data not shown). DNMT.TKO and J1 cells show similar morphology, stain positively for alkaline phosphatase activity, express pluripotency genes to similar levels and grow at comparable rates when co-cultured (Tsumura et al., 2006). Similarly to *Dnmt1*$^{-/-}$ ES cells, DNMT.TKO cells show reduced growth and increased rates of apoptosis following the induction of differentiation through LIF removal (Panning and Jaenisch, 1996; Tsumura et al., 2006).

We chose to disrupt HDAC activity using Trichostatin A (TSA) (Yoshida et al., 1990), an inhibitor of class I, II and IV HDACs such as those belonging to the NurD, NCoR or Ski complexes that interact with MBD-containing proteins (Huang et al., 2000). TSA has widely been used by multiple studies to alleviate the repressive activity of DNA methylation (Nan et al., 1998; Schübeler et al., 2000; Lyst et al., 2013). *Hdac1* knockout mESCs have also been used to

address the effect of HDAC inhibition on the repression of transposable elements, however only small changes in global histone acetylation were observed compared to those achieved through pan-HDAC inhibition with TSA (Brunmeir et al., 2010; Reichmann et al., 2012). A high level of redundancy between different HDAC enzymes likely explains this.

RNA sequencing (RNA-seq) was used to assess the gene expression changes in wild-type J1 and DNMT.TKO mESCs following incubation with TSA or a mock DMSO treatment (Figure 3.1). The results of this RNA-seq analysis are presented in this chapter. First, the gene expression differences between wild-type and DNMT.TKO cells will be described following which these will be compared to the alterations seen after TSA treatment in wild type and DNMT.TKO cells. As DNA methylation and HDACs play important roles in the transcriptional regulation of repeat elements, we examined changes in the abundance of mRNA originating from repetitive elements in the genome, with a special interest for mobile elements (LINEs, SINEs and LTRs).

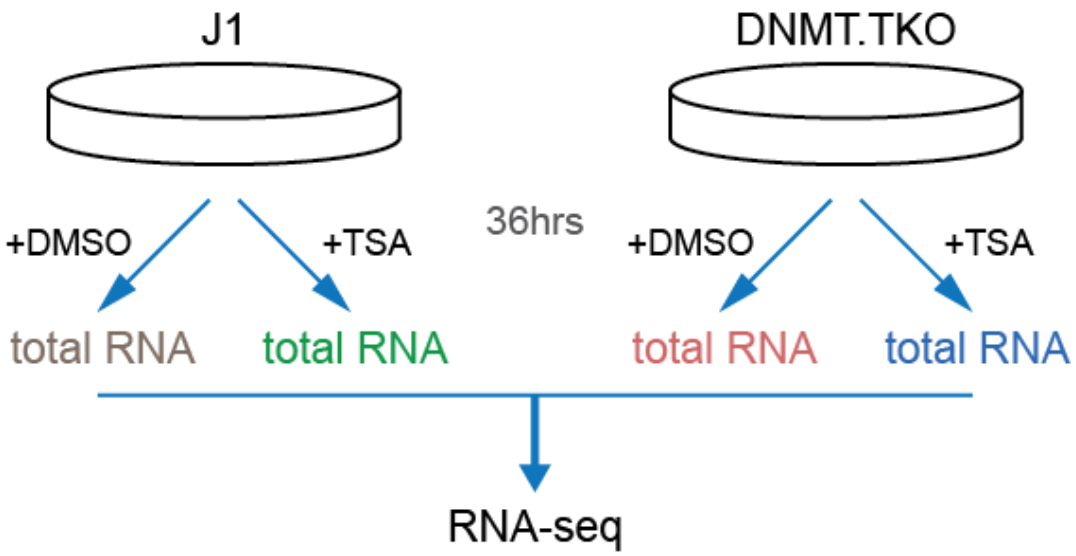


Figure 3.1: Schematic of the experimental approach used to investigate the transcriptional effects of DNA methylation loss and HDAC inactivation.

3.2 Gene expression profile of DNA methylation deficient cells compared to wild type cells

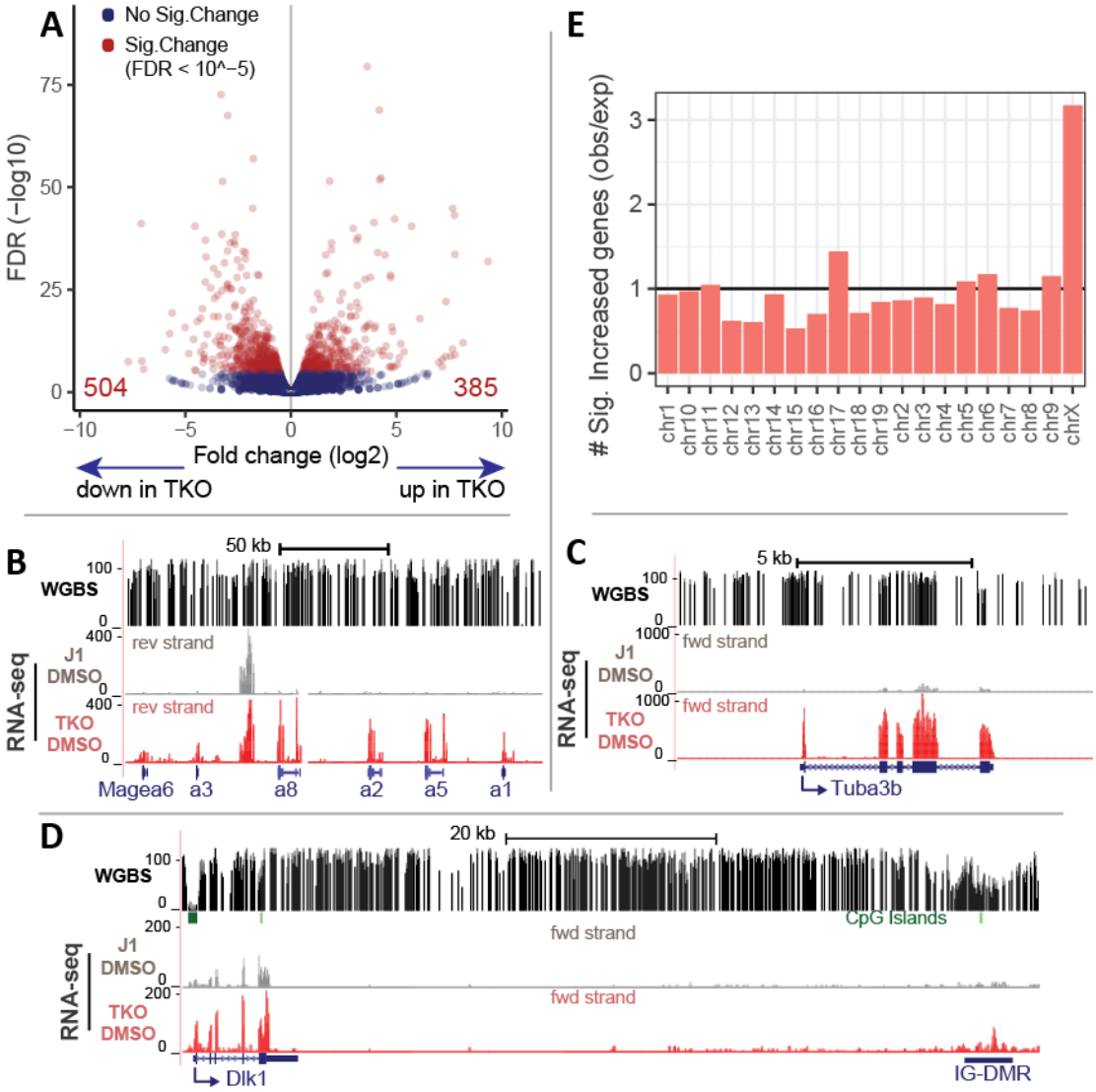
3.2.1 Identification of genes deregulated in DNMT.TKO versus wild-type cells

To identify transcriptional elements targeted by DNA methylation, total RNA was isolated from DNMT.TKO mES cells and their wild-type parental cell line, J1 mES cells. Following ribo-depletion, strand-specific library preparation was carried out enabling us to assign each read to its strand of origin. In order to detect non-coding transcriptional events, no polyA enrichment step was performed. Two biological repeats were performed for each experimental condition yielding on average 44.8 and 48.9 million 75bp paired-end reads for J1 and DNMT.TKO cells respectively (Table 3.1). Paired-end reads were aligned to the mm10 genome assembly and counts were generated for each gene using an UCSC annotation file.

Sample	Total read count	Overall alignment rate (%)	Total aligned mate-pairs	Mate-pairs with multiple alignments (%)	Uniquely aligned reads that map to an exon (%)	Number reads with > 20 alignments
J1.DMSO_REP1	43,231,428	85.4	33,730,128	20.3	64.69	606,280
J1.DMSO_REP2	46,480,210	85.8	35,768,352	19.1	62.38	661,705
J1.TSA_REP1	47,214,255	86	37,177,439	18.6	60.81	821,327
J1.TSA_REP2	47,055,901	86.6	36,591,351	18.5	62.32	1,039,607
TKO.DMSO_REP1	51,882,739	86	40,730,663	20.1	61.93	823,459
TKO.DMSO_REP2	45,921,222	84.2	34,660,303	19.7	59.26	661,129
TKO.TSA_REP1	45,494,078	88.7	36,918,590	21.5	60.81	1,847,368
TKO.TSA_REP2	43,231,428	84.2	38,288,866	21.6	60.13	2,117,320

Table 3.1: Total read counts and alignment metrics for RNA-seq samples

Differential gene expression analysis was carried out using the R package edgeR (Robinson et al., 2010; McCarthy et al., 2012). We identified 889 out of 15222 genes that showed significantly differential expression (FDR-adjusted $p<10^{-5}$) with 385 and 504 genes showing increased or decreased expression in TKO cells relative to wild type J1 cells (Figure 3.2A).



A. Volcano plot comparing the false discovery rate (FDR) significance value to the fold change in transcript level for every genes (n = 15222). Differential gene expression analysis was performed using the R package edgeR (Robinson et al., 2010); a FDR threshold of 10^{-5} was used to call significantly differentially expressed genes. **B-D.** UCSC genome browser snapshots showing examples of genes upregulated in DNMT.TKO versus wild-type cells. RNA-seq signal for DMSO-treated wild-type J1 and DNMT.TKO cells is displayed along with wild type CpG methylation obtained using whole genome bisulfite sequencing (WGBS) in E14 mESCs (Habibi et al., 2013). Only reads originating from the sense strand are shown. Coverage data from replicate samples are overlayed. RPKM normalisation was used to account for variations in library size. UCSC gene annotations and the position of the *Dlk1* Intergenic-DMR are shown. Mm10 coordinates – **B:** chrX:154,916,186-155,106,626 ; **C:** chr6:145,612,452-145,624,305 ; **D:** chr12:109,452,356-109,533,829. **E.** Bar graph representing the number of genes significantly upregulated per chromosome (Observed) relative to the expected number (Expected) based on the total number of genes per chromosome. The Observed/Expected ratios for each chromosome were calculated as follows: Observed = Number of genes on chromosome that are significantly upregulated in DNMT.TKO cells / Total number of genes significantly upregulated in DNMT.TKO cells. Expected = Total number of genes on chromosome / Total number of genes.

Amongst the genes found to be reactivated in DNMT.TKO cells were members of the *Magea* and *Rhox* gene clusters (Figure 3.2B) and other germ-line specific genes such as *Fkbp6*, *Nlrp4c* or *Tuba3b* (Figure 3.2C) known to be silenced by DNA methylation (Karimi et al., 2011) (Appendix - Table S1). Some known imprinted genes for example *Peg13*, *H19*, *Dlk1*, *Grb10*) (Xie et al., 2012) were also found to be derepressed in TKO cells confirming that their silencing relies on DNA methylation in mouse ES cells. (Figure 3.2D (*Dlk1)*). Genes located on the X chromosome were overrepresented in the list of genes significantly upregulated in DNMT.TKO cells (Figure 3.2E). Only for the X chromosome did the number of DNMT.TKO-upregulated genes deviate significantly from the expected number obtained by chance (p value < 0.001 according to a hypergeometric distribution based on total of 15222 genes)

Our data correlates with similar RNA-seq data published by Karimi *et al.* using the same cell lines (Karimi et al., 2011) (Figure 3.3A) and Domcke *et al.* using an independent derived DNMT.TKO mouse ES cell line (Figure 3.3B) (Domcke et al., 2015). Of the top 100 most significantly upregulated genes in TKO cells compared to J1, 60 were significantly upregulated in the Karimi *et al.* study (z-score >1 , 27 genes could not be compared) and 83 in the Domcke *et al.* study (FDR-adjusted $p<10^{-5}$, 10 genes could not be compared) in DNA methylation deficient cells. These observations gave us confidence that the significant gene expression changes we are seeing are associated to DNA methylation loss and reproducible by different groups.

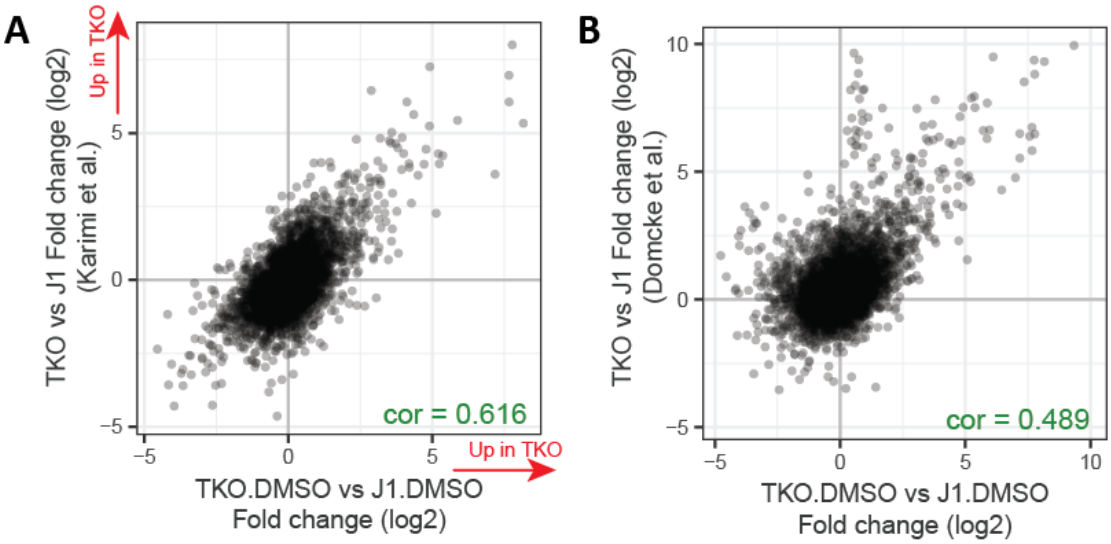


Figure 3.3: Comparison to publicly available RNA-seq data from wild type and DNMT.TKO mESCs. A. Scatter plot comparing gene expression changes in DNMT.TKO versus wild-type J1 as measured by the RNA-seq produced in our lab (x-axis) or that published by Karimi et al., 2011 using the same cell lines (y-axis) (GSE29413). Unlike us, Karimi *et al.* enriched for polyA-tailed RNA molecules and sequenced only one biological replicate. **B.** Scatter plot comparing gene expression changes in DNMT.TKO versus wild-type mESC cells as measured by the RNA-seq produced in our lab (x-axis) or that published by Domcke et al., 2015 (y-axis) (GSE30203). Domcke et al. used a distinct DNMT.TKO cell line, derived from HA36CB1/159-2 wild type mESCs. *cor* indicates Pearson correlation scores.

3.2.2 DNA methylation levels at and around genes deregulated in DNMT.TKO versus wild type cells

In order to appreciate the fraction of differentially expressed genes whose upregulation is a direct consequence of a loss of in DNA methylation we made use of public whole genome bisulfite sequencing (WGBS) data generated from E14 mESC grown in serum + LIF (Habibi et al., 2013). This methylome strongly resembles WGBS data published using a different serum-grown mES cell line (159-2, (Stadler et al., 2011)) (Pearson correlation 0.85 (from Habibi et al., 2013), we therefore assumed it also reflects the DNA methylation landscape in our WT J1 cells. We mainly used the Habibi *et al.* data for our analyses but when these were performed using the WGBS data from Stadler *et al.*, the same conclusions were obtained.

On average, genes that are significantly upregulated in TKO cells have elevated wild-type DNA methylation levels surrounding their transcription start site compared to all other genes (Figures 3.4A-B).

To assess the effect of DNA methylation loss at individual transcripts, we attempted to measure mean methylation across each promoter region. As DNA methylation has been shown to be able to obstruct activation of transcription from non-canonical transcription start sites (Brocks et al., 2017; Neri et al., 2017), we identified the annotated transcript that had the most significant change for each gene and focused on its TSS. We decided not to define promoter regions using the typical -1500bp/+500bp definition, but instead chose a narrower 620bp region surrounding the TSS. This value is the median length of accessible regulatory regions as identified through ATAC-seq in these cell lines (ATAC-seq data will be described and discussed in greater detail in the next chapter).

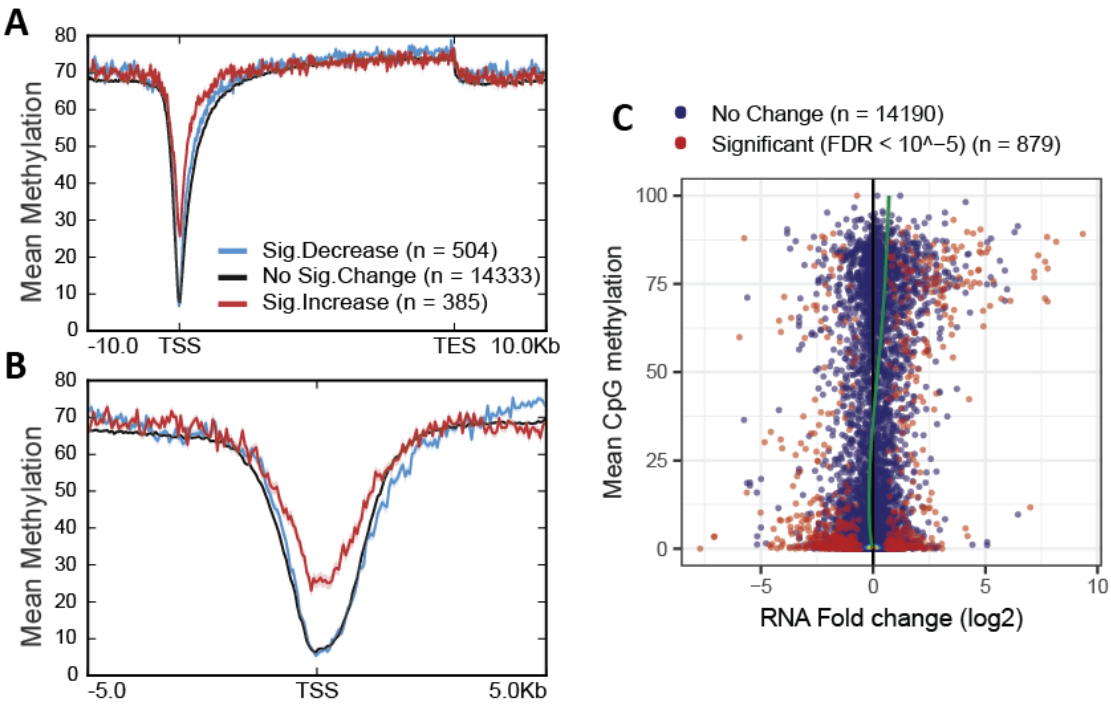


Figure 3.4: Wild type CpG methylation levels at genes that are differentially expressed in wild type and DNMT.TKO mESCs. A. Metaplot summarising mean CpG methylation levels across the gene bodies and the 10kb upstream and downstream. Genes were grouped by their differential expression status in DNMT.TKO versus J1 wild type cells. **B.** Metaplot summarising mean CpG methylation levels in the 10kb regions surrounding gene transcription start sites (TSS). Genes were grouped by their differential expression status in DNMT.TKO versus J1 wild type cells. CpG methylation values were averaged in 50bp windows. **C.** Scatter plot comparing the fold change in gene expression between DNMT.TKO and wild type cells to the average CpG methylation levels at the 620bp regions surrounding each gene's transcription start site. Only genes that have methylation data for more than one cytosine in the defined region were included (n = 15069). The green line represents a smoothened fit of the data modelled using the 'gam' method in the R package 'ggplot2'.

Whilst Figure 3.4C suggests that there is some relationship between methylation at the promoter region and change in expression after loss of DNA methylation, this is the case for relatively few genes. Significantly upregulated genes in TKO had a mean methylation of 31.6 around their TSS with 31% having a methylation value above 60. More reassuringly, 88 of the 100 most upregulated genes in TKO cells had high levels of methylation (≥ 60) around their TSS confirming that these are likely to be directly regulated by DNA methylation.

Unsurprisingly, promoters with high levels of DNA methylation have lower GC and CpG content compared to the majority of promoters (Figure 3.5) (Weber et al., 2007). Methylated promoters which were associated with TKO-specific expression had slightly elevated CpG and GC content compared to those that did not show an expression increase in TKO cells (One-sided Wilcoxon rank sum test) however these promoters rarely resembled CpG islands (Figure 3.5).

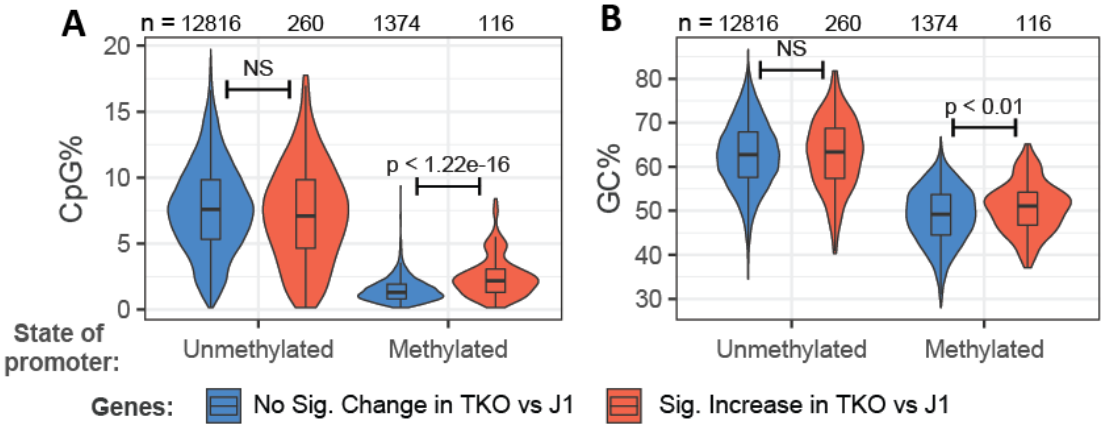


Figure 3.5: CpG and GC content at the promoters of genes grouped by their methylation status and differential expression status. A. Methylated promoters whose associated gene is upregulated in DNMT.TKO versus wild-type cells have a significantly higher CpG content than promoters of genes with no significant differential expression. **B.** Methylated promoters whose associated gene is upregulated in DNMT.TKO versus wild-type cells have a significantly higher GC content than promoters of genes with no significant differential expression. In **A** and **B**, promotes are defined as the 620bp region surrounding transcription start sites. Promoters were deemed methylated if they had an average CpG methylation > 0.6. One-tailed Mann-Whitney U tests were used to determine significant differences in the distributions CpG% or GC%: *ns* = non-significant (p-value >0.05).

3.3 The effect of HDAC inhibition on gene expression in wild-type and DNMT.TKO cells

With the ambition to gain insight into the relative contributions of DNA methylation and HDACs on transcriptional repression, we first sought to define the transcriptional changes that occurred upon HDAC inhibition in both WT cells and DNMT.TKO cells. Both cell lines were treated with 5ng/mL Trichostatin A (TSA, ~16nM) for 36 hours following which RNA-seq was performed on total RNA (Table 3.1). With this dosage, a robust increase in global histone acetylation was observed (Figure 3.6). The treatment did not result in increased lethality although cells presented with a slightly elongated shape compared to control treated mESCs, especially in the DNMT.TKO background.

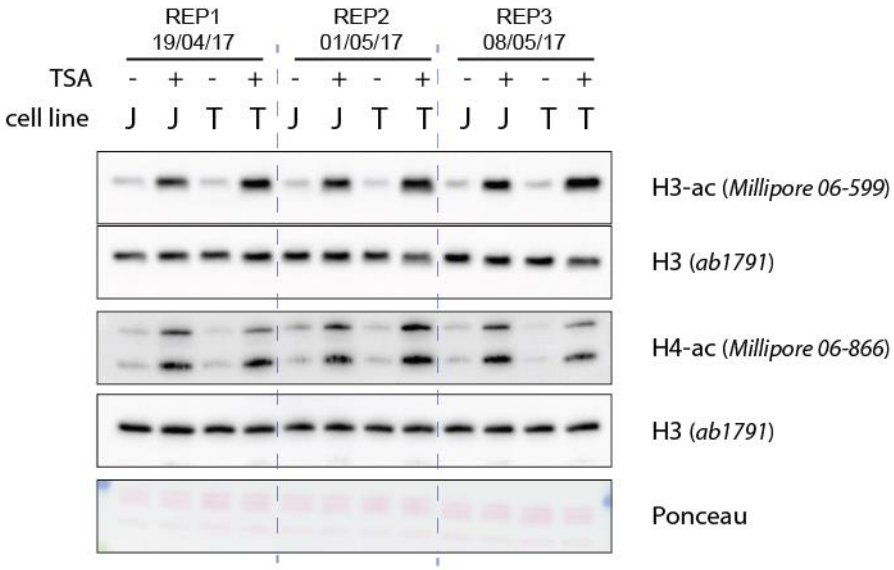


Figure 3.6: Immunoblot analysis of global histone H3 and H4 acetylation levels after Trichostatin A (TSA) treatment of mESCs. In preparation for immunoblotting, two independent SDS-PAGE gels were loaded with identical amounts of extracted histones. After electrophoresis and transfer, the two membranes were stained using antibodies towards pan-acetylated histone tails. Total histone H3 was detected to serve as a loading control for both membranes. Ponceau S staining is also shown for one of the membranes. Samples are derived from three biological replicate experiments. J: samples from wild type J1 mESCs; T: samples from DNMT.TKO mESCs. +: TSA treatment, -: mock DMSO treatment.

Treatment of J1 and DNMT.TKO cells with TSA led to the deregulation of 3060 and 4074 genes respectively compared to control treated cells (FDR-adjusted $p < 10^{-5}$). In both cell types, there was a greater number of upregulated than downregulated genes, which is in accordance with the proposed repressive function of HDACs (Figure 3.7).

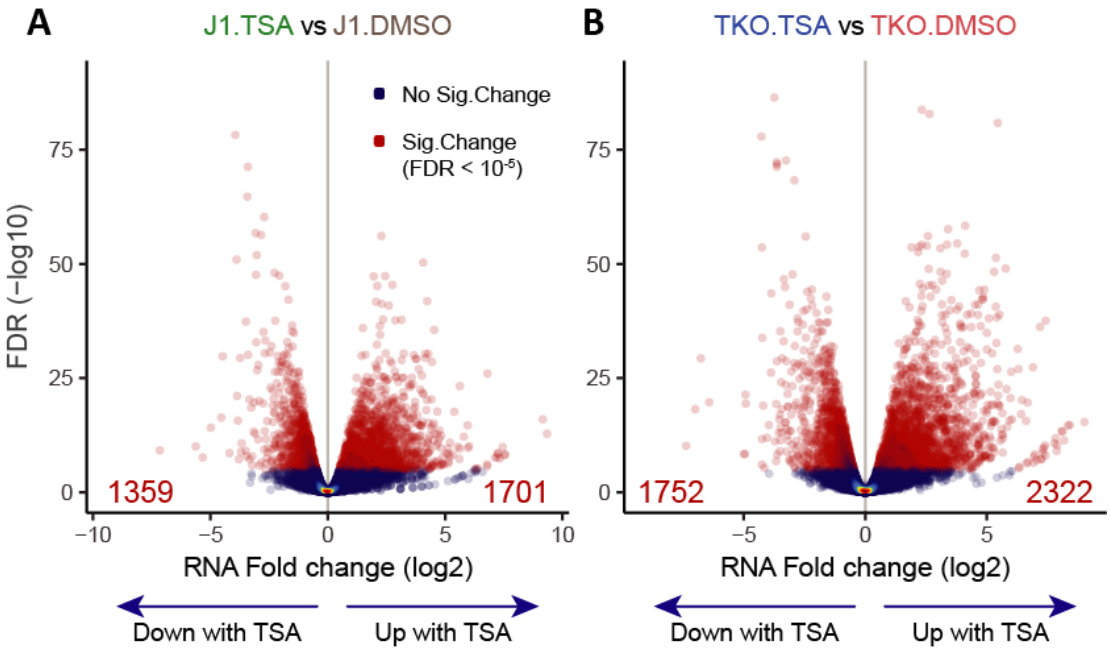


Figure 3.7: Differential gene expression analysis comparing TSA to DMSO treated mESCs. **A.** Differential gene expression analysis comparing TSA to DMSO-treated wild type J1 mESCs. **B.** Differential gene expression analysis comparing TSA to DMSO-treated DNMT.TKO mESCs. Volcano plots comparing the false discovery rate (FDR) significance value to the fold change in transcript level for every gene (n = 15222). Differential gene expression analysis was performed using the R package edgeR (Robinson et al., 2010); a FDR threshold of 10^{-5} was used to call significantly differentially expressed genes.

The genes that respond to TSA treatment appear to be similar in both cell lines with a strong correlation between the fold change values obtained for each gene in either cell type (Figure 3.8A, spearman correlation = 0.865). However more genes are identified as significantly upregulated when treating TKO cells with TSA as opposed to WT cells (Figure 3.8B). Furthermore, the genes that significantly increase in expression after TSA treatment in both

cell lines do so to a greater extent in the DNMT.TKO background (Figure 3.8C). The same can be said about genes that show a decrease in expression.

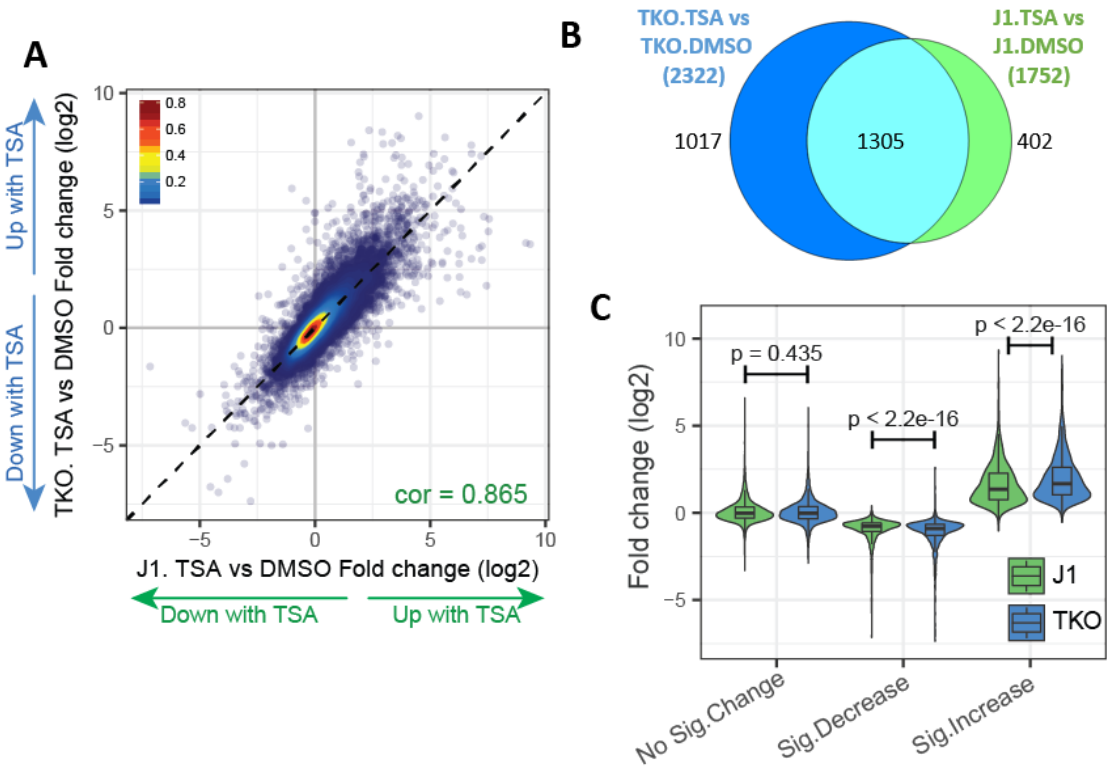


Figure 3.8: Comparison of the transcriptional effect of HDAC inhibition in wild type and DNMT.TKO cells. A. Scatterplot comparing the fold change in gene expression after TSA treatment in DNMT.TKO versus wild type J1 cells. The dashed line has a slope of 1 and an intercepts of 0. The Spearman correlation score is indicated. Fold change values were obtained through differential gene expression analysis carried out for 15222 genes **B.** Venn diagram representing the overlap in genes significantly upregulated upon TSA treatment in DNMT.TKO (blue) versus wild-type (green) cells. The number of genes in each group is indicated. **C.** Violin plots summarise the fold change in gene expression after TSA treatment in wild type J1 (green) or DNMT.TKO cells (blue) for genes grouped according to their differential expression status. The three groups include genes significantly upregulated in response to TSA in either cell line (Sig.Increase, n = 2717), genes significantly downregulated in response to TSA in either cell line (Sig.Decrease, n = 2040) or genes with no significant changes (No Sig.Change, n = 10465). One-tailed Wilcoxon matched-pairs signed rank test were used to determine significant differences in the distributions of gene expression fold changes. The two-tailed version of the test was used for the NoSigChange group.

These results suggest that the TSA-responsive genes are the same in WT and DNMT.TKO cells although the effect of HDAC inhibition in TKO.DMNT cells is quantitatively stronger. Consequently, DNA methylation does not seem to be necessary to direct the bulk of the TSA response.

3.4 Comparing the transcriptional effects of DNA methylation loss and HDAC inhibition

Multiple studies suggest that DNA methylation contributes to gene silencing by recruiting complexes that contain HDACs (Kokura et al., 2001; Nan et al., 1998; Yoon et al., 2003; Zhang et al., 1999). If this were the case, we would expect that genes derepressed following the loss of DNA methylation would also be activated by HDAC inhibition. We explore this hypothesis by comparing the transcriptional changes occurring with genetic removal of DNA methylation and HDAC inhibition with TSA.

The correlation in gene expression changes between these two treatments is weak (spearman correlation = 0.102) suggesting they have overall different transcriptional effects (Figure 3.9A). We found that 99 genes were significantly upregulated in both conditions, an overlap which is significant (p value < 10^{-16} according to a hypergeometric distribution based on total of 15222 genes) but still represents a minority of genes upregulated in either condition (Figure 3.9B).

We explored this further by plotting an additional gene set onto the Venn diagram, one that contains genes upregulated by TSA in DNMT.TKO cells (Figure 3.10A). Interestingly, just under half of the genes upregulated by either loss of HDAC or DNMT are also responsive to TSA in the absence of DNA methylation. It therefore appears that there are few genes whose repression relies solely on DNA methylation recruiting HDAC activity. Indeed, DNA methylation and HDAC inhibition do not seem to be working through the same pathway but rather seem to represent two levels of regulation. We can define three groups of genes: 1)

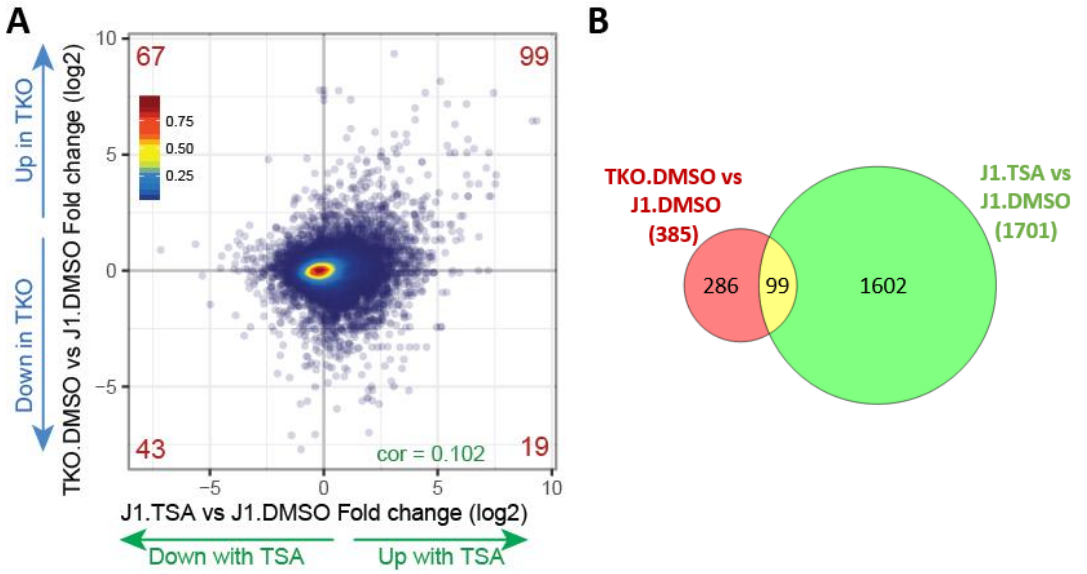


Figure 3.9: Comparison of the transcriptional effect of DNA methylation loss to that of HDAC inhibition in wild type cells. A. Scatterplot comparing the gene expression changes (log2 fold change) obtained after TSA treatment of wild type J1 cells or after removal of DNA methylation (DNMT.TKO versus WT cells). The numbers shown in the corner of each quadrant indicate the number of genes with the same significant change between both treatments. cor indicates the Pearson correlation score. **B.** Venn diagram representing the overlap in genes significantly upregulated upon TSA treatment in wild-type J1 cells (green) or genes significantly upregulated in DNMT.TKO compared to wild-type cells (red). The number of genes in each group is indicated.

genes that respond to DNA methylation loss independently of HDAC activity (Group 1, red population in Figure 3.10A, eg. *Dcdc2a* (Figure 3.11A)), 2) genes that respond to HDAC inhibition independently of DNA methylation (Group 2, blue, green and cyan populations in Figure 3.10A, eg. *Dgat2* (Figure 3.11B)) and genes that respond to either HDAC inhibition or removal of DNA methylation (Group 3, yellow, grey and purple populations in Figure 3.10A, eg. *Slc1a1, Dazl* and *Fkbp6* (Figures 3.11C-E)). Notably the genes belonging to group 3 tend to be more highly expressed after TSA treatment of TKO cells compared to untreated DNMT.TKO cells or TSA-treated J1 cells (Figure 3.10B). This fits with the suggestion that HDAC activity and DNA methylation have independent and additive effects. Expression changes at *Dazl* and *Fkbp6* illustrate this point (Figures 3.11D-E). In each case the effect of TSA adds to the change

that occurred with removal of DNA methylation (TSA-induced increase for *Dazl* (Figure 3.11D), TSA-induced decrease for *Fkbp6* (Figure 3.11E)).

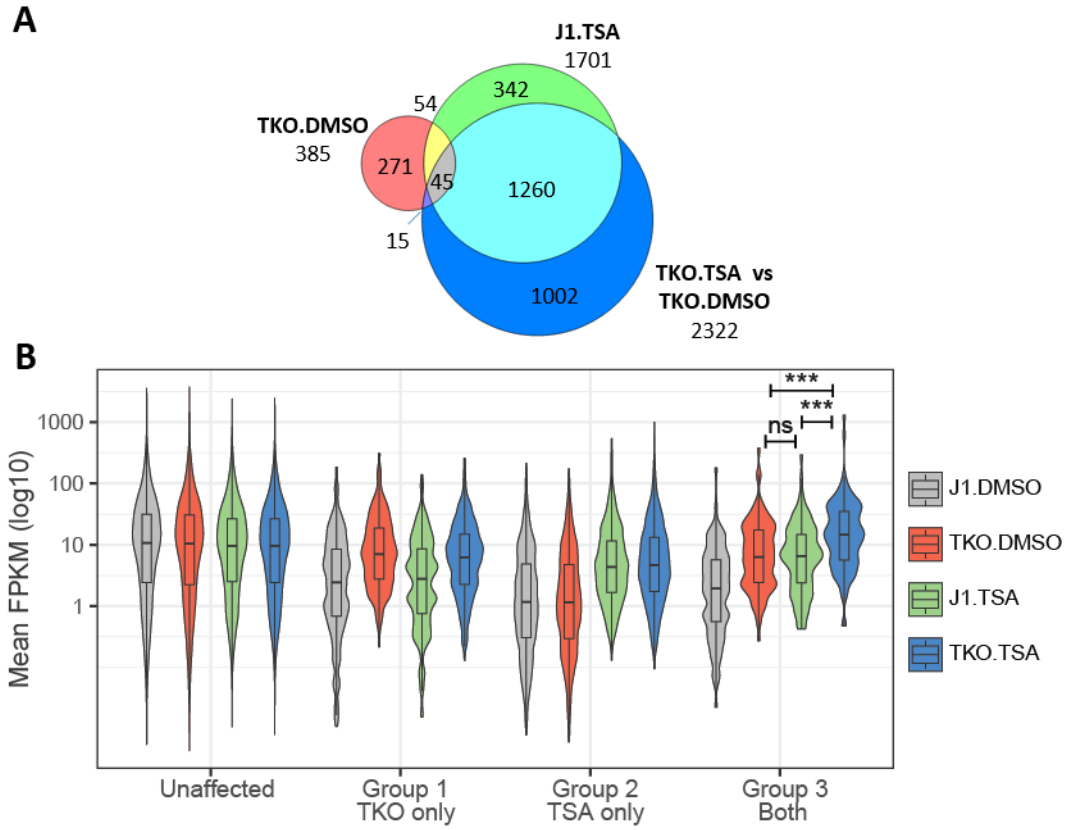


Figure 3.10: Comparison of the transcriptional effects of HDAC inhibition and DNA methylation loss. A. Venn diagram representing the overlap in genes significantly upregulated upon TSA treatment in wild-type J1 cells (green), genes significantly upregulated upon TSA treatment in DNMT.TKO cells (blue) or genes significantly upregulated in DNMT.TKO compared to wild-type cells (red). The number of genes in each group is indicated. **B.** Violin plots summarise gene expression in each experimental condition. Genes were grouped according to their differential expression status. For each gene, RNA-seq RPKM values from both biological replicate were averaged. The "unaffected" group includes genes that do not show any significant increase in expression when comparing any two conditions (n = 13235). Group 1 genes are upregulated in DNMT.TKO versus wild type cells but not upon TSA treatment in J1 cells (n = 286). Group 2 genes are upregulated upon TSA treatment of wild type but not in DNMT.TKO versus wild type cells (n = 1602). Group 3 genes are upregulated in response to either TSA treatment or DNA methylation removal (n = 99). Wilcoxon matched-pairs signed rank test were used to determine significant differences: *** p-value < 0.001 ; ns = non-significant (p-value >0.05).

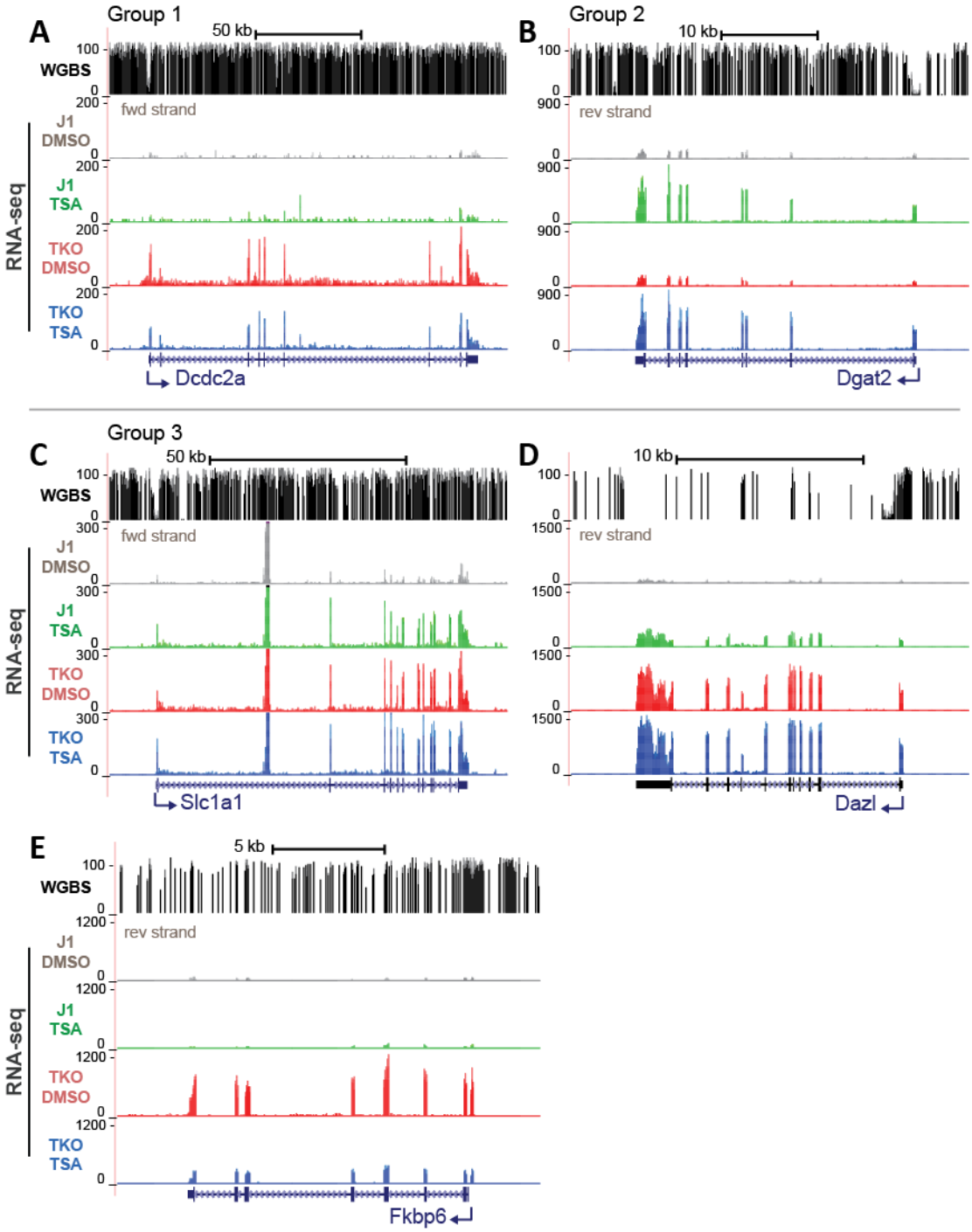


Figure 3.11: UCSC genome browser snapshots showing examples of genes belonging to groups 1, 2 or 3. RNA-seq signal for DMSO or TSA-treated wild-type J1 and DNMT.TKO cells is displayed along with wild type CpG methylation obtained using whole genome bisulfite sequencing (WGBS) in E14 mESCs (Habibi et al., 2013). UCSC gene annotations are shown. Only reads originating from the sense strand are shown. Coverage data from replicate samples are overlayed. RPKM normalisation was used to account for variations in library size. Mm10 coordinates – **A:** chr13:25,037,247-25,224,321 ; **B:** chr7:99,146,911-99,188,172 ; **C:** chr19:28,823,175-28,923,993 ; **D:** chr17:50,275,905-50,297,101 ; **E:** chr5:135,334,237-135,353,003

We were concerned that the genes in group 1 which are upregulated upon DNA methylation loss but not in response to HDAC inhibition represented genes which were not direct targets of DNA methylation but were instead affected indirectly. However, these genes (Group 1) and those that are also upregulated in TSA treated WT cells (Group3) showed similarly elevated levels of DNA methylation around their TSS (Figure 3.12). Furthermore the intersection between the different sets of differentially expressed genes is similar if we take into account all genes (Figure 3.10A) or only focus on genes with methylation in their promoter region (Figure 3.12B).

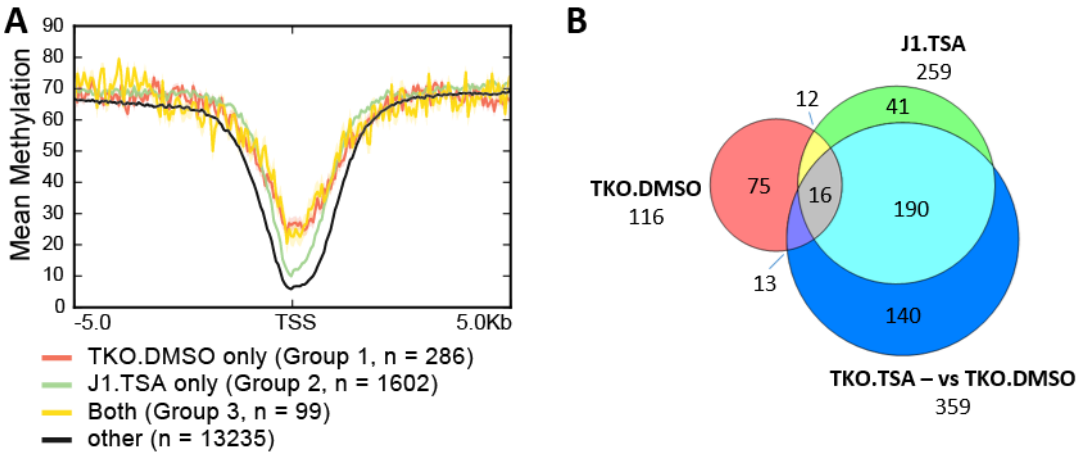


Figure 3.12: DNA methylation at the promoters of genes grouped according to their response to DNMT removal or TSA treatment **A.** Metaplot summarising mean CpG methylation levels in the 10kb regions surrounding gene transcription start sites (TSS). Genes were grouped according to their transcriptional response to DNMT removal or TSA treatment. "TKO.DMSO only" genes are significantly upregulated in DNMT.TKO cells versus wild-type but not with TSA treatment of wild-type J1 cells (Group 1). "J1.TSA only" genes are significantly upregulated upon TSA treatment of wild-type J1 cells but not DNMT.TKO cells compared to wild type (Group 2). Genes in the "Both" category are significantly upregulated by TSA treatment in wild-type cells or in DNMT.TKO cells compared to wild type (Group 3). Other genes do not show any significant increase in expression when comparing any two conditions. CpG methylation values were averaged in 50bp windows. Light shading indicates the area between the mean +/- the standard error of the mean. **B.** Venn diagram representing the overlap in genes significantly upregulated upon TSA treatment in wild-type J1 cells (green), genes significantly upregulated upon TSA treatment in DNMT.TKO cells (blue) or genes significantly upregulated in DNMT.TKO compared to wild-type cells (red). Only genes with promoter CpG methylation levels above 0.6 were included (average of all CpGs). The number of genes in each group is indicated.

As DNA methylation is thought to recruit HDAC inhibitors through MBD-binding proteins (Jones et al., 1998; Nan et al., 1998; Ng et al., 1999) and that these are known to associate with mCpG dense regions (Baubec et al., 2013; Hainer et al., 2016), we thought that genes that would be repressed through this mechanism would have elevated GC and CpG contents. When comparing the sequence composition of promoters from group 1 and 3, neither was found to have a particularly higher GC or CpG density than the other (Figure 3.13).

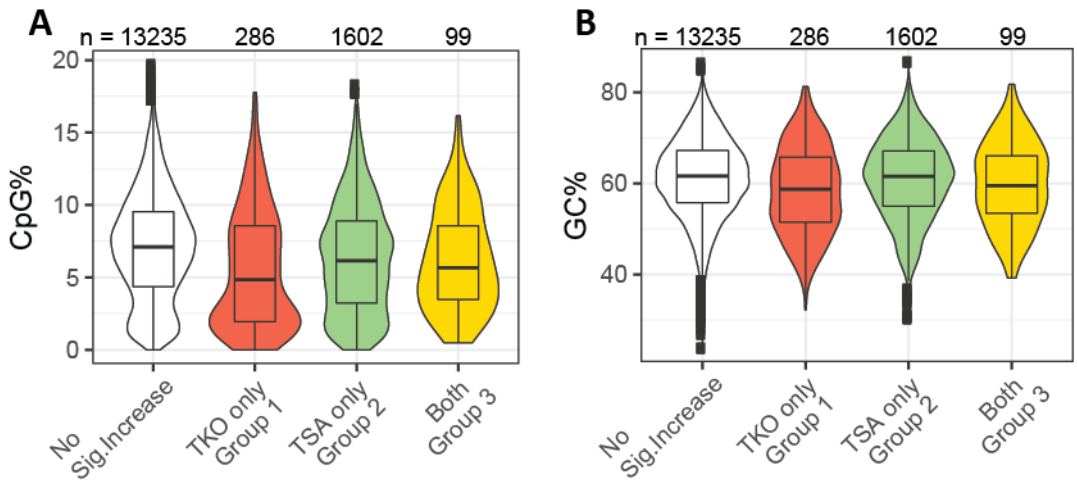


Figure 3.13: Sequence composition at the promoters of genes grouped according to their response to DNMT removal or TSA treatment. Violin plots summarise the percentage GC (**A.**) or CpG (**B.**) content at promoters of genes grouped according to their transcriptional response to DNMT removal or TSA treatment in mESCs. These groupings are identical to those used in the metaplot Figure 12A.

As observed in the previous section, there are more genes upregulated after TSA treatment in DNMT.TKO compared to WT cells. However, a large proportion of these also gain expression in wild-type cells albeit with a reduced fold change meaning that fewer reach the cut-off for significance testing. Differential gene expression analysis between TKO+TSA and J1+TSA cells was used to identify 162 genes that were significantly upregulated following TSA treatment in DNMT.TKO cells but not wild-type cells. 54 of these genes belonged to group 1 and were previously upregulated in untreated DNMT.TKO cells compared to WT. On the other

hand, 77 genes were highly expressed in response to TSA treatment in DNMT.TKO cells alone (eg. *BC051019* and *Nefm*, (Figure 3.14A-B)). For this subset of genes, removal of DNA methylation or HDAC inhibition led to small increases in expression but full reactivation was dependent on the removal of both repressive layers (Figure 3.14C). Interestingly 28 out of 77 of these genes reside on chromosome X, nine times more than expected by chance. We had

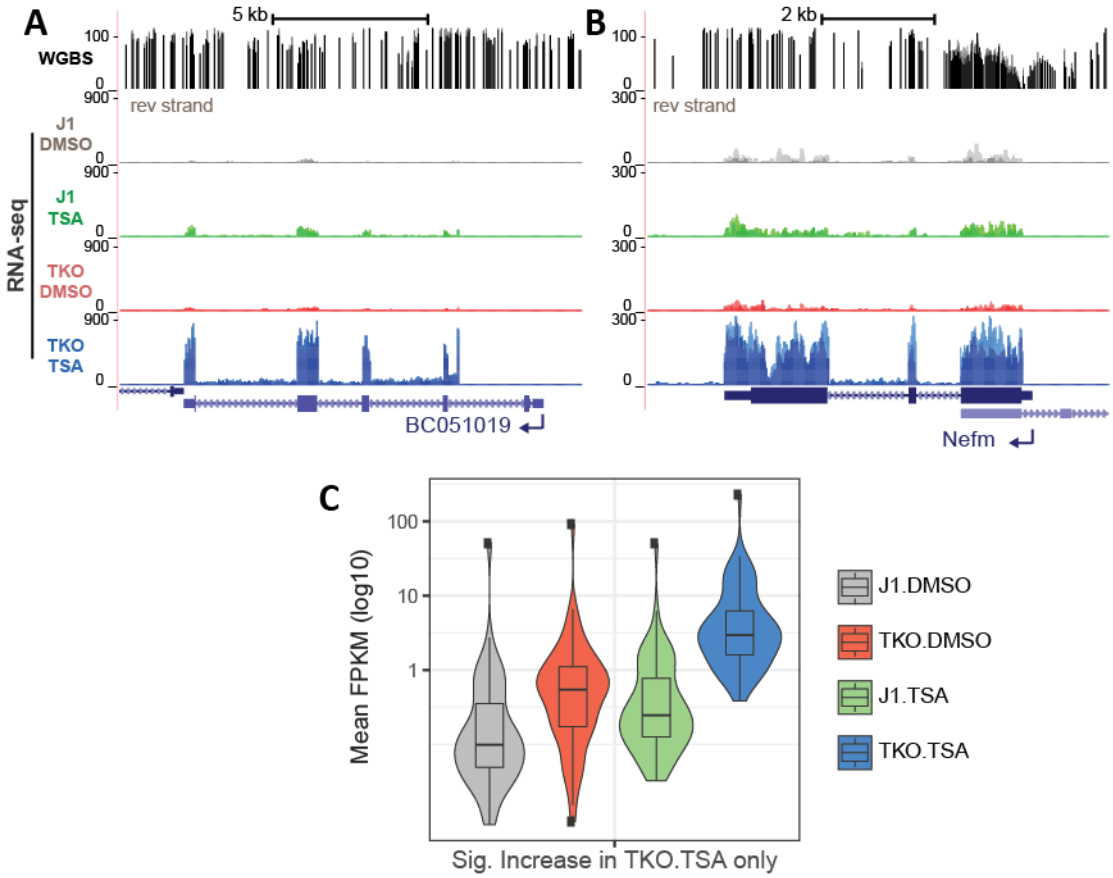


Figure 3.14: Genes upregulated in response to TSA in DNMT.TKO but not wild type cells. A. UCSC genome browser snapshots showing examples of genes upregulated in DNMT.TKO versus wild-type cells. RNA-seq signal for DMSO or TSA-treated wild-type J1 and DNMT.TKO cells is displayed along with wild type CpG methylation obtained using whole genome bisulfite sequencing (WGBS) in E14 mESCs (Habibi et al., 2013). Only reads originating from the sense strand are shown. Coverage data from replicate samples are overlayed. RPKM normalisation was used to account for variations in library size. UCSC gene annotations are shown. Mm10 coordinates – **A:** chr7:109,710,128-109,724,960 ; **B:** chr14:68,118,180-68,126,369 **C.** Violin plots summarise gene expression levels in each experimental condition for genes that were identified as significantly upregulate in DNMT.TKO cells treated with TSA (n = 77).

previously observed that genes located on the X chromosome are overrepresented in the group of genes that are significantly upregulated in DNMT.TKO cells compared to wild-type (Figure 3.2E). In contrast, there is no bias towards chromosome X for genes whose expression is significantly induced by TSA (Appendix - Figure S1).

3.5 Impact of DNA methylation loss and HDAC inhibition on the abundance of mRNAs originating from repetitive elements

Transposable elements in the genome are of a particular interest when it comes to DNA methylation and histone deacetylation with many studies describing a role for these epigenetic modifications in their transcriptional silencing, particularly in germ cell development, mESCs and somatic tissues (Walsh et al., 1998; Karimi et al., 2011; Reichmann et al., 2012; Maksakova et al., 2013; Barau et al., 2016).

Because transposable elements are heavily methylated in mES cells (Habibi et al., 2013; Seisenberger et al., 2012; Stadler et al., 2011), we wanted to determine whether these were activated in DNMT.TKO cells. To this aim, we counted the RNA-seq reads that mapped to the different types of repeat elements that are contained in the RepBase library (obtained from the UCSC database). This collection includes autonomous and non-autonomous transposable elements as well as DNA and RNA repeat elements (Jurka et al., 2005); each repeat type is classified first by class (eg. LINE, LTR, Simple Repeat) and then by family (eg. ERVK, ERVL are subfamilies of LTRs). For each separate entry, we obtained a total read count that we could use for differential analysis. We were most interested in the LINE and LTR classes of repeats with the latter comprising the ERV1, ERVK and ERVL families as these include members which are transcriptionally active in the mouse genome (Stocking and Kozak, 2008).

Differential analysis identified 92 out of 1544 repeat types that were significantly more transcribed in DNMT.TKO cells compared to WT J1 cells (FDR-adjusted p-value < 0.05, Figure 3.15A, Table 3.2, Appendix - Table S2). Most of these were LTR retrotransposons belonging

to the ERVK and ERV1 families (41 and 16 types of repeats significantly upregulated respectively, Figures 3.15B-D) including the IAP elements which are known to be particularly active upon reduction of DNA methylation (Walsh et al., 1998). Increased transcriptional activity from ERVK and ERV1 but not ERVL repeats in this DNMT.TKO mES cell line is in alignment with the results obtained from previous studies that have analysed transcription from TEs using microarray or RNA-seq methods (Karimi et al., 2011; Reichmann et al., 2012).

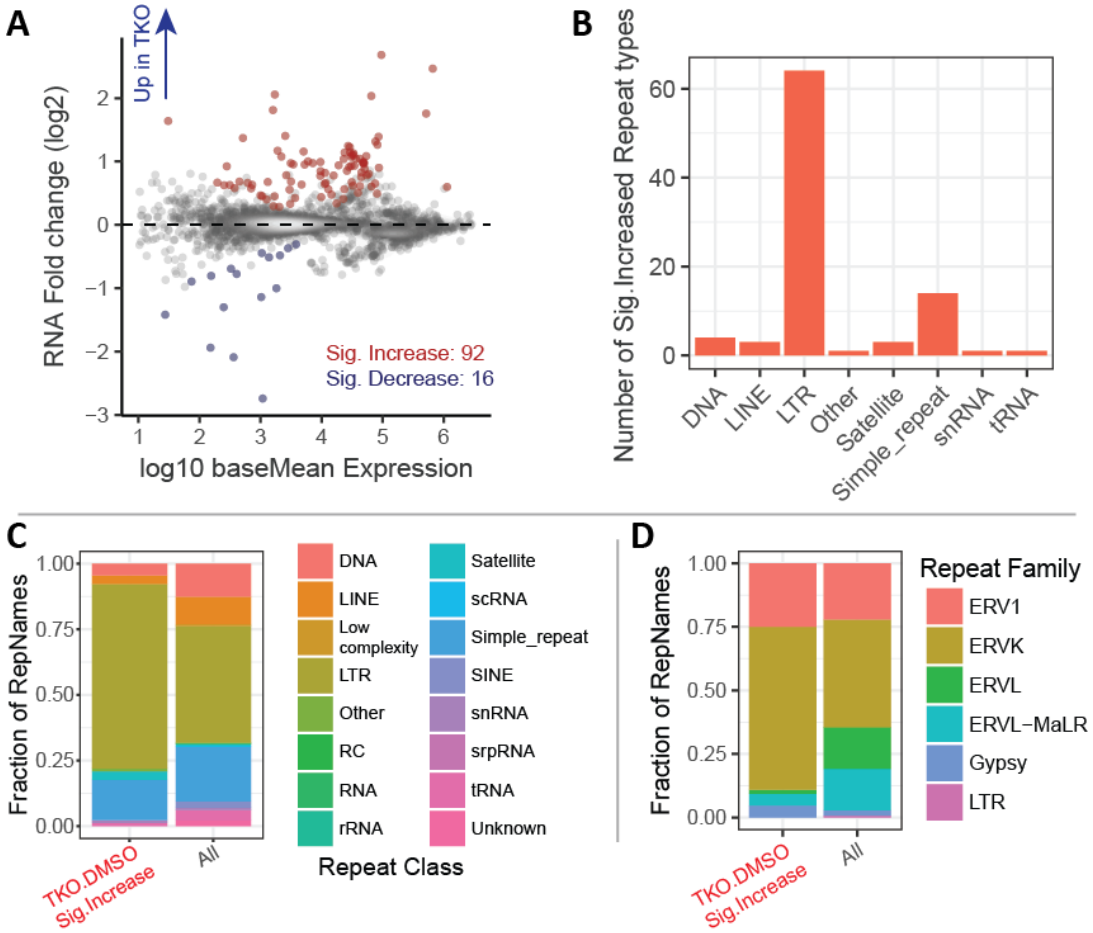


Figure 3.15: Differential analysis of transcription from repetitive elements in DNMT.TKO compared to wild-type cells. A. MA plot comparing for each type of repetitive element, the fold change in RNA-seq signal across all genomic copies to the baseMean RNA-seq signal in DNMT.TKO and wild type mESCs. RNA-seq signal was measured as the number of reads that map to one or more instances of the repeat element located across the genome. Differential analysis was performed using the R package DESeq2 (Love et al., 2014). Changes were deemed significant with an *FDR* < 0.05. **B.** Number of repeat elements types significantly upregulated in DNMT.TKO versus wild type J1 cells according to their class.

Figure legend continued on next page

Figure legend continued from previous page

C. Fraction of all repeat elements (n = 1544) or those significantly upregulated in DNMT.TKO versus wild type J1 cells (n = 92) that belong to each class **D.** Fraction of all LTR type transposable elements (n = 64) or those significantly upregulated in DNMT.TKO versus wild type J1 cells (n = 676) that belong to each family.

Repeat ID	Family	Total genomic space (bp)	Number of Copies	Mean CpG Percent	Estimated Repeat age (million years)	TKO.DMSO vs J1.DMSO Fold change (log2)	TKO.DMSO vs J1.DMSO FDR-adjusted p value
ORR1D2-int	ERVL-MaLR	575296	1361	0.5	114.247	1.044	1.28E-05
LTR86A2	ERVL	12303	79	0.5	142.048	2.057	8.18E-07
IAPEy-int	ERVK	2341002	755	1.7	32.698	2.683	1.53E-19
IAPA_MM-int	ERVK	37463	126	5.2	40.134	2.035	2.91E-24
IAPEz-int	ERVK	11883078	7319	2.4	27.078	1.778	6.24E-35
IAPLTR1_Mm	ERVK	551227	1492	5.4	15.995	1.757	9.36E-29
RLTR19A	ERVK	44831	153	0.6	77.023	1.404	5.02E-08
RLTR45-int	ERVK	918744	1317	1.1	50.890	1.264	2.60E-07
RLTR44E	ERVK	41391	122	3.1	48.357	1.247	2.07E-05
MMTV-int	ERVK	34354	35	0.5	118.571	1.227	5.69E-05
RLTR9C	ERVK	175420	477	0.7	68.890	1.157	2.98E-12
RLTR26	ERVK	139769	408	1.2	90.877	1.147	1.22E-04
RLTR20D	ERVK	206997	777	0.8	88.927	1.131	6.42E-05
RLTR16B_MM	ERVK	239283	635	1.1	72.527	1.101	5.45E-05
MLTR18A_MM	ERVK	109687	299	0.5	106.161	1.086	5.12E-05
RLTR10F	ERVK	136494	482	3.2	34.072	1.075	1.48E-05
MMERGLN-int	ERV1	1265344	856	1.2	115.346	2.467	8.82E-27
RLTR47_MM	ERV1	13906	45	1	77.404	1.371	6.03E-13
MMERGLN_LTR	ERV1	167469	435	2.7	22.526	1.317	1.36E-08
MuLV-int	ERV1	306109	392	0.6	134.114	1.141	2.74E-05

Table 3.2: Changes in RNA-seq coverage averaged over all copies of 20 LTR retrotransposon element types. The elements included are the top 20 significantly transcriptionally upregulated transposable element types in DNMT.TKO cells compared to wild-type cells. Differential analysis was performed using the R package DESeq2 (Love et al., 2014).

HDAC inhibition had a more widespread effect on the transcription of repeat elements, leading to the upregulation of 187 and 272 different repeat types in WT and DNMT.TKO cells respectively (Figure 3.16A-B). Here again, the most responsive class of repeats were LTR retotransposons (Figures 3.16C-D)) while significant increases were also observed for simple repeats, something we did not explore further. As opposed to the changes associated with DNA methylation loss, gains in expression were not restricted to the ERVK and ERV1 families but many members of the ERVL family were also affected (Figure 3.16E).

Consistent with what was observed when focusing on the expression of genes, the response to TSA is more robust in DNMT.TKO cells compared to WT cells (Figures 3.16C-E). Focusing on LTR retrotransposons, we compared the amplitude of the fold change for the three conditions (TKO.DMSO, TKO+TSA and J1+TSA) against the untreated WT cells (Figure 3.17A). By doing so we noticed that the different TE families are affected differently by DNA methylation and HDACs: ERVL elements are not de-repressed in DNMT.TKO cells but respond to TSA to approximately the same extent in both cell lines. On the other hand, members of the ERV1 and ERVK families show increased expression following TSA treatment or DNMT loss but removal of both repressive layers leads to a greater level of transcriptional activity (Figure 3.17B). Table 3.3 shows the top 30 significantly expressed transposable elements in TSA treated DNMT.TKO cells compared to wild-type and examples are given of both an ERVL and ERVK type transposable element in Figures 3.17C-D.

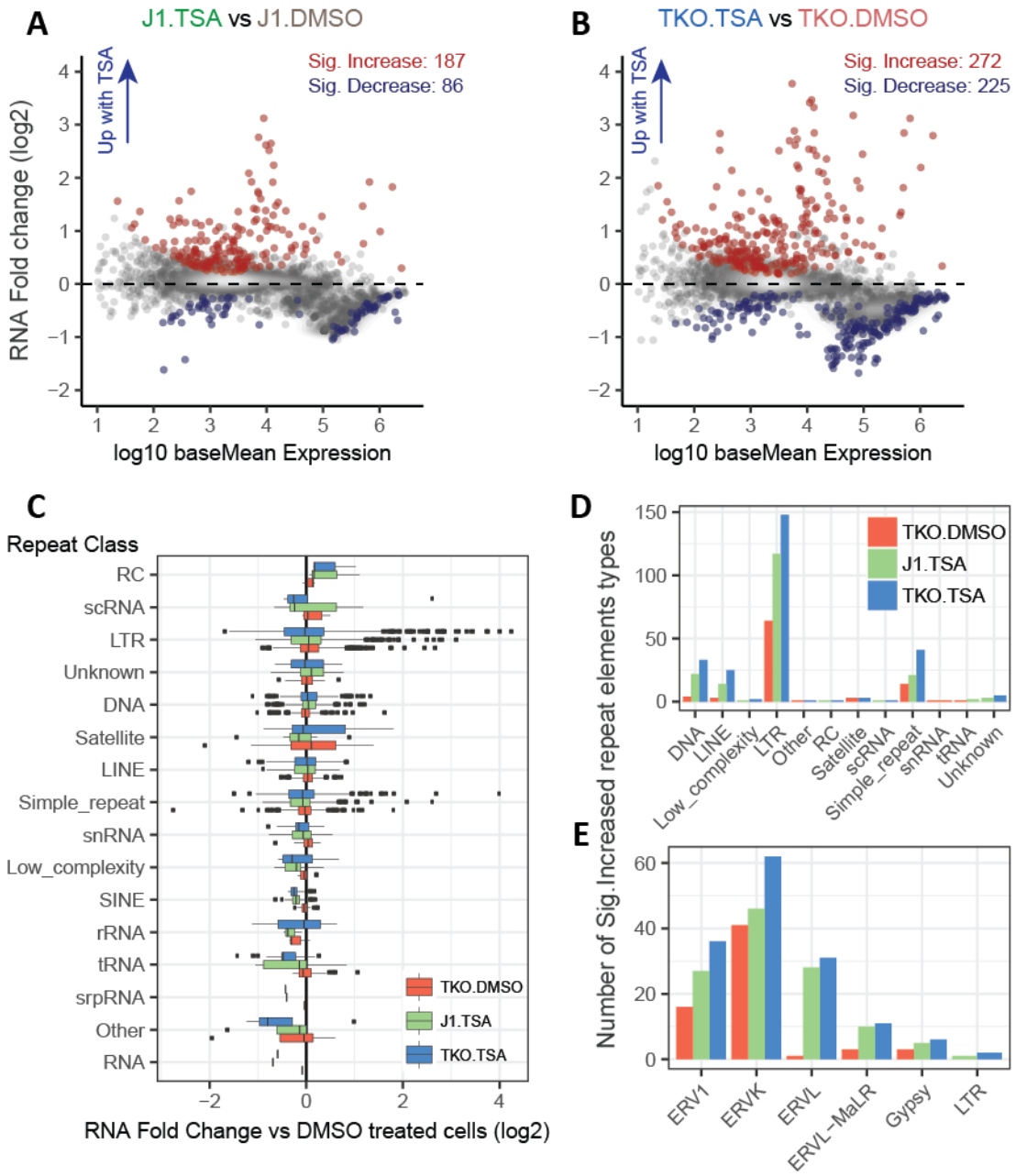


Figure 3.16: Differential analysis of transcription from repetitive elements comparing TSA to DMSO treated mESCs. A. MA plot comparing the fold change and baseMean RNA-seq signal in TSA and DMSO treated wild type J1 mESCs. **B.** MA plot comparing the fold change and baseMean RNA-seq signal in TSA and DMSO treated DNMT.TKO mESCs. **C.** Box plots summarise the fold change in RNA-seq signal for each type of repeat elements grouped by class (y-axis). Fold change values were determined for DNMT.TKO versus wild type J1 mESCs (red), TSA- versus DMSO- treated J1 cells (green) and TSA versus DMSO- treated DNMT.TKO cells (blue). **A-C.** RNA-seq signal was measured as the number of reads that map to one or more genomic copies of each repeat element type. Differential analysis was performed using the R package DESeq2 (Love et al., 2014). Changes were deemed significant with an *FDR* < 0.05. **D.** Number of repeat elements types per class that are significantly upregulated in DNMT.TKO versus wild type J1 cells (red) TSA- versus DMSO- treated J1 cells (green) or TSA versus DMSO- treated DNMT.TKO cells (blue). **E.** Number of LTR type transposable elements per family that are significantly upregulated in DNMT.TKO versus wild type J1 cells (red) TSA- versus DMSO- treated J1 cells (green) or TSA versus DMSO- treated DNMT.TKO cells (blue).

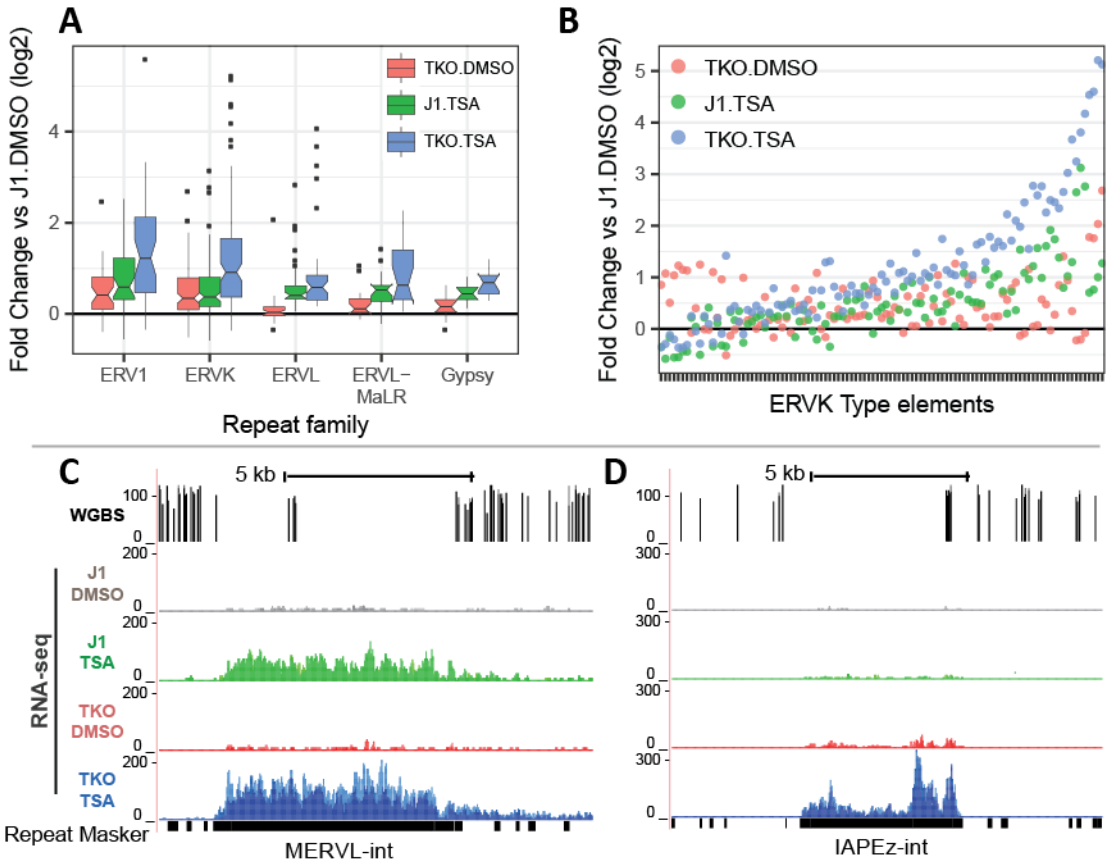


Figure 3.17: Comparison of the transcriptional effect of HDAC inhibition in wild type and DNMT.TKO cells at LTR transposable elements. A. Box plots summarise the fold change in RNA-seq signal for each LTR element type grouped according to their family (x-axis). Fold change values were determined for DNMT.TKO compared to wild type J1 mESCs (red), TSA-compared to DMSO- treated J1 cells (green) and TSA-treated DNMT.TKO cells compared to DMSO-treated DNMT.TKO cells (blue). **B.** RNA-seq fold change values for different ERVK-type retrotransposon elements. Different ERVK types are sorted along the x-axis according to their fold change in TSA-treated DNMT.TKO cells compared to DMSO-treated DNMT.TKO cells. In figures **A** and **B,** only repeat element types that are significantly upregulated when comparing any two conditions were included. **C.** UCSC genome browser snapshot showing an example of an ERVL MERVL-int element upregulated in response to TSA but not DNMT deficiency. Mm10 coordinates: chr1:54,888,355-54,899,905 . **D.** UCSC genome browser snapshot showing an example of an ERVK IAPEz-int element upregulated in response to TSA in DNMT.TKO cells only. Mm10 coordinates: chr1:26,345,077-26,358,809. **C-D.** RNA-seq signal for DMSO or TSA-treated wild-type J1 and DNMT.TKO cells is displayed along with wild type CpG methylation obtained using WGBS in E14 mESCs (Habibi et al., 2013). Reads originating from both strands are included. Coverage data from replicate samples are overlayed. The coverage tracks displayed include reads with up to five alignments (TopHat MAPQ score >1). UCSC repeat element annotations (Repeat Masker) are shown.

Repeat ID	Family	Number of copies	Mean CpG Percent	Estimated Repeat age (million years)	Expression Fold change versus J1.DMSO (log2)		
					TKO DMSO	J1 TSA	TKO TSA
MMERGLN-int	ERV1	856	1.2	115.346	2.47	1.92	5.59
LTR45C	ERV1	24	0.5	122.576	0.50	1.24	3.33
MMERGLN_LTR	ERV1	435	2.7	22.526	1.32	0.83	3.26
RLTR41C	ERV1	157	0.8	56.812	-0.27	2.52	3.14
RLTR5_Mm	ERV1	601	2	31.268	0.19	1.42	3.08
MER89	ERV1	614	0.7	134.750	0.10	2.24	2.94
MERV1_I-int	ERV1	1365	0.6	94.548	0.48	1.47	2.90
RLTR6C_Mm	ERV1	102	3	33.908	0.97	1.71	2.77
RLTR48C	ERV1	149	0.8	38.816	0.12	1.05	2.63
IAPA_MM-int	ERVK	126	5.2	40.134	2.03	1.00	5.21
IAPEy-int	ERVK	755	1.7	32.698	2.68	1.27	5.13
IAPLTR1_Mm	ERVK	1492	5.4	15.995	1.76	0.76	4.60
IAPEz-int	ERVK	7319	2.4	27.078	1.78	0.70	4.54
SRV_MM-int	ERVK	129	0.6	72.921	-0.08	2.76	4.17
ERVB2_1-I_MM-int	ERVK	83	0.6	83.385	-0.21	3.12	3.81
ERVB7_3-LTR_MM	ERVK	319	1.3	52.548	0.34	1.29	3.67
RLTR9B	ERVK	206	0.9	55.038	-0.22	2.65	3.24
IAPEY_LTR	ERVK	1410	5.1	42.551	0.94	1.02	3.03
RLTR31B_Mm	ERVK	349	1	64.965	0.04	1.22	2.77
MMERVK10C-int	ERVK	3230	1.7	50.880	0.57	0.99	2.77
RLTR19A	ERVK	153	0.6	77.023	1.40	0.80	2.65
RLTR12F	ERVK	60	0.3	89.060	0.07	1.57	2.59
RMER16B2	ERVK	329	0.5	62.382	-0.03	1.92	2.50
IAP1-MM_LTR	ERVK	420	2	35.859	0.27	1.74	2.46
RMER16A2	ERVK	273	0.5	68.816	-0.09	1.09	2.45
MMERVK10D3_I-int	ERVK	458	1.4	96.755	0.85	1.38	2.34
MERVL-int	ERVL	2600	0.6	19.820	0.30	2.83	4.05
LTR86A1	ERVL	92	0.7	138.607	-0.10	1.92	3.68
LTR86A2	ERVL	79	0.5	142.048	2.06	1.06	3.25
MT2_Mm	ERVL	2667	1.2	10.084	0.16	1.83	2.95

Table 3.3: Changes in RNA-seq coverage averaged over all copies of 20 LTR retrotransposon element types. The elements included are the top 30 significantly transcriptionally upregulated transposable element types in TSA treated DNMT.TKO cells compared to DMSO-treated wild-type cells. Differential analysis was performed using the R package DESeq2 (Love et al., 2014).

3.6 Discussion

3.6.1 Transcriptional landscape in DNA methylation deficient mouse ESCs

Using RNA-seq, we have compared the transcriptional landscape of DNA methylation deficient DNMT.TKO and J1 wild-type mouse ES cells. In accordance with previous studies that have analysed gene expression in the same cell lines using microarray or RNA-seq methods (Fouse et al., 2008; Karimi et al., 2011), we found that despite the complete loss of DNA methylation, only a limited number of genes display significant gains in expression. This indicates that there are few genes or genomic elements whose transcription is primarily regulated by DNA methylation in mESCs. Through differential gene expression analysis, we have identified 385 genes that are upregulated in DNMT.TKO cells compared to J1 cells. 69% of these genes do not have high levels of DNA methylation around their transcription start site in wild-type cells (Figure 3.4) making it difficult to attribute the changes in transcription at most significantly upregulated genes to a loss in DNA methylation. Many of these are likely to be indirect effects of DNA methylation removal. Alternatively their expression may be influenced by methylation changes at distal regulatory elements. Nevertheless, of the 100 genes whose expression is most increased in DNMT.TKO cells, 88 have high levels of DNA methylation surrounding their TSS. In addition, most of these 100 genes were found to be significantly upregulated in data obtained by another group in independent DNMT.TKO and wild-type ESCs (Domcke et al., 2015) thus validating them as targets of DNA methylation-mediated silencing.

Genes upregulated in the absence of DNA methylation include members of the *Rhox* gene cluster reported to be specifically expressed in the extraembryonic trophectoderm (Oda et al., 2006) and germline-specific genes such as members of the *Magea* cluster, *Fkbp6*, *Dazl*, *Asz1, Piwil1, Tex19.1* or *Tdrd1/9* which have previously been identified as being directly

regulated by DNA methylation (De Smet et al., 1999; Fouse et al., 2008; Karimi et al., 2011; Auclair et al., 2014; Sharif et al., 2016). Interestingly these genes are expressed and have important biological functions in gametes, cells that have undergone global demethylation (Guibert et al., 2012; Seisenberger et al., 2012). Particularly, *Fkbp6*, *Tdrd1/9*, *Asz1* and *Piwil1* are part of the piwi-interacting RNA (piRNA) machinery that is implicated in the methylation and repression of L1 and IAP elements in germ cells (Aravin et al., 2008; Shoji et al., 2009; Xiol et al., 2012). It is possible that the regulatory elements of these genes have evolved to be particularly sensitive to DNA methylation as a means to maintain DNA methylation and silencing of retrotransposons during stages of global DNA hypomethylation.

Of the genes significantly upregulated in DNMT.TKO compared to wild-type J1 cells, more than expected are localized to the X chromosome (Figure 3.2E). On one hand, this may be explained by the presence of several gene clusters (*Magea, Rhox and Xlr3/4*) on this chromosome which account for 18 of the 48 significantly upregulated genes. Genes within these clusters are likely to be regulated by a common DNA methylation-dependent mechanism (De Smet et al., 1999; Oda et al., 2006). On the other hand, the overrepresentation of genes on the X chromosome may reflect an enhanced use of epigenetic mechanisms in the regulation of X chromosome genes. Indeed DNA methylation along with other chromatin modification are necessary to maintain the selective inactivation of chromosome X in female somatic cells (Beard et al., 1995; Hansen et al., 1996).

3.6.2 DNA methylation and HDACs operate through independent pathways

Beyond our interest in determining the impact of DNA methylation deficiency on the transcriptome, this study was designed to characterise the extent to which the transcriptional repression mediated by DNA methylation and HDAC activities overlap. Based on previous studies, DNA methylation is thought to promote transcriptional repression in part through

the action of HDAC enzymes within multi-protein complexes that are recruited to regions with elevated methyl-CpG density via MBD-containing proteins (Jones et al., 1998; Nan et al., 1998; Zhang et al., 1999; Kokura et al., 2001). According to this model, genes upregulated in response to DNA methylation loss would also be expected to be de-repressed by the disruption of HDAC activity.

Comparison of the transcriptional changes that occur following the disruption of DNA methyltransferases to those observed upon TSA treatment revealed that most genes significantly upregulated in the absence of DNA methylation were not affected by HDAC inhibition (Figure 3.9). The same finding was made when focusing on genes with high level of promoter methylation (Figure 3.12). Furthermore, most genes that displayed significant gains in expression in both conditions were upregulated by TSA treatment regardless of the absence of DNA methylation (both in J1.TSA versus J1.DMSO cells and TKO.TSA versus TKO.DMSO cells, Figure 3.10) suggesting that HDACs and DNA methylation influence gene expression through independent pathways. In line with this, the removal of DNA methylation and HDAC inhibition have additive effects at common targets (Figure 3.10B), in some cases affecting transcription in opposite directions (eg. *Fkbp6*, Figure 3.11E).

Our data indicates that there are likely to be very few genes whose repression in mESCs is solely mediated by the recruitment of HDAC enzymes to methylated DNA via MBD proteins. As MBD-domain proteins preferentially occupy regions with a high density of methylated CpGs (Baubec et al., 2013; Hainer et al., 2016), one would expect that genes with densely methylated promoters would be targets to both TSA and DNA methylation. However, genes whose promoters are densely methylated and whose expression increases in DNMT.TKO cells, such as *Fkbp6* or *Tuba3b*, are not necessarily upregulated after TSA treatment (Figure 3.11B). Moreover, genes upregulated in DNMT.TKO cells have promoters with relatively low CpG density relative to the average gene promoter (Figure 3.5A) further suggesting that DNA

methylation-mediated repression of these genes in mESCs is generally independent of HDAC activity.

Our results agree with previous studies that have focused on transposable elements and suggested that the disruption of HDAC or DNMT activity have non-overlapping effects on their expression. In mouse 3T3 fibroblasts, both TSA and 5-aza-deoxycytidine treatments led to the transcriptional de-repression of VL30 ERV1 type repeats but only DNA methyltransferase inhibition resulted in IAP transcriptional activity (Brunmeir et al., 2010). Similarly *Hdac1* knockout mES cells were reported to have an increased transcription from RLTR45 ERVK elements but a decrease in expression from IAP elements (Reichmann et al., 2012). Brocks *et al.* 2017 treated a human lung cancer cell line (NCI-H1299) with either 5-aza-deoxycytidine or HDAC inhibitors. While both treatments caused the reactivation of an epigenetically silenced fluorescent reporter gene, this was achieved through different mechanisms: on one hand DNMT inhibition promoted transcription from this gene's previously hypermethylated CpG island promoter. On the other hand, various HDAC inhibitors caused transcriptional reactivation through the induction of a cryptic TSS (Brocks et al., 2017). In the same study, 5-aza-deoxycytidine treatment but not HDAC inhibition was reported to be accompanied by the appearance of various active histone modifications (H3K4me3, H3K9ac, or H3K27ac) at non-annotated transcription start sites throughout the genome (Brocks et al., 2017). Together, these observations along with our results indicate that DNA methylation mediates the repression of numerous genes or transposable elements in different cell lines through a mechanism independent of HDACs. Such mechanisms could include DNA methylation directly impeding the binding of transcription factors to DNA (Domcke et al., 2015; Maurano et al., 2015; Yin et al., 2017) (see Chapters 4 and 5) or involve chromatin modifiers that are not affected by TSA (eg. histone methyltransferases or sirtuins).

3.6.3 HDAC inhibition in DNMT.TKO compared to wild-type mESCs

Although the effect of TSA on the transcription of genes and repeat elements in WT or DNMT.TKO mES cells was similar (Figure 3.8A), there were some differences. Notably, the transcriptional changes were quantitatively more pronounced in DNMT.TKO cells compared to wild type cells resulting in the classification of a greater number of genes as significantly differentially expressed for a defined FDR threshold (Figure 3.8). In addition, the dual effect of DNA methylation loss and HDAC inhibition lead to the pronounced activation of a few single copy genes (Figure 3.14) and transposable elements (Figure 3.17 and Table 3.3) that were mostly unaffected by each treatment alone. Brocks *et al.* also observed such synergistic activity of DNA methylation loss and HDAC inhibition in their recent study: combinatorial treatment with 5-aza-deoxycytidine and the SB939 HDAC inhibitor induced a much greater transcriptional activation from LTR12 type transposable elements than either inhibitor alone (Brocks et al., 2017). Similar findings were made for several inactive genes whose CpG island-associated promoters were densely methylated in a colorectal carcinoma cell line: efficient reactivation of these genes was only achieved upon TSA treatment following pre-incubation of the cells with 5-aza-deoxycytidine (Cameron et al., 1999). The DNA methylation and HDAC-dependent mechanisms involved in restricting transcription from these elements are therefore likely to be conserved in different cell types from distinct mammalian species.

DNA methylation plays a role in transcriptional repression alongside other mechanisms that involve the methylation of lysine residues on histone tails (Karimi et al., 2011; Reichmann et al., 2012; Walter et al., 2016). These mechanisms are likely to be disrupted by HDAC inhibition. Indeed, a failure to remove acetyl marks from newly incorporated histones or those deposited by HATs will block the action of histone methyltransferases that target the same lysine residues. Increased sensitivity of DNMT.TKO cells to TSA treatment is likely to

reflect a greater dependence on these alternative mechanisms to maintain a normal transcriptional landscape in the absence of DNA methylation.

The crosstalk between the roles of DNA methylation and other histone modification such as H3K9me3 or H3K27me3 in transcriptional repression has been widely studied in mESCs. Most insight has been gained by studies that have focused on their role in the silencing of transposable elements. In mES cells, methylation of the H3K9 residue has an important role in the repression of LTR retrotransposons (Karimi et al., 2011; Matsui et al., 2010). H3K9 can be methylated through the enzymatic activity of SETDB1, SUV39H enzymes (H3K9me3, tri-methylation) or G9a/GLP (H3K9me2, di-methylation). The localisation of these enzymes to certain LTR elements relies in part on the KAP1 protein (Maksakova et al., 2013; Rowe et al., 2013). Notably, the KAP1 and SETDB1 proteins along with H3K9me3-modified histones are enriched at ERVK elements in mESCs and genetic abrogation of *Kap1* or *Setdb1* leads to increased expression from these retrotransposons (Karimi et al., 2011; Maksakova et al., 2013; Reichmann et al., 2012). ERVK are also the main TE family for which DNA methylation directly contributes to transcriptional silencing. Indeed, they are modestly expressed in DNMT.TKO cells (Figure 3.17 , (Karimi et al., 2011; Reichmann et al., 2012)) and in hypomethylated mESCs (Sharif et al., 2016; Walter et al., 2016). Simultaneous disruption of both DNA methyltransferases and SETDB1/KAP1 activity has been reported to lead to the synergistic activation of certain IAP elements suggesting that DNA methylation and H3K9 methylation represent two independent and compensatory repressive mechanisms whose activity converges at these elements (Karimi et al., 2011; Sharif et al., 2016). A similar observation was made when we have studied the intersection between DNA methylation and HDAC activity: IAP elements are fully activated in DNMT.TKO only when these are treated with TSA (Table 3.3) supporting our suggestion that HDAC inhibition interferes with the H3K9me3-mediated repression.

KAP1 and H3K9 methylation are involved in the transcriptional silencing of a wider range of repetitive elements than DNA methylation. For example, the transcriptional repression of ERVL class MERVL repeats in mESCs relies on the normal activity of G9a and KAP1 (Macfarlan et al., 2012; Maksakova et al., 2013). Disruption of this mechanism by HDAC inhibition potentially explains the aberrant expression of these repeats after TSA treatment in mouse ESCs in the presence or absence of DNA methylation (Table 3.3 and Figure 3.16D, (Macfarlan et al., 2012)).

Another histone modification associated with gene repression is H3K27 tri-methylation, catalysed by enzymes of the PRC2 complex. At CpG rich regions, the presence of DNA methylation and H3K27me3-modified histones are mutually exclusive (Brinkman et al., 2012). This has been attributed to the preferential occupancy of the PRC2 complex to methylation-free CpG-rich regions (Lynch et al., 2012), mediated by the interaction of CXXC-containing or Polycomb-like proteins with unmethylated DNA (Blackledge et al., 2014; Li et al., 2017). Importantly, in hypomethylated or DNA methylation deficient cells, the H3K27me3 mark is widely redistributed: new domains of H3K27me3 are formed at regions previously covered in DNA methylation in wild type cells while it is depleted from CpG rich regions where it is normally abundant (Brinkman et al., 2012; Reddington et al., 2013; King et al., 2016). In respect to the repression of retrotransposons, only RLTR4 and MuLV elements have been reported to be silenced in a PRC2-dependent manner in wild-type mouse embryonic stem cells (Leeb et al., 2010; Reichmann et al., 2012). However, in mES cells that were hypomethylated following a switch from serum-containing medium to medium containing 2i and vitamin C, H3K27me3 was reported to accumulate at LINE, MMERGLN and MERVL repeats. In addition, transcriptional silencing from MERVL elements in hypomethylated cells but not cells grown in serum was disrupted by the deletion of *Eed*, the gene encoding a core component of PRC2 (Walter et al., 2016). In mES cells that have adapted to a DNA methylation-free state, such as our DNMT.TKO cells, PRC2 and H3K27me3 therefore appear

to have additional roles in restricting the expression of a subset of transposable elements compared to wild-type cells. As HDAC inhibition potentially interferes with H3K27me3 deposition, a more widespread role for this modification in the absence of DNA methylation is likely to contribute to the increased sensitivity to TSA, which we have observed in DNMT.TKO relative to wild-type cells.

To summarise, our data suggests that DNA methylation is directly involved in the silencing of certain single-copy genes and retrotransposons, generally through mechanisms that are not disrupted by TSA in mouse embryonic stem cells. In the next chapters, we will attempt to gain insight into these mechanisms by investigating the impact of DNA methylation loss on accessibility and the occupancy of transcription factors. At several loci, and particularly at ERVK-type elements, repressive pathways independent of DNA methylation appear to be sufficient to prohibit full transcriptional activation although these can be blocked by HDAC inhibition. To evaluate the extent to which TSA is interfering with repressive processes that rely on histone H3K27 or H3K9 methylation, our RNA-seq data could be compared to transcriptomic data obtained from mES cells carrying inactivating mutations in *Kap1*, *Setdb1* (Karimi et al., 2011; Rowe et al., 2010) or genes encoding members of the PRC2 complex. Additionally, careful examination of the level and distribution of various histone modifications in our four conditions could help understand how they are affected DNA methylation loss and/or HDAC inhibition.

Chapter 4:

Investigating the contributions of DNA methylation and HDAC activity to the chromatin accessible landscape in mouse embryonic stem cells

4.1 Introduction

CpG methylation has commonly been regarded as a mechanism for transcriptional regulation. Interference with the DNA methyltransferase activity has revealed that DNA methylation participates in the transcriptional repression of certain tissue-specific genes, transposable elements and imprinted loci (Fouse et al., 2008; Borgel et al., 2010; Karimi et al., 2011; Auclair et al., 2014; Stelzer et al., 2016). In the previous chapter, we have described some of the differences in the transcriptional landscape of wild type and DNA methylation deficient mouse ES cells. In order to obtain insights into the mechanism by which DNA methylation loss leads to transcriptional changes we aimed to identify regulatory elements in the genome whose activity is influenced by the presence of DNA methylation.

Increased chromatin accessibility is a feature frequently used to detect functional regulatory elements in eukaryotic genomes. While the vast majority of genomic DNA is wrapped around regularly spaced nucleosomes, there are regions at which the binding of transcription factors (TF), other non-histone proteins or the presence of certain histone modifications can cause nucleosome displacement. These regions are described as being more accessible which, experimentally, manifests as a local increase in the sensitivity of the underlying DNA to chemical modification or digestion by nucleases such as DNaseI (Gross and Garrard, 1988). As such, DNaseI hypersensitivity has widely been used to identify and demarcate various types of genetic regulatory elements including promoters, enhancers, insulators and transcription factor binding sites (Stalder et al., 1980; Gross and Garrard, 1988). The recent use of DNaseI digestion coupled with high-throughput sequencing (DNase-seq), has allowed for the genome-wide mapping of DNaseI Hypersensitive Sites (DHS) in various tissues and cell lines (Boyle et al., 2008; Thurman et al., 2012). Notably, such studies have revealed the widespread existence of promoter distal regulatory elements whose cell-type specific activity can be monitored by chromatin accessibility profiling (Thurman et al., 2012).

Chromatin accessibility inversely correlates with the presence of DNA methylation. Both unmethylated CpG islands located at promoters as well as distal regulatory elements with low levels of CpG methylation are highly sensitive to DNaseI (Stadler et al., 2011). Furthermore, cell type specific increases in DNA methylation are strongly associated with reduced accessibility (Thurman et al., 2012). Nevertheless, the causal relationship between chromatin accessibility and DNA methylation has been difficult to define. Early functional studies found that experimentally elevating methylation levels at promoters, achieved through *in vitro* methylation followed by transfection (Keshet et al., 1986) or through the spreading of CpG methylation from an integrated provirus (Jähner and Jaenisch, 1985), could decrease DNaseI sensitivity. At the myogenin gene promoter, DNA methylation was found to be sufficient to impede the formation of a DHS in experimental conditions where either one, but not both of the SIX and MEF2 transcription factor binding sites are occupied. As a result, it was suggested that the presence of DNA methylation can restrain the premature activation of this gene until the appropriate combination of transcription factors are expressed (Palacios et al., 2010).

Despite the evidence suggesting that DNA methylation can suppress the formation of an accessible site, other studies have suggested that changes in DNA methylation are a consequence of changes in transcription factor occupancy and chromatin accessibility. Indeed, the binding of some transcription factors to previously methylated binding sites was sufficient to increase accessibility and decrease DNA methylation in mESCs (Stadler et al., 2011). Furthermore studies that compared DNA methylation, chromatin accessibility and transcription factor expression levels between cell types have suggested that DNA methylation is likely to be passively deposited at transcription factor binding sites as a secondary event to TF evacuation (Stadler et al., 2011; Thurman et al., 2012). The authors observed that in cell types where certain transcription factors are no longer expressed, regions containing their cognate recognition sequences increase in methylation.

In order to identify regulatory elements in the genome whose activity is influenced by the presence of DNA methylation we chose to map chromatin accessibility genome wide in DNMT.TKO versus wild type mES cells. Having generated RNA-seq data for DNMT.TKO and J1 cells we were interesting in identifying the regulatory regions whose changes in accessibility could participate in the transcriptional effects of DNA demethylation and contribute to the different sensitivities these cells display towards HDAC inhibition. Moreover, as DHSs typically occur at sites of DNA-protein interaction, differentially accessible sites in the two cell lines are likely to represent differential occupancy events by DNA methylation-sensitive transcription-factors. We expect that the identification of TF binding sites at these regions will help extend our understanding of how DNA methylation influences the binding of DNA methylation-sensitive transcription factors (also discussed in the following chapter). A study published while our research was being carried out, made use of a similar approach. The authors used DNase-seq to identify novel DHS sites in an independently derived DNMT.TKO cell line, and demonstrated that most of these changes resulted from differential binding of the NRF1 transcription factor (Domcke et al., 2015).

In addition to mapping the chromatin accessibility differences in wild type and DNMT.TKO cells, we aimed to identify the changes in accessibility that occur in response to HDAC inhibition. We anticipated that comparing the data obtained from TSA or mock treated cells would provide insight into the relationship between DNA methylation loss and HDAC inhibition in their effect on chromatin.

While DNase-seq has classically been used to map DHSs genome wide, we profiled chromatin accessibility using the Assay for Transposase-Accessible Chromatin and sequencing (ATAC-seq), a recently developed method which has several advantages over DNase-seq in terms of simplicity and timescale whilst providing comparable results (Buenrostro et al., 2013). In ATAC-seq, chromatin accessibility is measured by the sensitivity of genomic DNA to a

hyperactive Tn5 transposase enzyme. A "tagmentation" step during which nuclei, Tn5 and adapters are incubated allows for the simultaneous fragmentation and tagging of the genomic DNA with sequencing adaptors. In this manner the loss-prone gel purification and adaptor ligation steps necessary for DNase-seq are bypassed. At accessible loci, transposition at sites in close proximity to each other will create smaller DNA fragments compared to regions where Tn5 activity is restricted. These are preferentially amplified and subsequently sequenced meaning that accessible regions will show an enrichment in read coverage compared to background (Buenrostro et al., 2013).

In this chapter, I will present and discuss the results obtained by performing ATAC-seq on wild-type J1 and DNMT.TKO mouse embryonic stem cells with TSA or mock DMSO treatment (Figure 4.1). First, I will focus on a comparison of J1 and DNMT.TKO cells before including the data from TSA treated cells.

ATAC-seq sample preparation was jointly performed by myself and Hamish King, a former member of the Rob Klose lab. Library preparations were carried out by Hamish King.

4.2 Summary of ATAC-seq experiments

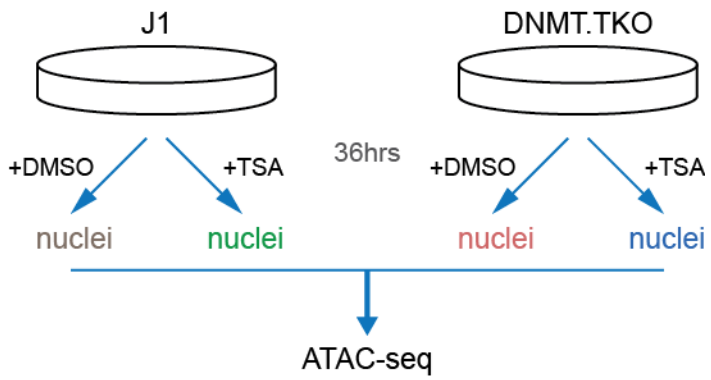


Figure 4.1: Schematic of the experimental approach used to investigate the effects of DNA methylation loss and HDAC inactivation on chromatin accessibility.

To assess the effects of DNA methylation loss and HDAC inhibition on genome-wide chromatin accessibility, nuclei isolated from wild-type J1 and DNMT.TKO mouse ES cells treated with DMSO or TSA were subjected to ATAC-seq in four biological replicate experiments (Figure 1). Tn5 transposase treatment was also performed on naked genomic DNA extracted from untreated J1 and DNMT.TKO cells (only one sample was sequenced for each). This served as a control for potential sequence bias in tagmentation and mapping. Amplified DNA fragments were sequenced on a HiSeq4000 instrument to generate 29-52 million 75bp paired-end reads. Read counts and alignment rates are presented in Table 4.1. Visual inspection of the data revealed high quality and a strong signal to noise ratio (Figure 4.2).

Sample	Total PE reads	Number of uniquely aligned fragments	Percentage of uniquely aligned fragments	Number of multiply aligned fragments	Percentage of multiply aligned fragments	Overall alignment (%)
ATAC_J1.DMSO_REP1	42,059,818	24,147,797	57.41	5,568,748	13.24	70.65
ATAC_J1.DMSO_REP2	40,802,365	24,418,764	59.85	5,213,630	12.78	72.62
ATAC_J1.DMSO_REP3	46,695,157	26,308,595	56.34	5,554,398	11.9	68.24
ATAC_J1.DMSO_REP4	35,410,344	20,100,001	56.76	4,630,521	13.08	69.84
ATAC_J1.TSA_REP1	49,048,271	28,250,192	57.6	7,330,027	14.94	72.54
ATAC_J1.TSA_REP2	40,261,683	24,599,450	61.1	5,252,673	13.05	74.15
ATAC_J1.TSA_REP3	43,202,070	24,893,137	57.62	5,305,652	12.28	69.9
ATAC_J1.TSA_REP4	47,911,819	27,012,373	56.38	6,765,411	14.12	70.5
ATAC_TKO.DMSO_REP1	46,392,280	26,670,591	57.49	6,263,205	13.5	70.99
ATAC_TKO.DMSO_REP2	36,432,184	20,766,129	57	4,893,727	13.43	70.43
ATAC_TKO.DMSO_REP3	41,513,659	23,218,184	55.93	4,939,570	11.9	67.83
ATAC_TKO.DMSO_REP4	44,012,434	25,348,374	57.59	5,739,221	13.04	70.63
ATAC_TKO.TSA_REP1	47,403,166	25,715,700	54.25	6,286,915	13.26	67.51
ATAC_TKO.TSA_REP2	36,279,809	20,576,728	56.72	4,962,908	13.68	70.4
ATAC_TKO.TSA_REP3	38,647,299	21,545,887	55.75	5,026,230	13.01	68.76
ATAC_TKO.TSA_REP4	36,748,337	21,235,191	57.79	5,089,744	13.85	71.64
ATAC.Input_J1	28,923,379	9,708,746	33.57	3,152,246	10.9	44.47
ATAC.Input_TKO	31,144,961	11,878,423	38.14	3,871,262	12.43	50.57

Table 4.1: Total read counts and alignment metrics for all ATAC-seq samples.

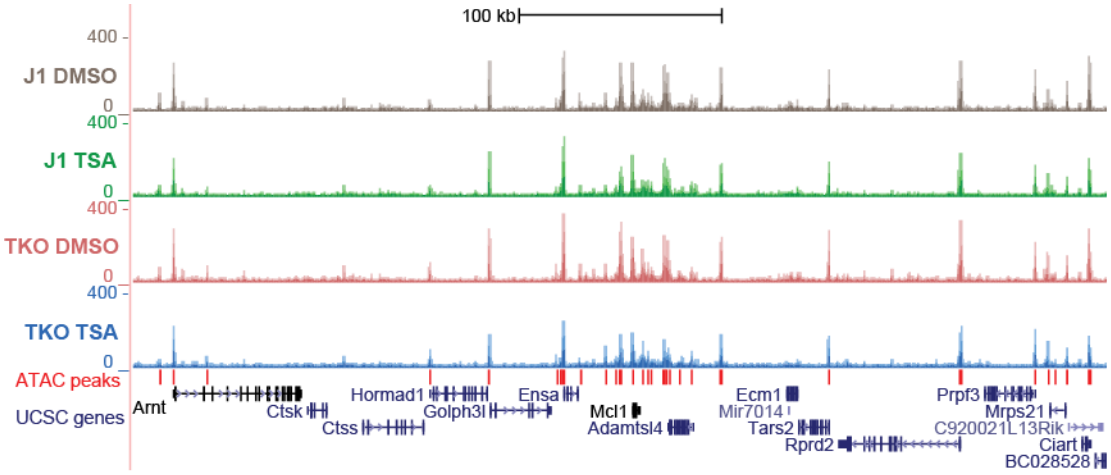


Figure 4.2: UCSC genome browser snapshot showing representative ATAC-seq data. ATAC-seq signal is shown for DMSO or TSA-treated wild-type J1 and DNMT.TKO cells. Coverage graphs were generated after merging ATAC-seq alignments from replicate samples and represent data normalised to library size. The position of ATAC-seq peaks called using Dpeak (DANPOS2 tools, Chen *et al.* 2013) with an FDR threshold of 10^{-40} are indicated. UCSC gene annotations are shown. Mm10 coordinates: chr3:95,416,873-95,899,203 .

4.3 Analysis of the changes in chromatin accessibility occurring upon loss of DNA methylation

4.3.1 Chromatin accessibility landscape in DNMT.TKO compared to wild type cells

In this section we will consider the differences in chromatin accessibility between DMSO-treated wild type and DNMT.TKO cells. Results from the ATAC-seq experiments on TSA-treated cells will be discussed later (Section 4.4)

Focusing on reads that mapped uniquely to the mouse mm10 genome, we identified accessible regions (also referred to as peaks) for each condition using information from replicate samples and the background input samples. Across a range of significance thresholds, a greater number of peaks were called for DMSO-treated DNMT.TKO cells compared to WT cells (Figures 4.3A-B). Going forwards, we settled on a set of peaks that had

a FDR adjusted p-value below 10^{-40} in at least one condition, the threshold we decided was most adequate to ensure peaks called showed enrichment over background without discarding regions that showed a reproducible but small increase in accessibility.

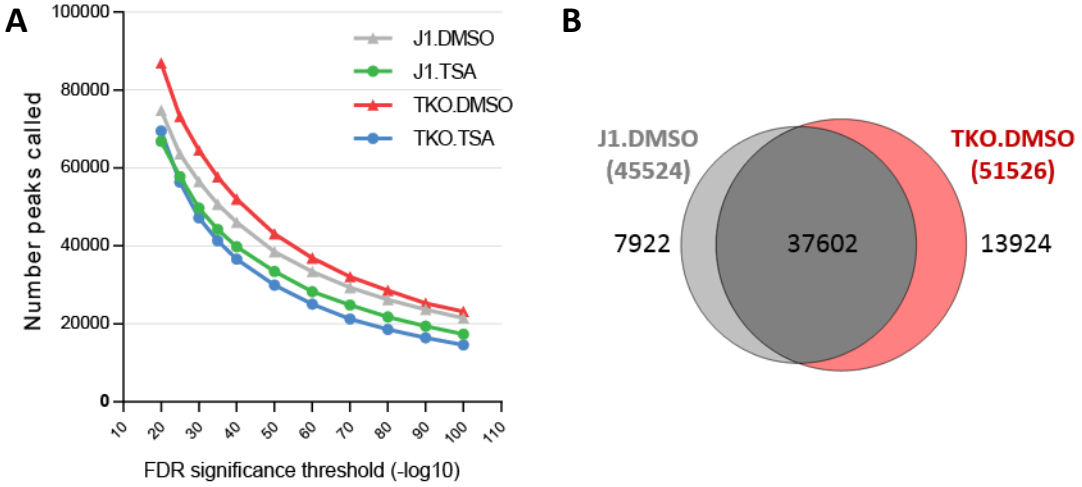


Figure 4.3: ATAC-seq peak calling in wild type J1 and DNMT.TKO cells. A. The number of ATAC-seq enriched regions (peaks) identified in DMSO or TSA-treated wild-type J1 and DNMT.TKO cells using the Dpeak peak-calling algorithm (DANPOS2 tools, Chen *et al.* 2013) for a range of FDR thresholds for significance. **B.** Venn diagram representing the overlap between the ATAC-seq enriched regions identified in DMSO-treated wild type J1 and DNMT.TKO cells using Dpeak with an FDR threshold of 10^{-40}.

To compare the read coverage between the two conditions we performed differential analysis at ATAC-seq peaks using the R package DiffBind (Stark and Brown, 2011). This led us to define sets of 3608 and 1466 regions which significantly gained and lost accessibility in DNMT.TKO cells compared to wild type (FDR > 0.01 and fold change > 1.5) (Figure 4.4).

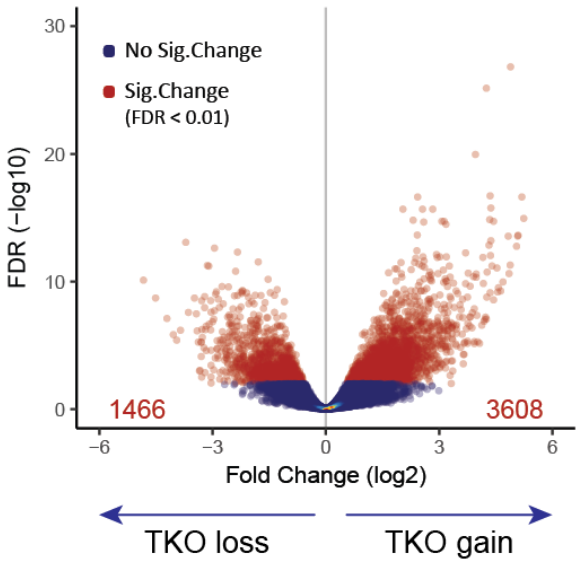


Figure 4.4: Volcano plot representing the false discovery rate (FDR) and fold change values obtained through differential analysis of ATAC-seq signal in mock-treated DNMT.TKO versus wild-type J1 cells. Differential analysis was performed on 59838 peak regions using the R package DiffBind (Stark and Brown, 2011). Significant changes were called with a FDR > 0.01 and fold change > 1.5. The numbers of regions showing significant changes in either direction are indicated in red.

Two examples of regions that gain accessibility in DNMT.TKO cells are presented in Figures 4.5A-B. In general, loci that show significant gains in accessibility in DNMT.TKO compared to WT cells have a lower read density compared to sites that are accessible in both cell lines (Figure 4.5C). In addition there was a small increase in signal at these regions compared to the surrounding genomic space even in WT cells (Figure 4.5D). This observation suggest that they contain elements in the underlying DNA sequence or have a particular chromatin state that can increase sensitivity to transposase. However in the majority of cases, DNA methylation is sufficient to obstruct the formation of an accessible chromatin state at these loci.

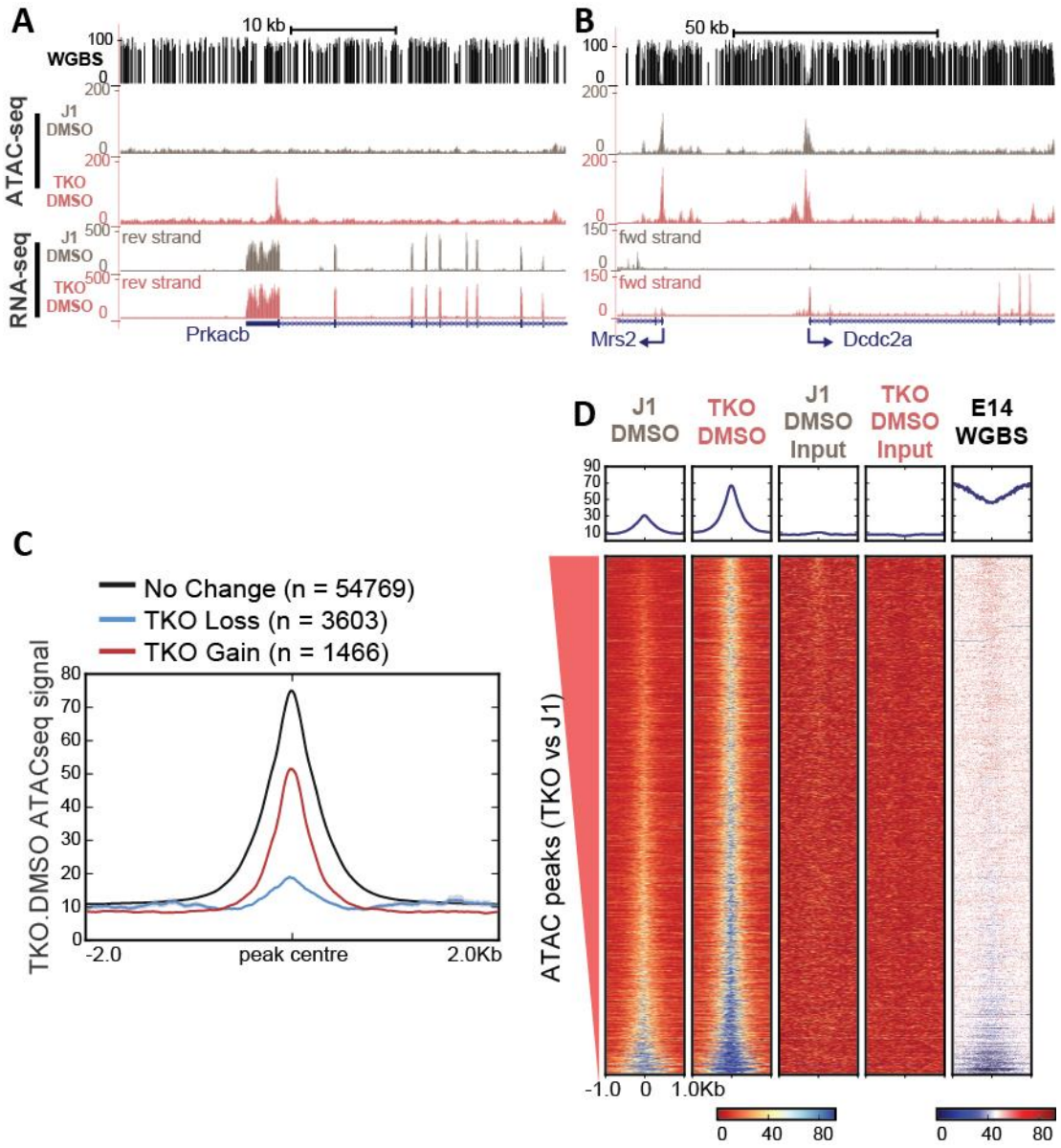


Figure 4.5: Profile of the genomic regions that show chromatin accessibility gains in DNMT.TKO compared to wild-type mESCs. A-B. UCSC genome browser snapshots showing examples of localised gains in chromatin accessibility in DNMT.TKO mES cells compared to wild-type J1. ATAC-seq signal for DMSO-treated J1 and DNMT.TKO cells is shown along with RNA-seq signal from the same cell lines and WGBS data from E14 mESCs (Habibi *et al*., 2013). Coverage graphs were generated after merging ATAC-seq alignments from replicate samples and represent data normalised to library size. UCSC gene annotations are shown. Mm10 coordinates - **A**: chr3:146,717,578-146,759,934 ; **B**: chr13:25,001,896-25,108,797 **C.** Metaplot comparing the ATAC-seq signal from DNMT.TKO cells at regions grouped according to their differential accessibility in DNMT.TKO versus wild type cells. Solid lines show mean ATAC-seq signal at *n* regions with significant gains (red) or losses (blue) in accessibility in DNMT.TKO cells or regions with no significant changes between the two cell lines (black). Light shading indicates the area between the mean +/- the standard error of the mean.

Figure legend continued on next page

Figure legend continued from previous page

D. Heatmaps showing the ATAC-seq or WGBS signal obtained at 3603 regions identified as having significantly increased accessibility in DNMT.TKO versus wild-type J1 cells. Regions were sorted according to the fold change in ATAC-seq signal. Regions span the 2kb surrounding the centre of ATAC-seq peaks and are split into 10bp bins. The colour keys for the heatmap are shown below; these indicate normalised ATAC-seq coverage (left) and mean CpG methylation (right). The profile plots above each heatmap show the average signal obtained across all regions.

In both cell lines most accessible regions localise to intergenic or intronic regions while 26% localise to gene promoter regions (here defined as TSS +/-1500bp). De novo accessible peaks in DNMT.TKO cells are less frequently found at promoters (7.5%) but tend to be localised at more distal sites (Figures 4.6). This is expected as the majority of gene promoters reside within CpG islands which are known to remain unmethylated across the majority of tissue types including ES cells (Hon et al., 2013; Stadler et al., 2011; Ziller et al., 2013). These are therefore unlikely to display chromatin accessibility changes upon DNMT knockout.

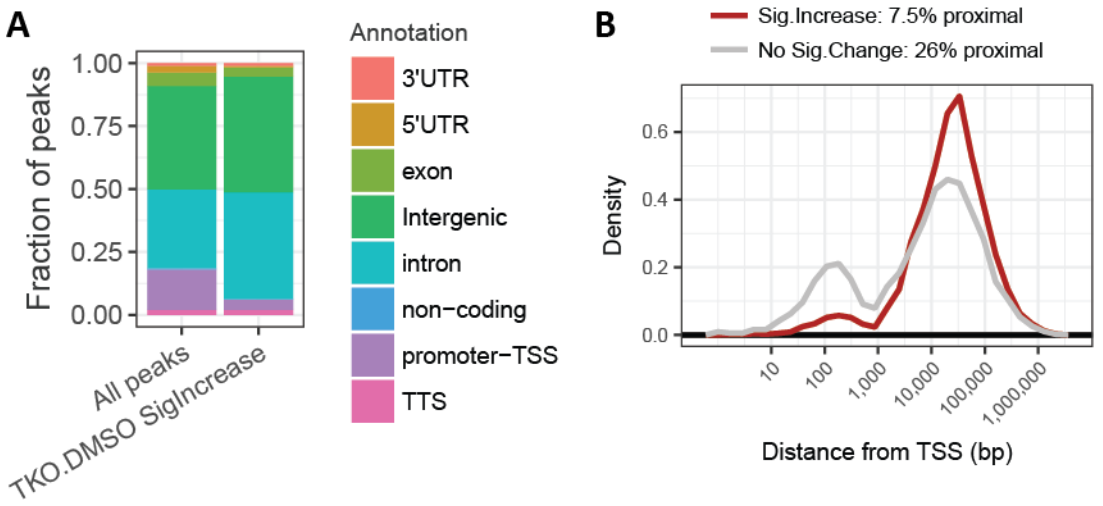


Figure 4.6: Genomic localisation of regions showing chromatin accessibility gains in DNMT.TKO compared to wild-type mESCs. A. The fraction of ATAC-seq peaks that localise to different genomic features. Data is shown for all ATAC-seq peaks (n = 59837) or those with significant gains in DNMT.TKO compared to wild type cells (n = 3608). Promoter-TSS is defined as the region from -1kb to +100bp relative to the TSS. The TTS (Transcription Termination Site) annotation is defined as the region from -100 bp to +1kb relative to a TTS. **B.** Density plots representing the distance of all ATAC-seq peaks (n = 59837) or DNMT.TKO-specific peaks (n = 3608) to the nearest transcription start site (TSS). The percentages of peaks that are proximal to the TSS (centre of peak within 1500 bp of TSS) are indicated.

4.3.2 Relationship between chromatin accessibility and DNA methylation

Using the same E14 mES WGBS dataset as used in the previous chapter (Habibi et al., 2013), we could show that regions that significantly gained accessibility in DNMT.TKO cells are regions with elevated levels of DNA methylation in WT cells (Figures 4.7A and 4.5D). This observation suggests that the increased sensitivity to transposase is likely to be dependent on the loss of DNA methylation. Of the 3288 DNMT.TKO-specific accessible regions for which there is a sufficient (≥5) coverage of bisulfite sequencing reads, 41% had an average methylation level above 60 (Figure 4.7B). Out of the 1000 most differentially accessible sites, the percentage of regions with an average methylation level above 60 rose to 64%. Although many accessible regions are methylated in WT cells, few of these show significantly decreased accessibility in DNMT.TKO cells implying that they do not depend on DNA methylation to maintain this state (Figure 4.7B).

DNMT.TKO–specific accessible sites have low CpG content (Figure 4.7C) and only 36 of these overlap with a methylated CpG island (CpG island intervals were obtained from UCSC mm10 genome annotation and were classified as methylated if they had a mean methylation level above 60). Therefore, site-specific gains in accessibility following DNA methylation loss are not restricted to CpG dense regions.

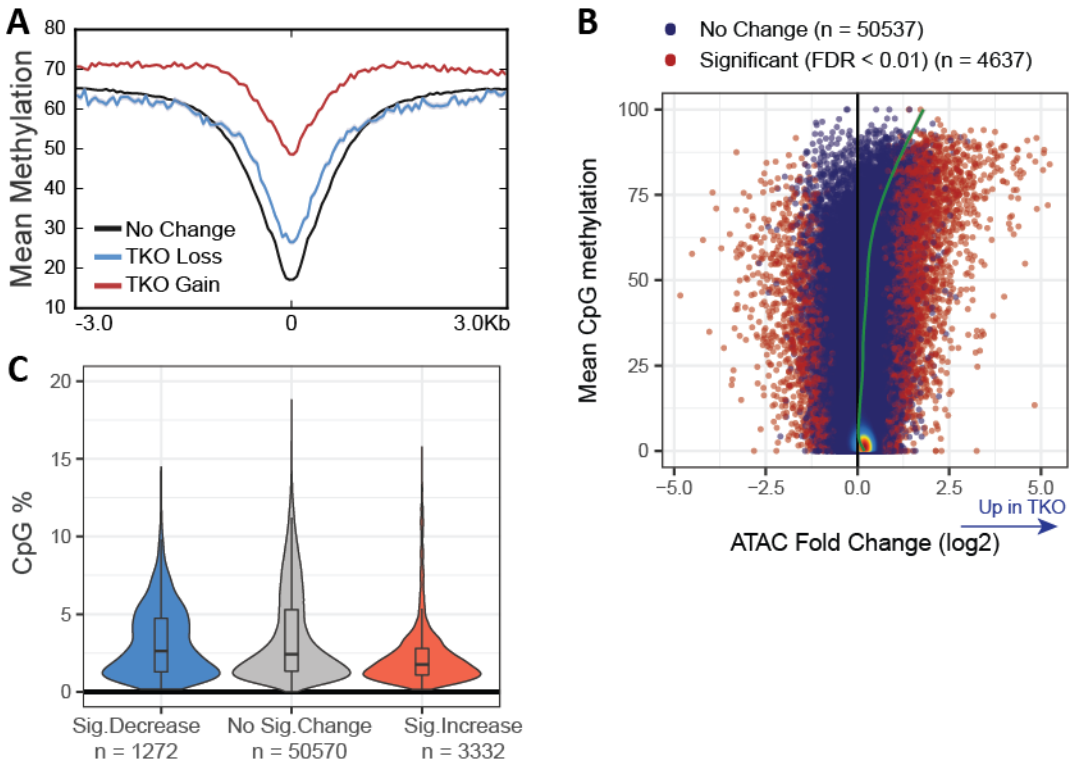


Figure 4.7: DNA methylation and CpG content at ATAC-seq peaks. A. Metaplot summarising mean CpG methylation levels in the 6kb regions surrounding ATAC-seq peaks grouped according to their differential accessibility in DNMT.TKO versus wild type cells. Solid lines show mean ATAC-seq signal at *n* regions with significant gains (red , n = 3603) or losses (blue , n = 1466) in accessibility in DNMT.TKO cells or regions with no significant changes between the two cell lines (black , n = 54769). Light shading indicates the area between the mean +/- the standard error of the mean. CpG methylation values were averaged in 50bp windows. **B.** Scatterplots comparing the change in ATAC-seq signal in DNMT.TKO versus wild type cells with the average CpG methylation levels at ATAC-seq peaks. Intervals showing significant or non-significant differential accessibility are represented by red and blue dots respectively. Methylation values are averaged for all CpG sites within the peak region. The green line represents a smoothened fit of the data modelled using the “gam” method in the R package ‘ggplot2’. **C.** Violin plots summarise the percentage CpG content at ATAC-seq peaks grouped according to their differential accessibility in DNMT.TKO versus wild type cells.

4.3.3 Motif enrichment at regions that are accessible in DNA methylation deficient but not wild type cells

As chromatin accessible sites are thought to represent transcription factor occupancy events (Neph et al., 2012; Thurman et al., 2012) we postulated that accessible sites specific to DNMT.TKO cells are likely to be formed at loci harbouring sequences that are recognised by DNA methylation-sensitive transcription factors. To investigate this, we screened a collection of 264 known vertebrate transcription factor motifs (HOMER database, Heinz et al., 2010) for enrichment at DNMT.TKO specific chromatin accessibility sites. For this analysis, we restricted the number of genomic intervals to those that were methylated and had at least a three-fold increase in accessibility compared to WT cells. Table 4.2 lists the top five known motifs that are enriched at these sites. The identification of the NRF1 motif as a top hit is consistent with the findings of Domcke *et al.*. In our data, this motif, containing two CpG sites, was present at 16-18% of novel accessible sites compared to ~40% in the DNMT.TKO-specific DNaseI hypersensitive sites that they defined (Domcke et al., 2015). Interestingly the NFY complex's cognate binding motif, which can contain a CpG outside of its core sequence, also appeared to be enriched at DNMT.TKO specific ATAC-seq peaks.

De novo motif discovery using the sequences underlying this same set of accessible sites identified a sequence resembling the YY1 recognition motif in addition to NRF1 and NFY motifs (Table 4.3).

Rank	Known Motif	HOMER motif database entry	Enrichment p-value	% of target sequences	% of background sequences
1		**NRF1(NRF)** MCF7-NRF1-ChIP-Seq (Unpublished)	1e-71	16.72	2.78
2		**NFY(CCAAT)**	1e-66	36.93	14.27
3		**NRF(NRF)**	1e-55	18.10	4.3
4		**FOXP1(Forkhead)** H9-FOXP1-ChIP-Seq (GSE31006)	1e-42	20.11	6.56
5		**Lhx3(Homeobox)** Neuron-Lhx3-ChIP-Seq (GSE31456)	1e-21	38.31	24.21

Table 4.2: Known motifs most enriched in DNMT.TKO specific ATAC-seq peaks (target sequences) relative to background sequences. Enrichment analysis of the 264 known vertebrate TF motifs in the HOMER database was performed using the script findMotifsGenome.pl (Heinz et al., 2010). The HOMER motif database entry provides information about the transcription factor's DNA-binding domain (DBD-family in brackets) and the ChIP-seq experiment the motif is derived from.

Rank	De novo Motif – top enrichment	Enrichment p-value	% of target sequences	% of background sequences	Best Match *Origin*	Similarity score
1		1e-158	27.51	3.21	**NFY(CCAAT)** *Homer*	0.929
2		1e-93	18.52	2.52	**NRF1(NRF)** *Homer*	0.964
3		1e-91	11.75	0.74	**MA0095.2_YY1** *Jaspar*	0.801
4		1e-84	14.07	1.44	**MA0031.1_FOXD1** *Jaspar*	0.936
5		1e-71	9.95	0.74	**PH0013.1_Cdx2** *Jaspar*	0.727

Table 4.3: *De novo* discovered motifs most enriched in DNMT.TKO specific ATAC-seq peaks (target sequences) relative to background sequences. De novo motif discovery and enrichment analysis was performed using the HOMER script findMotifsGenome.pl (Heinz et al., 2010). For each novel motif, the known TF motif that was most similar was reported (Best Match and respective similarity score).

4.4 Chromatin accessibility dynamics upon HDAC inhibition

4.4.1 Chromatin accessibility landscape of mES cells following TSA treatment

Using the same peak calling and differential analysis workflow, we examined the effect of HDAC inhibition on the chromatin accessibility landscape of WT or DNMT.TKO mES cells. To our surprise, TSA treatment in both WT and DNMT.TKO settings led to a reduction in the number of sites that were identified as accessible (Figures 4.8A-B and 4.3A). Differential analysis showed that a vast number of regions were appearing to lose accessibility in TSA treated wild type cells while few regions gained accessibility (Figure 4.8C). This observation was even more pronounced in DNMT.TKO in which 28% percentage of peaks showed a significant reduction in ATAC-seq signal (Figure 4.8D).

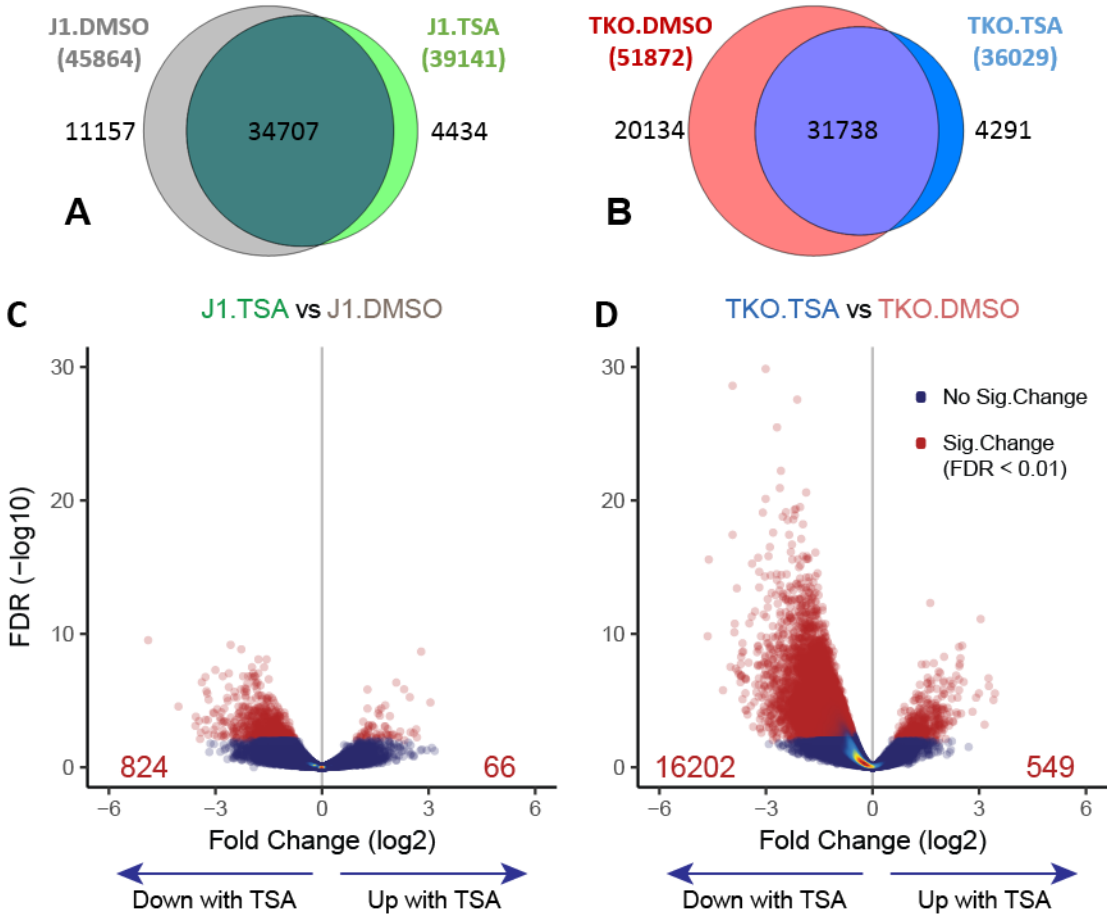


Figure 4.8: Comparison of the chromatin accessible landscape in TSA or DMSO treated mESCs

Figure legend continued on next page

Figure legend continued from previous page

A-B. Venn diagrams representing the overlap between the ATAC-seq enriched regions identified in TSA and DMSO-treated wild type J1 (**A**) or DNMT.TKO cells (**B**) using the Dpeak algorithm with an FDR threshold of 10^{-40} (DANPOS2 tools, Chen *et al.* 2013). **C-D.** Volcano plots representing the false discovery rate (FDR) and fold change values obtained through differential analysis of ATAC-seq signal from TSA or mock-treated wild-type J1 (**C**) or DNMT.TKO cells (**D**). The analysis was performed on 59838 peak regions. Significant changes were called with a FDR > 0.01 and fold change > 1.5. The numbers of regions showing significant changes in either direction are indicated in red.

In wild-type cells, peaks with decreased accessibility after TSA treatment were preferentially localised distal to the transcription start sites whereas in DNMT.TKO cells, the loss of signal was generalised to all peaks irrespective of their localisation (Figure 4.9). Examples of peaks with reduced accessibility are displayed in Figure 4.10. From these examples, it appears that in wild-type cells the decrease in accessibility affects select peaks whereas in DNMT.TKO cells the effect is widespread and all accessible regions have reduced signal.

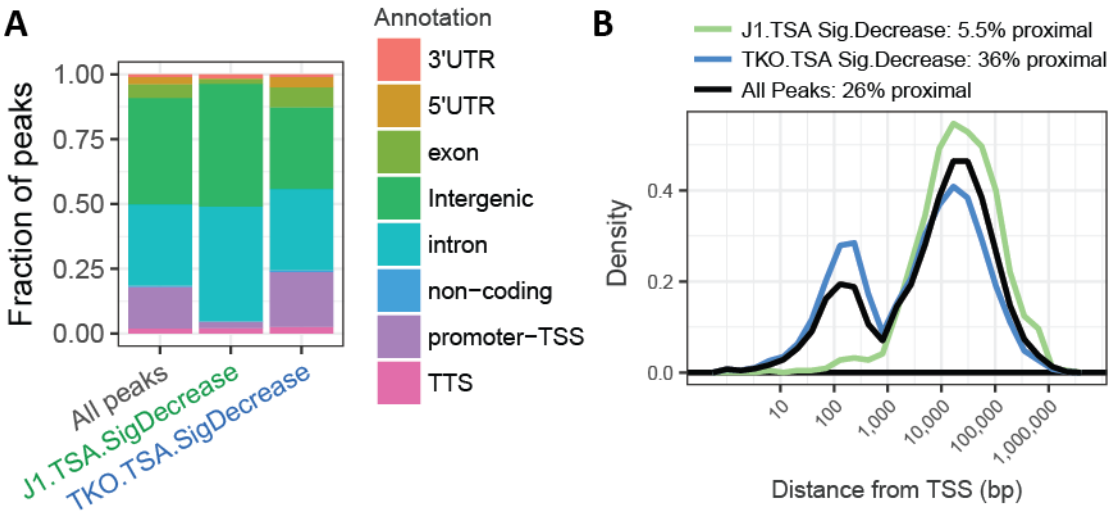


Figure 4.9: Genomic localisation of regions showing significant reductions in chromatin accessibility in TSA- compared to mock-treated mESCs. A. The fraction of ATAC-seq peaks that localise to different genomic features. Data is shown for all ATAC-seq peaks (n = 59837) or those with significant losses in TSA versus DMSO wild-type or DNMT.TKO cells (n = 824 and 16205 respectively). Promoter-TSS is defined as the region from -1kb to +100bp relative to the TSS. The TTS (Transcription Termination Site) annotation is defined as the region from -100 bp to +1kb relative to a TTS. **B.** Density plots representing the distance to the nearest transcription start site (TSS) for all ATAC-seq peaks (n = 59837) or peaks showing a significant reduction in signal after TSA treatment of wild-type (n = 824) or DNMT.TKO cells (n = 16205). The percentages of peaks that are proximal to the TSS (centre of peak within 1500 bp of TSS) are indicated.

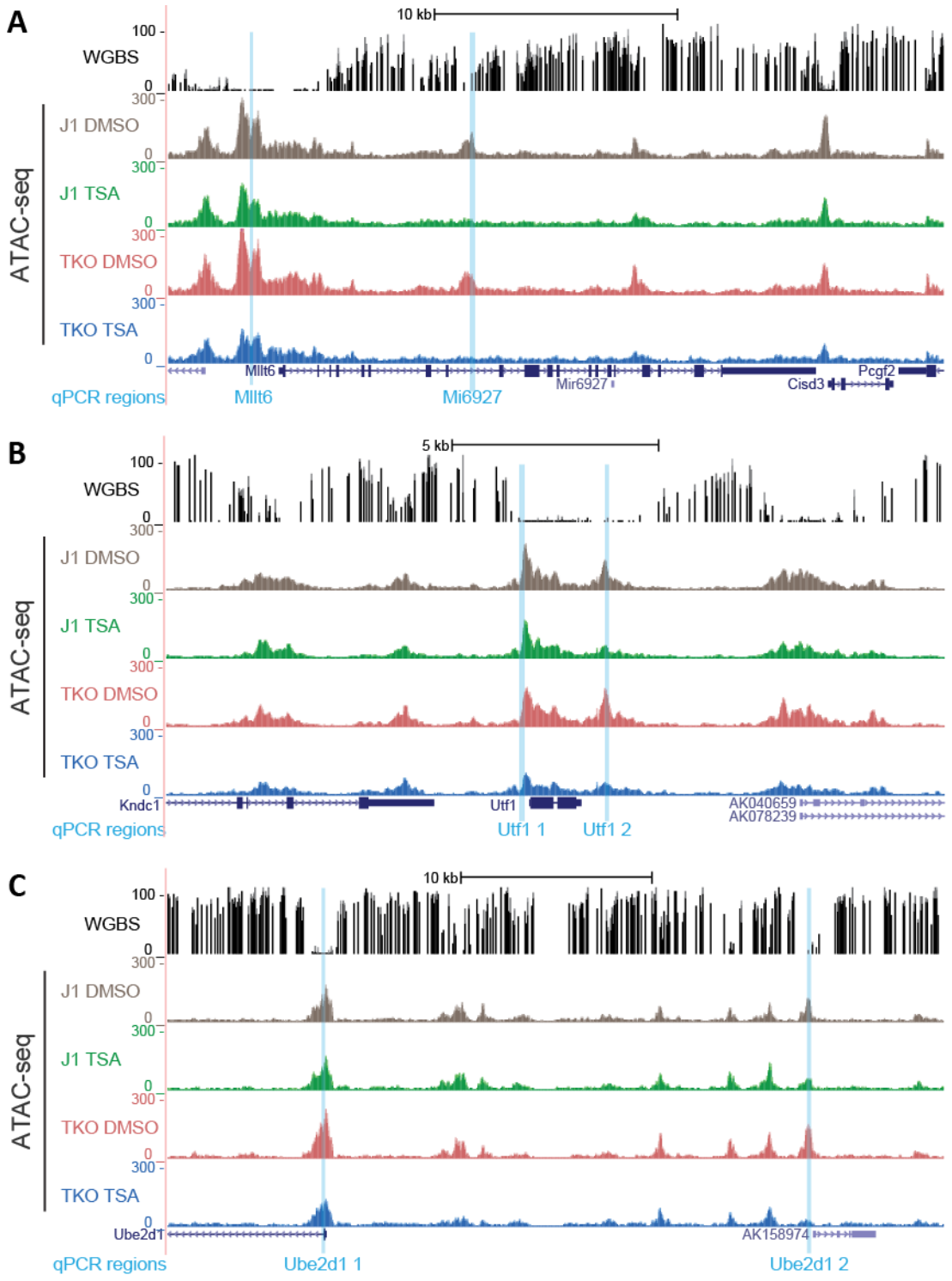


Figure 4.10: UCSC genome browser snapshots showing loci with reduced ATAC-seq signal in TSA treated cells. ATAC-seq signal is shown for DMSO or TSA-treated wild-type J1 and DNMT.TKO cells along with WGBS data from E14 mESCs (Habibi et al., 2013). Coverage graphs were generated after merging ATAC-seq alignments from replicate samples and represent data normalised to library size. The position of regions analysed by qPCR are indicated (see section 4.4.2). UCSC gene annotations are shown. Mm10 coordinates: **A**: chr11:97,658,837-97,690,704 ; **B**: chr7:139,935,057-139,953,911 ; **C:** chr10:71,276,948-71,317,574.

Despite this, there is a positive correlation between the fold change values obtained for each peak region when comparing the effect of TSA treatment in the two cell lines (Figure 4.11, Spearman correlation = 0.51). In both conditions, the median fold change shifts below zero. This observation suggests that the response to TSA is similar in both cell lines and the difference is quantitative rather than qualitative.

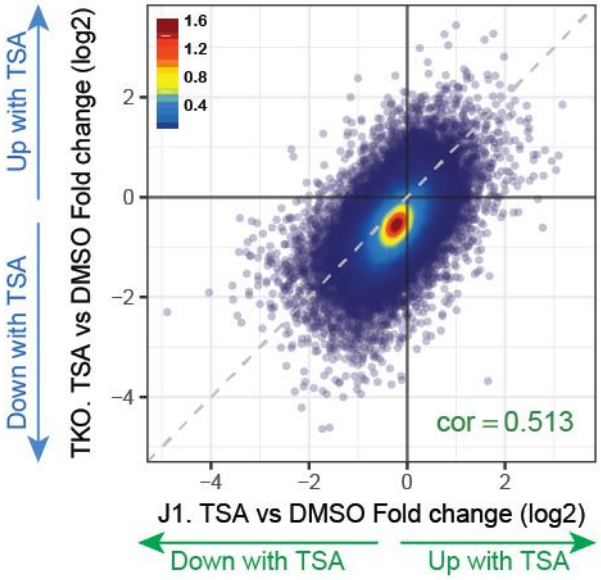


Figure 4.11: Scatterplot comparing the fold change in ATAC-seq signal after TSA treatment in DNMT.TKO versus wild type J1 cells. The dashed line has a slope of 1 and intercept of 0. The Spearman correlation score is indicated. Fold change values were obtained through differential analysis of the read coverage at 56705 ATAC-seq peaks.

4.4.2 Understanding the redistribution of ATAC-seq reads in TSA mESCs

The results of the differential analyses presented above are based on read count data that have been normalised to the total ATAC-seq library size. Consequently, the reduction of reads mapping to the peak intervals must be accompanied by a relative increase in signal in the background genome. In an attempt to understand where the reads were being redistributed to, we calculated the overall coverage for genomic regions falling into three mutually exclusive categories: transposase accessible regions (ATAC-seq peak set defined above), genomic intervals contained in the RepBase database of repetitive elements (excluding

sequences that overlapped with ATAC-seq peaks) and the remaining genomic space. The reason for separating out the repetitive regions was based on our previous observation that TSA treatment led to the widespread transcriptional activation of transposable elements (see Chapter 3). We hypothesised that an elevated sensitivity to Tn5 in these regions would cause an increase in reads mapping to these elements and a proportional reduction at conventional ATAC-seq peaks.

In agreement with the results above, there were fewer reads mapping to peaks in TSA-treated cells compared to untreated cells (Figure 4.12). On the other hand, an increase was seen both at Repbase elements as well as the remaining genomic space although the difference was most significant in the former.

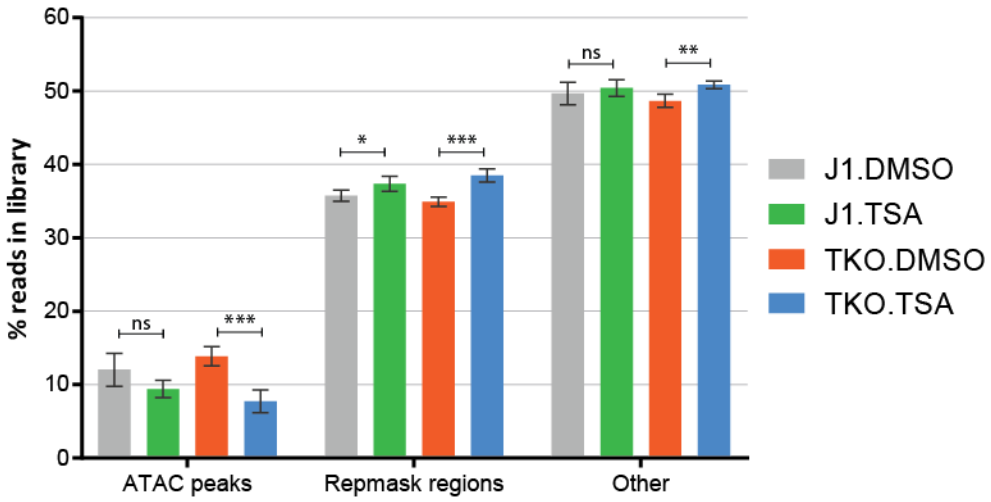


Figure 4.12: Percentage of reads from each ATAC-seq library that mapped to peaks, repetitive elements (Repmask regions) or the remaining genomic space (Other). Mean values from four biological replicates are plotted; error bars indicate standard deviation. Two-tailed Student t-tests were used to determine significant differences: * p-value < 0.05 ; ** p-value < 0.01 ; *** p-value < 0.001 ; *ns* = non-significant (p-value >0.05).

Due to the way in which the data was normalised, it is difficult to interpret this data in a biologically meaningful manner. With no fixed reference with which to estimate the number of sequenced fragments per cell we cannot determine whether the relative change in signal-to-noise ratio between samples represents a real decrease in accessibility at the peaks or an increase in the background.

In order to validate the sequencing data and as a method of addressing this issue, we carried out real-time PCR on the same ATAC libraries that were submitted for sequencing. We designed primers that would amplify small regions that showed a range of fold changes between TSA and mock treated cells (Table 4.4) as well as three background loci. Figure 4.13A shows the number of sequenced fragments that fully overlap these amplicons, allowing us to appropriately compare the real-time PCR and sequencing data.

qPCR region name	J1.TSA.vs.J1.DMSO Fold change	TKO.TSA.vs.TKO.DMSO Fold change
Sepsecs_1	1.04	-1.36
Tbp	-1.06	-1.58
Ube2d1_1	1.02	-1.64
Mllt6	-1.35	-1.82
Utf1_1	-1.22	-1.88
Utf1_2	-2.23	-2.19
Sepsecs_2	-2.20	-2.28
Mir6927	-3.68	-6.06
Ube2d1_2	-1.79	-6.96

Table 4.4: Fold change in accessibility at specific ATAC-seq peaks after TSA treatment. ATAC-seq coverage surrounding the Mllt6, Mir6927, Utf1_1, Utf1_2, Ube2d1_1 and Ube2d1_2 regions is shown in Figure 4.10.

To compare qPCR results between samples and remove variation due to unequal amounts of starting material in the reaction, it was necessary to normalise the data using values obtained at a reference locus whose accessibility level is the same between different

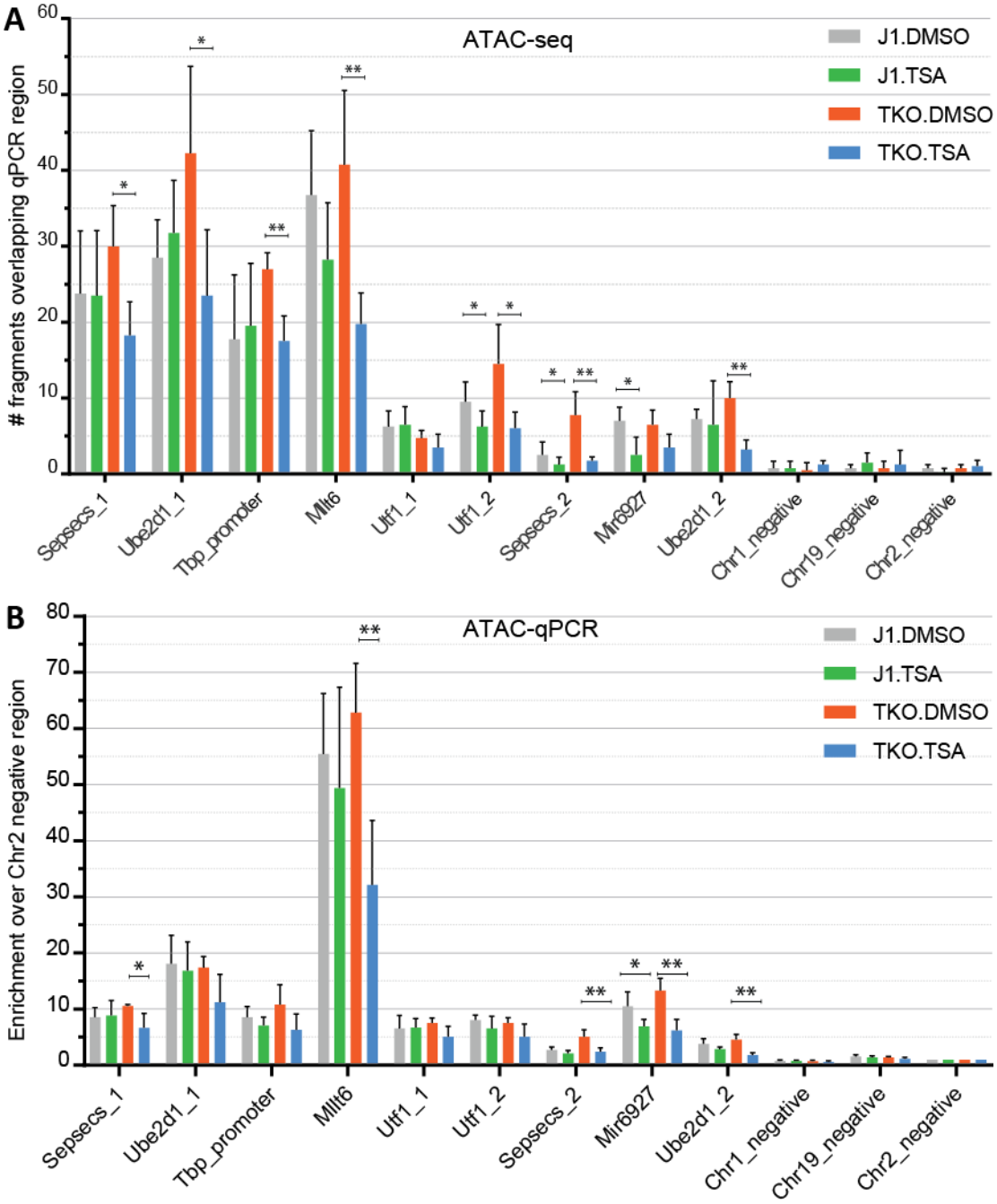


Figure 4.13: Quantification of ATAC fragments that overlap defined genomic loci by high-throughput sequencing (A) or qPCR (B). A. Raw number of sequenced fragments that fully overlap the regions amplified by the primers designed for qPCR. **B.** Enrichment for ATAC DNA fragments covering these same regions, relative to the Chr2_negative locus as measured by real-time PCR. **A-B.** Mean values from four biological replicates are plotted; error bars indicate standard deviation. Two-tailed Student t-tests were used to determine significant differences: * p-value < 0.05 ; ** p-value < 0.01 ; *** p-value < 0.001. Primer sequences and the region's coordinates are listed in Chapter 2 – Table 2.3.

conditions. As a reference we chose one of the three background loci. These were designed in gene desert regions that displayed no ATAC-seq or ChIP-seq signal and were not proximal to repetitive elements. We saw little reason to believe these sites would be more sensitive to Tn5 in TSA treated cells and therefore made the assumption that they were invariable.

The qPCR results confirm that TSA treated samples have reduced sensitivity to transposase compared to untreated cells matching what we observed with the sequencing data (Figure 4.13B). However the difference between the treated and untreated samples does not often reach the same significance value than in the sequencing data. The two background regions that were not chosen as reference points do not show any significant difference in enrichment. If our assumption that the selected reference region is invariable between samples is correct, this data suggests that the loss of accessibility seen at ATAC-seq peaks is representative of changes occurring in the cells and not simply a result of reads being redistributed away from peaks by the dominant effect of TSA on the background genome. Although confirming that the ATAC signal (ie. coverage) is reduced in TSA- compared to DMSO- treated cells at a selection of ATAC-seq peak regions relative to background regions, these qPCR results are not sufficient to conclude as to an absolute change in accessibility between the two conditions. Such quantitative comparison could only be achieved by calibrating the data to an external reference (for example exogenous spike-ins added in equal amounts to each sample prior to the ATAC assay, see section 4.7 and discussion sections 4.9.3 and 4.9.4).

4.4.3 Motif enrichment at regions with decreased chromatin accessibility in TSA-treated cells

ATAC-seq peaks that were significantly lost in wild-type cells treated with TSA showed enrichment for motifs recognised by the pluripotency factors OCT4 and SOX2 (Table 4.5). The enrichment for these motifs was still significant when comparing peaks with decreased signal against a background made up of sequences found at all accessible sites (Table 4.6) however this results should be taken lightly. Indeed, these motifs are frequently found at all accessible sites, even those not affected by TSA treatment.

Rank	Known Motif	HOMER motif database entry	Enrichment p-value	% of target sequences	% of background sequences
1		**OCT4-SOX2-TCF-NANOG (POU,Homeobox,HMG)** mES-Oct4-ChIP-Seq (GSE11431)	1e-137	32.77	5.07
2		**Sox2 (HMG)** mES-Sox2-ChIP-Seq (GSE11431)	1e-45	47.33	24.49
3		**Oct4 (POU,Homeobox)** mES-Oct4-ChIP-Seq (GSE11431)	1e-43	30.10	11.91
4		**Sox3 (HMG)** NPC-Sox3-ChIP-Seq (GSE33059)	1e-36	66.5	44.60
5		**Sox6 (HMG)** Myotubes-Sox6-ChIP-Seq (GSE32627)	1e-27	59.95	40.77

Table 4.5: Known motifs most enriched in the ATAC-seq peaks that are lost upon TSA treatment in wild-type cells (target sequences) relative to background sequences. Enrichment analysis of the 264 known vertebrate TF motifs in the HOMER database was performed using the script findMotifsGenome.pl (Heinz et al., 2010). The HOMER motif database entry provides information about the transcription factor's DNA-binding domain (DBD-family in brackets) and the ChIP-seq experiment the motif is derived from.

Rank	Known Motif	HOMER motif database entry	enrichment p-value	% of target sequences	% of background sequences
1		**OCT4-SOX2-TCF-NANOG (POU,Homeobox,HMG)** mES-Oct4-ChIP-Seq(GSE11431)	1e-27	33.51	11.99
2		**Sox3 (HMG)** NPC-Sox3-ChIP-Seq(GSE33059)	1e-13	63.4	44.36
3		**Sox10 (HMG)** SciaticNerve-Sox3-ChIP-Seq (GSE35132)	1e-11	57.22	39.88
4		**Sox2 (HMG)** mES-Sox2-ChIP-Seq (GSE11431)	1e-8	42.53	28.59
5		**Sox6 (HMG)** Myotubes-Sox6-ChIP-Seq (GSE32627)	1e-8	54.38	39.81
6		**Oct4 (POU,Homeobox)** mES-Oct4-ChIP-Seq (GSE11431)	1e-8	26.55	15.15

Table 4.6: Known motifs most enriched in the ATAC-seq peaks that are lost upon TSA treatment in wild-type cells (target sequences) relative to background sequences comprised of invariable ATAC-seq peaks. Enrichment analysis of the 264 known vertebrate TF motifs in the HOMER database was performed using the script findMotifsGenome.pl (Heinz et al., 2010). The HOMER motif database entry provides information about the transcription factor's DNA-binding domain (DBD-family in brackets) and the ChIP-seq experiment the motif is derived from.

Looking through the RNA-seq data reveals that the expression of *Pou5f1* (*Oct4)* as well as other genes encoding pluripotency factors are downregulated in response to TSA (Table 4.7). Reduced genomic occupancy of these proteins could therefore be in part responsible for the TSA-induced loss of accessibility we observe.

	Mean FPKM				Fold Change			FDR adjusted p-value		
Geneid	J1 DMSO	TKO DMSO	J1 TSA	TKO TSA	TKO vs J1	J1 TSA vs DMSO	TKO.TSA vs DMSO	TKO vs J1	J1 TSA vs DMSO	TKO.TSA vs DMSO
Pou5f1	630.350	650.668	284.080	242.835	1.032	0.451	0.373	8.69E-01	**3.42E-13**	**1.03E-19**
Sox2	161.292	195.160	130.667	146.907	1.210	0.810	0.753	1.40E-01	5.25E-02	6.74E-03
Klf4	8.933	15.444	5.057	7.674	1.728	0.565	0.498	5.74E-04	1.95E-04	**1.41E-06**
Myc	9.546	6.603	3.978	3.112	0.692	0.417	0.470	3.86E-02	**1.84E-08**	**2.33E-06**

Table 4.7: Gene expression changes for genes encoding the Yamanaka transcription factors (Takahashi and Yamanaka, 2006) as measured by RNA-seq. Significant changes are highlighted in red ((FDR-adjusted p-value<10^-5).

4.4.4 Chromatin accessibility changes associated with HDAC inhibition compared to DNA methylation loss.

One of our initial aims was to determine whether DNA methylation could be playing a role on chromatin accessibility through the action of HDACs. As implied by the limited number of regions that significantly gained accessibility following TSA treatment, this does not appear to be the case. Indeed novel accessible sites in DNMT.TKO cells do not show increased accessibility in wild-type cells treated with an HDAC inhibitor (Figure 4.14). Therefore a mechanism by which DNA methylation is recruiting HDAC proteins to preserve an inaccessible chromatin state seems unlikely.

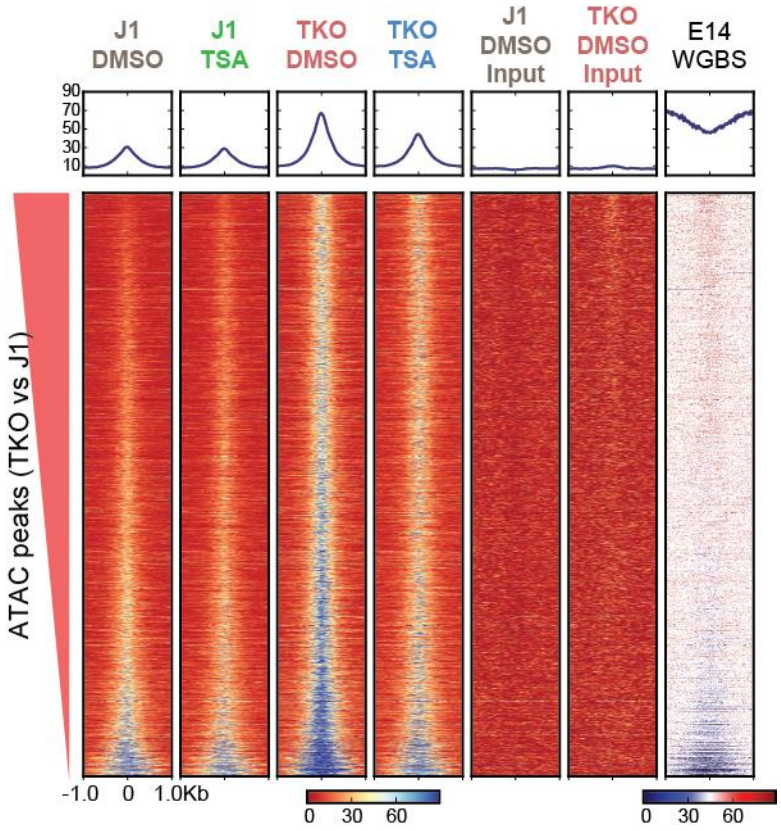


Figure 4.14: Heatmaps showing the ATAC-seq or WGBS signal obtained at 3603 regions identified as having significantly increased accessibility in DNMT.TKO versus wild-type J1 cells. Regions were sorted according to the fold change in ATAC-seq signal. Regions span the 2kb surrounding the centre of ATAC-seq peaks and are split into 10bp bins. The colour keys for the heatmap are shown below; these indicate normalised ATAC-seq coverage (left) and mean CpG methylation (right). The profile plots above each heatmap show the average signal obtained across all regions.

4.5 Exploring the relationship between changes in chromatin accessibility and gene expression

To investigate the impact of chromatin accessibility changes on gene expression we assigned each ATAC-seq peak to its nearest transcription start site and compared accessibility dynamics to gene expression differences (Figure 4.15). Whichever two conditions are compared the correlation between chromatin accessibility and expression of the neighbouring gene is poor (Spearman correlations for TKO.DMSO versus J1.DMSO, J1.TSA versus J1.DMSO or TKO.TSA versus TKO.DMSO are 0.12, 0.14 and 0.135 respectively). Most chromatin accessibility changes are not associated with transcriptional changes at the nearest transcription start site.

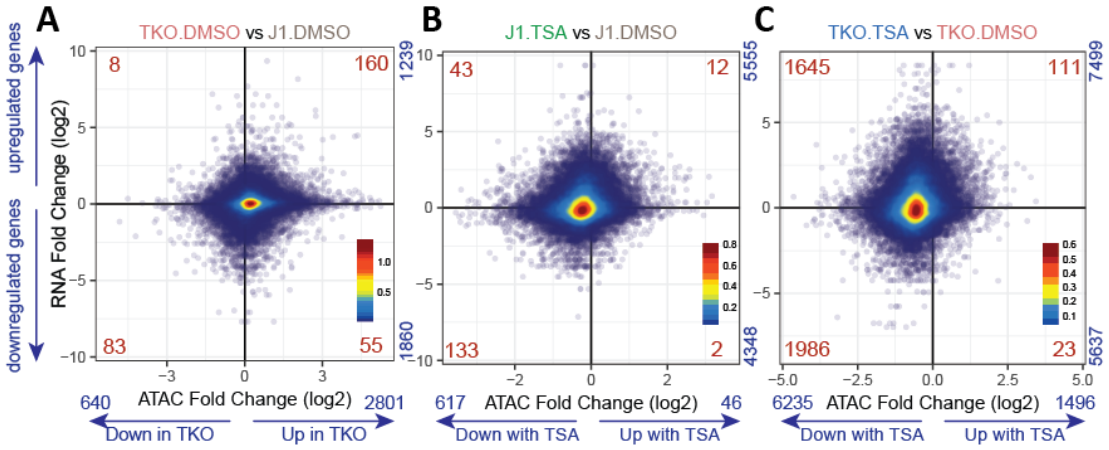


Figure 4.15: Scatterplots comparing the changes in accessibility and gene expression at ATAC-seq peaks and their closest gene. The numbers of ATAC-seq peaks associated with significant changes in both accessibility and gene expression are indicated in red in each quadrant. The total numbers of regions associated with significant changes in accessibility or gene expression are indicated in blue on the outside of the plots. The analysis was performed on 44206 peak-gene pairs involving 15100 unique genes. ATAC-seq peaks located on chromosome 15 were excluded for this analysis.

In DNMT.TKO cells, 160 site-specific gains in accessibility are linked with a significant increase in expression (78 are expected by chance according to the hypergeometric distribution) with 119 different genes involved. Of these 160 sites, 78 were methylated (mean methylation of

region > 60) and 83 were located within 1500bp of the TSS. We could not find any significant difference in the sequence composition of ATAC-seq peak regions that were associated with increased transcriptional activity (in terms of CpG and GC content or enrichment for transcription factor binding site motifs, data not shown). In summary, while it appears that DNA methylation can restrict gene expression by maintaining an inaccessible chromatin state, there are only rare elements at which this seems to be a primary repressive mechanism. Two examples are shown in Figure 4.16. Increased accessibility at the *Fkbp6* promoter in DNMT.TKO cells is associated with a gain in gene expression (Figure 4.16A) while at the *Nefm* promoter, increased accessibility is not sufficient for transcription; expression is only induced upon TSA treatment in the absence of DNA methylation (Figure 4.16B).

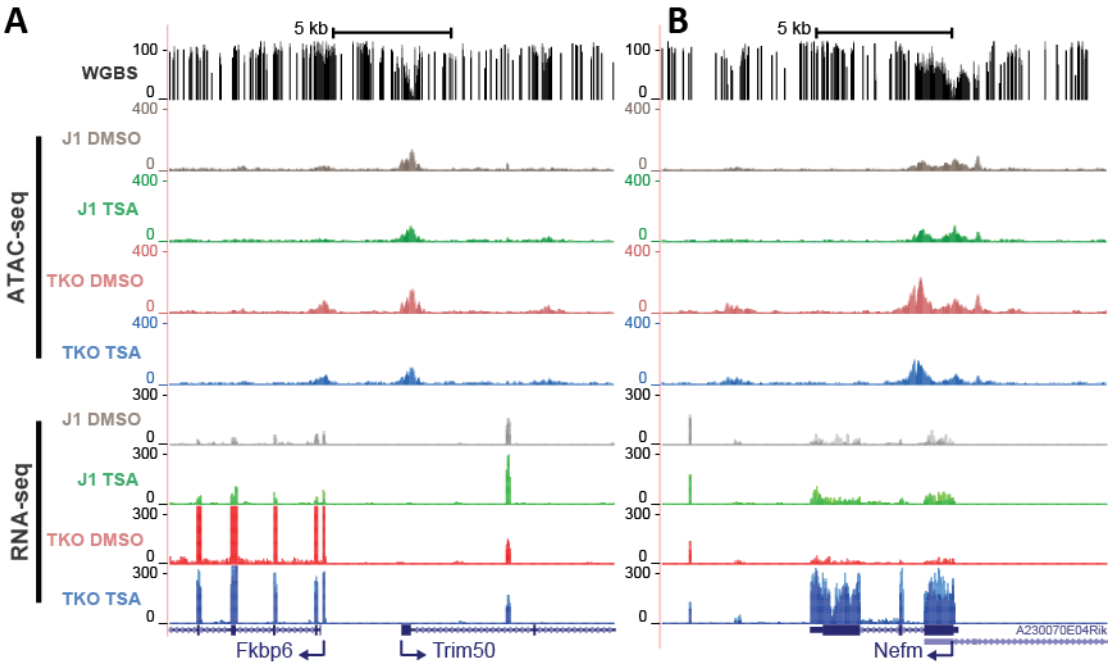


Figure 4.16: UCSC genome browser snapshots showing loci at which gains in accessibility in DNMT.TKO cells can be associated with changes in gene expression. ATAC-seq signal is shown for DMSO or TSA-treated wild-type J1 and DNMT.TKO cells along with RNA-seq signal from the same cell lines and WGBS data from E14 mESCs (Habibi et al., 2013). Coverage graphs were generated after merging ATAC-seq alignments from replicate samples and represent data normalised to library size. UCSC gene annotations are shown. Mm10 coordinates: **A**: chr5:135,343,481-135,362,247 ; **B**: chr14:68,114,085-68,130,464 .

4.6 Differential chromatin accessibility analysis at repetitive elements

Our RNA-seq data showed that both loss of DNA methylation and HDAC inhibition could lead to the transcriptional activation of numerous transposable elements, especially LTR elements. Here we explore whether these gains in expression are associated with changes in chromatin accessibility. We employed the same differential analysis tool used for the RNA-seq data to compare the number of ATAC-seq reads from different samples that map to each entry in the RepBase collection.

DNMT.TKO cells display significant gains in accessibility at a subset of 48 LTR retrotransposons (Figure 4.17A) but few other repetitive element types displayed significant changes. Consistent with the RNA-seq results, it was mostly elements belonging to the ERVK and ERV1 but not the ERVL family that were affected (Figure 4.17B). These retrotransposons appeared in the genome relatively recently based on the average substitution rate at individual instances (Figure 4.17C) and have a high CpG content (Figure 4.17D) fitting with a possible role for DNA methylation in regulating their activation. The vast majority of the significantly increased repeat elements did not gain accessibility after TSA treatment suggesting a role for DNA methylation independent of HDAC activity (Figure 4.17E).

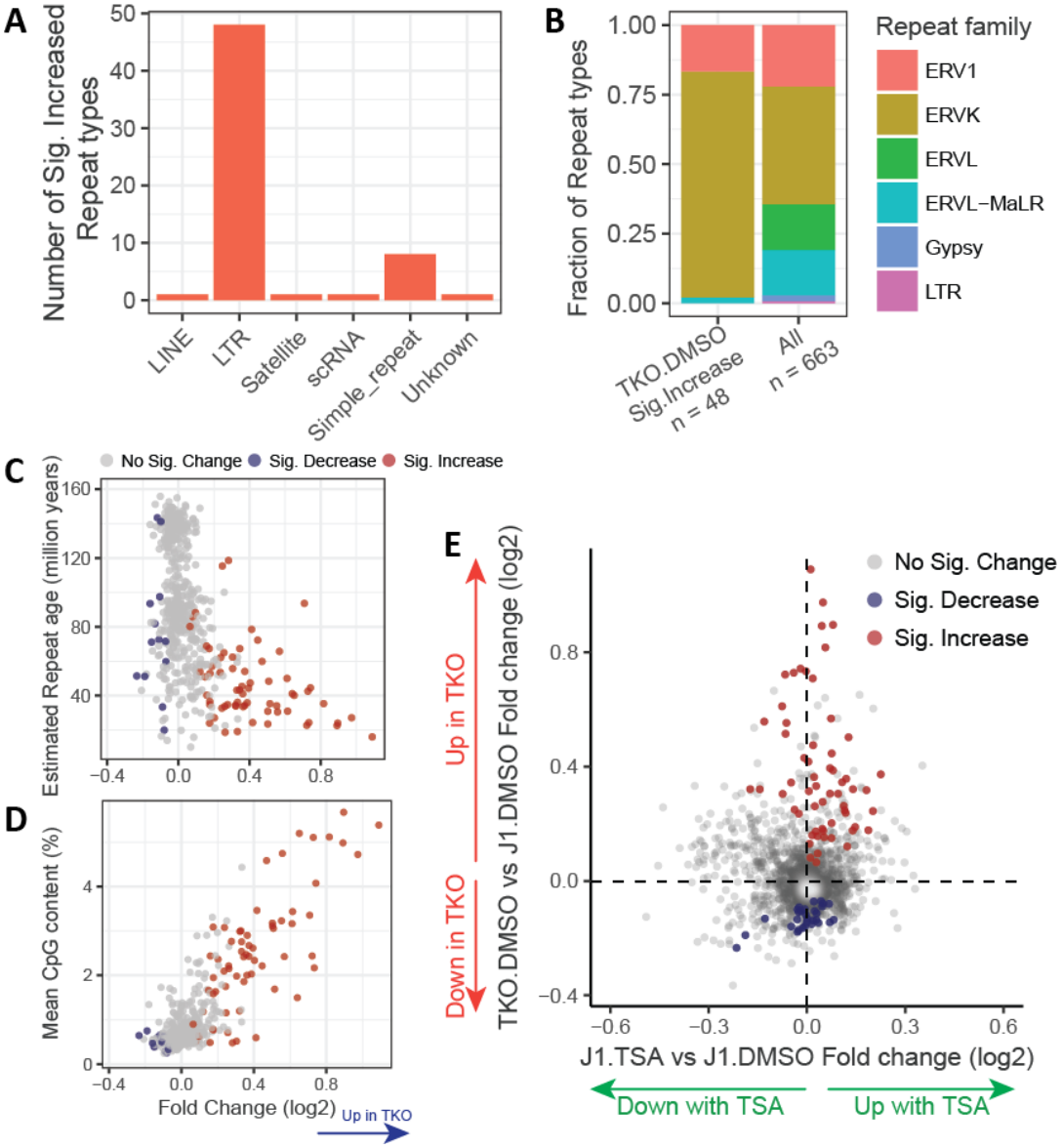


Figure 4.17: Differential analysis of ATAC-seq signal at repetitive elements in DNMT.TKO compared to wild-type cells. ATAC-seq signal was measured as the number of reads that map to one or more of the repeat element copies located across the genome. Differential analysis was performed using the R package DESeq2 (Love et al., 2014). **A.** Number of repeat elements types that show significantly increased signal in DNMT.TKO versus wild type J1 cells according to class. **B.** Fraction of LTR class elements that belong to each sub-family. **C.** Scatterplot comparing for each LTR repetitive element, the estimated repeat age to the fold change in ATAC-seq signal across all genomic copies. **D.** Scatterplot comparing for each LTR repetitive element, the mean CpG content to the fold change in ATAC-seq signal across all genomic copies. **E.** Scatterplot comparing the changes in ATAC-seq signal (log2 fold change) obtained after removal of DNA methylation (DNMT.TKO versus WT cells) or after TSA treatment of wild type J1 cells. In figures **C-E**, the repeat element types showing no significant change in ATAC-seq signal in DNMT.TKO versus wild type cells are represented by grey dots while those showing a significant increase or decrease in DNMT.TKO cells are represented by red or blue dots respectively. Differential analysis was performed using the R package DESeq2 (Love et al., 2014). Changes were deemed significant with an *FDR* < 0.05.

Surprisingly, a large proportion of the LTR types that showed significantly increased expression after DNA methylation depletion were not associated with a significant gain in accessibility (33 out of 45 LTR types, Figure 4.18A, blue points). Moreover only 12 of the 48 LTR elements that increased in accessibility were also transcriptionally upregulated in DNMT.TKO cells (Figure 4.18A gold and red points respectively). Interestingly, the DNMT.TKO-accessible LTR elements included a few of the subclasses that showed weak gains in expression in DNMT.TKO relative to WT cells but synergistic increases in expression following DNA methylation loss and HDAC inhibition (Figure 4.18B).

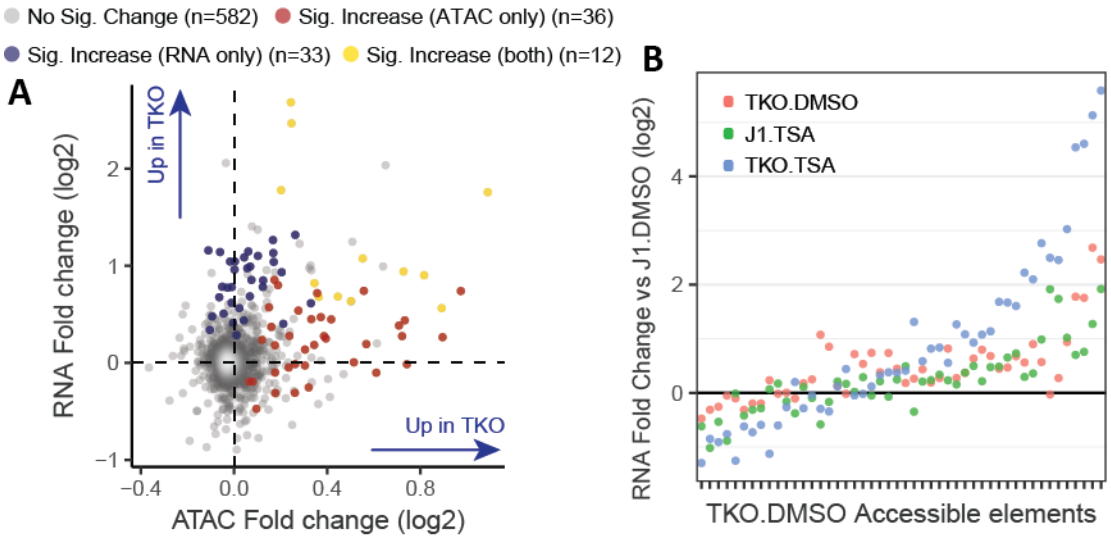


Figure 4.18: Relationship between changes in accessibility and transcription at LTR retrotransposons. A. Scatterplot comparing, for each type of LTR repeat element, the fold-change in ATAC-seq signal to the fold change in RNA-seq signal, both measured across all genomic copies. **B.** RNA-seq fold change values for the 48 LTR-class retrotransposon elements that show significant gains in ATAC-seq signal in DNMT.TKO versus wild type cells. Different LTR elements are sorted along the x-axis according to their fold change in TSA-treated DNMT.TKO cells compared to DMSO-treated wild type cells.

Compared to the changes seen in DNMT.TKO cells versus wild-type mESCs, TSA treatment led to fewer and less robust changes in accessibility (Figure 4.19). TSA treatment of DNMT.TKO cells caused more extensive changes in accessibility, including at some LTR retrotransposons (59 repeat types including 11 ERV1, 24 ERVK, 24 ERVL-family repeats) and L1 elements although the amplitude of these changes appears relatively low (Figure 4.19). Table 4.8 shows the top 30 significantly transcribed transposable elements in TSA treated DNMT.TKO cells compared to wild-type cells (comparable to Table 3.3 in the previous chapter) and specifies changes in RNA-seq and ATAC-seq signal for each. DNA methylation loss seems to have a more pronounced effect on accessibility than TSA.

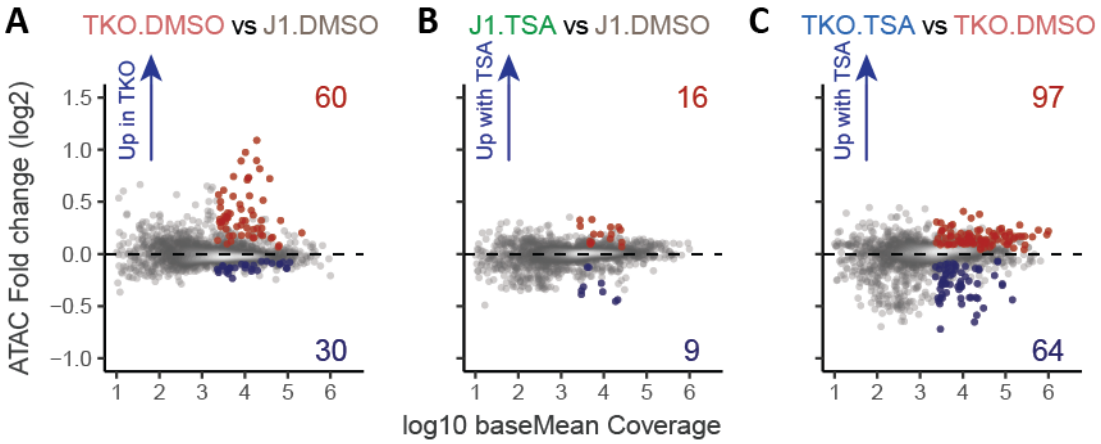


Figure 4.19: MA plots comparing for each type of repetitive element, the fold change in ATAC-seq signal across all genomic copies to the baseMean ATAC-seq signal. A. Differenial analysis was carried out to compare ATAC-seq signal in DNMT.TKO versus wild type mESCs (**A**), TSA versus DMSO treated wild type J1 mESCs (**B**) or TSA and DMSO treated DNMT.TKO mESCs (**C**). Repeat element types showing a significant change in ATAC-seq signal are represented by coloured dots. Differential analysis was performed using the R package DESeq2 (Love et al., 2014). Changes were deemed significant with an *FDR* < 0.05.

Repeat type	Repeat family	RNA Fold change (log2)			ATAC Fold change (log2)		
		TKO.DMSO	J1.TSA	TKO.TSA	TKO.DMSO	J1.TSA	TKO.TSA
MMERGLN-int	ERV1	2.47	1.92	5.59	0.25	0.11	0.48
LTR45C	ERV1	0.50	1.24	3.33	-0.10	-0.05	-0.05
MMERGLN_LTR	ERV1	1.32	0.83	3.26	0.26	0.11	0.46
RLTR41C	ERV1	-0.27	2.52	3.14	-0.08	-0.22	-0.30
RLTR5_Mm	ERV1	0.19	1.42	3.08	0.09	0.14	0.22
MER89	ERV1	0.10	2.24	2.94	-0.04	0.00	-0.03
MERV1_I-int	ERV1	0.48	1.47	2.90	-0.09	0.08	0.06
RLTR6C_Mm	ERV1	0.97	1.71	2.77	0.32	0.17	0.46
RLTR48C	ERV1	0.12	1.05	2.63	0.07	0.00	0.07
IAPA_MM-int	ERVK	2.03	1.00	5.21	0.65	0.16	1.08
IAPEy-int	ERVK	2.68	1.27	5.13	0.24	0.20	0.37
IAPLTR1_Mm	ERVK	1.76	0.76	4.60	1.09	0.01	1.24
IAPEz-int	ERVK	1.78	0.70	4.54	0.20	0.08	0.48
SRV_MM-int	ERVK	-0.08	2.76	4.17	-0.03	0.02	0.01
ERVB2_1-I_MM-int	ERVK	-0.21	3.12	3.81	-0.03	0.10	0.09
ERVB7_3-LTR_MM	ERVK	0.34	1.29	3.67	-0.18	0.03	-0.42
RLTR9B	ERVK	-0.22	2.65	3.24	0.09	-0.09	-0.06
IAPEY_LTR	ERVK	0.94	1.02	3.03	0.73	-0.04	0.58
RLTR31B_Mm	ERVK	0.04	1.22	2.77	0.03	-0.03	-0.02
MMERVK10C-int	ERVK	0.57	0.99	2.77	0.15	0.07	0.37
RLTR19A	ERVK	1.40	0.80	2.65	0.08	0.10	0.25
RLTR12F	ERVK	0.07	1.57	2.59	-0.26	0.08	0.09
RMER16B2	ERVK	-0.03	1.92	2.50	0.30	-0.07	0.37
IAP1-MM_LTR	ERVK	0.27	1.74	2.46	0.39	0.08	0.49
RMER16A2	ERVK	-0.09	1.09	2.45	0.16	-0.06	0.11
MMERVK10D3_I-int	ERVK	0.85	1.38	2.34	0.13	0.09	0.29
MERVL-int	ERVL	0.30	2.83	4.05	0.02	0.11	0.26
LTR86A1	ERVL	-0.10	1.92	3.68	0.04	-0.07	-0.01
LTR86A2	ERVL	2.06	1.06	3.25	-0.03	-0.01	0.10
MT2_Mm	ERVL	0.16	1.83	2.95	0.07	0.23	0.45

Table 4.8: Changes in RNA-seq and ATAC-seq coverage averaged over all copies of 30 LTR retrotransposon types. The elements included are the top 30 significantly transcriptionally activated transposable element types in TSA treated DNMT.TKO cells compared to DMSO-treated wild-type cells. ATAC-seq changes highlighted in red were found to be statistically significant using the R package DESeq2 (*FDR* < 0.05, Love et al., 2014).

We have previously seen that a significant fraction of reads within the ATAC-seq libraries were redistributed to repetitive elements in TSA- versus DMSO- treated DNMT.TKO cells (Figure 4.12). Pursuing this observation, we found that the number of reads mapping to LINE elements increased significantly upon TSA treatment in DNMT.TKO cells (p = 0.0067, one-sided Student's t-test) and to a greater extent than any other repeat element category (Figure

4.20). Differential analysis of reads mapping to different L1 element types revealed that 27 types of L1 elements significantly gained accessibility in TSA- compared to DMSO-treated DNMT.TKO cells (Appendix - Table S3). None of these showed significant changes in accessibility in TSA- compared to DMSO-treated wild types cells or in DNMT.TKO cells compared to wild type. As LINE1 elements represent ~20% of the genome (Consortium, 2002), an increase in accessibility at a majority of these sites could in part be responsible for the reduced signal we detect at conventional ATAC peaks in the TKO.TSA samples. However elevated accessibility at LINE elements following TSA treatment should be validated experimentally with orthogonal approaches.

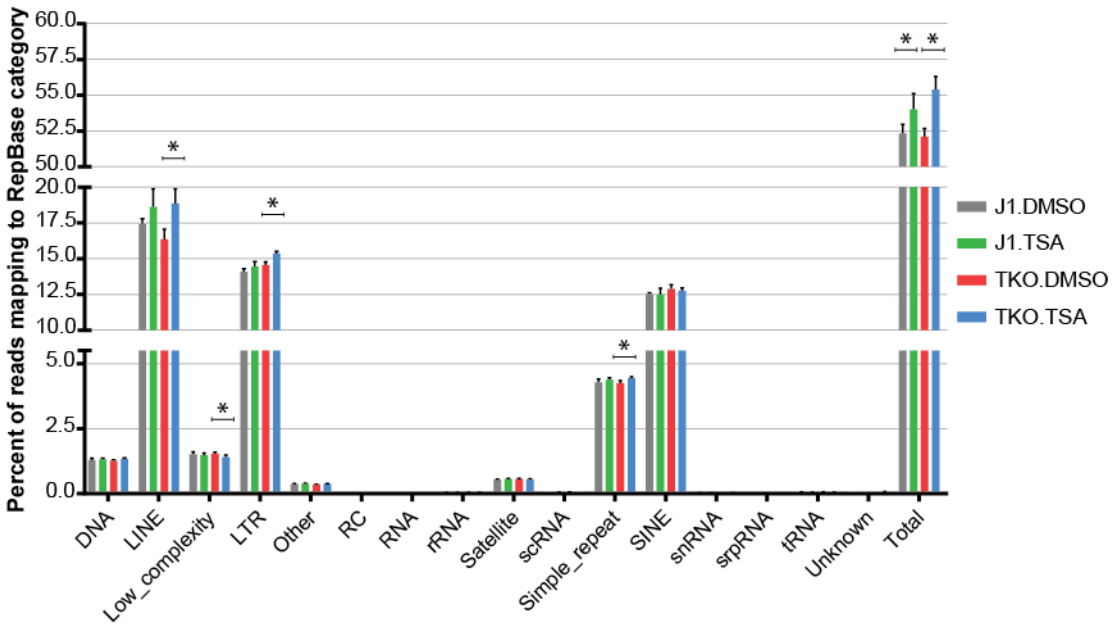


Figure 4.20: Distribution of ATAC-seq reads across different classes of repetitive elements. The number of reads that overlap each category was plotted as a percentage of the total library size. Mean values from four biological replicates are plotted; error bars indicate standard deviation. Two-tailed Student t-tests were used to determine significant differences between TSA and DMSO treated samples: * p-value < 0.05 ; ** p-value < 0.01 ; *** p-value < 0.001 ; non-significant differences are not indicated.

4.7 Genome-wide mapping of histone acetylation in wild type and DNMT.TKO mES cells in the presence TSA or DMSO.

Whilst histone acetylation increased globally in mESCs in response to TSA as monitored by immunoblotting (see Chapter 3 - Figure 3.6) it remained unclear whether acetylation levels were increasing equally throughout the genome or were restricted to certain genomic regions. Previously both HATs and HDACs were found to be associated with regions of elevated Pol2 binding and transcriptional activity (Wang et al., 2009). In E14 mouse ES cells, valproic acid caused an increase in H3K9 acetylation at regions that previously had intermediate levels of histone acetylation such as DHSs and active enhancers (Hezroni et al., 2011). On the other hand, a previous effort to map H3K9/14 acetylation dynamics after TSA treatment in primary human vascular endothelial cells showed that highly acetylated regions and many gene promoters lost acetylation upon HDAC inhibition (Rafehi et al., 2014). To aid our understanding of the gene expression changes and chromatin accessibility changes we observed in our system we decided to map histone acetylation levels genome-wide in WT and DNMT.TKO cells with or without TSA treatment.

We performed genome-wide mapping of acetylated histone 3 (H3ac) using native ChIP-seq. To allow for absolute quantitation of the ChIP signal, we used a reference exogenous genome that was “spiked-in” on a per-cell basis (referred to as a calibrated ChIP-seq) (Bonhoure et al., 2014; Orlando et al., 2014; Egan et al., 2016). To be able to compare between samples in a traditional ChIP-seq experiment the total read count is used for normalization (reads per million sequenced, RPM). This permits the detection of qualitative but not quantitative differences in the distribution of a histone modification or DNA binding protein. Particularly, if the levels of a histone modification increase (or decrease) globally after perturbation of a chromatin-modifying enzyme, the signal-to-noise ratio will not necessarily vary and no difference will be perceived using the RPM methodology (Figure 4.21A). However, with a fixed amount of exogenous genome added to each sample, signal can be normalized to this

external reference allowing for accurate quantification of the differences between each sample (Figure 4.21B).

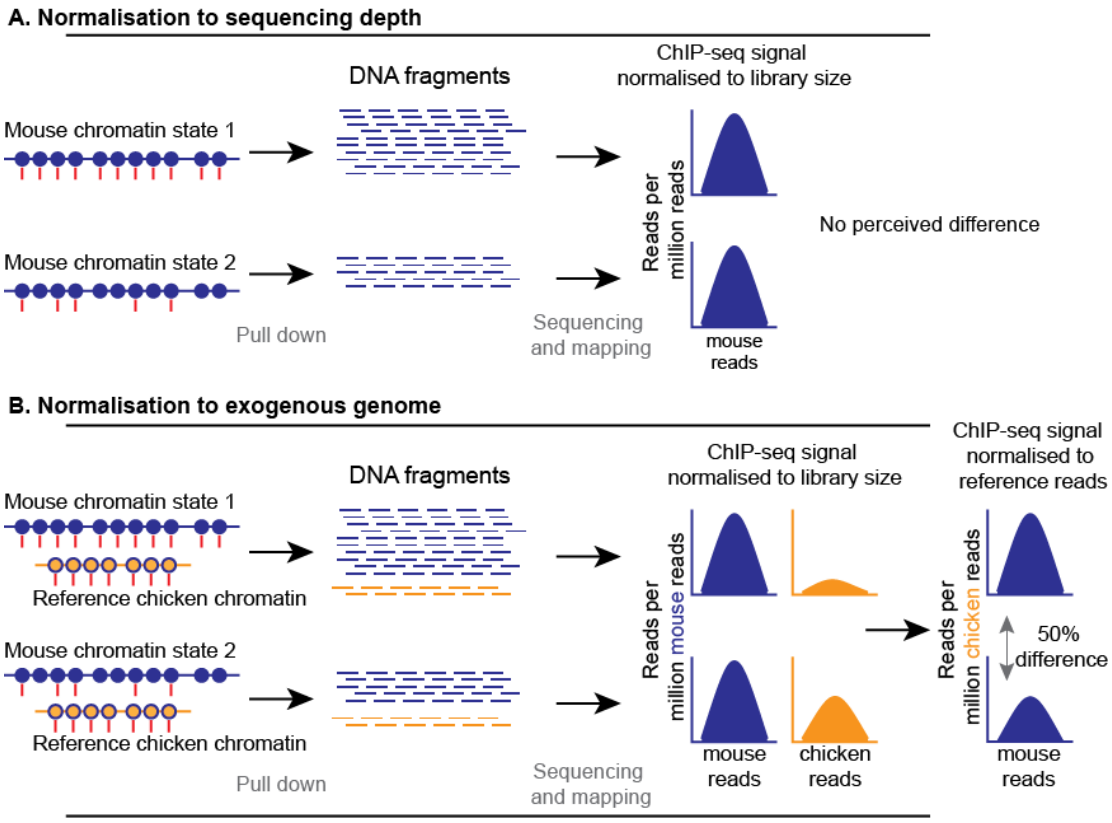


Figure 4.21: Normalisation and interpretation of ChIP-seq data A. Schematic representation of a typical ChIP-seq experiment aiming to interrogate the genomic distribution and abundance of a histone modification in two theoretical experimental conditions. In the first condition (state 1), nucleosomes (blue circles) are fully modified (red bars) while in the second (state 2), only half the nucleosomes are modified. After immunoprecipitation, sequencing and mapping, ChIP-seq signal at a particular genomic locus is measured as the read coverage normalised to the total number of reads in the library. This normalisation obscures the difference between the two chromatin states. **B.** Schematic representation of a ChIP-seq experiment carried out on the same two samples, but where a fixed amount of exogenous chicken chromatin is added to each sample. After immunoprecipitation, sequencing and mapping, an increase in reads mapping to the chicken genome as a percentage of the total read count is seen in the second sample. ChIP-seq signal at a particular mouse genomic locus can be measured as the read coverage normalised to the total number of sequencing reads that map to the chicken genome. This normalisation reveals a 50% difference in ChIP-seq signal between the two chromatin states. This figure was adapted from Orlando *et al.,* 2014.

For our experiment, we mixed a fixed number of chicken fibroblast DF1 cells with DNMT.TKO or WT cells, treated with TSA or DMSO-vehicle before proceeding with the ChIP protocol (Figure 4.22). Preliminary analyses using publicly available sequencing data of the chicken genome (European Nucleotide Archive accession SRR3407542) suggested that only 0.54% of 75bp paired-end chicken reads mapped to the mouse mm9 genome using bowtie2 (Langmead and Salzberg, 2012) with default settings (alignment rate to the chicken genome was 93.17%). Conversely, 0.65% of mouse reads from an input ChIP-seq sample mapped to the galGal5 genome indicating that the mouse and chicken genomes are distinct enough to allow for the correct assignment of sequencing reads to the correct genome. As histone proteins and modifications are conserved in vertebrates, suggesting that H3ac ChIP will be equally successful in mouse and chicken cells, these results assured us that DF1 chicken fibroblast cells were suitable for use as exogenous spike-in in our calibrated ChIP-seq experiments. Adding the exogenous cells at the beginning of the experiment has the benefit of normalizing for variations throughout the experiment for example during chromatin digestion or immunoprecipitation. Two biological ChIP replicas were sequenced for this experiment. Total read counts, alignment metrics to the mm10 or galGal4 genomes and ratio of galGal4:mouse reads for all samples are presented in Table 4.9.

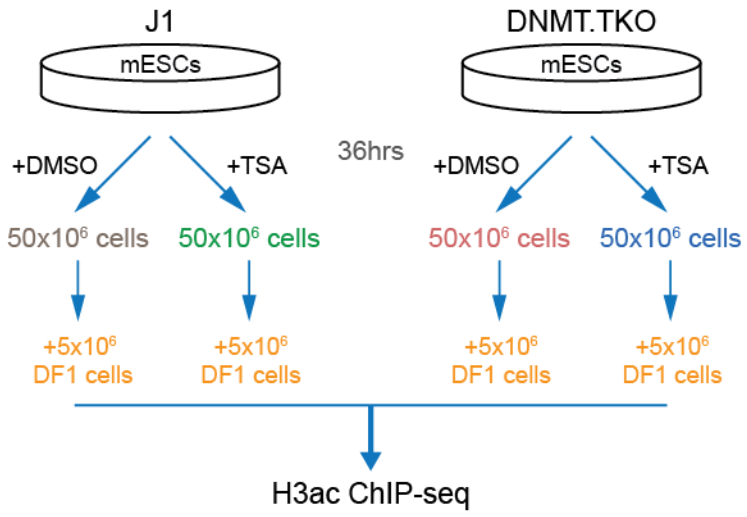


Figure 4.22: Schematic of the experimental approach used to investigate the effects of DNA methylation loss and HDAC inactivation on histone H3 acetylation.

Sample	Total PE reads	Overall alignment for uniquely mapped reads (%)	Number of reads aligning uniquely to mm10	Number of reads aligning to uniquely to galGal4	Ratio of galGal4 to mm10 reads (%)	Downsampled number of reads aligning to galGal4 after normalisation to input galGal4:mm10 ratio	Ratio of galGal4 to mm10 reads after downsampling (%)
H3ac_J1.DMSO_REP1	44,044,587	84.75	67,270,526	3,375,531	5.02	2,973,838	4.42
H3ac_J1.TSA_REP1	48,568,812	84.75	77,009,096	846,951	1.10	846,951	1.10
H3ac_TKO.DMSO_REP1	41,784,839	85.83	63,051,980	3,514,149	5.57	3,172,401	5.03
H3ac_TKO.TSA_REP1	38,683,109	84.81	61,773,782	586,373	0.95	517,826	0.84
Input_J1.DMSO_REP1	43,447,413	76.56	62,300,900	2,152,686	3.46	1,896,428	3.04
Input_J1.TSA_REP1	42,468,072	76.97	61,573,557	1,874,011	3.04	1,874,011	3.04
Input_TKO.DMSO_REP1	39,423,719	75.68	56,354,973	1,900,678	3.37	1,715,569	3.04
Input_TKO.TSA_REP1	47,340,659	76.56	66,970,404	2,309,001	3.45	2,038,560	3.04
H3ac_J1.DMSO_REP2	44,837,586	86.02	67,563,485	3,826,776	5.66	3,826,776	5.66
H3ac_J1.TSA_REP2	38,280,334	84.59	60,709,915	627,616	1.03	617,157	1.02
H3ac_TKO.DMSO_REP2	41,003,758	86.61	61,737,747	4,177,216	6.77	3,919,965	6.35
H3ac_TKO.TSA_REP2	38,057,918	85.02	61,666,541	491,099	0.80	396,959	0.64
Input_J1.DMSO_REP2	41,775,357	76.69	60,361,695	1,804,565	2.99	1,804,565	2.99
Input_J1.TSA_REP2	43,129,608	76.08	61,555,568	1,872,201	3.04	1,840,740	2.99
Input_TKO.DMSO_REP2	38,550,388	75.39	54,606,462	1,739,366	3.19	1,632,684	2.99
Input_TKO.TSA_REP2	43,513,376	76.24	61,341,340	2,268,169	3.70	1,834,467	2.99

Table 4.9: Total read counts, alignment metrics to the mm10 or galGal4 genomes and ratio of galGal4:mouse reads for all H3ac calibrated ChIP-seq and Input samples.

As was expected based on prior knowledge of acetylated histones distribution in the genome (Wang et al., 2008), acetylated histones were enriched at gene transcription start sites and regions of accessible chromatin (Figure 4.23). Using an external reference turned out to be key in interpreting the data from this experiment. As shown in Figure 4.24 and exemplified for a 200kb region (Figure 4.25), TSA caused a global increase in H3ac signal that would have gone mostly undetected without calibration. In fact, when normalizing to the library size, TSA appears to cause a reduction in H3ac signal at peaks (Figures 4.24A-B and 4.25).

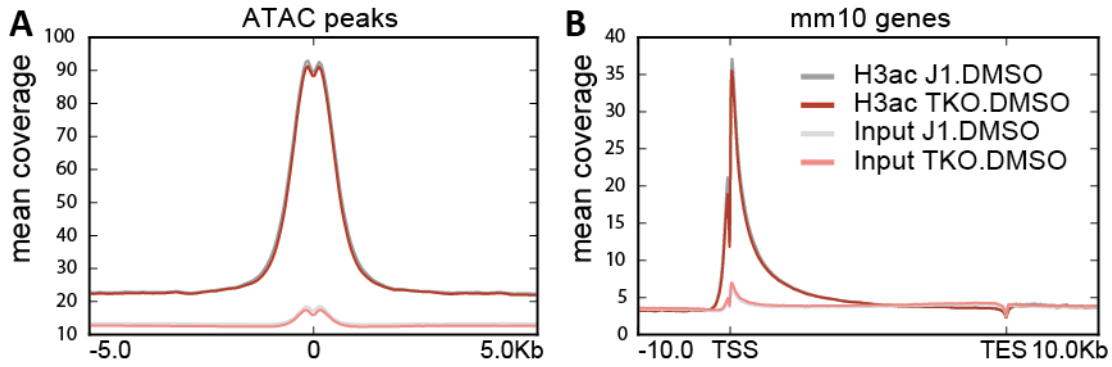


Figure 4.23: H3ac enrichment relative to genomic features. A. Metaplot showing the average H3ac and Input ChIP-seq signal in the 10kb regions surrounding the centre of ATAC-seq peaks (n = 59838, 50bp windows) for DMSO-treated wild type (red) or DNMT.TKO (grey) cells. Alignments from replicate samples were merged and coverage normalised to sequencing depth. **B.** Metaplot showing the average H3ac and Input ChIP-seq signal obtained over gene bodies and the 10kb upstream and downstream for DMSO-treated wild type (red) or DNMT.TKO cells (grey). All 15222 annotated genes were scaled to the same size. Alignments from replicate samples were merged and coverage normalised to sequencing depth.

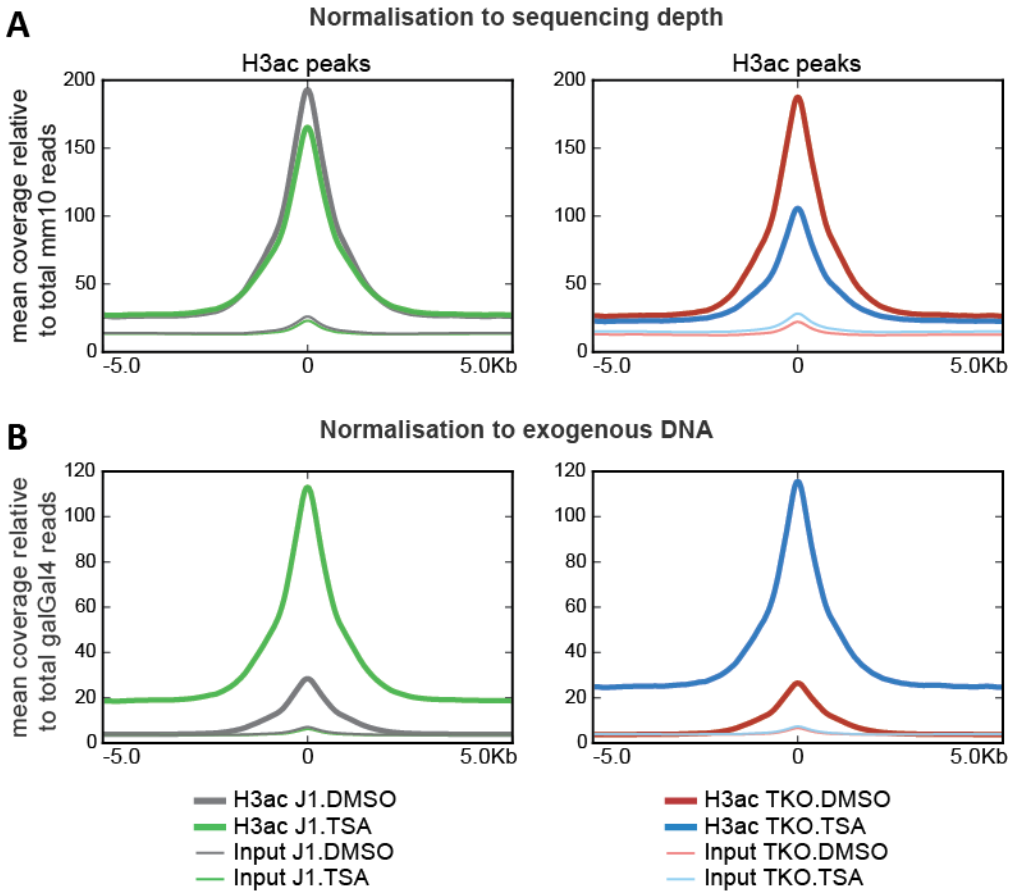


Figure 4.24: Metaplots showing the average H3ac or input ChIP-seq signal measured in the 10kb regions surrounding the centre of H3ac ChIP-seq peaks (n = 28608) using different normalisation approaches. Left, data from wild type cells treated with DMSO or TSA. Right, data from DNMT.TKO cells treated with DMSO or TSA. **A.** Alignments from replicate samples were merged and coverage normalised relative to the total number of reads mapping to the mm10 genome. **B.** Alignments from replicate samples were merged and coverage normalised relative to the number of reads in each sample that mapped to the galGal4 genome.

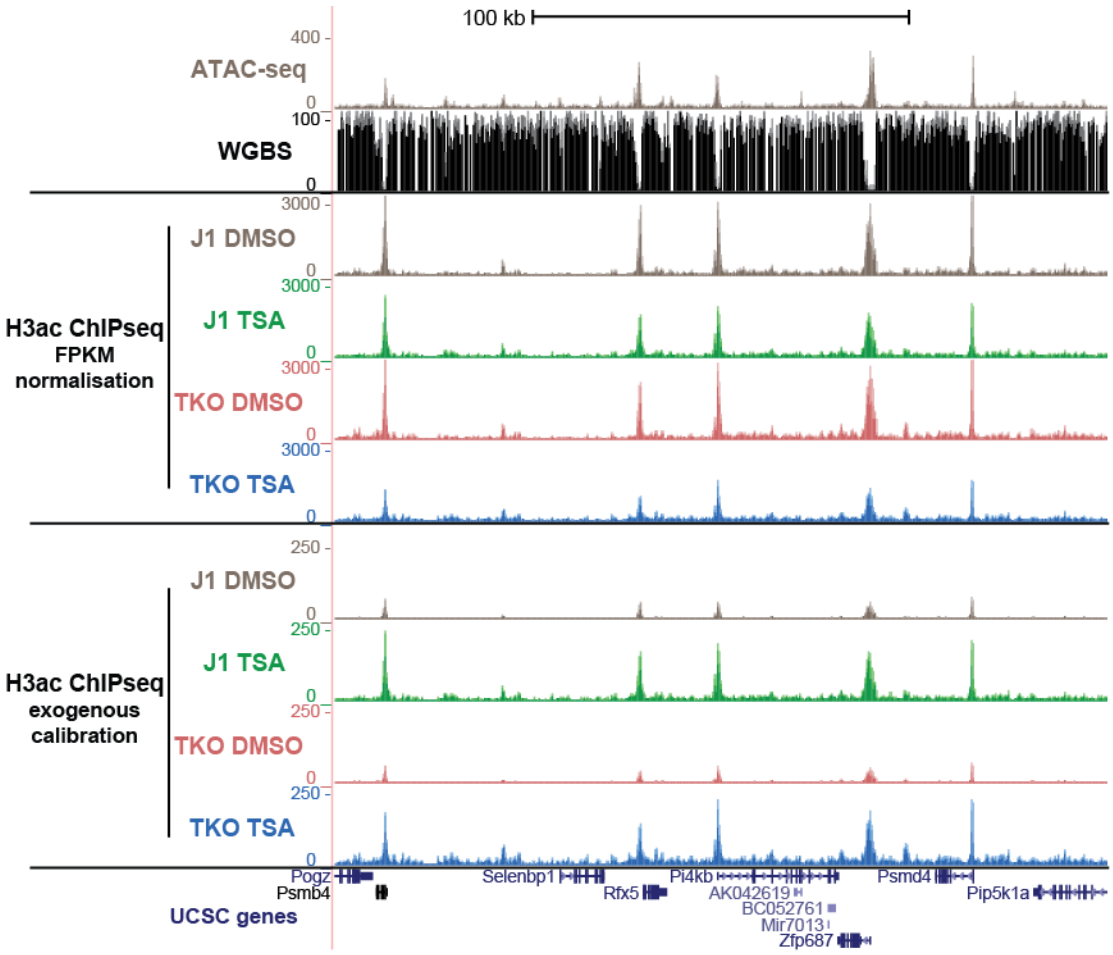


Figure 4.25: UCSC genome browser snapshots illustrating how the use of different normalisation methods changes our visualisation of H3ac ChIP-seq signal. H3ac ChIP-seq signal is shown for DMSO or TSA-treated wild-type J1 and DNMT.TKO cells. Coverage graphs were generated after merging alignments from replicate samples. In the top four H3ac ChIP-seq tracks, read coverage at each base pair was normalised to the total number of mm10 reads in the library (RPKM normalisation). In the lower tracks, the number of reads originating from the exogenous galGal4 genome in each library was used for normalisation. UCSC gene annotations are shown along with WGBS data from E14 mESCs (Habibi et al., 2013) and ATAC-seq data from wild type J1 cells. Mm10 coordinates: chr3:94,876,915-95,075,953 .

Peak calling using non-normalised data consistently led to the identification of fewer peaks in the TSA treated samples compared to the control-treated samples (Figures 4.26A-C). Together with the observation that TSA-treated cells, particularly DNMT.TKO, have a decreased signal-to-noise ratio (Figure 4.27A) these results suggests that the increase in acetylation is unequal between previously hyper-acetylated regions and the other genomic regions.

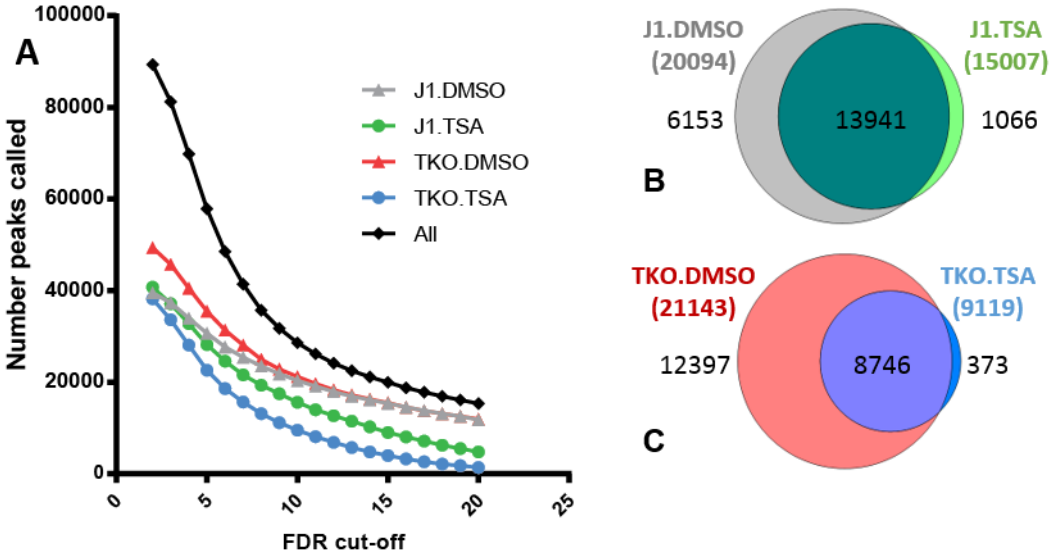


Figure 4.26: H3ac ChIP-seq peak calling. A. The number of H3ac enriched regions (peaks) identified in DMSO or TSA-treated wild-type J1 and DNMT.TKO cells using the MACS2 callpeak algorithm (Zhang et al., 2008) for a range of FDR thresholds for significance. Only peaks identified in both biological replicates were included **B-C.** Venn diagram representing the overlap between the H3ac enriched regions identified in TSA and DMSO-treated wild type J1 (**B**) or DNMT.TKO cells (**C**) using MACS2 with an FDR threshold of 10^{-10}.

Differential analysis was performed using data scaled to the external reference at H3ac enriched regions (H3ac peaks) as well as randomly selected regions outside of peaks (shuffled gene body intervals). In J1 cells, the signal was on average 4.1x and 5.1x higher after TSA treatment at peaks and shuffled regions respectively. In TKO.DNMT cells, the signal at peaks was on average 4.9x increased compared to 9x increased at shuffled regions (Figure 4.27B). This analysis further suggests that TSA leads to a greater increase in acetylation outside of previously hyperacetylated regions. Moreover, this increase at regions with low starting levels

of acetylation is more robust in cells depleted of DNA methylation. Notably, TSA treatment led to the formation of very few novel hyperacetylated loci. (Figure 4.26B).

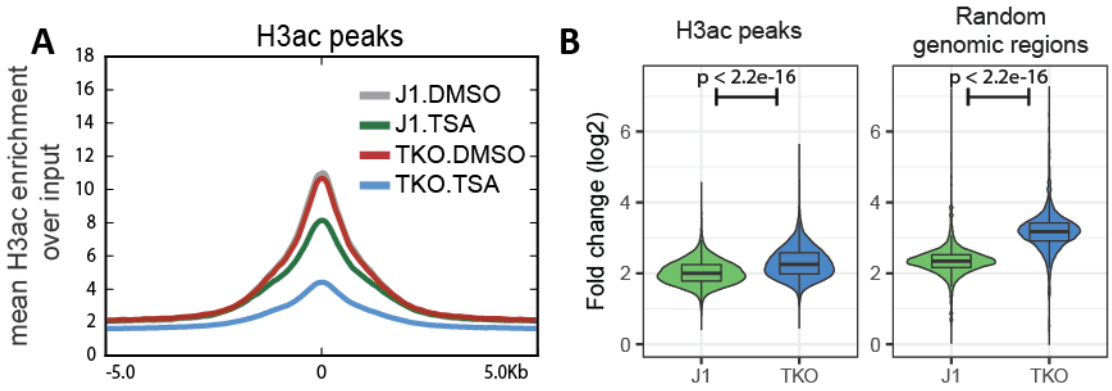


Figure 4.27: Changes in the profile of the H3ac landscape after TSA treatment of mESCs. A. Metaplot showing the average H3ac ChIP-seq signal from DMSO or TSA-treated wild-type J1 and DNMT.TKO cells in the 10kb regions surrounding the centre ChIP-seq peaks (n = 28608). For each 10kb region, RPKM-normalised H3ac ChIP-seq signal within 50bp windows was divided by the signal from the corresponding Input ChIP-seq samples. **B.** Violin plots summarise the fold change in H3ac signal after the TSA treatment of wild type J1 (green) or DNMT.TKO cells (blue) at ChIP-seq peaks (n = 28608) or random genomic regions (n = 22964). The genomic locations of these regions were obtained by random permutation of the genomic locations of mm10 genes. Shuffled features were not allowed to overlap with H3ac ChIP-seq peaks. Fold-change values were obtained through differential analysis using the R package DeSeq2 (Love et al., 2014) with the data scaled to the number of reads originating from the galGal4 genome (see Chapter 2 - Section 2.9.6.4). One-tailed Wilcoxon matched-pairs signed rank test were used to determine significant differences in the distributions of fold change values.

4.8 Copy number variation of chromosome 15 in DNMT.TKO versus wild type J1 cells.

In the process of analysing the ATAC-seq dataset we noticed that most of the chromatin accessibility peaks mapping to chromosome 15 had reduced signal in DNMT.TKO compared to WT cells. This trend was opposed to that seen for all other chromosomes (Figure 4.28A). Investigating this further, we realised that ~1.5x more reads mapped to chromosome 15 in wild-type cells compared to the DNMT.TKO mutants indicative of a copy number gain (Figure 4.28B). Based on our visualisation of the data and the position of peaks showing reduced signal along the entire chromosomal sequence we believe that the wild-type cells used in this experiment have an additional copy of chromosome 15. This was discovered at a late stage of the analysis after having used the same J1 cell line batches for our transcription factor and histone ChIP-seq experiments (Chapters 4 and 5). Sequencing data from these experiments also show increased coverage of chromosome 15.

This was only noticed when analysing the ATAC-seq data because the tagmented genomic DNA used as an input was produced from a different batch of cells that appear to have a normal karyotype (Figure 4.28B). The peak calling and differential analysis steps, which rely on normalisation to this “background” data, therefore miscalculated the signal across chromosome 15. Where appropriate all peaks mapping to Chr15 were removed from our analyses, this is indicated in the figure legends. Some implications of this finding will be discussed in the next chapter.

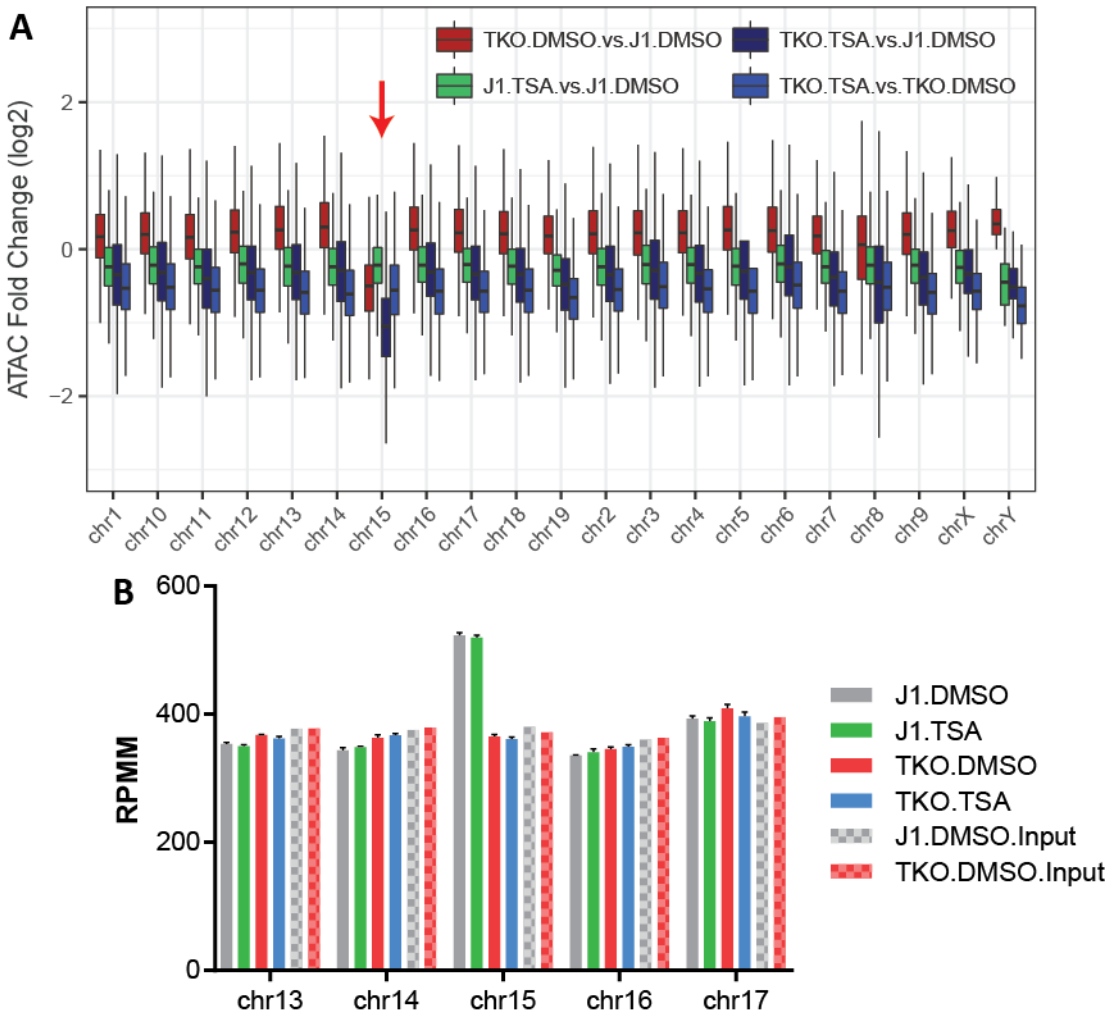


Figure 4.28: Chromosome 15 copy number variation. A. Boxplots summarise the fold change in ATAC-seq signal obtained at all peaks that co-localise to the same chromosome. Peaks that locate to chromosome 15 appear to have an overall reduction in accessibility when comparing data from the TKO.DMSO (or TKO.TSA) versus J1.DMSO samples. This was strikingly different to the trend seen at peaks localising to the remaining chromosomes. **B.** For each sample, the number of reads mapping to five different chromosomes was counted. Plotted are the read counts normalised to the library size and the chromosome length (RPMM = Reads Per Megabase per Million mapped reads). For non-input samples, mean RPMM from four biological replicates is plotted; error bars indicate standard deviation.

4.9 Discussion

4.9.1 Chromatin accessibility landscape in DNA methylation deficient mouse ESCs

In an attempt to identify DNA methylation-sensitive regulatory elements and transcription factor binding sites, we profiled chromatin accessibility in wild-type and DNMT deficient mES cells using ATAC-seq. Considering that over 80% of CpGs in the genome of mESCs are methylated (Stadler et al., 2011; Habibi et al., 2013) and potential transcription binding sites are highly prevalent, relatively few increases in accessibility were detected upon DNA methylation loss (6% of all sites that were detected to be accessible in J1 or DNMT.TKO mES cells) suggesting that DNA methylation does not have an instrumental role in modulating chromatin accessibility throughout the genome.

Nevertheless we identified 3608 genomic sites that displayed significant gains in chromatin accessibility in DNMT.TKO compared to wild-type cells with those showing the greatest fold change being at sites that are methylated in wild type cells (Figures 4.7A-B). These represent loci at which DNA methylation plays a role in restraining the formation of a Tn5 accessible state. Mostly found outside of genic promoters (Figure 4.6), these regions have the potential to be distal regulatory elements such as enhancers and insulators or promoters of LTR transposable elements.

A potential mechanism by which DNA methylation is maintaining these elements in an inaccessible state is through promoting the occupancy of MBD-domain proteins and activity of their associated HDAC enzymes (Nan et al., 1998; Kokura et al., 2001; Lyst et al., 2013). However this is unlikely to be the case at DNMT.TKO specific ATAC-seq peaks, as these sites have a lower CpG content than accessible loci common to both cell lines (Figure 4.7C) and few overlap CpG islands whereas MBD-domain proteins preferentially localise to CpG-rich regions (Baubec et al., 2013; Hainer et al., 2016). Moreover, DNMT.TKO-specific ATAC-seq peaks are

not affected by HDAC inhibition (Figure 4.14) further arguing against a mechanism by which DNA methylation's role relies on the recruitment of HDACs by MBD proteins.

A more likely alternative is that DNMT.TKO specific ATAC-seq peaks represent novel binding of methylation-sensitive transcription factors. Increased occupancy of other chromatin modifying complexes containing CXXC domains may also be involved. That the absence of DNA methylation is associated with the differential occupancy of transcription factors is supported by the work by Domcke *et al.* along with results from experiments that will be presented in the following chapter. In the study published by Domcke *et al.*, methylation sensitive DHS sites were identified in an independently derived DNMT.TKO cell line by using DNase-seq. Similar to our observations, DNMT.TKO-specific DHS were mostly promoter distal and in CpG poor regions arguing against a role of MBD proteins (Domcke et al., 2015). Over 40% of these sites could be attributed to novel binding of the NRF1 protein to its binding motif in the absence of CpG methylation. This was validated by ChIP-seq experiments performed in cells with varying levels of DNA methylation. ETS (5'-CGGAA-3') and E-box (5'-CACGTG-3') motifs were also enriched at DNMT.TKO-specific compared to wild-type DHS (present at 20-30% of novel DHS) in their study indicating that the genomic localisation of TFs such as GABPA and c-Myc, MAX or MYCN that bind these sequences could also be influenced by DNA methylation. We did not find any highly significant enrichment for ETS and E-box motifs in the ATAC-peaks specific to our DNMT.TKO cell line however we did identify motifs recognised by the NFY complex and the YY1 protein (Tables 4.2 and 4.3).

Although a CpG dinucleotide is not included within the core NFY binding motif the presence of a CpG directly adjacent to the central CCAAT sequence may still influence the DNA-protein interaction. Indeed, the CCAATCpG sequence is known to occur in ERV-9 LTR retrotransposons which display methylation-sensitive transcriptional activity in human cells. Furthermore, EMSA experiments indicate that NFY binding to this sequence is reduced when the motif-

adjacent CpG is methylated (Hu et al., 2017). YY1 is also known to occupy ERV elements, notably IAP elements in mouse embryonic stem cells (Guallar et al., 2012). *In vitro*, the interaction between YY1 and sequences within the IAP LTRs has been shown to be reduced by DNA methylation (Satyamoorthy et al., 1993). As some IAP LTRs show gains in ATAC-seq signal in DNMT.TKO cells (Table 4.8) it seems plausible that in wild-type cells, DNA methylation is restricting the access of the YY1 and NFY transcription factors to binding sites within these elements. Further work is necessary to support this idea, beginning with providing evidence for the DNA methylation sensitive binding of these transcription factors *in vivo*. In the following chapter the differential genomic occupancy of several transcription factors, including YY1, upon loss of DNA methylation will be described.

4.9.2 The relationship between changes in chromatin accessibility and changes in gene expression

Most of the changes in accessibility that we detected were not associated with changes in mRNA expression: only a minority of DMNT.TKO site-specific gains in accessibility were linked with increased expression from neighbouring genes (Figure 4.15A). Additionally, LTR transposable elements types that overall showed significant increase in ATAC-seq signal did not display significant gains in mRNA abundance (Figure 4.18A). On average, the DNMT.TKO specific ATAC-seq peaks are smaller than those present in both cell lines suggesting that these are the result of relatively weak protein-DNA interactions that do not contribute to the formation of stable complexes capable of enhancing transcription. Moreover, the loci that lacked an association between accessibility and transcription may represent sites with no regulatory function or having a regulatory role on transcriptional elements other than the closest gene. The analysis shown in Figure 4.15A does not take into account that some of these novel ATAC-seq sites overlap transposable elements and may regulate their transcriptional activity instead of a neighbouring gene. At other promoter distal elements, RNA-seq may not be sensitive enough to capture the production of unstable transcripts. For

example, this may be the case for non-coding enhancer RNAs that are known to be unstable and less abundant than mRNAs (Kim et al., 2010; Pefanis et al., 2015). In a study by Neri *et al.*, methods to enrich for capped RNA were necessary to detect non-canonical transcription initiation sites that appeared in gene bodies upon *Dnmt3b* knockout (Neri et al., 2017).

Finally, alternative mechanisms may be present to limit the transcriptional activity of unfavourable elements despite the increase in accessibility in DNA methylation deficient cells. This is particularly the case for ERVs, for which repression in hypomethylated cells is achieved through pathways that depend on histone modifications, particularly H3K9me3 and H3K27me3 (Karimi et al., 2011; Walter et al., 2016). Additionally, YY1 is known to act as both an activator and repressor, with its repressive function shown to be dependent on its interaction with an HDAC (Yang et al., 1996; Weill et al., 2003). YY1 occupancy at TEs in the absence of DNA methylation could contribute in maintaining a silenced transcriptional state.

Our RNA-seq results (Chapter 3) together with data published by others (Cameron et al., 1999; Coffee et al., 1999; Brocks et al., 2017) show that combining DNA demethylation and HDAC inhibition leads to the synergistic activation of specific genes (eg. *Nefm* in Figure 4.16, Figure 3.14) as well as ERVK type transposable elements (Figure 4.18B). As discussed in Chapter 3, we believe this reflects a role for HDACs and DNA methylation in two independent repressive pathways whose activity converge to silence these elements. Our ATAC-seq data reveals that removal of the DNA methylation but not TSA treatment can facilitate the formation of chromatin accessible regions, possibly driven by factors whose affinity is reduced in wild-type cells by methylation of CpG residues. While this is not sufficient for transcriptional activity, subsequent HDAC inhibition appears to be sufficient to disrupt the remaining repressive mechanisms (eg. *Nefm*, Figure 4.18B and Table 4.8) be they dependent on histone modifications or the binding of repressor proteins.

4.9.3 Chromatin accessibility landscape in mESCs following HDAC inhibition

In addition to mapping changes in chromatin accessibility resulting from the loss of DNA methylation, another aim of this experiment was to assess the role of HDACs in shaping the chromatin accessible landscape in mouse embryonic stem cells. In K562 cells, HDAC inhibition using SAHA or sodium butyrate was previously reported to cause localised changes in accessibility. Particularly noticeable were gains in DNaseI sensitivity at sites containing PU.I binding motifs (Frank et al., 2016). Contrastingly, in our experiment, HDAC inhibition had a global effect on the chromatin landscape with few local gains in accessibility. Indeed, TSA treated cells appeared to have decreased accessibility at previously Tn5 hypersensitive regions but increased accessibility throughout the genome particularly at LINE and LTR transposable elements (Figures 4.12 and 4.20). In the absence of DNA methylation, these differences were strikingly more pronounced (Figure 4.8) suggesting that HDACs have a greater responsibility in limiting chromatin mobility across the genome of DNMT.TKO cells. Interestingly, in K562 cells, the presence of H3K27me3 and enrichment for CBX2/CBX8 (components of polycomb group complexes) were identified as being the most informative factors to predict PU.I sites that gain accessibility upon HDAC inhibition (Frank et al., 2016). H3K27me3 is dramatically redistributed in hypomethylated cells: instead of being constrained to CpG islands, the modification is deposited throughout the genome (Brinkman et al., 2012; Reddington et al., 2013; King et al., 2016). If regions with enriched H3K27me3 are more likely to increase in accessibility in response to HDAC inhibition, the widespread distribution of this histone mark in DNMT.TKO may explain their stronger response to TSA compared to wild type mESCs in terms of increased chromatin accessibility in the background genome.

The interpretation of our ATAC-seq analysis is confounded by our normalisation method that measures changes in read density relative to the library size. To what extent the differences we detect before and after TSA treatment reflect the changes undergone in cells is difficult to

assess, particularly in DNMT.TKO cells. While hypersensitive sites undoubtedly have reduced accessibility relative to the surrounding genome, is accessibility reduced in absolute terms?

Decreased expression of the core pluripotency transcription factor *Pou5f1* (*Oct4*) and enrichment for its cognate recognition motifs at sites of reduced accessibility (Tables 4.5, 4.6 and 4.7) provides a rational to explain why DHS sites may be lost. A list of accessible chromatin loci that are abolished upon OCT4 depletion in mESCs has been determined (King and Klose, 2017). Comparison between the regions that show significantly reduced accessibility upon HDAC inhibition or OCT4 knockout could help elucidate to what extent decreased OCT4 occupancy plays a role in the TSA mediated reduction in chromatin accessibility.

Our real-time PCR results support the observation that TSA induces a reduction in accessibility at ATAC-seq peaks relative to selected background loci (Figure 4.13). Nevertheless, the validity of the changes we observe between samples relies on the assumption that the regions we chose for normalisation do not have different sensitivities to Tn5 upon TSA treatment. Ultimately, alternative experimental approaches should be used to demonstrate the quantitative loss of accessibility at ATAC-seq peaks. For locus specific analyses, an alternative method we could use involves measuring the sensitivity of certain regions to increasing concentrations of DNaseI and comparing this between samples (McMorrow et al., 2000; Bottardi et al., 2003). The interpretation of the results from any experiment remains challenged by our inability to pick reference loci proven to be unaffected by TSA treatment. Spiking-in exogenous material appears to be the obvious solution to allow for calibrated quantification.

4.9.4 The use of calibrated approaches to monitor epigenomic changes upon TSA treatment

A calibrated ChIP-seq approach was used to monitor changes in H3 acetylation throughout the genome in response to TSA treatment. With ChIP-seq signal normalised to the sequencing depth, H3ac appeared to be reduced at previously hyperacetylated sites upon HDAC inhibition (Figures 4.24A-B). To some extent, this resembled observations made in independent studies in which the loci most highly enriched in acetylated histones in E14 mouse cells or vascular endothelial cells displayed decreased signal after HDAC inhibition (Hezroni et al., 2011; Rafehi et al., 2014). Incidentally, in mouse ES cells, valproic acid treatment led to the localised increase in H3K9ac at regions that previously had intermediate levels of histone acetylation. These overlapped with DHSs, active enhancers and were marked by increased occupancy of the histone acetyltransferase p300. The authors suggested that HDAC inhibition induced gains in histone acetylation were dependent on local enrichment of HAT proteins (Hezroni et al., 2011). However, in our system this does not seem to the case; using the exogenous calibration we could deduce that H3 acetylation was increased throughout the genome in response to TSA, and to a relatively lesser extent at previously hyperacetylated loci (Figures 4.24C-D and 4.27). This experiment highlights the extent to which the use of calibrated approaches can modify the interpretation of ChIP-seq type data in situations where changes are anticipated at more than a few genomic sites.

Parallels can be drawn between the TSA-induced changes in H3ac-seq and ATAC-seq signal when these are measured using methods that normalise to sequencing depth. For both accessibility and H3 acetylation, loss of coverage at previously enriched regions is associated with increased signal from the background genome. Incorporating a calibration strategy to the ATAC-seq experiment is likely to facilitate our interpretation of the data, as was the case for H3ac ChIP-seq.

Chapter 5:

The role of DNA methylation in genome-wide transcription factor occupancy in mouse embryonic stem cells

5.1 Introduction

In mammals, DNA methylation is widely regarded as playing a role in the binding of transcription factors to specific sequences in the genome. As introduced in Section 1.5.4, the antagonism between DNA methylation and the binding of certain transcription factors to their cognate sequences has been proposed as a mechanism by which DNA methylation could modulate transcription (Tate and Bird, 1993). Early reports demonstrated that methylation of CpG sites in purified DNA probes was sufficient to reduce the binding of certain transcription factors *in vitro* (Iguchi-Ariga and Schaffner, 1989; Prendergast and Ziff, 1991). This model was further supported by the negative association between transcription factor occupancy and increased levels of DNA methylation that has been observed in numerous cell types (Stadler et al., 2011; Thurman et al., 2012; Hon et al., 2013; Ziller et al., 2013). However, growing appreciation that transcription factors could promote DNA demethylation upon binding (Brandeis et al., 1994; Stadler et al., 2011; Feldmann et al., 2013) has put into question the causal relationship between DNA methylation and transcription factor binding. While there is evidence that binding of some transcription factors can contribute to defining the DNA methylation genomic landscape (Stadler et al., 2011; Thurman et al., 2012; Feldmann et al., 2013), the extent to which DNA methylation shapes transcription factor binding genome-wide has not fully been elucidated and is only beginning to be studied (Domcke et al., 2015; Maurano et al., 2015). At the time when our experiments were conceived and begun, genome-wide transcription factor binding data in the absence of DNA methylation was only available for CTCF (Stadler et al., 2011).

Increasingly comprehensive and quantitative *in vitro* experiments have revealed that DNA methylation influences the affinity of a majority of DNA-binding proteins to sequences containing CpG dinucleotides (Hu et al., 2013; Kribelbauer et al., 2017; Spruijt et al., 2013; Yin et al., 2017). In addition to reducing the affinity of certain transcription factors for their

cognate recognition sequences (as is the case for ETS or bHLH family proteins for example), DNA methylation was found to increase the affinity of other transcription factors to non-cognate binding sequences (eg. p53, KLF4 and POU domain-containing proteins that do not normally bind to CpG-containing sequences) (Hu et al., 2013; Kribelbauer et al., 2017; Yin et al., 2017). However, it remains unclear whether the influence of CpG methylation on DNA-protein interactions *in vitro* translates into an effect on target selection in a cell. Evidence that DNA methylation modulates the genome-wide occupancy of many of these transcription factors *in vivo* is limiting.

To investigate the contribution of DNA methylation in the selection of transcription factor binding sites *in vivo* we chose to compare the genomic occupancy of a panel of transcription factors in wild type versus DNMT.TKO mouse embryonic stem cells. Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) was used to profile transcription factor binding genome-wide. Integration with whole genome bisulfite sequencing data from wild-type mESCs allowed us to determine sites at which differential protein occupancy is associated with the absence of DNA methylation. This approach complements our ATAC-seq data (Chapter 4) and published DNase-seq experiments performed in independent DNMT.TKO versus wild-type mES cells (Domcke et al., 2015). While these were used to indirectly identify transcription factors whose genome-wide occupancy was affected by the loss of DNA methylation, the ChIP-seq method allowed us to directly assess the genomic localisation of candidate transcription factors.

In this chapter I will briefly explain the process that was adopted to select candidate transcription factors, following which I will describe and discuss ChIP-seq data that was obtained for seven transcription factors in wild type and DNMT.TKO mESCs.

5.2 Selection of candidate transcription factors and ChIP-seq optimisation

We began this study by defining a panel of prospective transcription factors for which to perform genome-wide occupancy profiling using ChIP-seq. We chose to focus on proteins that contain a CpG dinucleotide in their recognition motif. Candidate transcription factors were picked that had previously been shown to preferentially bind to unmethylated DNA *in vitro*, although some had been found to bind DNA irrespective of methylation status *in vitro* and/or *in vivo* (Table 5.1 and Appendix - Table S4). We also included factors for which methylation was reported to promote or be necessary for binding to specific sequences (KLF4, CEBPB) (Hu et al., 2013; Mann et al., 2013; Spruijt et al., 2013; Yin et al., 2017). Finally, we attempted to choose transcription factors from different families in order to appreciate whether various classes of DNA binding domains were affected differentially by DNA methylation. Having selected a large panel of 38 transcription factors (Appendix - Table S4), we subsequently narrowed down on factors for which ChIP-seq experiments had previously been performed presuming that we would increase our chances of obtaining interpretable data. We ultimately selected a set of ten transcription factors (CEBPB, CREB, GABPA, E2F1, KLF4, MAX, MYC (c-Myc), NRF1, SP1 and YY1) (Table 5.1) and proceeded to optimise the ChIP-seq protocol.

Transcription Factor	Family or DNA binding domain type	CpG present in consensus recognition motif?	Evidence for DNA methylation sensitive binding
MAX	bHLH	**YES** (Ebox element)	None
MYC (c-Myc)	bHLH	**YES** (Ebox element)	*In vitro* (Prendergast and Ziff, 1991)
CEBPB	Basic leucine zipper	**NO** (CCAAT element) **YES** (CRE element)	*In vivo* and *in vitro* (Mann et al., 2013)
CREB1	Basic leucine zipper	**YES** (CRE element)	*In vitro* (Spruijt et al., 2013)
E2F1	E2F winged helix	**YES** (5'-TTTCC**CG**C-3')	*In vitro* (Campanero et al., 2000)

NRF1	NRF1/Ewg family	**YES** (5'-G**CG**CATG**CG**C-3')	*In vitro* (Kumari and Usdin, 2001)
SP1	SP/KLF zinc finger protein	**YES** (5'-C**CG**CCC-3')	Mixed (Clark et al., 1997; Harrington et al., 1988; Guo et al., 2012)
YY1	Gli-Kruppel zinc finger	**YES** (5'-C**CG**CCATNTT-3')	*In vitro* (Satyamoorthy et al., 1993; Kim et al., 2003)
GABPA (NRF2)	ETS family	(5'-c**G**GAAT-3')	*In vitro* (Lucas et al., 2009)
KLF4	Kruppel-like zinc finger	**NO**	*In vivo* and *vitro* (Hu et al., 2013)

Table 5.1: Candidate transcription factors selected for chromatin immunoprecipitation experiments in this study. ChIP-seq data in wild-type and DNMT.TKO cells was obtained for the underlined proteins. bHLH: Basic Helix-loop-helix factor; CRE element: 5'-TGACGTCA-3'; Ebox element: 5'-CACGTG-3'.

For each transcription factor, qPCR was used to test for successful enrichment at known binding regions after chromatin immunoprecipitation. Known binding sites to use as positive control loci were picked based on publicly available ChIP-seq data within unmethylated regions while negative background regions were designed within gene deserts. After optimising the ChIP protocol, we obtained promising ChIP-qPCR results for CEBPB, GABPA, MAX, NRF1 and SP1 that showed reproducible enrichment at multiple known binding sites compared to background loci (data not shown). However, we failed to detect any enrichment at known binding sites for CREB1, c-Myc and E2F1 under a variety of conditions (data not shown). For two factors, YY1 and KLF4, higher levels of enrichment were achieved at known binding sites when an additional crosslinking step was carried out (Figures 5.1A-B). Specifically, prior to formaldehyde crosslinking, cells were first incubated in the presence of disuccinimidyl glutarate (DSG) or ethylene glycol bis(succinimidyl succinate) (EGS), homobifunctional crosslinkers with two reactive NHS-ester groups (Mattson et al., 1993). In light of our ChIP-qPCR results, DSG was used as a second crosslinker in subsequent YY1 or KLF4 ChIP-seq experiments.

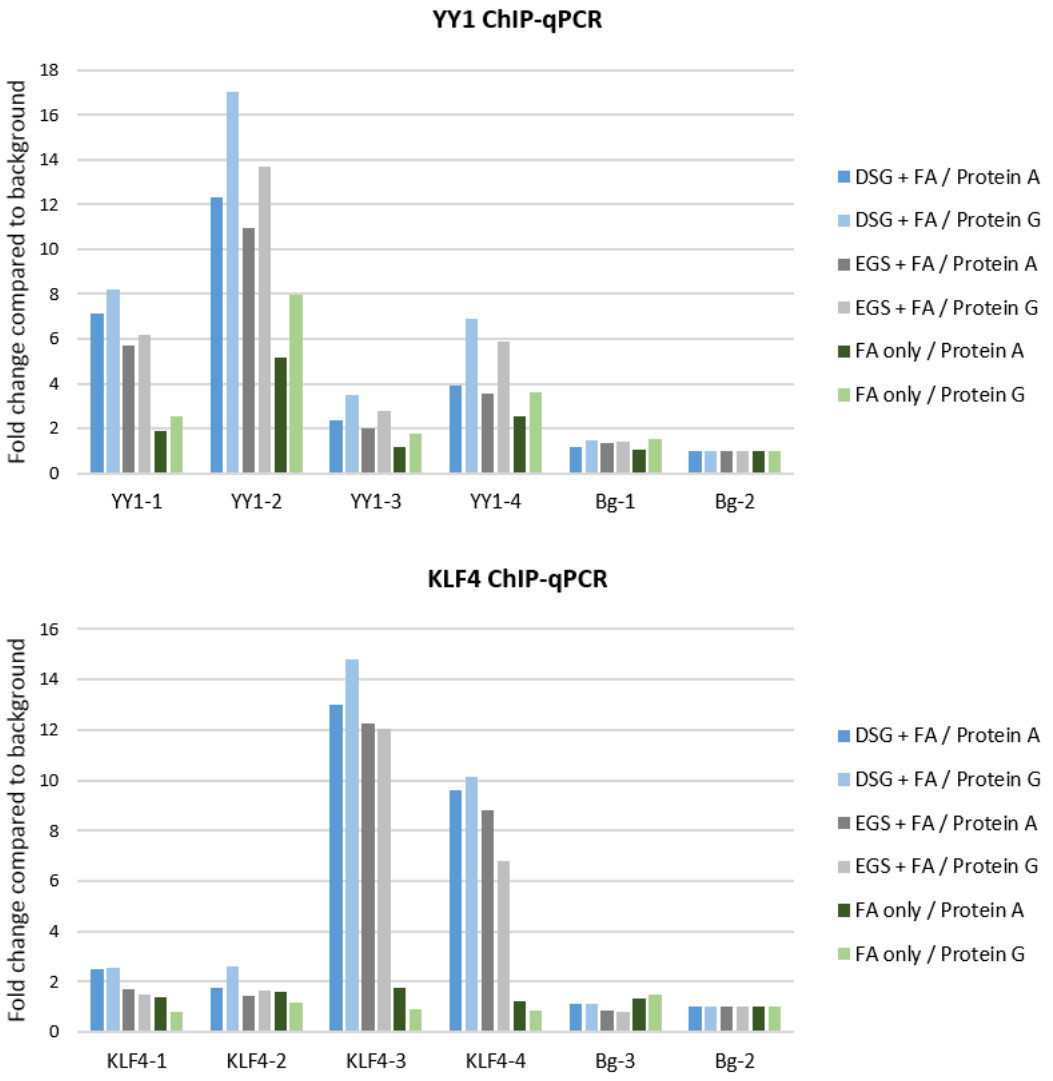


Figure 5.1: Real-time PCR analysis of KLF4 and YY1 ChIP DNA using different crosslinking and pull-down conditions. Wild type mESCs were crosslinked with EGS + formaldehyde, DGS + formaldehyde or formaldehyde alone. Crosslinked and solubilised chromatin was immunoprecipated with an antibody against YY1 (**A**) or KLF4 (**B**) and immune complexes captured with Protein A or Protein G agarose beads. After, purification, ChIP DNA was analysed by performing real-time PCR using primer pairs designed to amplify genomic DNA underlying four putative binding regions (YY1-1:4 and KLF4-1:4) and two background regions (Bg-1:3). For each sample, relative DNA amounts were calculated from Ct values, normalised first to input and then to the Bg-2 background locus. Only one biological replicate was performed. DSG: disuccinimidyl glutarate; EGS: ethylene glycol bis(subbinimidyl succinate); FA: formaldehyde.

We decided to proceed with the high-throughput sequencing of ChIP material that we obtained for seven transcription factors (CEBPB, GABPA, KLF4, MAX, NRF1, SP1 and YY1) from both wild type J1 and DNMT.TKO cells in biological triplicate. DNA fragment size distribution and enrichment at known regions were verified before carrying out library preparation and the libraries were sequenced on an Illumina HiSeq4000 platform. To estimate the non-specific binding of the target antibodies, we also sequenced ChIP DNA obtained using two different antibodies not directed to any known antigen. These consisted of a normal rabbit IgG isolated from naïve rabbit serum (Cell Signalling Technologies #2729; ChIP-seq sample name rAb_Igg) and an unspecific mouse monoclonal IgG1 antibody (Cell Signalling Technologies #5415; ChIP-seq sample name mAb_IgG1_G3A1) (All antibodies used for ChIP experiments are listed in Chapter 2 - Table 2.2). Input ChIP material from J1 and DNMT.TKO cells crosslinked with formaldehyde alone or with both DSG and formaldehyde was used as a control for sonication and sequencing biases. To minimise the amount of sequencing to be performed, only one biological sample was sequenced for each of these controls (Table 5.2).

Table 5.2 – continued on next page

ChIP-seq Sample	Total paired-end reads	overall alignment (%)	% uniquely aligned fragments	% multiply-aligned fragments	# uniquely mapped reads after filtering
CEBPB_J1_REP1	52,512,210	97.68	72.88	23.36	76,173,251
CEBPB_J1_REP2	56,050,813	97.78	73.37	22.85	80,714,135
CEBPB_J1_REP3	60,948,818	97.81	72.84	23.53	87,102,976
CEBPB_TKO_REP1	58,291,139	97.8	72.87	23.42	83,972,382
CEBPB_TKO_REP2	54,313,214	97.94	72.53	24.02	76,251,885
CEBPB_TKO_REP3	60,555,220	97.87	73.49	22.97	88,147,935
GABPA_J1_REP1	46,216,554	97.07	76.6	18	72,620,466
GABPA_J1_REP2	53,022,130	97.18	76.02	19.21	83,105,996
GABPA_J1_REP3	58,944,212	96.76	76.13	18.83	87,057,420
GABPA_TKO_REP1	41,882,116	96.64	73.64	20.2	64,297,311
GABPA_TKO_REP2	64,863,967	97.17	75.04	20.01	97,907,092
GABPA_TKO_REP3	58,705,368	96.64	74.31	20.89	84,083,539
KLF4_J1_REP1	48,214,399	96.74	73.5	19.24	73,857,824
KLF4_J1_REP2	47,109,175	97.27	77.77	17.09	75,045,785
KLF4_J1_REP3	47,315,250	96.62	75.16	17.73	73,902,983
KLF4_TKO_REP1	50,694,412	95.68	71.51	19.49	74,123,809

KLF4_TKO_REP2	49,542,336	97.71	75.39	19.2	76,343,772
KLF4_TKO_REP3	42,825,087	96.78	73.28	19.42	65,342,596
MAX_J1_REP1	60,528,828	96.89	74.28	21.1	87,202,265
MAX_J1_REP2	58,198,902	97.47	75.32	20.5	86,649,433
MAX_J1_REP3	58,366,972	97.14	75.23	20.2	86,139,866
MAX_TKO_REP1	67,340,276	96.52	73.51	21.47	96,085,687
MAX_TKO_REP2	59,817,535	96.77	74.47	20.51	86,639,858
MAX_TKO_REP3	56,149,755	96.99	72.34	20.85	82,166,226
NRF1_J1_REP1	55,862,657	97.22	75.23	20.44	82,624,343
NRF1_J1_REP2	55,275,800	96.29	75.43	18.06	77,736,036
NRF1_J1_REP3	49,858,499	97.27	74.42	19.93	74,853,595
NRF1_TKO_REP1	47,983,956	93.33	71.16	20.52	67,864,851
NRF1_TKO_REP2	52,325,079	96.4	74.02	19.42	74,121,075
NRF1_TKO_REP3	56,917,110	97.22	73.87	19.84	84,383,080
SP1_J1_REP1	47,468,483	97.6	75.35	20.19	72,431,400
SP1_J1_REP2	50,796,943	97.06	75.46	19.23	76,496,078
SP1_J1_REP3	55,657,945	95.3	73.16	20.43	80,279,269
SP1_TKO_REP1	48,450,135	97.5	73.22	22.33	72,776,111
SP1_TKO_REP2	61,680,630	96.93	74.24	20.07	92,713,702
SP1_TKO_REP3	65,487,094	96.16	73.29	20.7	91,672,303
YY1_J1_REP1	48,838,028	97.85	75.2	20.22	74,146,059
YY1_J1_REP2	51,330,547	96.28	73.01	17.61	79,231,470
YY1_J1_REP3	48,465,141	96.6	73.78	19.07	74,514,434
YY1_TKO_REP1	46,967,645	97.47	73.19	22.18	70,370,156
YY1_TKO_REP2	50,504,537	96.88	72.5	20.68	74,991,372
YY1_TKO_REP3	52,117,603	96.87	73.1	19.77	77,526,359
rAb_Igg_J1	57,258,729	97.58	75.07	20.35	96,718,086
rAb_Igg_TKO	46,695,021	97.47	73.83	21.4	80,277,056
mAb_IgG1_G3A1_J1	47,959,622	96.11	73.99	19.68	75.838.937
mAb_IgG1_G3A1_TKO	51,824,416	96.37	73.94	19.18	86.893.323
Input_DSG_J1	54,810,829	95.86	62.46	25.61	75,586,261
Input_DSG_TKO	59,261,696	93.76	53.74	26.9	76,143,033
Input_J1	57,048,153	97.06	71.18	22.93	83,937,664
Input_TKO	49,262,687	97	68.84	25.44	71,137,730
Max	67,340,276	97.94	77.77	26.9	97,907,092
Median	52,767,170	96.995	73.805	20.285	77,631,198
Mean	53,650,353	96.8472	73.2832	20.7254	79,803,966
Min	41,882,116	93.33	53.74	17.09	64,297,311

Table 5.2: Total read counts and alignment metrics for all transcription factor ChIP-seq samples.

5.3 Transcription factor occupancy in wild type versus DNA methylation deficient cells

5.3.1 Identification of TF differential binding sites using ChIP-seq data

An average of 53.6 million, 75bp paired-end reads were sequenced for each sample (range: 41.9 - 67.3 million; median: 52.8 million; Table 5.2). These were aligned to the mouse mm10 genome assembly, duplicate reads were removed and multiply mapping reads discarded (see details in Chapter 2 – Section 2.9.2). Visualisation of the data showed strong signal-to-noise ratios for the transcription factor ChIP-seq samples (Figures 5.2A-B).

For each of the seven transcription factors, binding sites were identified genome-wide using the Dpeak peak-calling algorithm (DANPOS2 toolkit, Chen et al., 2013) which takes into account information from replicate samples and uses the input sample as background. Peak intervals defined for WT and DNMT.TKO cells were then concatenated and merged to obtain a list of all possible binding sites. At these genomic intervals, replicate samples from the same cell line clustered together and correlated strongly with each other indicating that there was more variation from biological rather than experimental sources (Figure 5.3). For GABPA ChIP-seq, one of the DNMT.TKO biological replicate samples clusters separately from the other two with these displaying higher correlation scores with the WT samples. An explanation for this observation will be proposed in an upcoming section.

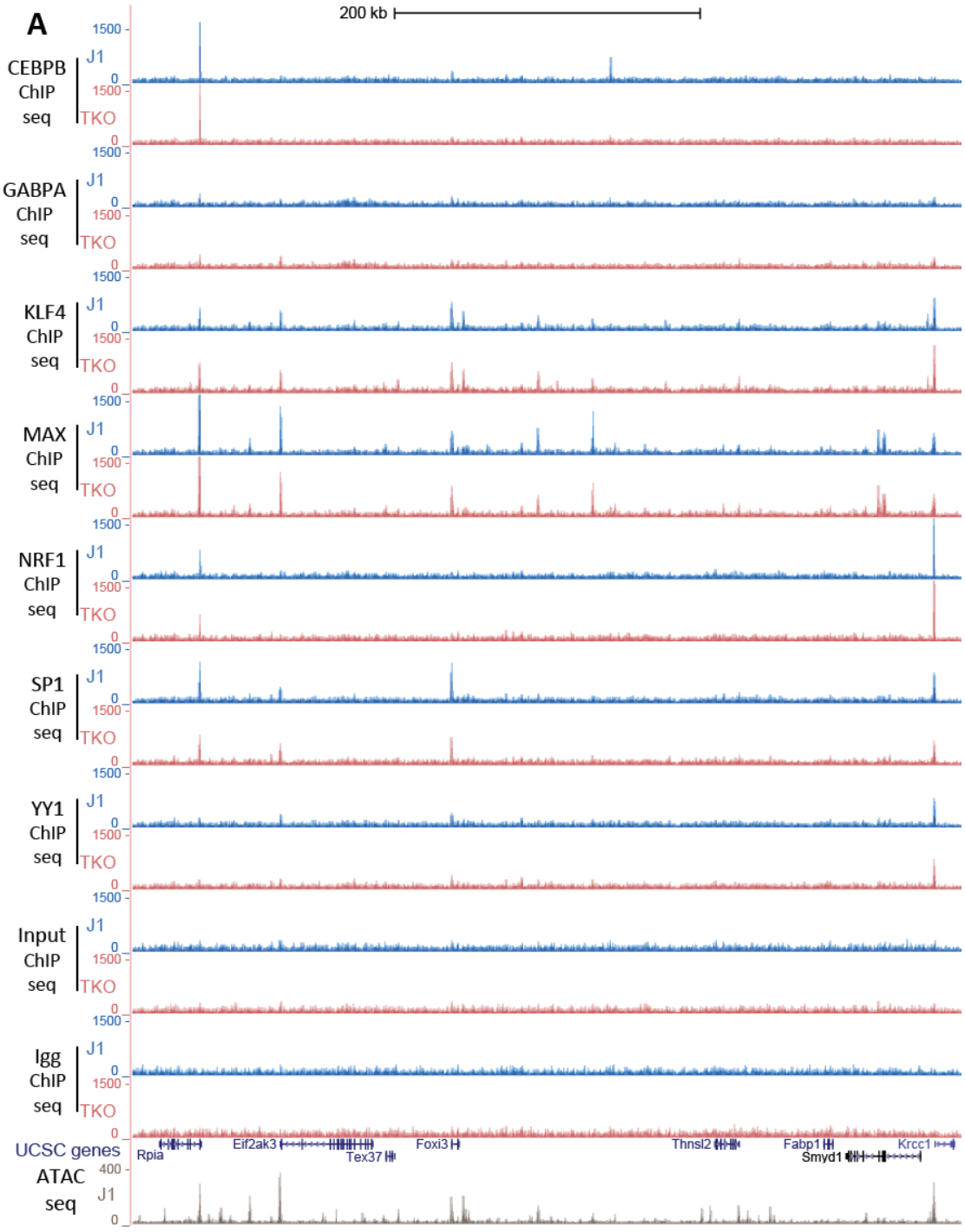


Figure 5.2

Figure continued on next page

Figure continued from previous page

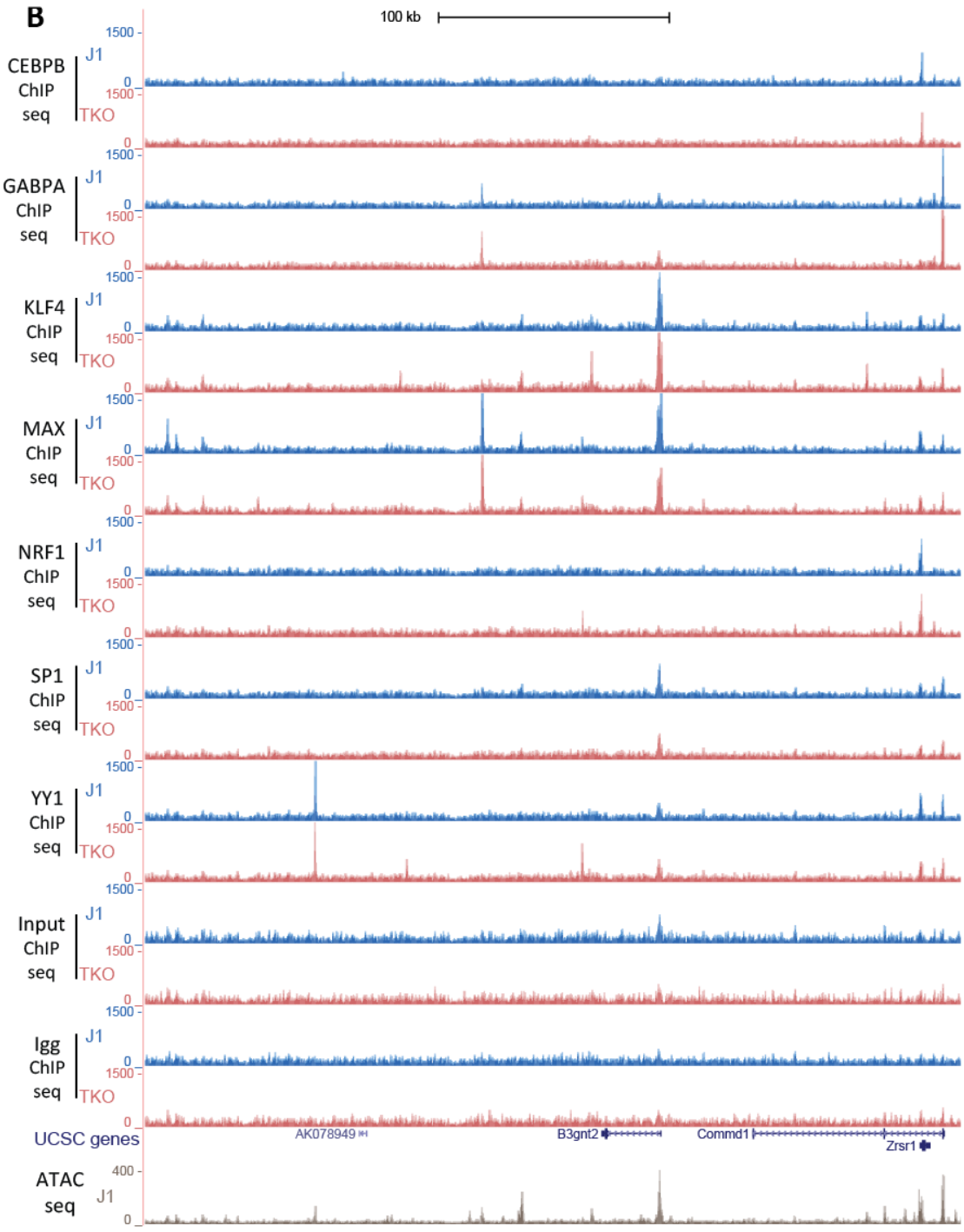


Figure 5.2: UCSC genome browser snapshots showing representative transcription factor ChIP-seq data. ChIP-seq signal for wild type (J1) and DNMT.TKO cells is displayed for each transcription factor as well as rAb-Igg and Input controls. UCSC gene annotations are shown along with ATAC-seq signal from DMSO-treated wild-type cells. Coverage graphs were generated after merging ChIP-seq or ATAC-seq alignments from replicate samples and were normalised to library size. Mm10 coordinates - **A**: chr6:70,748,013-71,289,372 ; **B**: chr11:22,637,385-22,990,744 .

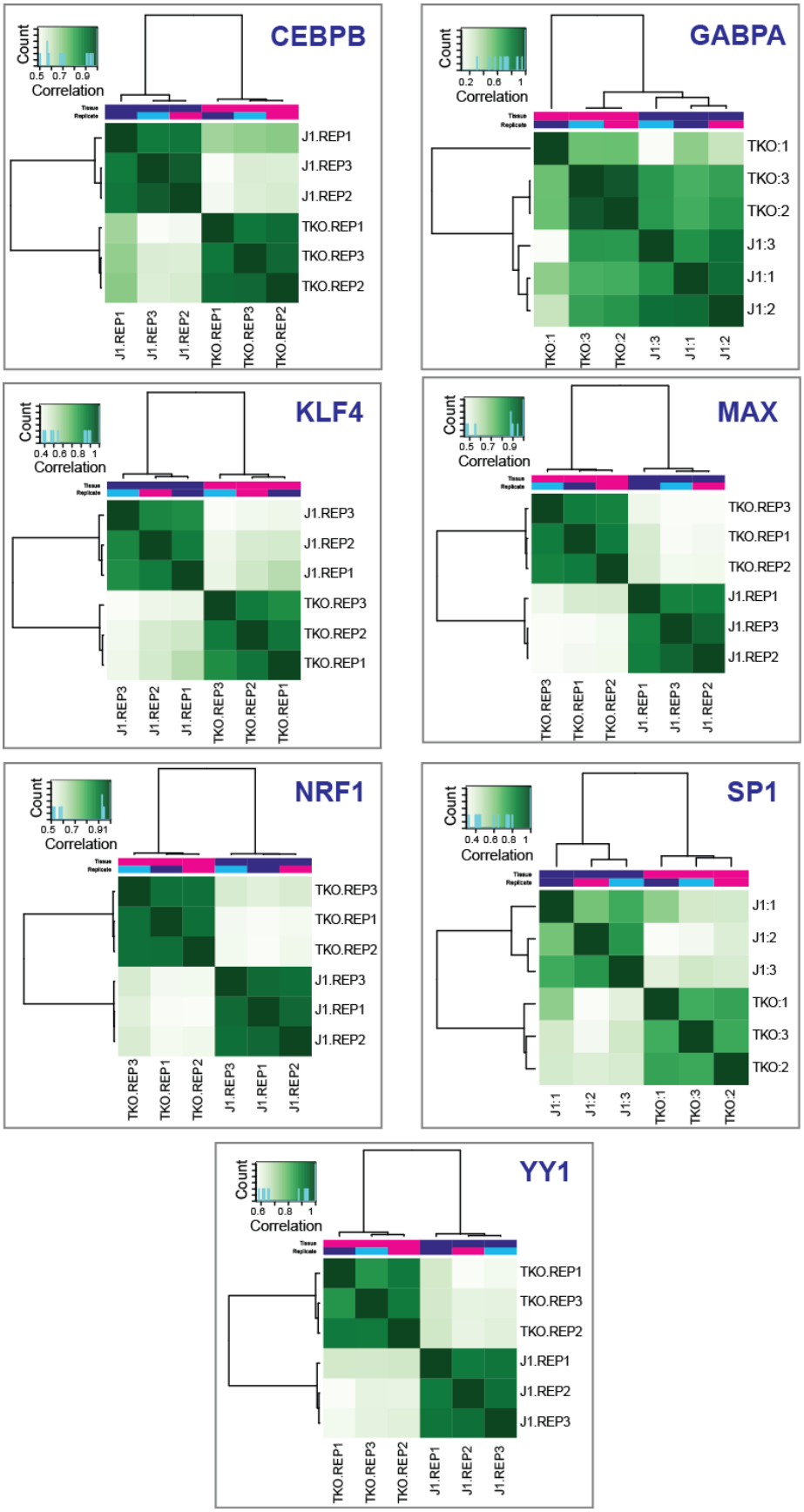


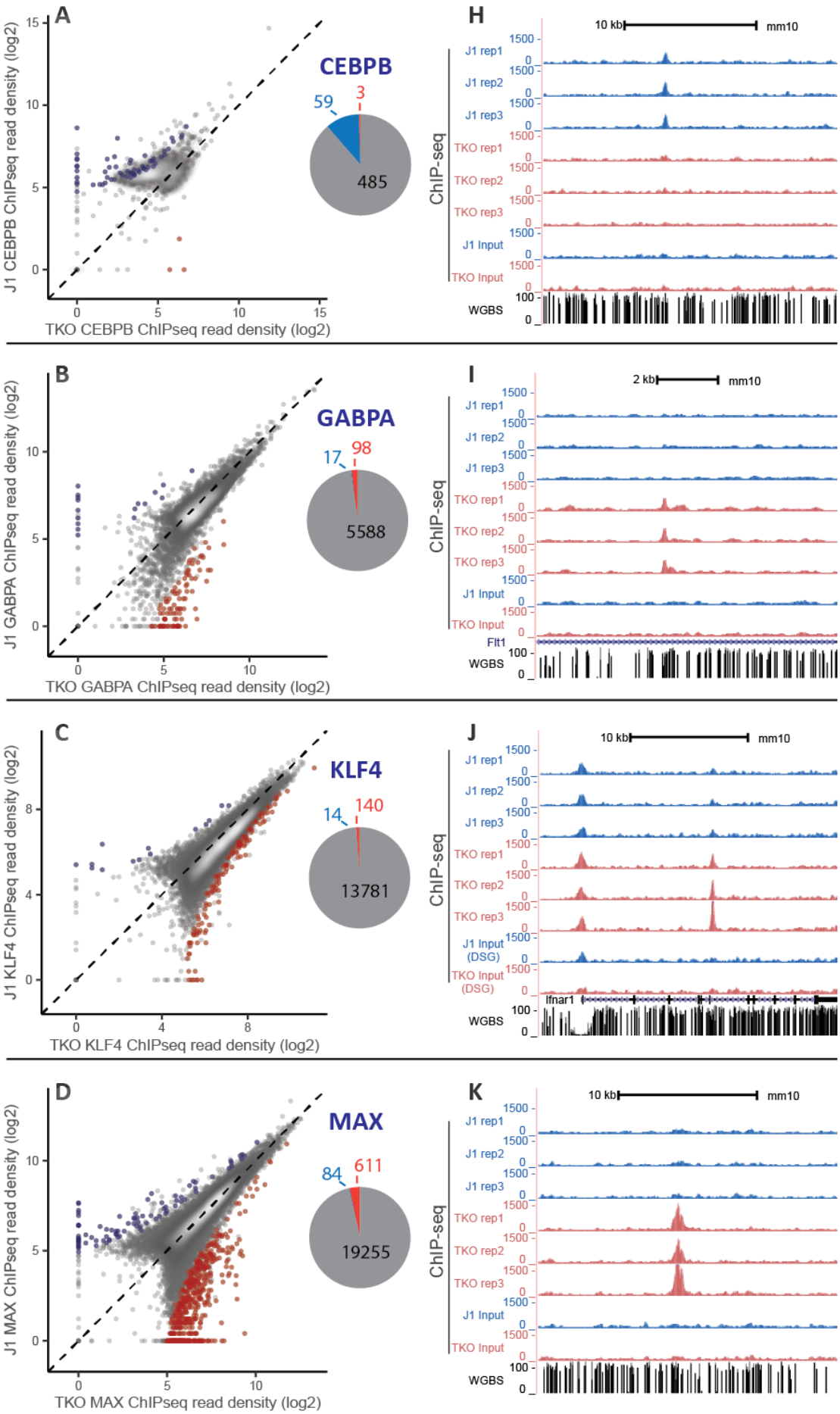
Figure 5.3: Correlation heatmaps comparing the ChIP-seq signal obtained for biological replicate samples at all identified binding sites per transcription factor.

Figure legend continued on next page

Figure legend continued from previous page
At every binding region identified for CEBPB, GABPA, KLF4, MAX, NRF1, SP1 or YY1, ChIP signal was calculated as the RPKM of the TF ChIP-seq reads divided by RPKM of the input control reads. Pearson correlation scores were calculated for each pair of biological replicas and these values were used for clustering and the color scale. A color key is drawn at the top left corner of each heatmap. Correlation heatmaps were generated in R using the *dba.plotHeatmap()* function of the DiffBind package (Stark and Brown, 2011). J1: samples from wild type mESCs; TKO: samples from DNMT.TKO mESCs.

Using the R package DiffBind (Stark and Brown, 2011), differential analysis was carried out to identify peaks that had significant differential binding in WT and DNMT-TKO cells (Figures 5.4A-G). The number of regions that significantly gain or lose binding in DNMT.TKO compared to WT cells are summarised in the form of pie charts (Figures 5.4A-G). GABPA, KLF4, MAX, NRF1 and YY1 all had an increased number of binding sites in DNMT.TKO cells compared to WT while relatively few regions had significantly reduced occupancy. On the other hand, CEBPB appeared to lose binding sites in DNMT.TKO cells compared to WT, an observation that fits with previous reports suggesting that it has greater affinity for methylated sequences (Mann et al., 2013). Examples of TKO- or WT-specific binding events are presented in Figures 5.4H-N. Fewer than ten significant differential binding sites were identified for SP1 suggesting that this transcription factor's occupancy is unaltered by global loss of DNA methylation.

As was pointed out in the previous chapter, the wild-type J1 cells used for these experiments harbour an additional copy of chromosome 15 compared to DNMT.TKO cells (see Section 4.8). However, in these ChIP-seq experiments the chromosomal abnormality does not affect the differential analysis of peaks located on chromosome 15 as the difference in coverage at these loci, resulting from copy number variation, is corrected for by the input samples that are derived from the same batch of cells.


A
J1 CEBPB ChIPseq read density (log2)
TKO CEBPB ChIPseq read density (log2)
CEBPB
59
3
485
H
10 kb
mm10
J1 rep1
J1 rep2
J1 rep3
TKO rep1
TKO rep2
TKO rep3
J1 Input
TKO Input
ChIP-seq
WGBS
B
J1 GABPA ChIPseq read density (log2)
TKO GABPA ChIPseq read density (log2)
GABPA
17
98
5588
I
2 kb
mm10
Flt1
C
J1 KLF4 ChIPseq read density (log2)
TKO KLF4 ChIPseq read density (log2)
KLF4
14
140
13781
J
10 kb
mm10
J1 Input (DSG)
TKO Input (DSG)
Ifnar1
D
J1 MAX ChIPseq read density (log2)
TKO MAX ChIPseq read density (log2)
MAX
84
611
19255
K
10 kb
mm10

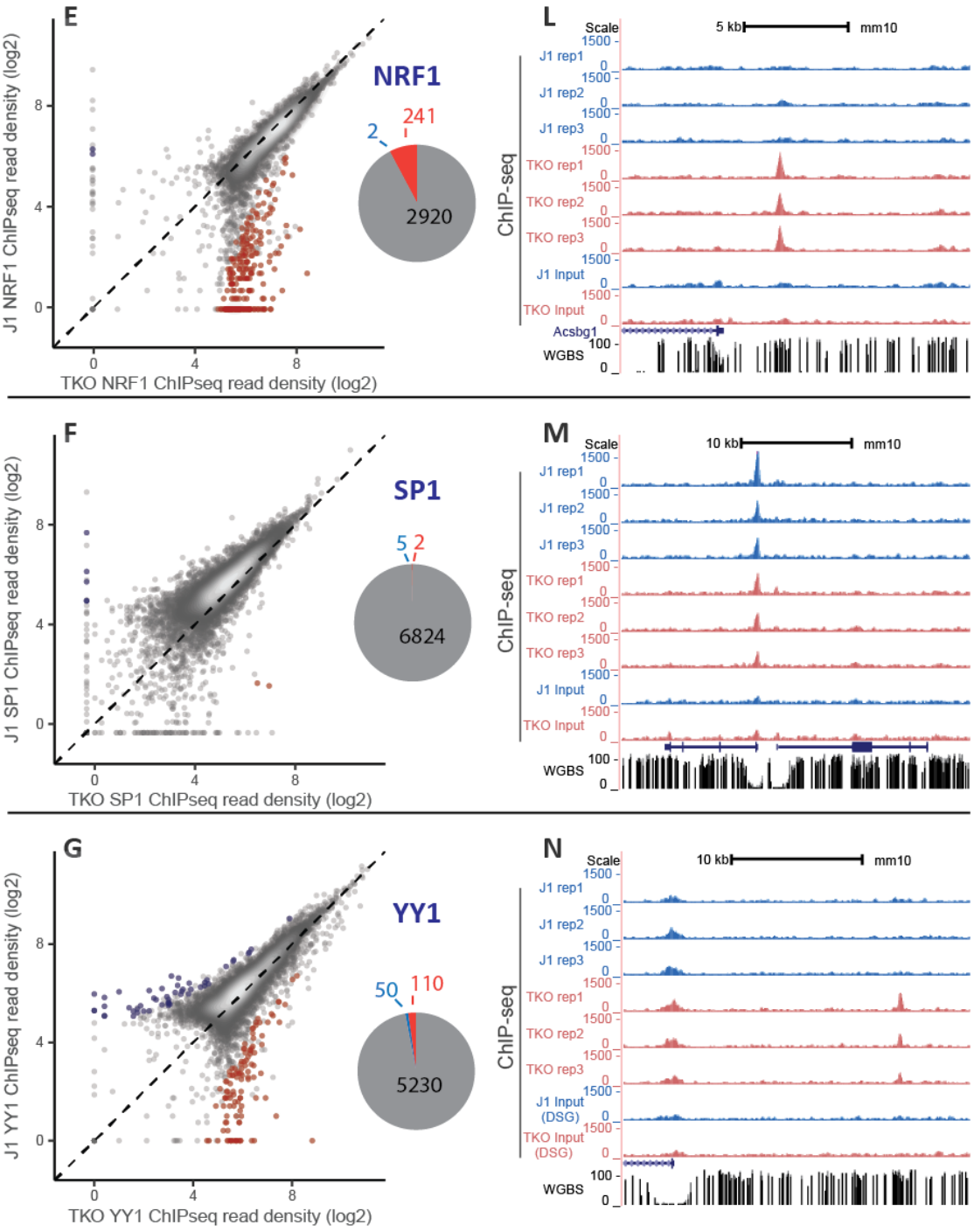


Figure 5.4: Differential analysis of transcription factor binding affinities in wild-type or DNMT.TKO cells. A-G scatterplots: At every binding region identified for CEBPB, GABPA, KLF4, MAX, NRF1, SP1 or YY1, normalised ChIP-seq signal is plotted for samples from wild type (J1) versus DNMT.TKO (TKO) mESCs. The numbers of ChIP-seq peaks showing no significant differential binding (grey), significantly increased occupancy in wild type cells (blue) or significantly increased occupancy in DNMT.TKO cells (red) are indicated in the pie charts to the right. Regions with significant differential occupancy were defined using a fold change cut-off of 1.5x and a p-value threshold of 10^{-3} (CEBPB, GABPA, SP1, YY1) or 10^{-5} (KLF4, MAX, NRF1) after multiple-testing correction.

Figure legend continued on next page

H-N: UCSC genome browser snapshots of regions differentially occupied by CEBPB, GABPA, KLF4, MAX, NRF1 and YY1 in wild type versus DNMT.TKO cells. For SP1, the snapshot displays a binding site with no significant differential binding. FPKM normalised ChIP-seq signal is displayed for all biological replicates per TF as well as Input controls. Wild type CpG methylation obtained using whole genome bisulfite sequencing (WGBS) (Habibi et al., 2013) is also shown. Mm10 coordinates – **H:** chr6:71,051,136-71,073,294 ; **I:** chr5:147,655,643-147,665,482 ; **J:** chr16:91,481,727-91,507,034 ; **K:** chr2:144,732,066-144,753,538 ; **L:** chr9:54,657,051-54,673,602 ; **M:** chr12:4,829,264-4,860,807 ; **N:** chr5:21,051,973-21,078,320 .

It should be mentioned that the significance threshold used to call differential peaks for the CEBPB, GABPA, SP1 and YY1 ChIP-seq data was lower compared to that used for the other datasets (FDR-corrected p-value < 10^{-3} versus < 10^{-5}). The intention was to capture a greater number of differential regions that could be used for downstream analyses and, in the case of SP1, to test whether lack of differential binding sites was due to excessive stringency. Visual inspection of the differential binding sites identified using these less stringent thresholds confirmed that the differences were reproducible between replicates.

As can be seen in Figure 5.4C, many of the peaks that show significantly increased KLF4 binding in DNMT-TKO cells also have high ChIP-seq signal in WT cells (high values on the y-axis). Moreover, all peaks appear to have a higher density of reads in DNMT.TKO cells (deviation away from the dashed line). RNA-seq data suggests that *Klf4* is more highly expressed in DNMT.TKO compared to WT cells (Table 5.3). The implication is that increased levels of KLF4 in DNMT.TKO causes a global increase in ChIP-seq signal at KLF4 peaks. Therefore many of the identified differentially bound sites possibly represent sites with increased occupancy of KLF4 due to higher protein concentration in the nucleus (assuming alterations in mRNA levels translate to protein abundance) rather than *de novo* binding sites resulting from DNA methylation loss. The same phenomena may apply to CEBPB; the *Cebpb* gene is expressed 1.6 fold higher in WT cells compared to DNMT.TKO. However we cannot be confident of this difference in expression as is it not statistically significant (FDR-corrected p-value = 0.13, Table 5.3).

	mean FPKM		Fold Change	FDR
Gene	**J1.DMSO**	**TKO.DMSO**	**TKO.DMSO vs J1.DMSO**	**TKO.DMSO vs J1.DMSO**
Cebpb	9.125	5.663	0.621	1.26E-01
Gabpa	52.382	60.438	1.154	2.80E-01
Klf4	8.933	15.444	1.728	5.74E-04
Max	32.716	35.711	1.091	4.75E-01
Nrf1	21.105	21.363	1.012	9.41E-01
Sp1	43.120	47.993	1.113	3.17E-01
Yy1	40.458	40.286	0.996	9.86E-01

Table 5.3: Differential gene expression analysis for the genes encoding the transcription factors for which ChIP-seq data has been generated.

Furthermore, SP1 ChIP-seq signal also seems to be globally elevated in WT cells compared to DNMT.TKO (Figure 5.4F). Although there is no difference in *Sp1* expression in WT or DNMT.TKO cells according to the RNA-seq data, the gene localises to chromosome 15, which is present at a higher copy number in the WT cells used for these ChIP-seq experiments (see Section 4.8). Nevertheless, the finding that SP1 occupancy is unaltered between the two cell lines still applies despite the variation in protein level, as one would expect DNMT.TKO- or J1-specific binding sites to show higher fold change values.

5.3.2 Genomic distribution of differentially bound sites

Peaks that show increased GABPA, MAX, NRF1 or YY1 binding in DNMT.TKO are predominantly located at intronic or intergenic regions distal from transcription start sites and rarely at promoters (Figure 5.5). This contrasts with the genome-wide distribution of these factors which are otherwise frequently present at gene promoters (Figures 5.5A-G). On the other hand, differentially bound KLF4 peaks follow the same distribution as all KLF4 peaks relative to transcription start sites or genomic annotations (Figures 5.5C and 5.5J). Less than a quarter of all CEBPB peaks are located in proximity to a promoter and those that have significantly decreased binding in DNMT.TKO cells tend to be located distal to transcription start sites (Figures 5.5A and 5.5H).

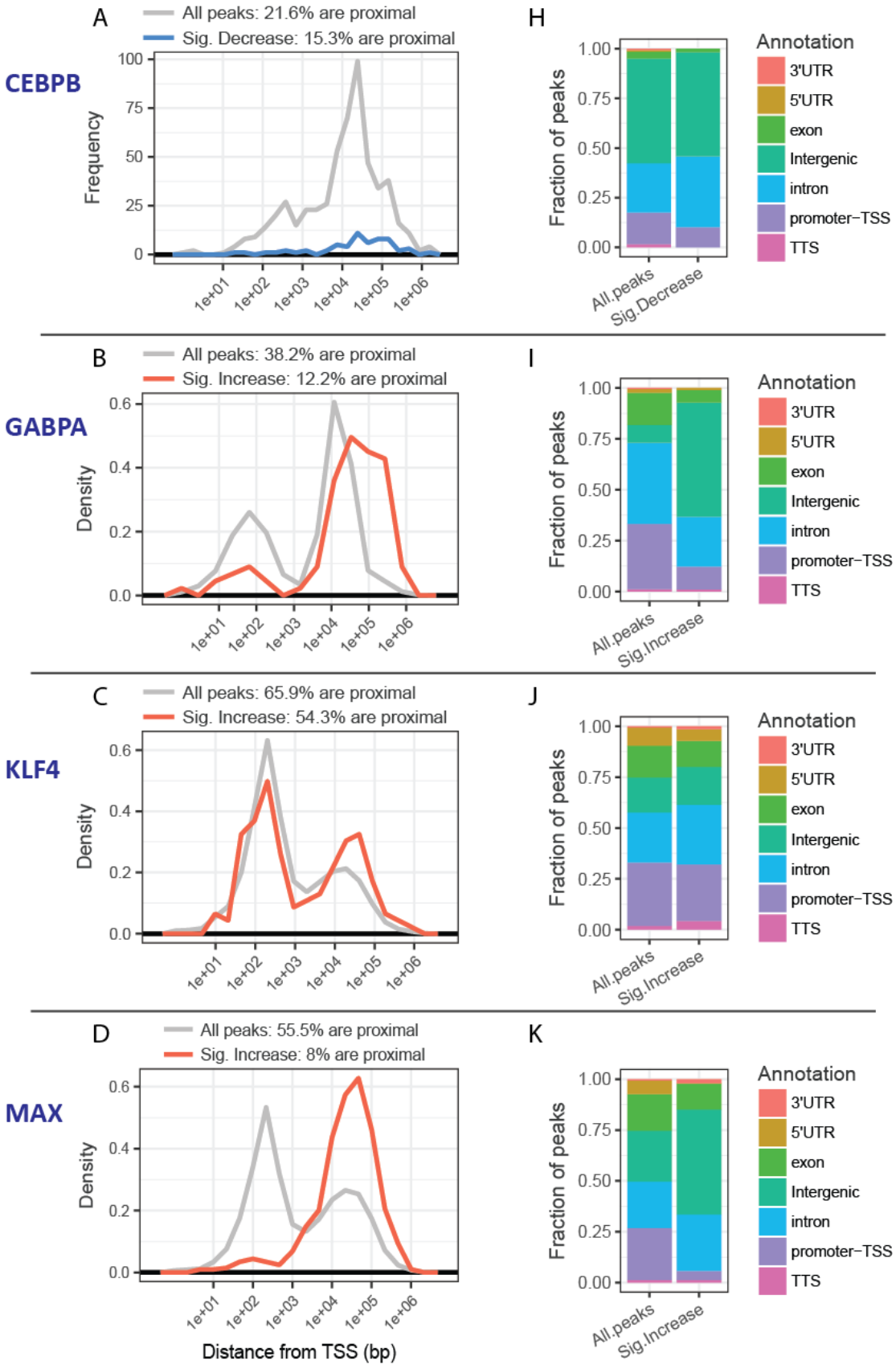


Figure 5.5

Figure continued on next page

Figure continued from previous page

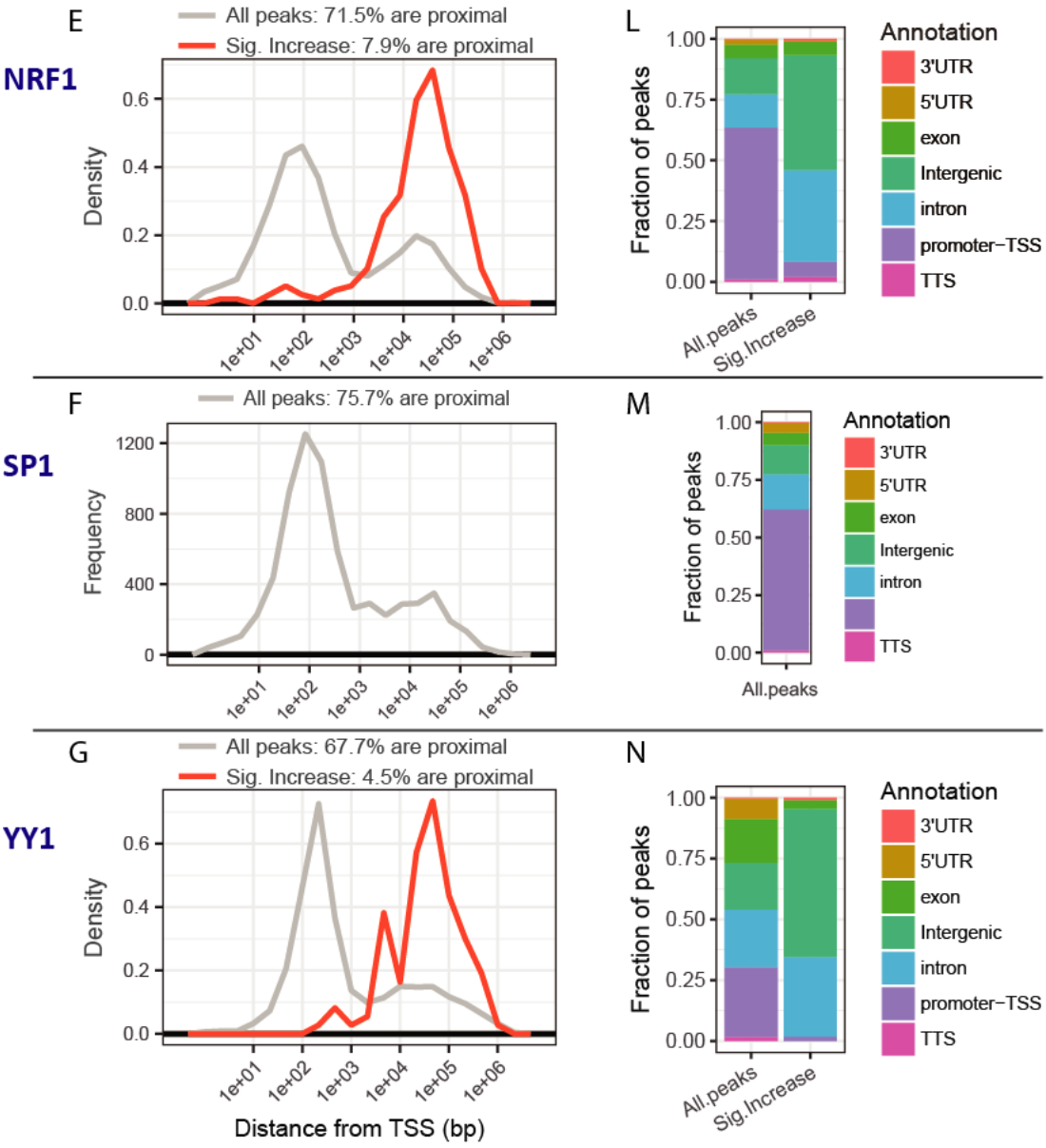


Figure 5.5: Genomic localisation of regions showing differential transcription factor occupancy in wild-type versus DNMT.TKO cells. A-G: Frequency or density plots showing the distance of all ChIP-seq peaks (n = 547 (**A**), 5703 (**B**), 13935 (**C**), 19950 (**D**), 3163 (**E**), 6831 (**F**), 5390 (**G**), grey), wild-type specific peaks (n = 57 (**A**), blue), or DNMT.TKO-specific peaks (n = 98 (**B**), 140 (**C**), 611 (**D**), 241 (**E**), 110 (**G**), red) to the nearest transcription start site (TSS). The percentages of peaks that are proximal to the TSS (centre of peak within 1500 bp of TSS) are indicated above the plots. **H-N:** Fraction of all ChIP-seq peaks, wild-type specific peaks or DNMT.TKO-specific peaks overlapping with mutually exclusive genomic annotations. The number of peaks in each dataset is the same as for figures 5A-G. Promoter-TSS is defined as the region from -1kb to +100bp relative to the TSS. The TTS (Transcription Termination Site) annotation is defined as the region from -100 bp to +1kb relative to a TTS.

5.3.3 The relationship between differential transcription factor binding and DNA methylation

We used publicly available whole-genome bisulfite sequencing data (Habibi et al., 2013) to investigate whether differential binding sites can be related to changes in DNA methylation state between the two cell lines. For each ChIP-seq peak, the DNA methylation level was averaged across the region and this was compared to the fold change in ChIP-seq signal (Figure 5.7). This analysis revealed that most GABPA, MAX, NRF1 and YY1 DNMT.TKO specific peaks were highly methylated in wild type cells (70.01%, 92.73%, 82.35% and 88.99% respectively; mean methylation > 0.6) (Figure 5.7B, 5.7D-E and 5.7G). In addition the difference in methylation between DNMT.TKO-specific sites and binding regions common to both cell lines is highly significant for these transcription factors although to a lesser extent for GABPA (p-value = 9.086e-06, one-tailed Mann-Whitney U test; Figure 5.8). This data suggests that DNA methylation restricts these factors from binding to certain regions in WT cells.

On average, methylation levels were slightly elevated at the KLF4 peaks which show significantly increased enrichment in DNMT.TKO cells (5.8) however there was no clear relationship between DNA methylation and ChIP-seq fold change (Figure 5.7C). 17.78% of novel binding sites had high levels of DNA methylation compared to 29.73% of common peaks suggesting that loss of DNA methylation in DNMT.TKO cells is unlikely to explain the observed differences in KLF4 binding.

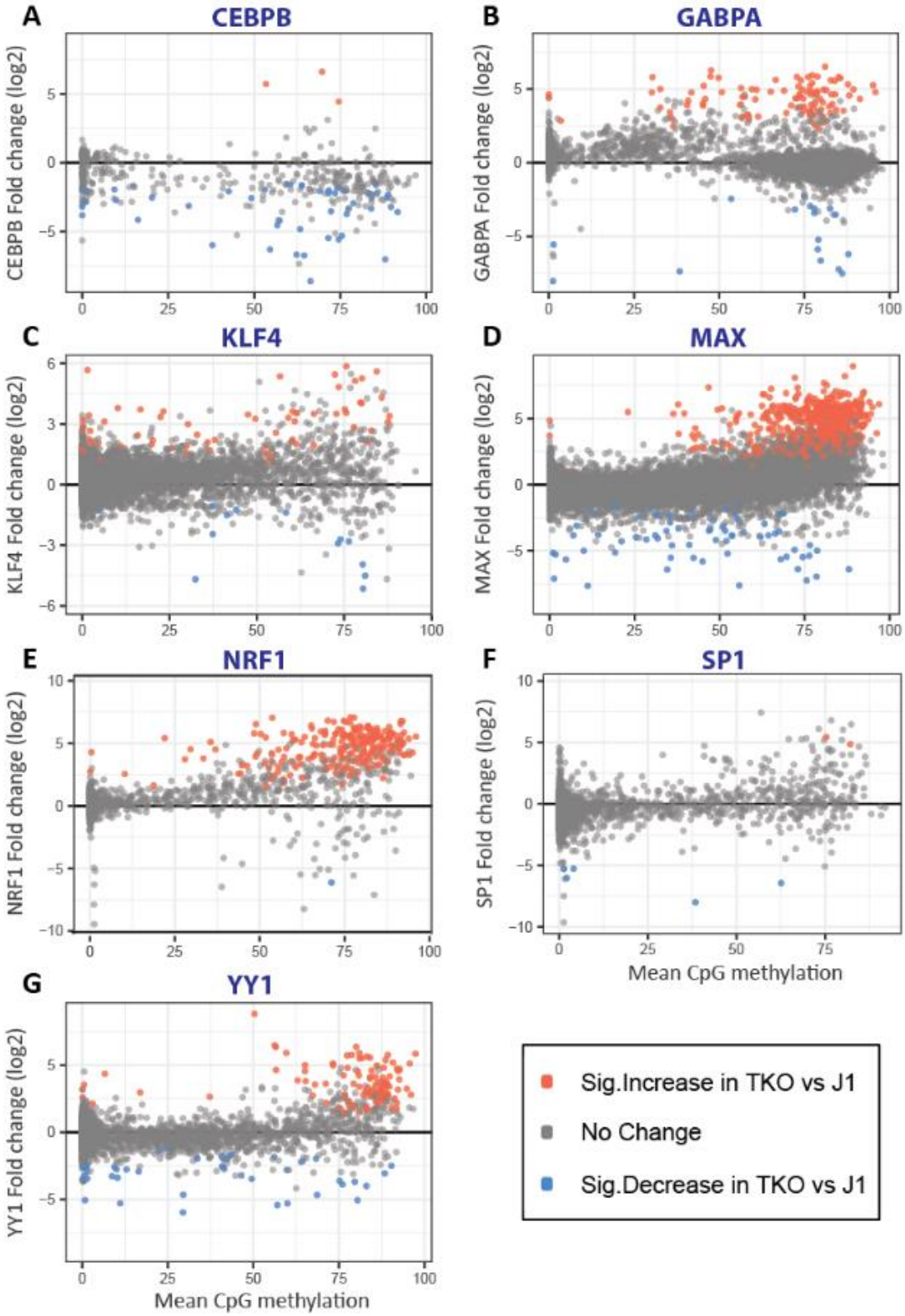


Figure 5.7: Scatterplots comparing the change in TF occupancy in wild type (J1) versus DNMT.TKO (TKO) cells with the average CpG methylation levels at ChIP-seq peaks. ChIP-seq peaks showing no significant differential binding, significantly increased occupancy in wild type cells or significantly increased occupancy in DNMT.TKO cells are represented by grey, blue or red dots respectively. Methylation values are averaged for all CpG sites within the peak region.

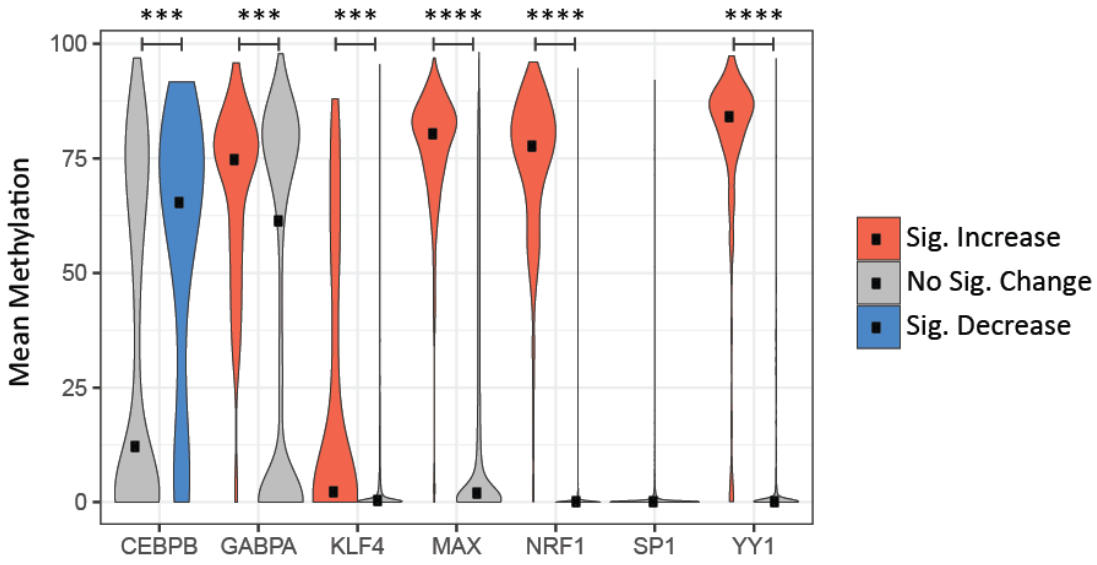


Figure 5.8: Violin plots showing the distribution of CpG methylation levels at transcription factor ChIP-seq peaks grouped according to their differential occupancy in wild type and DNMT.TKO cells. Methylation values are averaged for all CpG sites within each peak region. Data is shown for ChIP-seq peaks equally occupied by CEBPB, GABPA, KLF4, MAX, NRF1, SP1 or YY1 in wild-type and DNMT.TKO cells (No Sig. Change, grey); showing significantly decreased CEBPB occupancy in DNMT.TKO versus wild type cells (Sig. Decrease, blue) or showing significantly increased GABPA, KLF4, MAX, NRF1 or YY1 occupancy in DNMT.TKO versus wild type cells (Sig. Increase, red). Black dots indicate the median methylation level across the group of peaks. One-tailed Mann-Whitney U tests were used to determine significant differences in the distributions of CpG methylation levels: *** p-value < 0.001 ; **** p-value < 2.2e-16.

Likewise, there is no straightforward relationship between CEBPB binding and the average DNA methylation level of the peak regions (Figures 5.7A and 5.8). 61% of regions that display significantly decreased binding in DNMT.TKO cells are highly methylated, against 38.7% for all other peaks. Whether or not some CEBPB binding sites are dependent on DNA methylation warrants for further investigation of the sequence motifs that underlie these peaks.

Before further exploring whether or not the novel binding events in DNMT.TKO cells are associated with the methylation of the relevant transcription factor's consensus binding sequences, we will address an unexpected observation relating to the GABPA ChIP-seq data.

5.4 Identification of broad GABPA enriched domains over gene bodies.

As shown in Figures 5.7B and 5.8, there are many highly methylated regions bound by GABPA that do not display significant differential binding. In fact, visualisation of the coverage graphs revealed that GABPA not only associates with DNA in a localised manner, forming the narrow peaks typically seen in transcription factor ChIP-seq data. Increased signal compared to background was also observed in broad regions as exemplified for two separate loci (Figure 5.9). These broad domains were unique to GABPA and were not observed in ChIP-seq data obtained from Igg controls.

Inspection of the data revealed that the boundaries of these broad domains coincided with the edges of transcribed genes, particularly those with high levels of transcription. Indeed when plotting GABPA ChIP-seq signal for genes ranked and split by expression level, the most highly transcribed genes showed the highest level of GABPA occupancy (Figure 5.10 and Appendix – Figure S2). This observation was not restricted to genes that had a GABPA peak in their promoter region (data not shown).

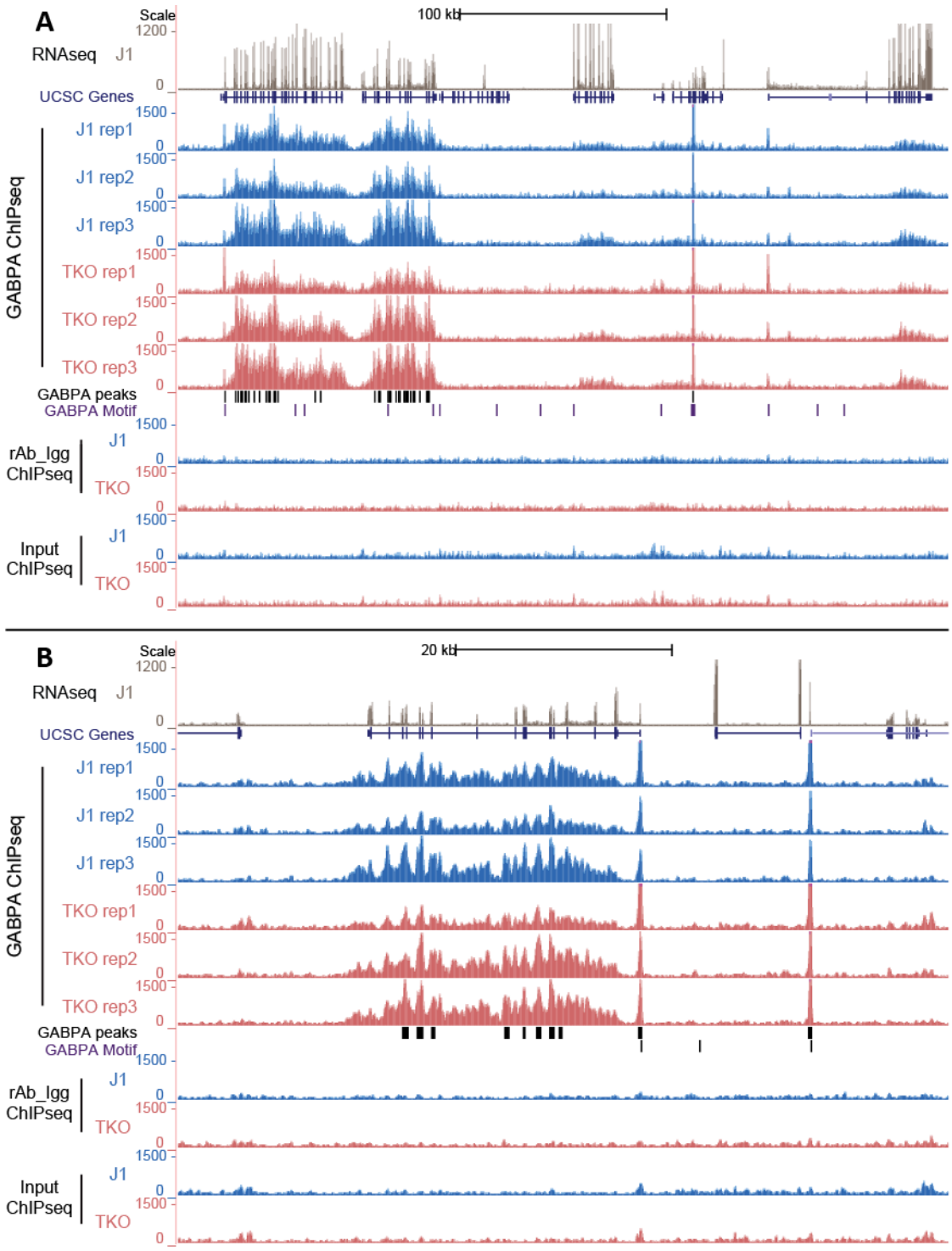


Figure 5.9: UCSC genome browser snapshots showing representative examples of broad GABPA-enriched regions. RPKM normalised ChIP-seq signal in wild type (J1) and DNMT.TKO (TKO) cells is displayed for all GABPA biological replicates as well as rAb-Igg and Input control samples. UCSC gene annotations are shown along with RNA-seq signal from DMSO-treated wild-type cells (both strands, merged replicas, RPKM normalisation). Also shown are the positions of GABPA-peaks called using the Dpeak algorithm (DANPOS2, Chen et al., 2013) and 5'-CGGAA-3' sequences (GABPA motif). Mm10 coordinates - **A**: chr2:90,657,079-91,027,199 ; **B**: chr3:95,813,032-95,884,290.

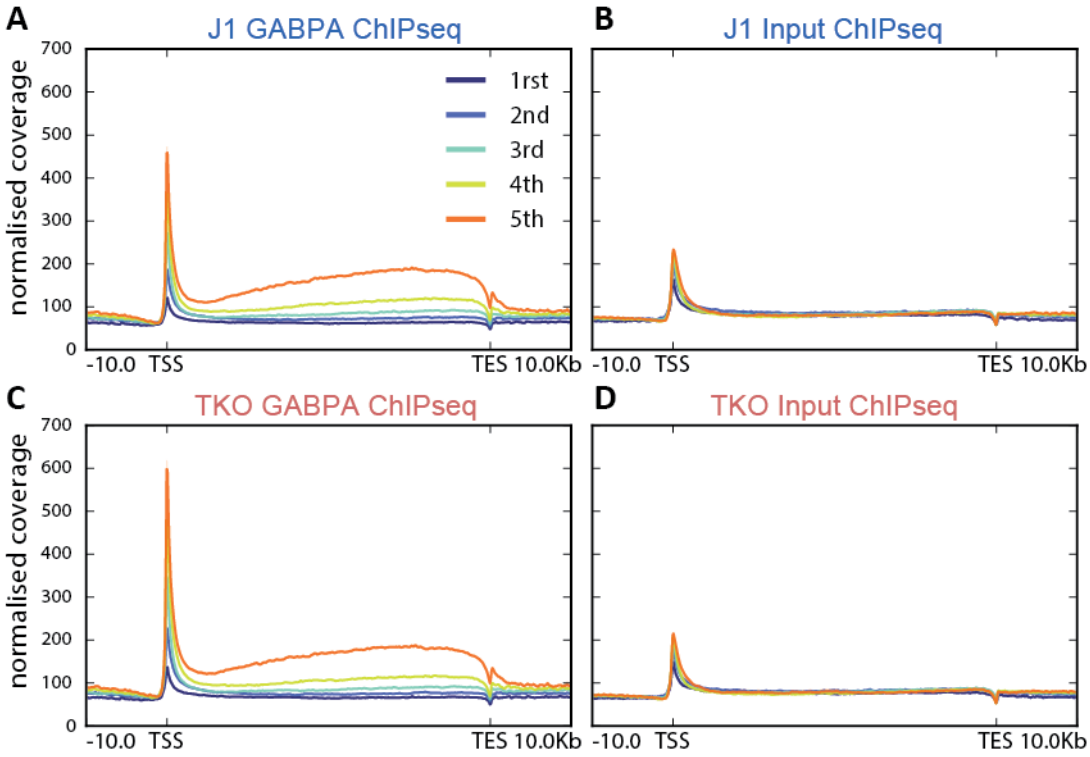


Figure 5.10: Metaplots showing GABPA or Input ChIP-seq signal at genic regions categorised by transcript expression level. Each line represents an equal number of transcripts ranked according to their expression level as measured by the mean RNA-seq FPKM values obtained across both wild-type (J1, figures **A-B**) or DNMT.TKO (TKO, figures **C-D**) biological replicates. RNA-seq FPKM values were calculated using fragments mapping to exons only. GABPA ChIP-seq signal, obtained after merging alignments from replicate samples, is plotted in figures **A** and **C**, Input ChIP-seq signal is plotted in figures **B** and **D**. 5^{th} : Top 20% transcribed genes (5^{th} quintile), 4^{th}: 20%-40%, 3^{rd}: 40%-60%, 2^{nd} : 60%-80%, 1rst : Bottom 20% transcribed genes (1rst quintile). Each category includes 3044 genes

The biological validity and relevance of these broad GAPBA enriched regions was not investigated further in this study. While the Dpeak peak-calling algorithm is designed to identify shorter enriched regions rather than larger domains (Chen et al., 2013), many such "peaks" were located within the broader GABPA enriched regions positioned at gene bodies (see Figure 5.9 for examples). These regions and therefore the peaks that fall within them are typically highly methylated. In order to distinguish between the peaks situated within broad domains and localised, sequence-specific GAPBA binding sites we classified all identified peaks based on the presence of a canonical GABPA binding sequence site (the ETS motif 5'-CGGAA-3') (Lucas et al., 2009). 59% of GABPA peaks were found not to contain a GABPA

binding motif with the majority of these having high levels of DNA methylation and no significant difference in coverage between WT and TKO cells (Figure 5.11). For further analyses, we therefore focused on GAPBA peaks overlapping an ETS motif as these are likely to represent direct binding sites between the protein and DNA. The inclusion of peaks located within broad GABPA enriched regions in the clustering analysis presented in Figure 5.3B may explain the clustering of some DNMT.TKO ChIP-seq replicate samples with the WT samples. The DNMT.TKO replicate that clusters apart from the other seems two display slightly lower enrichment levels within gene bodies relative to the other samples (based on visualisation of data).

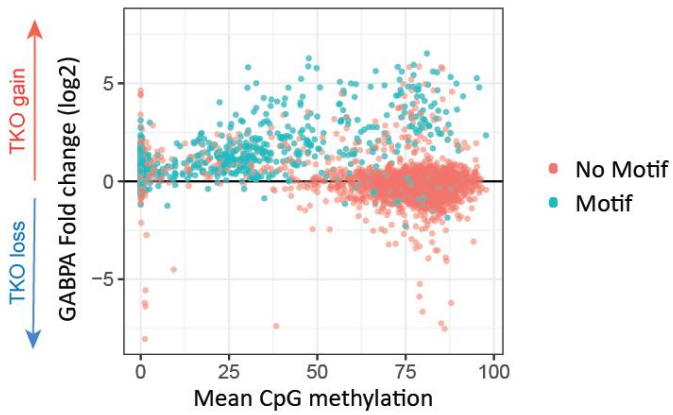


Figure 5.11: Scatterplots comparing the change in GABPA occupancy in wild type (J1) versus DNMT.TKO (TKO) cells with average CpG methylation at ChIP-seq peaks. GABPA ChIP-seq peaks harbouring a 5'-CGGAA-3' sequence are represented by turquoise dots (n = 2106) and other peaks by a red dot (n = 3383).

5.5 Analysis of the transcription factor binding motifs bound upon loss of methylation.

5.5.1 Identifying transcription factor binding motifs and their methylation status

CpG methylation could impede the DNA binding capacity of transcription factors either indirectly, by recruiting DNA methylation binding proteins and causing steric hindrance or by directly affecting the interaction between the protein and its recognition sequence. To determine whether differential occupancy of our transcription factors between WT and DNMT.TKO cells can be linked to methylation of their consensus binding site we explored the sequence underlying the differential binding regions more carefully.

The *de novo* binding motif discovery function of the HOMER tool suite (Heinz et al., 2010) was used to generate position weight matrix (PWM) models depicting the sequences frequently present at peak locations for each transcription factor. Table 5.4 presents the PWM motif most enriched in each peak set compared to a collection of background regions generated automatically and corrected for sequence composition (see methods in Section 2.9.7). The novel motif identified for each transcription factor shows strong similarity to its known binding motif, determined from previous studies, or a motif that was identified for a member of the same class of DNA binding proteins (rightmost column in Table 5.4). The MAX PWM motif was further optimised using HOMER software restricting its size to 8bp. The PWM was shortened to eliminate the influence of the peripheral nucleotides surrounding the core E-box motif CACGTG when screening for the presence of this motif in downstream analyses. Indeed, certain CACGTG instances would be discarded when using the longer PWM, which was not justified by the limited additional information it contained. The most significantly enriched sequence motif underlying KLF4 peaks was 5'-CTSAACAATAGA-3' (where S represents a G or C nucleotide), a motif similar to the Sox15 motif. However, this motif was only found at 1.36% KLF4 target regions so we focused on the second most enriched motif

that was more frequently present at target regions and similar to the recognition sites of the KLF family of proteins.

Factor	De novo Motif – top enrichment	Enrichment p-value	% of target sequences	% of background sequences	Best Match Origin (similarity score)
CEBPB		1e-526	36.56	0.04	**CEBP(bZIP)** ThioMac-CEBPb-ChIP-Seq(GSE21512) *Homer* **(0.922)**
GABPA		1e-753	38.77	9.57	**MA0076.2_ELK4** *Jaspar* **(0.972)**
KLF4		1e-348*	57.84	41.00	**KLF5(Zf)** LoVo-KLF5-ChIP-Seq(GSE49402) *Homer* **(0.975)**
MAX		1e-1134	34.30	13.99	**c-Myc(bHLH)** mES-cMyc-ChIP-Seq(GSE11431) *Homer* **(0.985)**
NRF1		1e-2111	82.99	9.26	**NRF(NRF)** *Homer* **(0.960)**
SP1		1e-771	58.40	24.48	**POL003.1_GC-box** *Jaspar* **(0.931)**
YY1		1e-1234	52.26	10.78	**MA0095.2_YY1** *Jaspar* **(0.955)**
MAX (opt)		1e-963	47.88	25.89	**c-Myc(bHLH)** LNCAP-cMyc-ChIP-Seq(Unpublished) *Homer* **(0.990)**

Table 5.4: *De novo* discovered motifs most enriched in transcription factor ChIP-seq peaks (target sequences) relative to background sequences. De novo motif discovery and enrichment analysis was performed using the HOMER script findMotifsGenome.pl (Heinz et al., 2010). For each novel motif, the known TF motif that was most similar was reported (Best match and similarity score on a scale of 0 to 1). For GABPA, all peaks identified with Dpeak were used regardless of their location within gene bodies or sequence composition. * The novel motif listed for KLF4 is actually the second most enriched at target versus background sequences. The HOMER motif database entry provides information about the transcription factor's DNA-binding domain (DBD-family in brackets) and the ChIP-seq experiment the motif is derived from.

Binding motifs had previously been defined for this set of transcription factors and the PWMs were included in the HOMER database. To validate that the binding sites we determined for these transcription factors harbour sequences matching their cognate recognition sequences, we tested whether the previously characterised binding motifs were enriched for in ChIP-seq peaks. These known motifs are shown in Table 5.5 along with the frequency at which they are found in peak or background regions for the relevant transcription factor.

Factor	Known Motif	enrichment p-value	% of target sequences	% of background sequences	HOMER motif database entry
CEBPB		1e-162	51.92	7.68	**CEBP(bZIP)** ThioMac-CEBPb-ChIP-Seq(GSE21512) *Homer*
GABPA		1e-375	50.48	24.88	**GABPA(ETS)** Jurkat-GABPa-ChIP-Seq(GSE17954) *Homer*
KLF4		1e-303	35.13	21.37	**Klf4(Zf)** mES-Klf4-ChIP-Seq(GSE11431) *Homer*
MAX		1e-1044	38.09	17.37	**Max(bHLH)** K562-Max-ChIP-Seq(GSE31477) *Homer*
NRF1		1e-1872	84.73	12.39	**NRF1(NRF)** MCF7-NRF1-ChIP-Seq(Unpublished) *Homer*
SP1		1e-684	66.64	33.37	**Sp1(Zf)** Promoter *Homer*
YY1		1e-963	34.99	5.39	**YY1(Zf)** Promoter *Homer*

Table 5.5: Most enriched known motifs in transcription factor ChIP-seq peaks (target sequences) relative to background sequences. Enrichment analysis of the 264 known vertebrate TF motifs in the HOMER database was performed using the script findMotifsGenome.pl (Heinz et al., 2010). For all factors except GABPA, the motif shown is the most enriched compared to others entries in the collection. This GABPA motif is ranked 5th most enriched, however the top 10 motifs enriched at GABPA peaks are all highly similar ETS factor motif.. For GABPA, all peaks identified with Dpeak were used regardless of their location within gene bodies or sequence composition.

In all cases these were the motifs most significantly enriched at target regions relative to other motifs in the HOMER collection except for the GABPA PWM, which was the 5th most enriched. Nevertheless, the top 15 known motifs enriched at GABPA peaks are highly similar ETS factor motifs containing the core 5'-CGGAA-3' sequence (data not shown). These findings reassured us in the success of our ChIP-seq experiments and gave us confidence that the peaks that contain sequences matching the motifs listed in table 5 are *bona fide* binding sites for these transcription factors. The fraction of peaks containing a match to their respective PWM that show differential binding between WT and DNMT.TKO cells is represented in a pie chart for each transcription factor (Figure 5.12). The trends in differential genomic occupancy that we had previously seen for each transcription factor (increased occupancy for GABPA, MAX, NRF1 and YY1, decreased occupancy for CEBPB and few differential changes for KLF4 and SP1, see Figure 4) are also observed when focusing on the ChIP-seq peaks that are most likely to represent direct protein-DNA interactions.

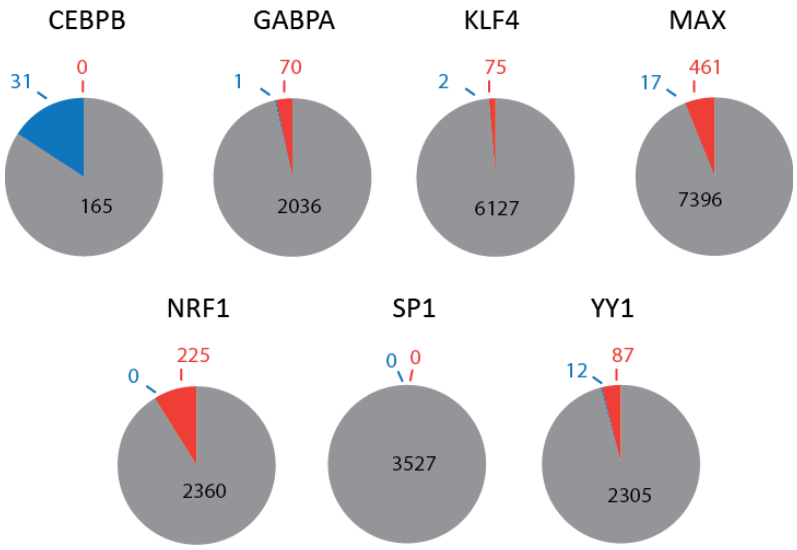


Figure 5.12: Pie charts showing the proportion of ChIP-seq peaks harbouring a TF binding sequence that are differentially occupied in DNMT.TKO versus wild-type cells. The numbers of ChIP-seq peaks showing no significant differential binding (grey), significantly increased occupancy in wild type cells (blue) or significantly increased occupancy in DNMT.TKO cells (red) are indicated. For each transcription factor, only ChIP-seq peak regions that harbour a sequence matching the relevant TF motif (Table 5.4) are included.

For each transcription factor, we isolated all sequences that were located at peak regions that matched the relevant PWM motif and determined the methylation status of the CpG nucleotides. This allowed us to confirm that the DNMT.TKO-specific binding events of GABPA, MAX, NRF1 and YY1 are associated with the methylation of their binding sites (Figure 5.13B, D-E and G): the motifs present at these loci have significantly higher methylation levels as compared to motifs bound both in WT and DNMT-TKO cells (p-value < 2.2e-16, one-tailed Mann-Whitney U tests; Figure 5.14).

The lack of a significant difference in methylation between common and differentially bound KLF4 motifs (p-value = 0.13, Mann-Whitney U test, Figure 5.14) supports the conclusion that DNA methylation is not involved in directing the occupancy of this transcription factor. KLF4 was previously reported to bind an oligonucleotide containing the sequence 5'-TCCmCpGCCC-3', but not its unmethylated form in a protein microarray based assay (Hu et al., 2013). Only 13 such methylated sequences were identified at KLF4 peaks, none of which were present at a binding site devoid of the unmethylated KLF4 motif and none were associated with differential binding. 512 KLF4 peaks contained the shorter 5'-mCpGCCC-3' sequence included in the motif above, however none of these peaks were significantly lost in DNMT.TKO cells. The interaction of KLF4 with methylated DNA is therefore unlikely to direct KLF4 genomic localisation in mouse embryonic stem cells.

Previous studies have reported that CEBPB also has a preference for sequences that contained methylated CpGs (Mann et al., 2013; Yin et al., 2017). In our data, 64% of CEBPB peaks that included a 5'-TTGCGCAA-3' sequence were methylated at the CpG dinucleotide and many of these showed decreased binding in DNMT-TKO compared to WT cells although only 28 out of 128 were significant (Figure 5.13A). This suggests that CEBPB can bind both methylated and unmethylated versions of its motif. Further validation would be required to determine whether binding to the J1-specific loci is dependent on methylation of this motif.

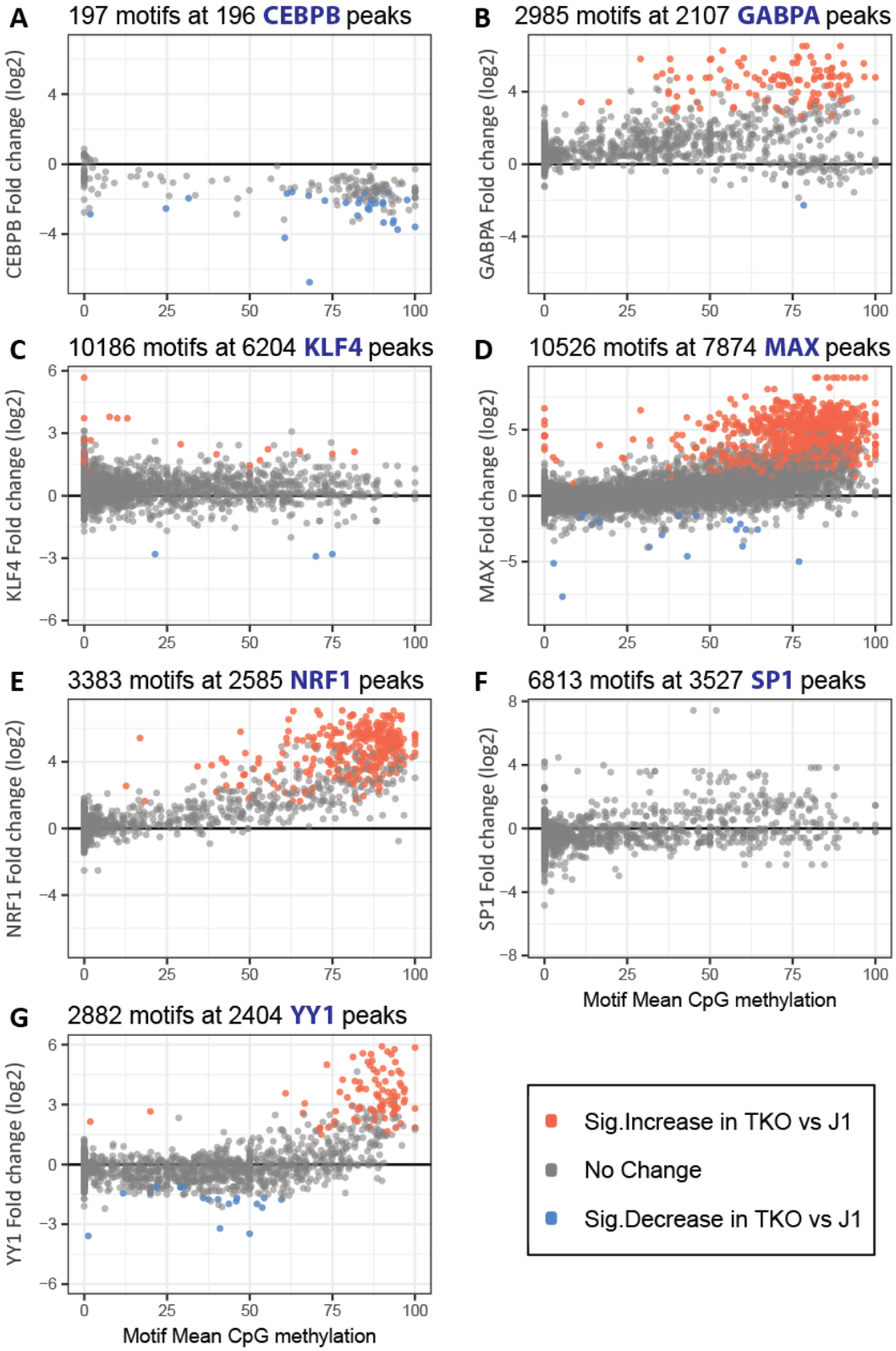


Figure 5.13: Scatterplots comparing the change in TF occupancy in J1 versus DNMT.TKO cells at ChIP-seq peaks with CpG methylation of the underlying TF binding site. ChIP-seq peaks showing no significant differential binding, significantly increased occupancy in wild type cells or significantly increased occupancy in DNMT.TKO cells are represented by grey, blue or red dots respectively. Methylation values are taken for CpG sites within sequences that match the transcription-factor specific PWMs as they are listed in Table 5.4.

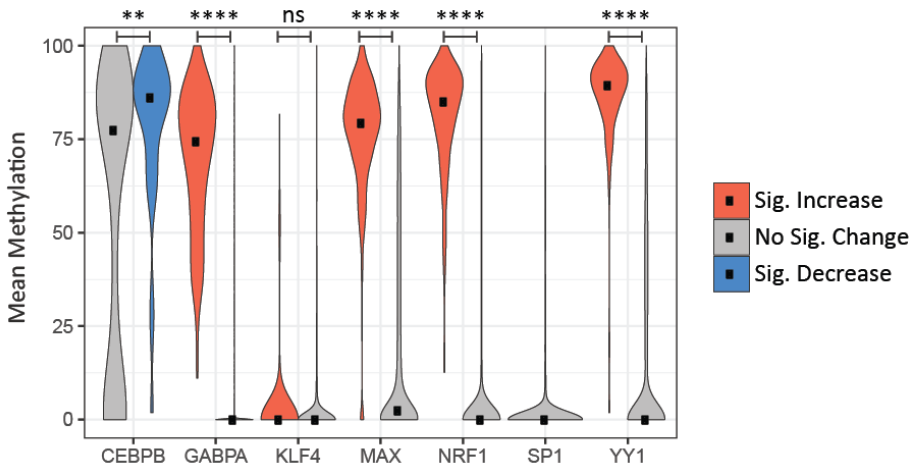


Figure 5.14: Violin plots showing the distribution of CpG methylation levels at TF binding sites that underlie ChIP-seq peaks. ChIP-seq peaks are classified according to their differential occupancy in wild type (J1) and DNMT.TKO (TKO) cells. Data is shown for ChIP-seq peaks equally occupied by CEBPB, GABPA, KLF4, MAX, NRF1, SP1 or YY1 in wild-type and DNMT.TKO cells (No Sig. Change, grey); showing significantly decreased CEBPB occupancy in DNMT.TKO versus wild type cells (Sig. Decrease, blue) or showing significantly increased GABPA, KLF4, MAX, NRF1 or YY1 occupancy in DNMT.TKO versus wild type cells (Sig. Increase, red). Black dots indicate the median methylation level across the group of regions. One-tailed Mann-Whitney U tests were used to determine significant differences in the distributions of CpG methylation levels: ** p-value < 0.01 ; *** p-value < 0.001 ; **** p-value < 2.2e-16; *ns* = non-significant (p-value > 0.05).

In this experiment, diminished binding could also result from a possible reduction in nuclear CEBPB. Finally Mann *et al.* found that CEBPB could bind 5'-(mC/T)GATGCAA-3' sequences in coordination with ATF4 *in vitro* and in primary mouse dermal fibroblasts (Mann et al., 2013) however we did not find any binding sites that contained the 5'-CGATGCAA-3' sequence while 28 harboured the 5'-TGATGCAA-3' sequence. We can therefore provide no further evidence for the binding of CEBPB to regions containing methylated 5'-CGATGCAA-3. Particularly, this interaction may depend on the induction of ER stress and the translocation of ATF4 to the nucleus (Mann et al., 2013) and thus not occur in our model.

5.5.2 Divergence of binding sequences from the consensus motif

As suggested by our data, the genome-wide occupancy of the transcription factors GABPA, MAX, NRF1 and YY1 in wild-type mouse ES cells is restricted by the methylation of a subset of

their binding motifs. We hypothesised that this subset of binding sites contained low affinity recognition sites that diverged from the consensus. If this were the case, it could imply a role for DNA methylation in focusing the binding of transcription factors towards sequences for which they have a highest affinity.

To investigate deviation from the consensus binding site within different subsets of TF peaks, we relied on a system used by the HOMER suite of tools (Heinz et al., 2010) to score short DNA sequences according to their similarity to a PWM. Any given sequence can be scored against a particular PWM by summing up the log-odds probability for each nucleotide to be present at a specific position in the sequence. At each position, presence of a nucleotide with a high probability in the PWM increases the score while a negative value is added to the score if a disfavoured nucleotide is present. To be identified as a match to the consensus motif, the given sequence must have a score greater than a pre-established detection threshold. Lower thresholds for detection threshold represent PWMs that allow for more degeneracy in sequence composition. Detection thresholds for the motifs discovered in our ChIP-seq data were calculated by the HOMER algorithm in a manner that ensures the highest possible enrichment for this motif in peak versus background regions (Heinz et al., 2010).

Within each peak in the CEBPB, GABPA, KLF4, MAX, NRF1, SP1 and YY1 datasets, the sequence most similar to the respective PWM was identified and its score was plotted against the fold change in TF occupancy at this region between WT and DNMT-TKO cells (Figures 5.15 and 5.16). Generally, DNMT.TKO-specific GABPA, MAX and NRF1 or WT-specific CEBPB peaks contained sequences with higher motif scores than those found in binding sites common to both cell lines. This indicates that these sequences are more highly similar to the consensus binding motifs. The YY1 PWM model allows for a relatively high level of degeneracy meaning that both sequences that include or are devoid of a CpG dinucleotide can be characterised as

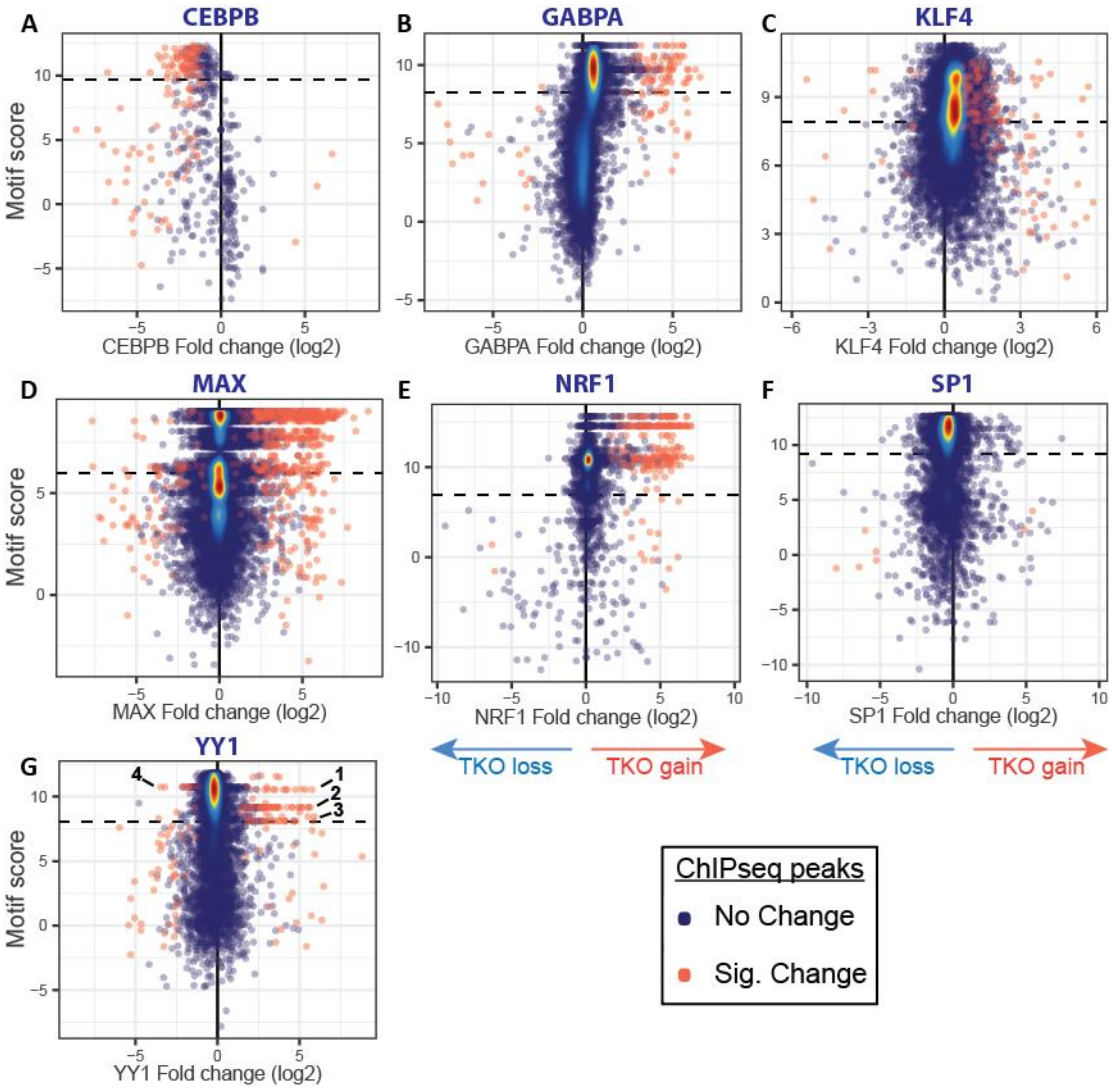


Figure 5.15: Similarity of sequences underlying ChIP-seq peaks to transcription factor binding motifs. A-G: For every CEBPB, GABPA, KLF4, MAX, NRF1, SP1 or YY1 ChIP-seq peak, the underlying sequence was scanned for the maximum log-odds score based on the transcription factor's PWM (listed in Table 5.4). The scatterplots compare this log-odds score with the change in TF occupancy between wild type (J1) and DNMT.TKO (TKO) cells at the same peak region. A high motif score indicates the presence of a sequence underlying the peak that is more similar to the transcription factor's binding motif. The dashed lines indicate the log-odds score threshold above which a sequence is said to match the PWM. ChIP-seq peaks showing significantly differential occupancy in DNMT.TKO versus wild type cells are represented by a solid red dot. **G.** Several specific sequences are frequently present underlying YY1 ChIP-seq peaks. These peaks share the same maximum log-odds score therefore appear as a line in reference to the y-axis. The log-odds scores most frequently found in the sequences underlying YY1 ChIP-seq peaks correspond to the following sequences: 5'-CAAGATGGCGCC-3' (1), 5'- CAAGATGGCGCT-3' (2) , 5'-TAAGATGGCGCT-3' (3), 5'-CAAGATGGCGAC-3' (4) respectively. These are labelled on the scatterplot.

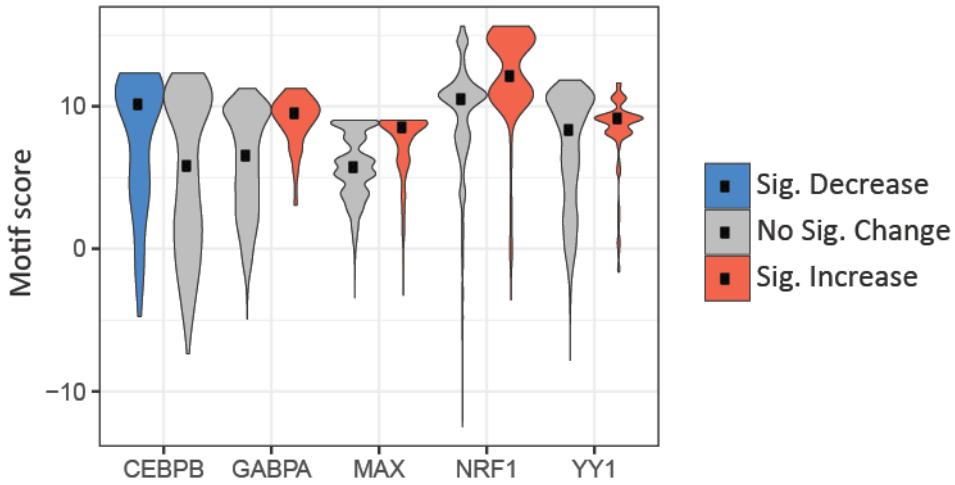


Figure 5.16: Violin plots summarising the distribution of maximum log-odds scores found at different subsets of ChIP-seq peaks. For each transcription factor, ChIP-seq peaks are grouped based on their differential occupancy in wild-type and DNMT.TKO cells. Black dots indicate the median motif score. From left to right, the number of regions included in each grouping is as follows: 59, 485 (CEBPB), 5588, 98 (GABPA), 19255, 611 (MAX), 2920, 241 (NRF1), 5230, 110 (YY1).

matching this motif. As expected from the high levels of methylation found at DNMT.TKO-specific YY1 binding sites Figure 5.13G), YY1 motifs present at these particular sites contain a CpG dinucleotide. Furthermore, these sites contain specific versions of the YY1 motif almost all of which contain the central 5'-AAGATGGCGC-3' sequence (labelled 1-3 on Figure 5.15G). This particular sequence is present at regions annotated as ERVK-type retrotransposons, a finding we will expand upon below. The sequence most frequently found underlying YY1 peaks is 5'-CAAGATGGCGAC-3' (label 4 on Figure 5.15G) and the majority of YY1 binding sites containing this motif overlap the 5' of L1 LINE transposable elements. In mouse embryonic stem cells, 5mC levels are low at the 5' ends of LINE elements (Ficz et al., 2011; Seisenberger et al., 2012) suggesting that YY1 is binding unmethylated motifs at these loci.

Further confirming that the loci bound by GABPA, MAX or NRF1 specifically in DNMT.TKO are likely to represent high-affinity binding sites, these peaks frequently contain multiple copies of the motif (Figure 5.17B-D). On the other hand, YY1 and CEBPB peaks tend to contain only one copy of the TF motif (Figures 5.17A and E).

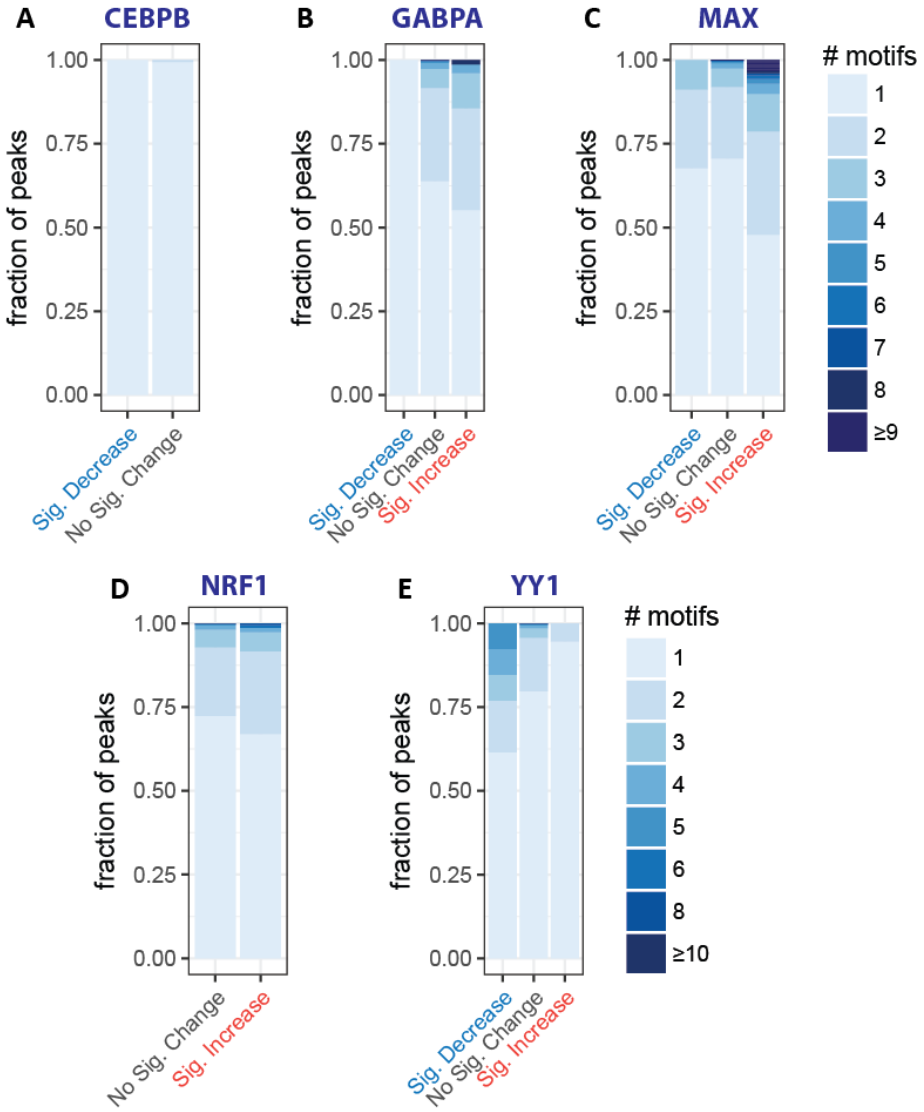


Figure 5.17: Stacked bar charts indicate the fraction of ChIP-seq peaks that harbour one or multiple sequences matching the recognition motif for each transcription factor. For each transcription factor, ChIP-seq peaks are grouped based on their differential occupancy in wild-type and DNMT.TKO cells. "Sig. Increase" and "Sig.Decrease" refer to peaks that display significantly increased or decreased occupancy in DNMT.TKO compared to wild-type cells respectively. For each transcription factor, peaks were scanned for the presence of sequence motifs that match their respective PWM as listed in Table 5.4. From left to right, the number of regions included in each grouping is as follows: **A**: 31, 165; **B**: 1, 2036, 70; **C**: 17, 7396, 461; **D**: 2360, 225; **E**: 12, 2305, 87.

5.5.3 Exploring transcription factor binding site selectivity genome-wide after loss of DNA methylation.

Our ChIP-seq data has led to the identification of DNA methylation sensitive binding events involving GABPA, MAX, NRF1 or YY1 and sequence motifs for which they have high-affinity. However, these novel binding events affect only a small fraction of all potential binding sites located throughout the genome. Despite the genome-wide loss of DNA methylation, barely 1% of approximately 25,000 possible genomic NRF1 binding sites that are methylated in WT cells are stably bound in DNMT.TKO cells. Similarly, as little as 0.2% of ~40,000 GABPA or ~42,000 YY1 and 0.5% of ~230,000 MAX potential binding sites are occupied upon loss of DNA methylation. Evidently, the contribution of DNA methylation to directing genome-wide occupancy of these transcription factors is limited even if we suppose an under-estimation of the number of binding events.

There are many other factors which have been reported to play a role in directing transcription factors to specific genomic locations such as the cooperative binding with other DNA binding proteins that recognise nearby sequence elements, chromatin modifications and accessibility, nucleosome positioning or composition of surrounding sequences (reviewed in (Slattery et al., 2014)). Additionally the transition from a transient binding event to a stable interaction can rely on the recruitment of cofactors, chromatin modifying complexes or transcriptional machinery (Slattery et al., 2014).

Full characterisation of the elements that guide the selection of binding sites or the formation of stable protein-DNA complexes for this set of transcription factors is beyond the scope of this study. Our results indicate that lack of DNA methylation facilitated their binding to certain sites. In an attempt to understand why these loci were selected as opposed to other genomic regions which contained a potential binding site we compared the CpG and GC composition of regions surrounding the TF binding motifs between occupied and unoccupied loci.

Following this we searched for the enrichment of secondary motifs at bound regions with the hope of identifying putative factors that may cooperate with the binding of our transcription factors.

For the transcription factors that we considered, the DNMT.TKO-specific binding sites are located in regions with generally reduced CpG and GC content compared to loci bound in both cell lines (Figure 5.18); the differences between regions bound specifically in DNMT.TKO and those bound in both cell lines are all highly significant ($p<2.2e-16$ one-sided Student's t-test; comparing grey and red violin plots). This agrees with the observation that cell-type specific peaks are predominantly located distal from promoters (Figure 5.5) which tend to be GC and CpG rich. Furthermore, the GC and CpG composition of sequences surrounding motifs with increased binding in DNMT.TKO cells is slightly but significantly higher than around unoccupied motifs (comparing red and green violin plots in Figure 5.18). Whether this difference in sequence composition influences the transcription factors' affinity for these sites is unclear.

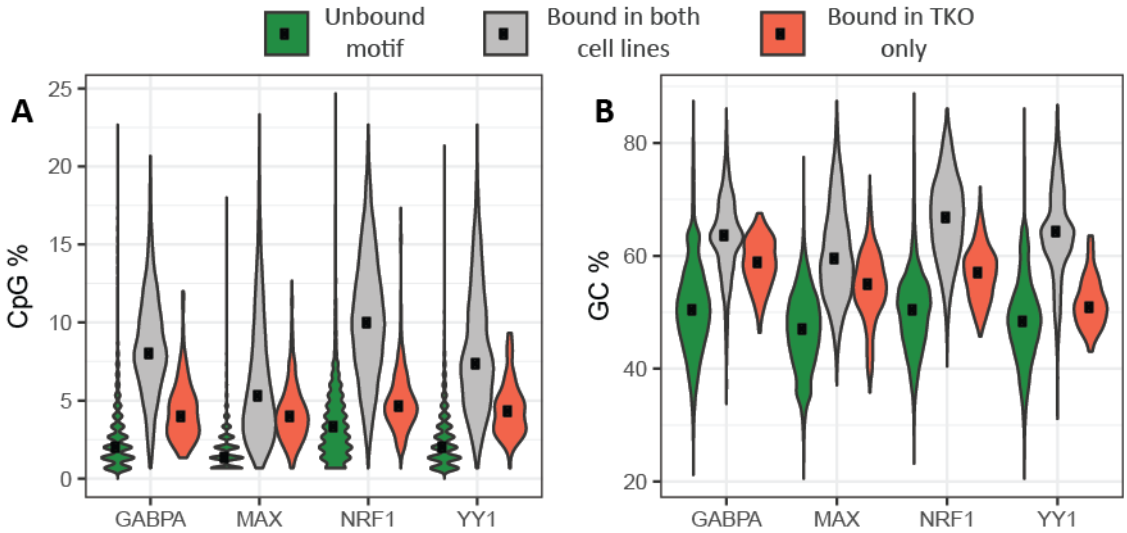


Figure 5.18: Local CpG and GC content for all genomic sequences that match the GABPA, MAX, NRF1 or YY1 PWM motifs. The genome was scanned for the presence of the GABPA, MAX, NRF1 and YY1 binding motifs as they are listed in Table 5.4. For each match, the per cent CpG and GC content for the surrounding 150bp was obtained. Matches were grouped on whether they overlapped a DNMT.TKO-specific ChIP-seq peak (red), a ChIP-seq showing no significant differential binding (grey) or no ChIP-seq peak of the corresponding transcription factor (green). Violin plots summarise the CpG (**A**) or GC (**B**) sequence composition for each subset of loci. Black dots indicate the median CpG or GC content. From left to right, the number of regions included in each grouping is the same in **A.** and **B.** and is as follows: 49226, 2999, 87 (GABPA motif); 282551, 9738, 838 (MAX motif); 28130, 3157, 264 (NRF1 motif); 49993, 2813, 88 (YY1 motif).

The observation that multiple binding motifs are present at a fraction of transcription factor peaks (Figure 5.17) indicates that interactions between proteins of the same family may be involved in the creation of stable protein-DNA complexes. Nevertheless, the close proximity of a second motif is not sufficient in itself for the formation of a stable binding complex as a large number of unbound motifs throughout the genome also appear with a second motif within 150bp (Figure 5.19A). Interestingly, when visualising the data, we noticed that DNMT.TKO specific GABPA peaks frequently have two copies of the ETS motif on the same strand. Furthermore, both motifs were preferentially situated 20-50bp from each other (Figure 5.19B) implying that the orientation and location of sequence motifs may be important to the cooperative binding of multiple GABPA molecules.

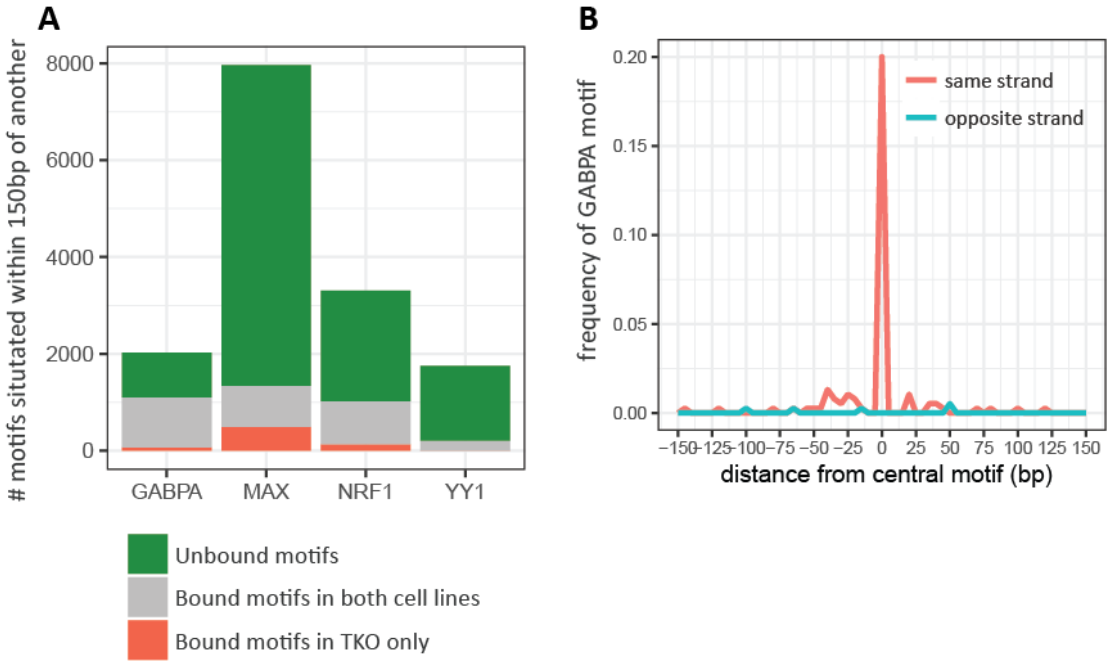


Figure 5.19: Position of transcription factor binding motifs relative to one another. A. Bar chart indicates the number of matches to the GABPA, MAX, NRF1 or YY1 binding motifs located within 150bp of another matching sequence. **B.** 70 GABPA peaks that show significant preferential occupancy in DNMT.TKO compared to wild-type cells were re-centered onto the GABPA motif. The density plot indicates the distribution of GABPA motifs relative to the peak centre within the surrounding 300bp, split into 5bp bins. Motifs were separated based on their presence on the same strand or opposite strand relative to the central GABPA motif.

In order to identify other DNA binding protein motifs that may guide the binding of MAX, NRF1, GABPA or YY1 to only a subset of their potential recognition sites we screened for the enrichment of known motifs. Target regions included motifs specifically bound in DNMT.TKO cells while the background regions consisted of all other motif instances that were methylated in wild-type cells. For each known motif in the HOMER database, the enrichment p-value was plotted against the frequency at which it appears in target regions (Figure 5.20). Only for YY1 was there a significant enrichment for motifs that did not resemble its own consensus binding site. For example the 5'-YTGTTTAC-3' sequence bound by FOX (Forkhead box) proteins was present at 59-73% of DNMT.TKO-specific peaks and the Nkx6.1 motif 5'-GKTAAGR-3' at 84% (Figure 5.20D).

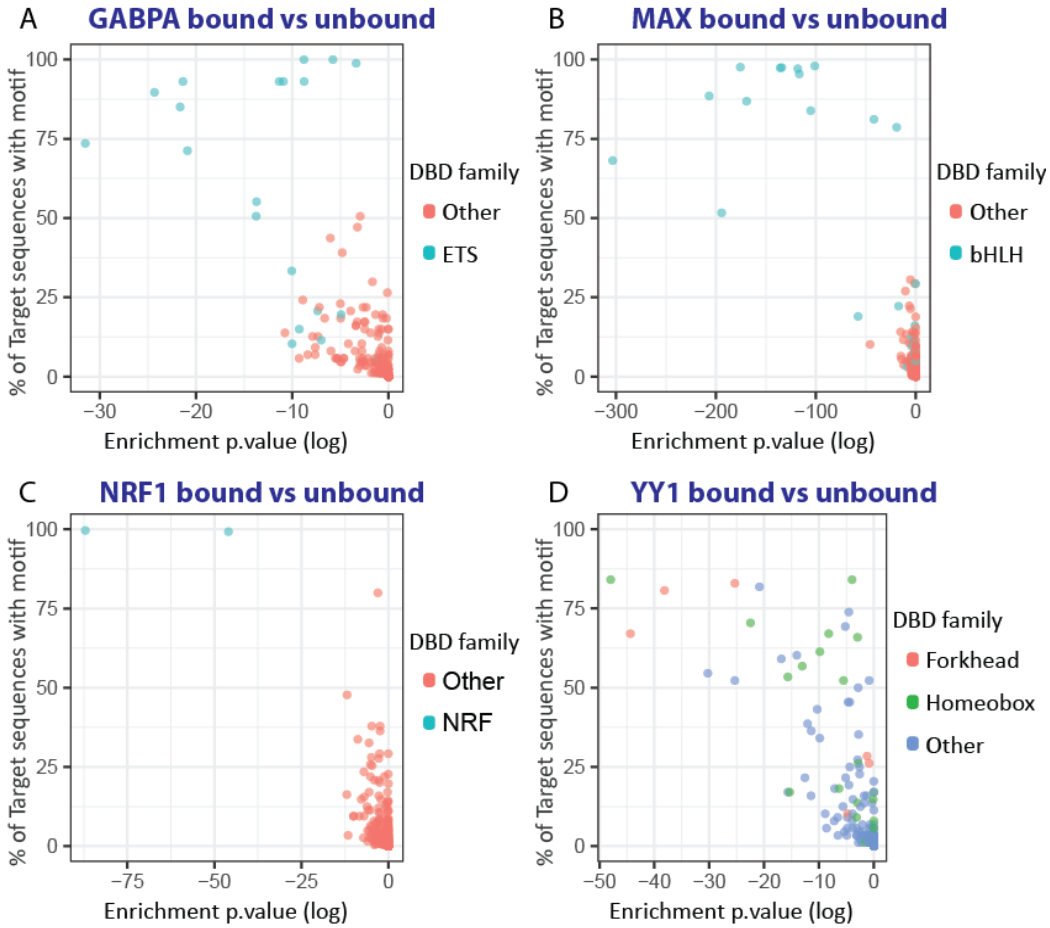


Figure 5.20: Motif enrichment analysis at occupied versus unoccupied regions containing transcription factor binding motifs. Every genomic match identified for the GABPA, MAX, NRF1 or YY1 binding motifs was annotated as bound or unbound if it overlapped a DNMT.TKO-specific ChIP-seq peak or was unoccupied according to our ChIP-seq data, respectively.

Figure legend continued on next page

Figure legend continued from previous page

Only matches with a CpG methylation value above 0.6 were considered for motif enrichment analysis. For motif enrichment analysis, target regions included regions specifically bound in DNMT.TKO cells while the background regions consisted of all other motif instances. For each of the 264 known vertebrate TF motif in the HOMER database, the enrichment significance (p-value) was plotted against the frequency at which it appears in target regions. Points are coloured based on the transcription factor family that is known to bind each specific motif.

Although there was no significant enrichment for the NRF1 motif at novel MAX binding sites and vice-versa, there were 44 instances where MAX and NRF1 were co-localised at the same loci (Figure 5.21A). An example is shown in Figure 5.21B. Although these peaks represent a minority of DNMT.TKO-specific MAX or NRF1 peaks, the finding serves as an example of potential cooperative binding between transcription factors facilitating site-selectivity. Whether the binding of these transcription factors is dependent on the presence of the other cannot be addressed without further experimentation.

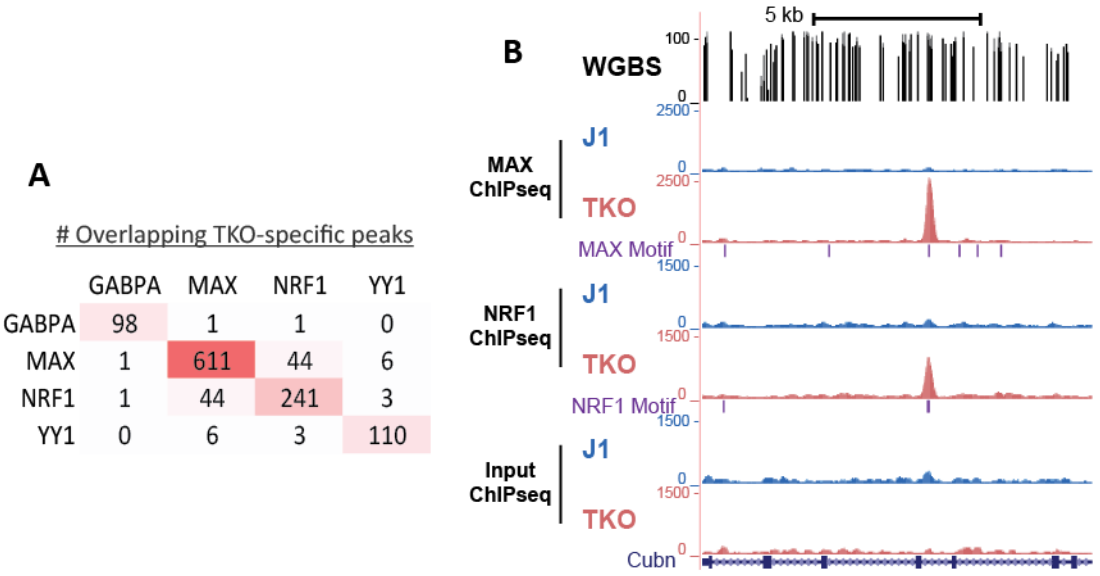


Figure 5.21: Co-localisation of DNMT.TKO-specific transcription factor ChIP-seq peaks. A. Matrix indicating the number of overlapping DNMT.TKO-specific ChIP-seq peaks for different pairs of transcription factors. **B.** UCSC genome browser snapshot of a locus within the *Cubn* gene (chr2:13,283,181-13,283,410) occupied by MAX and NRF1 in DNMT.TKO cells specifically. ChIP-seq signal for wild type (J1) and DNMT.TKO cells is displayed for both transcription factors as well as Input controls. Wild type CpG methylation obtained using whole genome bisulfite sequencing (WGBS) (Habibi et al., 2013) is shown along with the positions of sequences matching the MAX or NRF1 PWM motifs and UCSC gene annotations. ChIP-seq coverage data was generated after merging ChIP-seq alignments from replicate samples and was normalised to library size.

5.6 The impact of transcription factor occupancy at DNMT.TKO-specific binding sites on chromatin accessibility and gene expression

To assess the impact of novel transcription factor binding sites that are related to loss of DNA methylation we made use of our ATAC-seq and RNA-seq data from WT and DNMT.TKO cells to integrate protein occupancy, chromatin accessibility and gene expression data.

5.6.1 Relationship between transcription factor binding and chromatin accessibility

For each transcription factor, differential analysis was carried out to measure the difference in ATAC-seq signal at each peak region. Significantly increased GABPA, KLF4, MAX, NRF1 and YY1 binding in DNMT.TKO cells was associated with significant gains in accessibility both when considering all identified peaks or only those that contain the binding sequence motif (Figure 5.22).

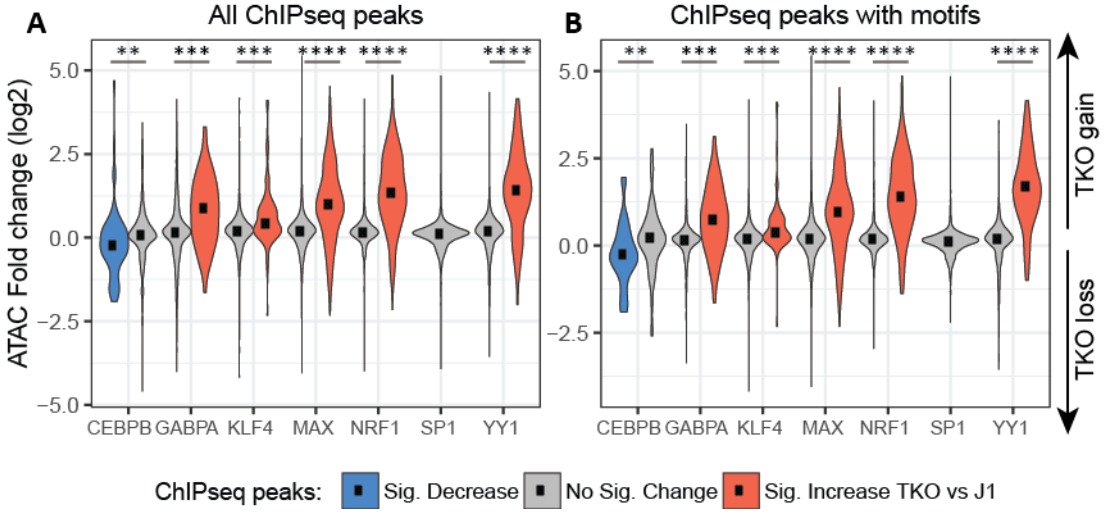


Figure 5.22: Quantitation of ATAC-seq fold change at transcription factor ChIP-seq peaks grouped according to their differential occupancy in wild-type (J1) versus DNMT.TKO cells. Data is shown for ChIP-seq peaks equally occupied by CEBPB, GABPA, KLF4, MAX, NRF1, SP1 or YY1 in wild-type and DNMT.TKO cells (No Sig. Change, grey); showing significantly decreased CEBPB occupancy in DNMT.TKO versus wild type cells (Sig. Decrease, blue) or showing significantly increased GABPA, KLF4, MAX, NRF1 or YY1 occupancy in DNMT.TKO versus wild type cells (Sig. Increase, red). Black dots indicate the median ATACseq fold change. In **A.** all peaks called using the Dpeak algorithm (DANPOS2 toolkit, Chen et al., 2013) were included (from left to right n = 59 , 485 (CEBPB), 5588 , 98 (GABPA), 13781, 140 (KLF4), 19255, 611 (MAX), 2920, 241 (NRF1), 6824 (SP1), 5230, 110 (YY1)). In **B.** the analysis was restricted to ChIP-seq peaks that harboured a TF binding motif (from left to right n = 32 , 168 (CEBPB), 2134 , 76 (GABPA), 7692, 91 (KLF4), 9011, 508 (MAX), 2398, 227 (NRF1), 3989 (SP1), 2712, 92 (YY1)). One-tailed Mann-Whitney U tests were used to determine significant differences in ATAC-seq fold change across peaksets: ** p-value <0.01 ; *** p-value < 0.001 ; **** p-value < 2.2e-16.

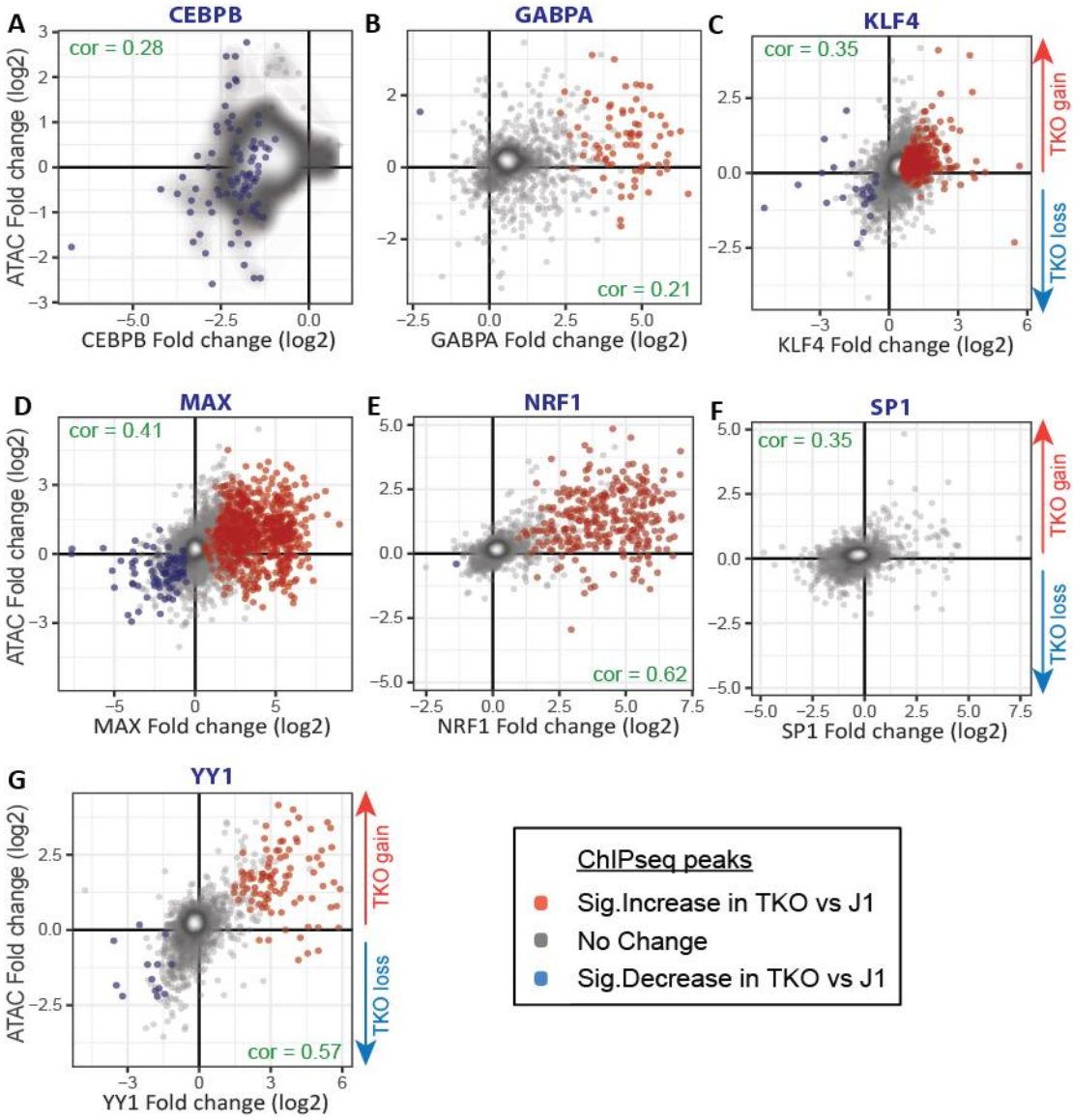


Figure 5.23: Scatterplots comparing the changes in ATAC-seq signal and TF occupancy (ChIP-seq signal) in wild type (J1) versus DNMT.TKO (TKO) cells at ChIP-seq peaks. ChIP-seq peaks showing no significant differential binding, significantly increased occupancy in wild type cells or significantly increased occupancy in DNMT.TKO cells are represented by grey, blue or red dots respectively. Only ChIP-seq peaks that harboured a binding motif for the corresponding transcription factor (see Table 4) were included. *cor* reflects the Pearson's correlation coefficient.

For all transcription factors, there was a positive correlation between differential binding and differential accessibility (Figure 5.23). The correlation coefficients were highest for NRF1 and YY1 while for GABPA and CEBPB the correlation is less convincing of a genuine relationship (Figure 5.23). Out of the 3603 ATAC-seq peaks that have significantly higher signal in DNMT.TKO cells (as determined in Chapter 4), 436 overlap with a DNMT.TKO-specific binding site for GABPA, MAX, NRF1 or YY1 (37, 246, 82 and 71 regions respectively). These findings suggest that novel binding of transcription factors at previously methylated sites can lead to an increase in chromatin accessibility. Nevertheless the inverse causal relationship cannot be excluded. The ATAC-seq signal at regions of DNMT-TKO specific TF occupancy was reduced compared to that measured at regions occupied by transcription factors in both cell lines (Figure 5.24) similar to that observed for DNMT.TKO specific ATAC-seq peaks (Chapter 4 - Figure 4.5).

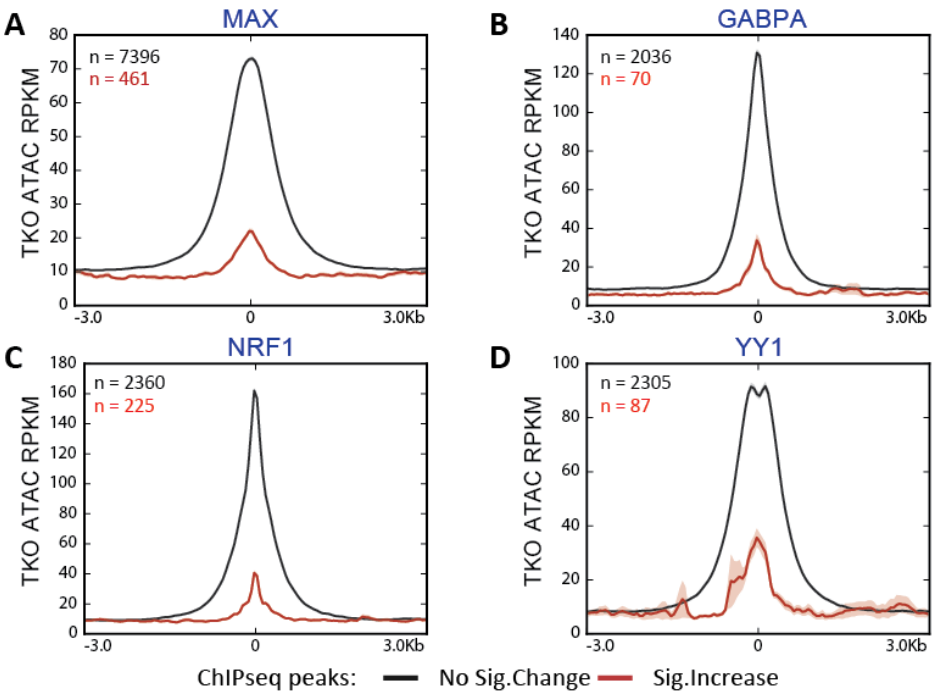


Figure 5.24: Metaplots comparing ATAC-seq signal at transcription factor binding sites with or without significantly increased occupancy in DNMT.TKO cells. For each transcription factor, ChIP-seq peaks were separated based on their differential binding status. Black and red solid lines show mean ATAC-seq signal at common (no significant differential occupancy) and DNMT.TKO specific ChIP-seq peaks respectively. ATAC-seq signal from DMSO-treated DNMT.TKO cells (merged replicas, RPKM normalisation) was averaged in 50bp windows across *n* number of regions. Light shading indicates the area between the mean +/- the standard error of the mean. Position 0 indicates the centre of the peaks. Only ChIP-seq peaks that harboured a binding motif for the corresponding transcription factor (see Table 5.4) were included.

5.6.2 Relationship between transcription factor binding and gene expression

To investigate whether novel transcription factor binding events in DNA methylation deficient cells have an impact on transcription we examined the expression levels of the genes closest to these peaks. Only novel GABPA peaks were associated with a significant increase in the expression of the closest genes when comparing RNA-seq data from DNMT.TKO versus WT cells (p-value = 3.771e-11, one-tailed Wilcoxon matched-pairs signed rank test) (Figure 5.25). However only five of the 48 genes closest to a DNMT.TKO-specific GABPA binding site were significantly more expressed in DNA methylation deficient cells (*Fkbp6*, *Gpat2*, *Hus1b*, *Sema5b* and *1700019B21Rik*). Of of the top 100 genes most significantly upregulated in DNMT.TKO compared to WT cells, 16 could be associated with the *de novo* binding of one of the seven transcription factors. (Table 5.6). Figures 5.26A and 5.26B present two UCSC genome browser snapshots showing the binding of NRF1 and MAX/GABPA to the *Gm595* and *Fkbp6* promoters respectively, two genes whose expression is specific to DNMT-TKO cells.

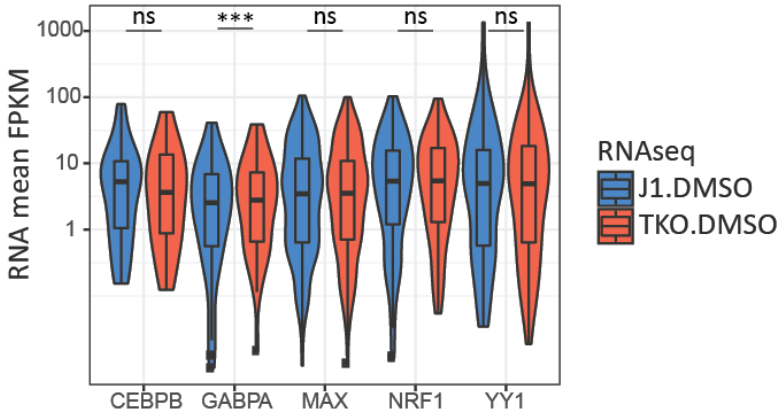


Figure 5.25: Quantitation of gene expression for genes nearest to the ChIP-seq peaks that are preferentially occupied in DNMT.TKO cells. The mean RNA-seq FPKM value obtained across both wild-type (J1.DMSO) or DNMT.TKO (TKO.DMSO) biological replicates was used to measure gene expression. Only ChIP-seq peaks that harboured a binding motif for the corresponding transcription factor (see Table 5.4) were included. Two-tailed Wilcoxon matched-pairs signed rank test were used to determine significant differences in the distributions of gene expression levels: *** p-value < 0.001 ; *ns* = non-significant (p-value >0.05). 25, 48, 356, 162 and 60 genes were analysed for CEBPB, GABPA, MAX, NRF1 and YY1 respectively.

	RNA-seq					
Gene	**J1 DMSO RPKM**	**TKO DMSO RPKM**	**Fold Change**	**FDR**	Novel Transcription factor binding site	Distance from peak to TSS (bp)
1700019B21Rik	0.000	1.647	708.32	4.04E-11	GABPA	404,659
Gm595	0.012	1.069	256.43	2.47E-18	NRF1	482
Asz1	0.234	7.090	125.42	5.20E-41	MAX and NRF1	61
Gm6329	0.033	1.344	66.82	3.94E-09	MAX and NRF1	255
4933425L06Rik	0.036	0.483	63.63	5.46E-06	NRF1	-52
Nlrp4c	0.501	4.500	49.83	7.67E-21	MAX	29,319
Tktl2	0.442	3.494	49.79	1.71E-20	MAX and NRF1	305
Nxf2	0.092	2.451	25.30	1.25E-28	MAX	14
Gpat2	1.580	11.131	13.855	1.69E-18	GABPA	96
Ctcfl	0.302	3.460	12.77	6.03E-28	MAX	120
Slc1a1	0.887	5.521	10.85	8.46E-15	NRF1	43,464
Fkbp6	1.573	28.527	8.68	1.27E-65	GABPA and MAX	44
Kcnt1	0.053	0.395	6.52	2.01E-09	MAX	23,104
Gpr146	0.052	0.418	6.48	1.22E-07	MAX	3,051
Hus1b	0.956	6.278	5.28	5.75E-16	GABPA	14
Zfp112	0.235	2.022	2.11	6.26E-19	NRF1	5,870

Table 5.6: Genes belonging to the top 100 most significantly upregulated in DNMT.TKO (TKO) versus wild type J1 cells and that are located in proximity to a TKO-specific transcription binding site.

These are evidently examples of a rare phenomenon, in most cases the binding of these transcription factors does not associate with transcriptional changes of a nearby gene. Nevertheless, the numbers we have reported are likely to be a slight under-estimation. Indeed, through visual inspection of the data, we have identified a handful of loci which appear to be differentially occupied by TFs but are not recognised as such with our peak calling and differential analysis parameters. For example, several promoters within the *MageA* cluster of genes show DNMT.TKO specific GABPA enrichment and gene expression although no ChIP-seq peaks were called here using our methodology (Figure 5.26C).

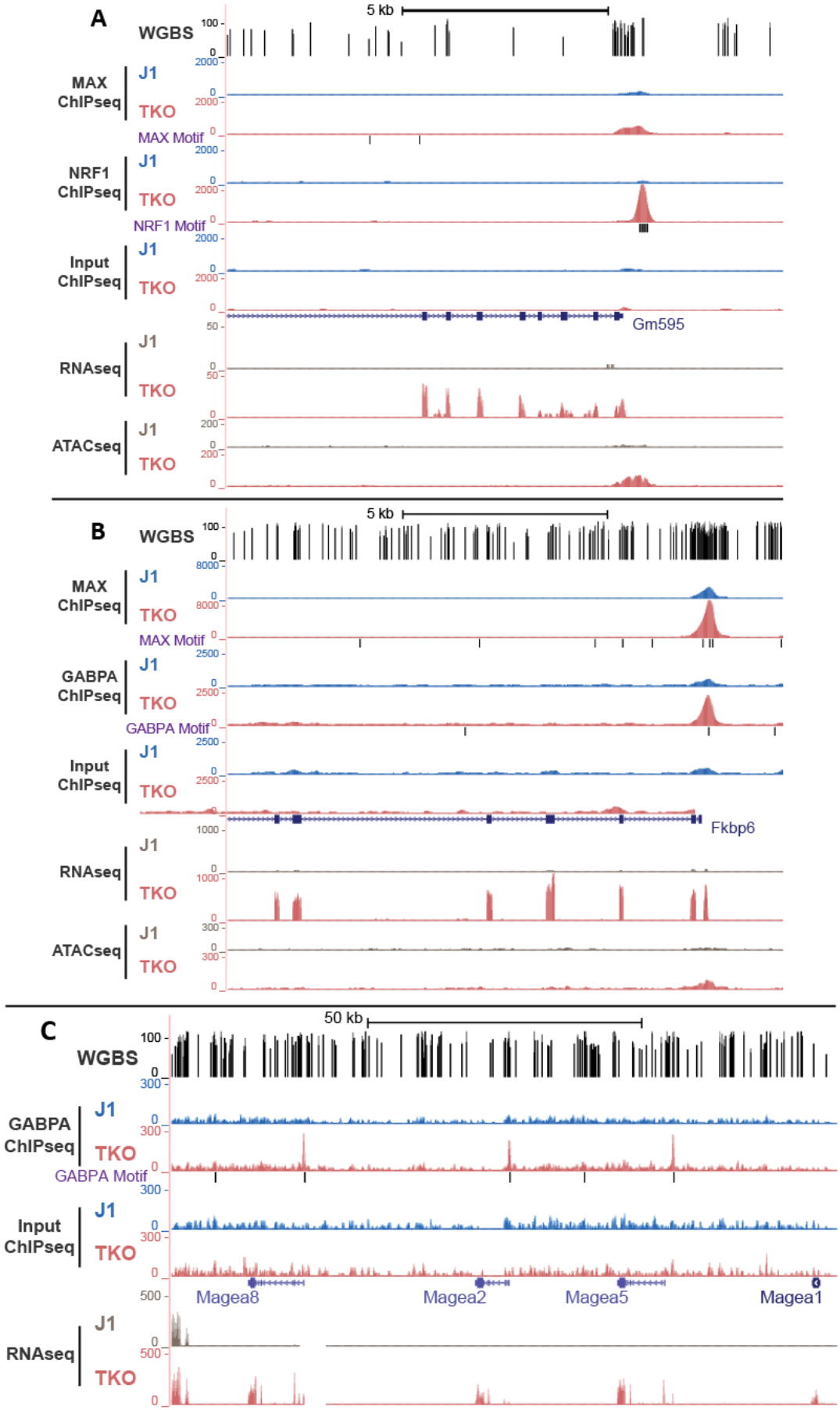


Figure legend on next page

Figure 5.26: UCSC genome browser snapshots showing examples of genes whose DNMT.TKO-specific expression is associated with a novel transcription factor binding event in the absence of DNA methylation. ChIP-seq signal for wild type (J1) and DNMT.TKO cells is displayed for MAX, NRF1 or GABPA transcription factors as well as Input controls. Wild type CpG methylation obtained using whole genome bisulfite sequencing (WGBS) (Habibi et al., 2013) is shown as well as RNA-seq and ATAC-seq signal from DMSO-treated wild-type cells. Also indicated are the positions of sequences matching the GABPA, MAX or NRF1 PWM motifs and UCSC gene annotations. ChIP-seq, RNA-seq or ATAC-seq coverage graphs were generated after merging alignments from replicate samples and was normalised to library size. For the RNA-seq, only reads originating from the sense strand are shown. Mm10 coordinates: **A**: chrX:48,868,107-48,881,594 ; **B**: chr5:135,338,336-135,351,856 ; **C:** chrX:154,971,779-155,093,112 .

5.7 Transcription factor binding at transposable elements in WT and DNMT.TKO cells

Analysis of ATAC-seq from DNMT.TKO and wild-type mouse ES cells revealed that some transposable elements display increased sensitivity to Tn5 transposition after loss of DNA methylation, particularly ERVK type retrotransposons. While only a few of these elements are transcribed in DNMT.TKO cells, they are more prone to transcriptional activation by HDAC inhibition (see Chapters 3 and 4). Using our ChIP-seq data, we attempted to discern whether the spurious binding of transcription factors was involved in the de-repression of these transposable elements.

Some of the peaks bound by GABPA, MAX, NRF1 or YY1 that were identified as DNMT.TKO specific were found to overlap transposable elements (Figure 5.27). Notably, 78.2% of DNMT.TKO-specific YY1 peaks overlapped with a LTR type element. In addition 84 DNMT.TKO-specific MAX peaks (16.6%), 37 TKO-specific NRF1 (16.3%) and 11 GABPA (14.5%) peaks also mapped to retrotransposons while a few binding sites for each factor overlapped with a LINE or SINE element (23.7%, 8.7%, 7.9% and 6.5% of GABPA, MAX, NRF1 and YY1 DNMT-TKO peaks respectively). Table 5.7 lists the LTR elements most frequently bound by GABPA, MAX, NRF1 or YY1 most of which are ERVK type elements. Fifteen out of the 21 retrotransposons

for which we have site-specific evidence of MAX, NRF1 or YY1 binding show significantly increased ATAC-seq signal in DNMT.TKO cells (column with green-coloured cells).

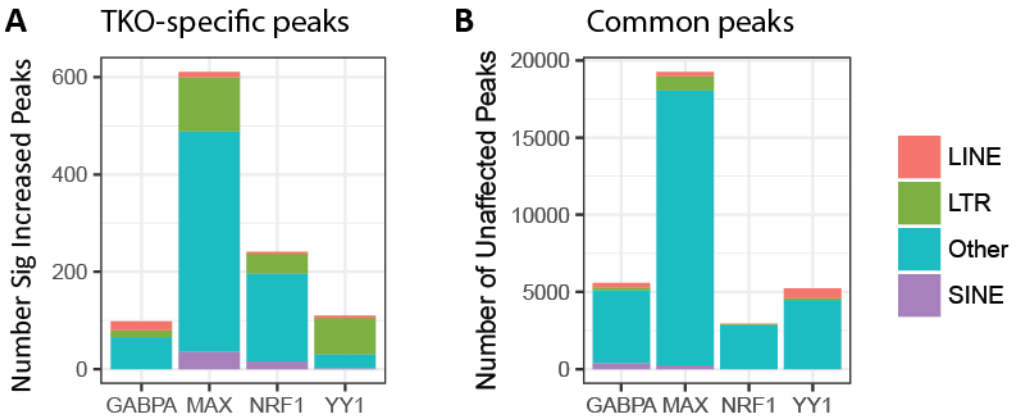


Figure 5.27: Bar charts indicate the number of ChIP-seq peaks that overlap a transposable element feature. A. GABPA, MAX, NRF1 or YY1 ChIP-seq peaks that show significantly increased occupancy in DNMT.TKO compared to wild type cells. **B.** GABPA, MAX, NRF1 or YY1 ChIP-seq peaks that show no significant differential occupancy. Only ChIP-seq peaks that harboured a binding motif for the corresponding transcription factor (see Table 5.4) were included.

For every LTR element type in the RepBase collection, we calculated the fraction of copies that contained a potential binding site for our set of transcription factors based on the motifs shown in Table 5.4 (Figure 5.28). Few LTR elements harbour CEBPB or GABPA motifs and for those that do; this is the case at a minority of copies only. On the other hand, the E-box bound by MAX, the NRF1 motif and the YY1 motif are highly conserved amongst all copies of a subset of retrotransposons elements. For example, 97% and 98% of the 1492 IAPLTR1_Mm copies contain a NRF1 and YY1 motif respectively. 94% of IAPLTR2a2_Mm repeats have a MAX element. This data is supportive of a potential role for MAX, NRF1 or YY1 in regulating the transcription of these recently integrated and conserved ERVK transposable elements. For the following analysis, we focused on these three transcription factors.

LTR subtype				ChIP-seq peaks overlapping a transposable element				% copies harbouring a TF binding sites			Significant increase in DNMT.TKO versus wild-type (Differential analysis)				
Name	Family	Estimated Repeat age (million years)	# Genomic Copies	GABPA	MAX	NRF1	YY1	MAX	NRF1	YY1	ATAC	MAX	NRF1	YY1	RNA
RMER19B	ERVK	80.2	5640	0	26	0	0	58.6	0.2	1.1					x
RLTR27	ERVK	93.6	1621	0	8	11	2	0.7	13.4	27.1					
IAPEY2_LTR	ERVK	30.5	1062	0	6	1	12	16.9	39.9	68.4					x
RLTR44C	ERVK	44.3	198	0	4	10	0	47.5	36.9	0.5					
IAPEY_LTR	ERVK	42.6	1410	0	1	0	12	39.1	0.7	66.3					x
IAPEY3_LTR	ERVK	44.5	553	0	0	0	12	3.4	49.0	59.7					x
IAPLTR2b	ERVK	27.2	1017	0	4	0	6	85.4	77.1	76.0					
RLTR44B	ERVK	72.3	275	0	6	3	0	61.5	8.4	30.2					x
RLTR10F	ERVK	34.1	482	0	0	0	8	22.4	65.4	45.9					
RLTR44E	ERVK	48.4	122	0	4	4	0	54.1	36.1	0.8					
ORR1E	ERVL-MaLR	129.0	22742	6	1	0	0	11.2	0.1	3.6					x
RLTR44A	ERVK	65.7	259	0	7	0	0	57.9	8.9	43.2					x
RLTR46	ERVK	59.9	46	0	1	0	4	6.5	69.6	65.2					
MTE2a	ERVL-MaLR	119.8	15148	0	2	2	0	6.0	0.4	2.1					
MTEa	ERVL-MaLR	122.8	20061	1	3	0	0	7.8	0.2	3.4					
ORR1D2	ERVL-MaLR	114.5	15061	3	1	0	0	7.3	0.1	4.3					x
RLTR10E	ERVK	50.0	227	0	0	0	4	7.0	64.8	34.4					
IAPEz-int	ERVK	27.1	7319	0	0	3	0	19.2	26.2	21.8					x
IAPLTR2a2_Mm	ERVK	24.3	1541	0	3	0	0	94.9	15.5	69.4					x
ORR1F	ERVL-MaLR	130.7	23086	0	3	0	0	12.2	0.1	2.7					x
RLTR51B_Mm	ERVK?	98.4	662	0	1	2	0	20.2	0.8	6.5					

Table 5.7: LTR elements for which at least three copies overlap with a DNMT.TKO-specific transcription factor binding event. A coloured cell in one of the five right-most columns is indicative of a significant change in ATAC-seq (green), MAX, NRF1, YY1 ChIP-seq (blue) or RNA-seq (orange) signal across all copies of a LTR element. The presence of an "X" in the RNA-seq column is indicative of a significant increase in signal in TSA versus mock treated DNMT.TKO cells. Changes were deemed significant with an *FDR* < 0.05 using the R package DESeq2 (Love et al., 2014).

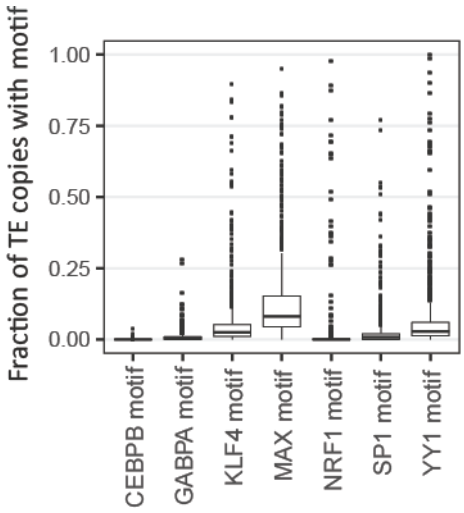


Figure 5.28: Boxplots representing for each LTR transposable element type, the fraction of copies that harbour a CEBPB, GABPA, KLF4, MAX, NRF1, SP1 or YY1 binding motif.

Using the same analysis pipeline that was used to characterise differential mRNA expression and accessibility of repeat elements, we identified transposable elements at which there was a differential enrichment for MAX, NRF1 or YY1 ChIP-seq reads between wild-type and DNMT.TKO cells. Few retrotransposon types showed a significant increase in ChIP-seq signal for these transcription factors in DNMT.TKO cells (Figure 5.29 and Table 5.7). Repeat elements for which most copies harbour an NRF1, MAX or YY1 motif do not have significantly increased signal for these transcription factors (Figure 5.29).

Further information about the importance of these transcription factors in mediating the chromatin accessibility and increased transcriptional potential for ERVK retrotransposons could be obtained by comparing protein enrichment at TE copies that do or do not contain a TF binding site. We would presume that only copies harbouring an appropriate binding site would be bound and display increased accessibility and possibly transcriptional activity. However, due to the difficulty in mapping the precise origin of sequencing reads amongst all copies of these highly conserved repetitive elements, this type of analysis was not attempted.

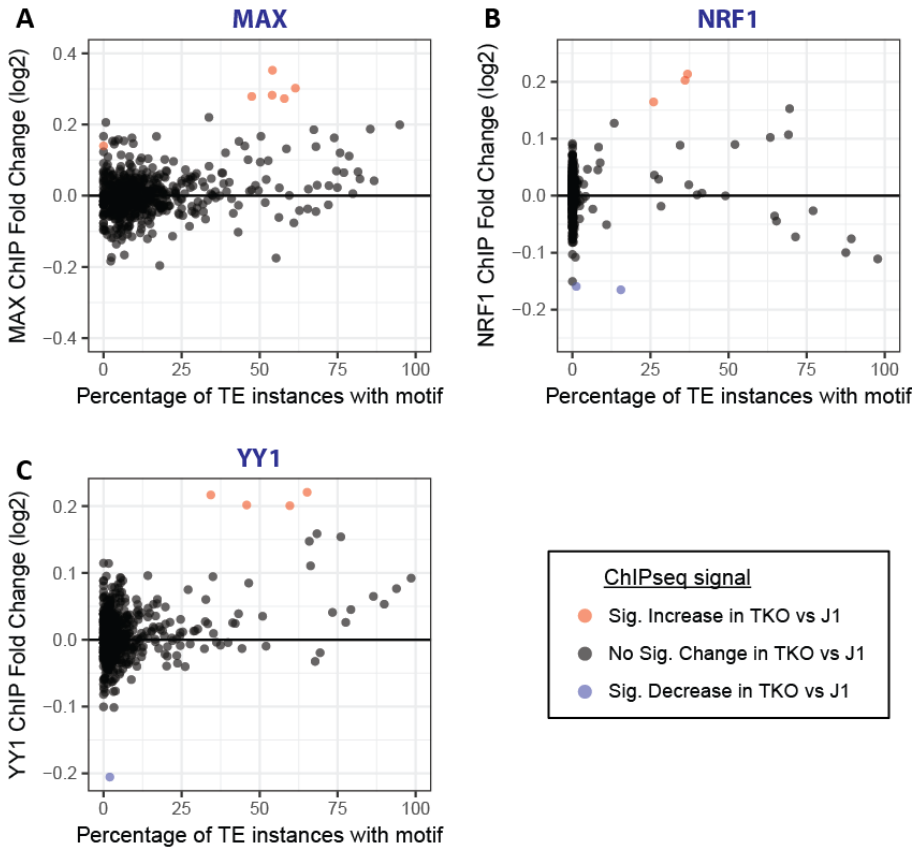


Figure 5.29: For each LTR transposable element type, the change in overall TF occupancy was compared against the percentage of LTR copies that harbour a binding motif. Transcription factor occupancy was measured as the number of ChIP-seq reads that map to one or more of the LTR instances located across the genome. Changes were deemed significant with an *FDR* < 0.05 using the R package DESeq2 (Love et al., 2014)

The 10 ERVK subtypes that, on average, are associated with a significant increase in MAX, NRF1 or YY1 occupancy in DNMT.TKO cells also display increased chromatin accessibility (Table 5.7). Three were also more highly transcribed in DNA methylation deficient cells. These preliminary observations based on correlative data suggest that DNA methylation may participate in keeping young repeat elements repressed by directly inhibiting the binding of certain TFs.

5.8 Discussion

Summarising experimental evidence collected over the last 30 years, DNA methylation has been described as broadly having three different effects on the binding of transcription factors to CpG-containing sequences (Harrington et al., 1988; Höller et al., 1988; Iguchi-Ariga and Schaffner, 1989; Prendergast and Ziff, 1991; Macleod et al., 1994; Yoon et al., 2003; Quenneville et al., 2011). The binding of a first group of transcription factors to their cognate recognition sequences has been found to be largely unaffected by cytosine methylation (Harrington et al., 1988; Höller et al., 1988; Feldmann et al., 2013). A second category of transcription factors show reduced affinity to CpG containing sequences when these are methylated (Iguchi-Ariga and Schaffner, 1989; Prendergast and Ziff, 1991) while the opposite was found to be the case for a third group of DNA-binding proteins: for these, cytosine methylation appeared to be required for the interaction with certain CpG containing sequences (Quenneville et al., 2011; Hu et al., 2013).

In recent years, several studies have made use of comprehensive oligonucleotide and transcription factor libraries to investigate the relative affinity of many transcription factors for methylated or unmethylated DNA sequences *in vitro* (Hu et al., 2013; Spruijt et al., 2013; Kribelbauer et al., 2017; Yin et al., 2017). This has enabled the classification of most vertebrate transcription factors into one of the groups described above.

To further investigate DNA methylation's contribution in shaping the genome-wide occupancy of transcription factors *in vivo*, we generated ChIP-seq data for seven candidate transcription factors in DNMT.TKO mutant mES cells and corresponding wild-type cells. In line with the findings from *in vitro* studies, we found that the absence of DNA methylation had three different effects on the localisation of these transcription factors. The genomic distribution of KLF4 and SP1 was mostly unchanged between wild type and DNMT.TKO cells.

On the other hand, DNA methylation loss was associated with an increase in the genomic occupancy of GABPA, MAX, NRF1 and YY1 while it was linked to a reduction in the localisation of CEBPB in the genome. In the following sections, we will discuss the results obtained for each transcription factor in more depth in and in light of previous knowledge.

5.8.1 Genomic occupancy of SP1 and KLF4 appeared mostly unaffected by the absence of DNA methylation.

SP1 is a member of the SP/XKLF family of transcription factor characterised by possessing a DNA-binding domain made up of three C_2H_2-type zinc fingers (Philipsen and Suske, 1999). SP1 is ubiquitously expressed and binds the GC box motif (5'-CCCCGCCC-3', see Table 4.5) commonly found at the promoters of housekeeping genes. It functions as a transactivator acting through homo-oligomerisation and interaction with the general transcription factors of the pre-initiation complex (Wierstra, 2008). Early on it was recognised that SP1 binding to DNA probes containing GC boxes was unaffected by *in vitro* DNA methylation (Harrington et al., 1988; Höller et al., 1988). Furthermore, the integration of intact SP1 binding sites into the genome was demonstrated to promote local demethylation and a role for SP1 was suggested in protecting CpG islands from methylation (Brandeis et al., 1994; Macleod et al., 1994; Siegfried et al., 1999). The *Sp1* gene is localised to chromosome 15 and therefore present at a higher copy number in J1 compared to DNMT.TKO cells. This raises the possibility that SP1 abundance is relatively reduced in DNMT.TKO cells which could explain the generally lower enrichment that is observed at SP1 ChIP-seq peaks in DNMT.TKO cells (Figure 5.4). Nevertheless, our analysis suggests that there are very few significant differences in SP1 occupancy in the two cell lines. The absence of novel SP1 binding sites in DNMT.TKO compared to wild-type cells therefore supports the view that DNA methylation itself does not act as a barrier to SP1 binding in mouse embryonic stem cells, in line with previous *in vitro* work (Harrington et al., 1988; Höller et al., 1988).

However, an alternative explanation for our data is that all potential genomic sites at which stable SP1 binding can occur are previously occupied and unmethylated in wild-type cells and therefore unaffected by DNMT knockout. As a result we cannot dismiss the possibility that aberrant methylation may displace SP1 protein binding. In fact, a few studies have identified specific genes whose transcriptional activation by SP1 is repressed in human cancer cell lines by DNA methylation of GC box elements in their promoter regions. EMSA experiments suggested that SP1 binding to CpG island sequences was abrogated by DNA methylation and 5-aza-2'-deoxycytdine treatment could restore SP1 binding *in vivo* (Guo et al., 2012; Tian et al., 2015). Experiments in which GC box elements are targeted for re-methylation in DNMT.TKO mESCs could provide insight concerning the ability for DNA methylation to disrupt SP1 binding *in vivo*.

The DNA binding domain of KLF4 bears structural similarities to that of SP1 and primarily recognises and binds a sequence similar to the GC box (5'-GGGGTGGGGC-3', also called a GT box, see Table 4.5) that does not contain a CpG dinucleotide (Philipsen and Suske, 1999). Nevertheless, using screening methods aimed in identifying proteins whose affinity to oligonucleotides was sensitive to CpG methylation, two independent studies reported that, in addition to its canonical motif, KLF4 could bind the CpG containing motif 5'-CpGCCC-3' in a methylation-dependent manner (Hu et al., 2013; Spruijt et al., 2013). Binding to endogenous CCmCpGCC sites in human H1 ESCs was confirmed by ChIP coupled with bisulfite sequencing (Hu et al., 2013). Furthermore, a KLF4 mutant that has abrogated binding to methylated but not unmethylated sequences failed to transactivate genes involved in the mobility of glioblastoma cells suggesting that these interactions have biological significance (Wan et al., 2017). The ability for KLF4 to bind methylated sequences is likely to facilitate its ability to reprogram somatic cells into induced pluripotent stem cells in conjunction with other Yamanaka factors (Takahashi and Yamanaka, 2006). In our data we found little evidence that DNA methylation was contributing to KLF4 binding in mESCs as very few sites showed

significantly reduced binding in DNMT.TKO cells (Figure 5.4) none of which contained a CpGCC motif that was methylated in wild type cells. This finding may be confounded by the observation that DNMT.TKO cells have elevated expression of *Klf4* (Table 5.3), which may allow KLF4 to bind despite the absence of DNA methylation. In fact, many KLF4 binding sites displayed increased enrichment in DNMT.TKO compared to wild type cells although only rarely were changes in methylation found to be responsible for increased occupancy (an example is shown in Figure 5.4J). Together our data does not support a major role for DNA methylation in shaping KLF4 occupancy genome wide in mouse embryonic stem cells.

5.8.2 CEBPB displays reduced genomic occupancy in DNMT.TKO cells

In contrast to the other transcription factors whose genomic occupancy we have profiled in wild-type and DNMT.TKO cells, CEBPB binding was decreased at a number of loci in the absence of DNA methylation. Previous studies have reported that CEBPB has greater affinity for methylated versus unmethylated 5'TTGCpGCAA-3' sequences (Mann et al., 2013; Yin et al., 2017). In wild type mES cells, we identified CEBPB ChIP-seq peaks that harboured both unmethylated and methylated versions of this sequence suggesting that CEBPB can bind both *in vivo*. Although CEBPB binding to several of the methylated motifs was significantly reduced in DNMT.TKO cells it remains difficult to assess whether loss of DNA methylation is responsible for this. Indeed, our interpretation is confounded by the fact that all CEBPB ChIP-seq peaks display reduced enrichment in DNMT.TKO cells, most likely due to reduced *Cebpb* gene expression. In addition the small number of high-confidence CEBPB binding sites we detect in mouse embryonic stem cells (< 200 ChIP-seq peaks were identified that contained the CEBPB cognate binding motif) limits the power of our analysis.

5.8.3 DNA methylation loss was associated with an increase in the genomic occupancy of GABPA, MAX, NRF1 and YY1.

Through our ChIP-seq experiments we have shown that in the absence of DNA methylation, there is an increase in the number of binding sites for the transcription factors GABPA, MAX, NRF1, and YY1 (Figure 5.4). Our analyses indicate that the loci that are preferentially bound by these transcription factors in DNMT.TKO versus wild-type cells harbour sequence motifs which they are known to bind with high affinity (Figures 5.15 and 5.16). In addition, these motifs display high levels of DNA methylation in wild-type cells (Figure 5.13 and 5.14) suggesting that DNA methylation is directly impeding the interaction between these factors and their cognate motifs under wild-type conditions. The motifs preferentially bound in DNMT.TKO cells were in regions with relatively low CpG contents (Figure 5.18) arguing against a role for MBD proteins in the displacement of these transcription factors in wild-type cells. I will discuss the results obtained for each of these four transcription factors with a consideration for published studies that have examined their genomic localisation.

5.8.3.1 GABPA

GABPA belongs to the ETS family of transcription factors that contain a conserved Ets winged helix-turn-helix DNA binding motif and broadly recognise 5'-(C)GGA(A/T)-3' ETS motifs (Rosmarin et al., 2004). GABPA is essential for mESC viability and early embryogenesis in mice and is required for the transcriptional activation of genes involved in the mitochondrial respiratory chain and cell cycle progression (Ristevski et al., 2004; Ueda et al., 2017). Mutations that disrupt or create novel GABPA binding sites have been reported to cause aberrant silencing of tumour suppressor genes or oncogene over-expression respectively (Huang et al., 2013; Zhang et al., 2017). Careful regulation of GABPA binding is therefore likely to be important to maintain appropriate rates of cell proliferation. DNA methylation may contribute to this regulation.

The methylation of CpG sites within the ETS motif has been demonstrated to disrupt GABPA binding to oligomers *in vitro* and at several gene promoters *in vivo* (Nickel et al., 1995; Yokomori et al., 1995; Lucas et al., 2009). Although we did not find the ETS motif to be enriched at DNMT.TKO-specific ATAC-seq peaks (see Tables 4.2 and 4.3 in Chapter 4), previous studies have reported their enrichment at novel DHS sites formed in an independent DNMT.TKO mES cell line (Domcke et al., 2015) or at transcription initiation sites formed in hypomethylated *Dnmt3b*$^{-/-}$ mESCs (Neri et al., 2017) suggesting that DNA methylation could be responsible for restricting GABPA binding throughout the genome and important for limiting aberrant transcriptional events. Our ChIP-seq data provides evidence that DNA methylation does indeed protect certain genomic loci containing ETS motifs from GABPA binding and transactivation (Figures 5.12, 5.13 and 5.26, Table 5.6) although such events are rare and represent only a small fraction of all ETS sites.

We found that a third of GABPA binding sites had multiple ETS motifs (Figure 5.17B). Most GABPA ChIP-seq peaks in the Jurkat human cell line were also found to contain two ETS motifs (Valouev et al., 2008) implying this to be a common feature of GABPA interactions with DNA. Furthermore, the multiple ETS motifs were frequently positioned in the same orientation and on the same strand, including at sites uniquely bound in DNMT.TKO cells (Figure 5.19B). This is likely a reflection of the manner in which GABPA interacts with DNA. Two GABPA proteins associate with two GABPB proteins to form a $GABPA_2B_2$ heterotetramer that binds DNA. As GABPB does not possess a DNA binding module, it is the two ETS domains that are responsible for sequence-selectivity (Batchelor et al., 1998). Moreover, $GABPA_2B_2$ was reported to bind preferentially to sequences with two elements separated by 2 to 3 helical turns (~20-30bp) and aligned on the same face of the DNA helix. Positioning of the motifs on opposite sides of the DNA helix weakened the stability of the $GABPA_2B_2$/DNA complex and its transcriptional potency (Yu et al., 1997). The importance of the spacing and orientation of ETS motifs is

therefore likely to contribute significantly to the selective binding of GABPA to only a small fraction of ETS motifs even in hypomethylated cells.

Our data also revealed enrichment for GABPA at gene bodies, especially those with high transcriptional activity (Figures 5.9 and 5.10). To our knowledge this has not been observed previously; independent GABPA ChIP-seq experiments that have made use of the same primary antibody do not display such broad domains of enrichment (Birney et al., 2007; Hollenhorst et al., 2011; Perdomo-Sabogal et al., 2016). Although these intriguing results could be indicative of an association between GABPA and the transcriptional machinery in mESC cells, they could also be explained by unspecific binding of the antibody. Experiments using independent antibodies or different approaches should be carried out to explore this observation further.

5.8.3.2 MAX

MAX is a member of the basic helix-loop-helix leucine zipper (bHLHZ) family of transcription factors. It is most well characterised as the obligate binding partner for MYC, encoded by the proto-oncogene *c-myc*, with which it binds the palyndromic E-box motif (5'-CACGTG-3') (Blackwood et al., 1992; Sabò and Amati, 2014). Unlike MYC, MAX can also homodimerise or form heterodimers with other bHLHZ proteins meaning that MAX ChIP-seq peaks are likely representative of both MYC-independent as well as MYC-dependent binding sites.

According to our data from both cell lines, MAX occupancy is widespread throughout the genome although over half of the identified binding sites do not harbour an E-box motif (Tables 5.4 and 5.5). This is comparable with the results obtained by previous efforts to map Myc occupancy genome-wide, which showed that a vast proportion of Myc/Max binding sites lacked an E-box motif (Guo et al., 2014). In fact Myc protein occupancy has been documented to extend to almost all open or accessible DNA regions with little selectivity for the underlying DNA sequence particularly when it is overexpressed in cancer cells, a phenotype described as

Myc "invasion" (Lin et al., 2012; Sabò and Amati, 2014; Sabò et al., 2014; Allevato et al., 2017). Nevertheless, when present at low levels, Myc preferentially localises at the promoters of genes containing E-box motifs and is directly responsible for their transcriptional activation (Sabò et al., 2014). Considering these observations, it has been suggested that low affinity interactions between Myc/Max and Pol2 or other general transcription factors facilitate Myc/Max localisation at promoters while high-affinity binding to E-box motifs allow for the stabilisation of protein/DNA complexes at target genes (Guo et al., 2014; Sabò et al., 2014). The implication for our data is that, while the presence of an E-box motif underlying a MAX peak is likely to be indicative of a direct protein-DNA interaction we cannot dismiss the possibility that MAX binding is secondary to an increase in accessibility or the presence of a protein with affinity to Myc.

Most novel MAX binding sites in DNMT.TKO cells contain strong E-box motifs (Figures 5.15 and 5.16) suggesting that they represent direct binding of MAX to DNA as opposed to regions of Myc/MAX invasion. Methylation-sensitive binding of Myc or the related N-Myc factor to oligos *in vitro* has been reported (Prendergast and Ziff, 1991; Perini et al., 2005) but only this year was it shown that MAX binding to E-box motifs was significantly reduced by CpG methylation (Yin et al., 2017). Our finding that DNMT.TKO cells display increased occupancy of MAX at previously methylated E-box motifs suggests that DNA methylation can also directly interfere with MAX binding *in vivo* in wild-type mESCs. Interestingly, a number of MAX ChIP-seq peaks that are constant between the two cell lines contain methylated E-box motifs in the underlying sequence (Figure 5.13). Although this seems to disagree with the proposed negative effect of DNA methylation on MAX binding, a likely explanation is that protein-protein interactions with other DNA-binding factors drive MAX occupancy at these sites.

It remains unclear whether MAX occupancy at DNMT.TKO specific ChIP-seq peaks is dependent on its interaction with Myc or other bHLHZ partners. This could be determined by

profiling the genome-wide occupancy of Myc and other MAX binding partners in DNMT.TKO cells. Interestingly, the MAX/MGA heterodimer has been reported to participate in the localisation of the non-canonical PCGF6-PRC1 complex to the promoters of germ-cell related genes and contribute to their silencing in mESCs (Endoh et al., 2017). This opens the possibility that novel binding of MAX/MGA to the hypomethylated promoters of germ-cell related genes in DNMT.TKO could maintain their repression in the absence of DNA methylation. However there is little evidence for this hypothesis in our data; on the contrary we identify several germ-cell specific genes at which increased MAX occupancy is associated with transcriptional activation (*Asz1, Fkbp6, Nlrp4c, Nxf2*, *Tktl2,* see Table 5.6).

5.8.3.3 NRF1

NRF1 is a transcriptional activator that binds DNA as a homodimer and is involved in the expression of a wide range of genes involved in mitochondrial biogenesis (Virbasius et al., 1993; Gugneja and Scarpulla, 1997; Satoh et al., 2013). The NRF1 cognate recognition motif was identified as enriched in DNMT.TKO-specific Tn5-transposase accessible regions (Chapter 4 - Tables 4.2 and 4.3) and ChIP-seq confirmed NRF1 occupancy was increased in DNMT.TKO cells at sites that were previously methylated in WT cells (Figures 5.13 and 5.14). The impact of DNA methylation loss on the localisation of NRF1 in mESCs has been determined independently (Domcke et al., 2015). In line with our findings, the NRF1 motif was identified as being enriched in sequences underlying DHS sites specific to DNMT-depleted cells. Its occupancy at these regions was confirmed by ChIP-seq. Additional experiments that were performed by Domcke *et al.* provided further insight into the competition between DNA methylation and NRF1 binding. For instance, the culturing of mES cells in 2i-containing media followed by a transition to serum-containing media lead to *de novo* methylation of NRF1 binding sites and caused eviction of the NRF1 protein (Domcke et al., 2015).

NRF1 has been characterised as being a pioneer transcription factor, able to bind to its cognate sequence in heterochromatinised DNA and form a DNaseI hypersensitive site (Sherwood et al., 2014). Our data, together with the work published by Domcke *et al.* suggest that DNA methylation remains a barrier to NRF1 binding. Thus, NRF1 binding can still be regulated through epigenetic modifications. The comparative enrichment of NRF1 but not GABPA or MAX binding motifs at DNMT.TKO specific ATAC-seq peaks may reflect its ability to act as a pioneer transcription factor as opposed to the other factors that may not be able to bind sequences within non-permissive chromatin. Notably, MYC was reported to be a settler transcription factor unable to bind heterochromatinised DNA (unlike the other Yamanaka reprogramming factors (OCT4, SOX2, KLF4)) (Soufi et al., 2015). In our study, we have identified a number of NRF1 binding sites that co-localise with MAX ChIP-seq peaks. One implication of these findings is that NRF1 could facilitate MAX binding to a subset of E-box elements in the absence of DNA methylation. To pursue this idea, it may be necessary to measure MAX binding in cells depleted of NRF1 or in which NRF1 binding to co-occupied loci is compromised (achieved through engineering precise DNA methylation or genetic modifications).

5.8.3.4 YY1

YY1 is a ubiquitously expressed transcription factor of the GLI-Krüppel family of zinc-finger transcription factors and is essential for mouse development (reviewed by (Gordon et al., 2005). It's name Yin-Yang 1 derives from the fact that it displays both repressive or activating transcriptional effects that can be modulated by the presence of the E1A adenovirus derived co-activator (Shi et al., 1991). In mESCs YY1 was suggested to act predominantly as a transcriptional activator as it localised to active enhancers and promoters in a manner that was facilitated by its low-affinity interaction with RNA (Sigova et al., 2015). Our data also showed that YY1 preferentially bound to promoters (Figure 5.5) though we did not analyse

the association between YY1 binding and transcriptional activity systematically and cannot provide evidence for its role as a co-activator.

Comparing ChIP-seq data from wild type and DNMT.TKO cells, we found that loss of DNA methylation promotes YY1 binding to a subset of LTR retrotransposons that harbour a specific CpG-containing YY1 motif (Figures 5.13, 5.14, 5.15, 5.27). Previously, *in vitro* methylation of constructs that harboured IAP LTR elements positioned upstream of a reporter gene was found to disrupt YY1 binding and transactivation potential upon transfection into F9 embryonic carcinoma cells (Satyamoorthy et al., 1993). DNA methylation at IAP LTR elements was consequently suggested to protect from YY1-mediated transactivation however our data only partly supports this idea. Indeed, YY1 binding is associated with increased chromatin accessibility (Figure 5.23) however these are not generally associated with increased expression (Figure 5.25 and visual inspection of RNA-seq data surrounding the ChIP-seq peaks). Of the four LTR types for which there is an overall enrichment of YY1 ChIP-seq reads in DNMT.TKO compared to wild type cells, only RLTR10F shows significantly increased mRNA abundance (Table 5.7).

On the other hand IAPEY3_LTR elements show a pronounced increase in transcription after HDAC inhibition in DNMT.TKO cells. This was also the case for IAPEY_LTR and IAPEY2_LTR elements. Overall, YY1 ChIP-seq reads mapping to these elements were not found to be significantly enriched in DNMT.TKO compared to wild type cells however, YY1 was found to bind individual copies of these LTR types (Table 5.7). As YY1-mediated repression has been reported to be dependent on HDAC activity (Yang et al., 1996; Weill et al., 2003) this opens up the possibility that YY1 is acting to maintain these IAP LTR elements in a silenced state in DNMT.TKO cells in a TSA-sensitive manner. Further analysis and experiments will be necessary to elucidate the role of YY1 as a transcriptional repressor or activator at novel binding sites in DNMT.TKO. For example the effect of *Yy1* knockdown on the transcriptional

activity of YY1-bound LTR elements could help us in determining if YY1 repression is necessary to maintain these silenced in DNMT.TKO cells.

5.8.3.5 General discussion of the significance of DNA methylation's role in the occupancy of the four methylation-sensitive transcription factors

The biological significance of the DNA methylation-sensitive binding events that we have described remains to be explored. Few of the transcription factor binding sites that appeared in DNMT.TKO cells were associated with an increase in transcription from the nearest gene (Figure 5.25). Those novel binding sites that were linked to a significant increase in gene transcription tended to localise to gene promoters (Table 5.6). These included the promoters of a subset of genes whose expression is normally restricted to the germline (*Asz1, Fkbp6, Gm595, Gpat2, Nlrp4c, Nxf2*, *Tktl2, MageA cluster, 1700019B21Rik, 4933425L06Rik*). DNA methylation was previously known to be the primary mechanism by which these genes are silenced in mouse embryonic stem cells and adult tissues (Fouse et al., 2008; Borgel et al., 2010; Auclair et al., 2014). Our results suggest that this is achieved by DNA methylation directly hindering the occupancy of ubiquitously expressed transcription factors to their cognate recognition sequences.

Notwithstanding these results, few DNA methylation-sensitive differential binding events were promoter proximal or associated with a change in gene transcriptional activity. Most DNMT.TKO specific binding sites localised distal to genic transcription start sites (Figure 5.5), in line with the changes that were observed in chromatin accessibility (Chapter 4, Domcke et al., 2015), and were not associated with transcriptional activity. Although many of these sites overlap with transposable elements and may play a role in their regulation, further analysis is necessary to establish whether the novel binding sites in other parts of the genome have regulatory function. Insight could be gained through searching WGBS and DNase-seq data from various tissues or cell types to determine whether these loci show differential

methylation or chromatin accessibility state during development, an indication that they have could possess a regulatory role. An alternative is that these binding sites are non-functional and only occupied in aberrantly hypomethylated cells.

A widespread role for DNA methylation in impeding transcription factor binding to non-functional sequences is unlikely in light of our ChIP-seq and ATAC-seq results. The absence of DNA methylation led to the creation of novel binding sites at only a minority of the motifs bound by GABPA, MAX, NRF1 and YY1. There are several explanations for this observation one of which being that many of the sequences in the genome that match the TF recognition motifs may be inherently low affinity-binding sites. For example GABPA preferentially binds two ETS motifs with specific spacing and orientation constraints (Yu et al., 1997). Many potential binding sites within the genome may not fit the optimal specifications. Curtailing a list of all bona fide binding sites to which these transcription factors are capable of occupying *in vivo* is necessary to fully appreciate the contribution of DNA methylation in restricting superfluous transcription factor binding events throughout the genome. As transcription factor occupancy is likely to vary between cell types depending on the presence of cofactors, this could be achieved by generating ChIP-seq maps for transcription factors in a variety of different tissues or cell lines. Such an approach was used by Maurano *et al.* to establish a complete list of biologically valid CTCF binding sites within the mouse genome (Maurano et al., 2015).

Of final note, we cannot completely exclude the possibility that DNA methylation plays a role in limiting rare and transient interactions between transcription factor and weak binding motifs. Our ATAC-seq and ChIP-seq experiments may not be sensitive enough to detect binding events that occur infrequently within a population of cells. Considering that cells can be propagated despite vastly reduced DNA methylation (Rhee et al., 2002; Tsumura et al., 2006; Ying et al., 2008) we can be confident that such a function for DNA methylation would not be critical for normal cellular processes.

Chapter 6:

General Discussion

DNA methylation has long been proposed to play a role in the transcriptional regulation of mammalian genomes. The canonical view is that DNA methylation at promoters, enhancers and repeat elements contributes to transcriptional repression (see Introduction Section 1.4). Two key mechanisms have been put forward to explain how DNA methylation impacts gene regulation however their contribution *in vivo* has not been fully elucidated. The first of these involves the recruitment of protein complexes containing HDAC enzymes to sites of methylated DNA via MBD-domain proteins which is thought to lead to the formation of a chromatin conformation non-permissive to transcription. In the second proposed mechanism, DNA methylation directly modulates the occupancy of sequence-specific transcription factors to their cognate binding sites in the genome.

In this thesis, we aimed to investigate the contribution of these two mechanisms in mouse embryonic stem cells, a model cell line in which DNA methylation can be completely removed without compromising cell viability (Tsumura et al., 2006). We followed two broad lines of enquiry: the first was to determine the relationship between DNA methylation and HDAC activity in their role on gene expression and chromatin accessibility. The second was to examine the impact of DNA methylation loss on the genome-wide occupancy of transcription factors.

6.1 Comparison of the repressive roles of DNA methylation and HDACs

According to certain models, DNA methylation has been suggested to promote the occupancy of MBD-domain proteins whose associated HDAC activity contributes to transcriptional repression (Nan et al., 1998; Kokura et al., 2001; Lyst et al., 2013). The preferential binding of MBD-domain proteins to mCpG-rich regions has been well documented (Meehan et al., 1989; Hendrich and Bird, 1998). In addition, several experiments have demonstrated that HDAC inhibition or disruption of the interaction between MBD proteins and their co-repressive

complexes abrogates the transcriptional repression associated with DNA methylation in reporter gene assays (Nan et al., 1998; Lyst et al., 2013). Nevertheless, at several loci, disruption of histone deacetylation enzymatic activity was found not to recapitulate a loss of DNA methylation (Cameron et al., 1999; Coffee et al., 1999; Brunmeir et al., 2010; Reichmann et al., 2012). In addition, mice where the genes encoding MBD proteins are knocked out display mild phenotypes compared to *Dnmt* knockout mice (Martín Caballero et al., 2009). This underlined the need to establish the extent to which HDAC activity mediates DNA methylation-dependent repression.

Our first objective was to examine to what extent DNA methylation relies on HDAC activity for its role as a transcriptional repressor. To this purpose, we compared the effects of DNA methylation loss and HDAC inhibition on gene expression genome-wide. We found a relatively small overlap between the genes that were significantly upregulated in response to DNA methylation loss or HDAC inhibition. Most genes whose expression increased in the absence of DNA methylation, did not show significant gains in expression upon TSA treatment (Figure 3.9). In addition, many of the genes that responded to either a perturbation in DNA methyltransferase or HDAC activity were found to be upregulated by TSA in DNMT.TKO cells (Figure 3.10). This indicated that their sensitivity to TSA was not dependent on the presence of DNA methylation. Likewise, mRNA expression from retrotransposons belonging to the ERVK and ERV1 families that were upregulated in DNA methylation deficient cells (Figure 3.15, Table 3.2 and Karimi et al., 2011) was significantly increased by TSA irrespective of the presence of DNA methylation. These results suggest that there are few genes whose repression depends solely on the recruitment of HDAC enzymes by DNA methylation.

As described in the Chapter 3, our data supports a model by which DNA methylation and HDACs represent two mainly independent repressive pathways in mESCs whose activity converges at certain elements. This is supported by our and other's observations that the

removal of DNA methylation and HDAC inhibition leads to the upregulation of mostly different sets of genes and repeat elements (Chapter 3) (Brunmeir et al., 2010; Reichmann et al., 2012). Furthermore, we found that at many common targets, the effects of DNA methylation loss and TSA treatment was additive (Figures 3.10 and 3.11). However, at IAP elements and a few single-copy genes, the DNA-dependent and HDAC-dependent pathways appear to act redundantly meaning that one mechanism is able to maintain a repressive environment in the absence of the other. This phenomenon has also been observed in non-embryonic stem cells suggesting that the mechanism involved are not restricted to mouse embryonic stem cells (Cameron et al., 1999; Coffee et al., 1999; Brocks et al., 2017).

We gained further insight into the mechanism of DNA methylation-mediated repression from our ATAC-seq data. In agreement with data published by Domcke *et al.,* we found that DNA methylation is able to protect numerous genomic regions from increases in chromatin accessibility (Figures 4.7 and 4.17, Domcke et al., 2015) Regions that gained accessibility in DNMT.TKO cells did not show increased accessibility following TSA treatment in wild-type cells indicating that this function of DNA methylation is independent of HDAC activity. These loci tended to have low CpG density while the majority of CpG islands did not show increased accessibility. This argued against a role for MBD proteins in the maintenance of chromatin inaccessibility by DNA methylation. Indeed MBD-proteins are most likely to play a role at mCpG-rich regions to which they are preferentially localised (Hendrich and Bird, 1998; Baubec et al., 2013; Hainer et al., 2016). The enrichment of several transcription factor motifs in the sequences underlying DNMT.TKO-specific ATAC-seq peaks indicates that novel binding of transcription factor binding may be involved in the formation of chromatin accessible regions in DNMT.TKO cells. In our third results chapter we presented evidence that MAX, GABPA, NRF1 and YY1 occupy novel binding sites in DNMT.TKO cells and that this is associated with gains in accessibility and occasionally gene activation.

In contrast, there is little evidence from our data that HDAC enzymes interfere with the binding of transcription factors in the same way as DNA methylation. Indeed, few individual loci display gains in accessibility upon TSA treatment in mES cells similar to those seen upon loss of DNA methylation. However, profiling of transcription factor occupancy in the absence or presence of TSA would be necessary to substantiate this claim.

From our results, it is difficult to define the direct function of HDACs on gene expression and chromatin accessibility as TSA treatment is likely to have consequences on many different cellular processes and indirect effects are likely to arise. For example, our RNA-seq analysis indicates that TSA treatment in mES cells leads to a reduction in *Pou5f1 (Oct4)* gene expression (Table 4.7). This leads us to question whether the reduction in accessibility we see at numerous loci containing OCT4 binding motifs following TSA treatment (Tables 4.5 and 4.6) is a direct effect of abrogated HDAC activity or differential occupancy of Oct4. In addition to affecting histone modifications, HDAC inhibition potentially affects the function of certain transcription factors that are known to be non-histone substrates for HDAC enzymes with unforeseen consequences (Glozak et al., 2005).

HDAC inhibition in mESCs has previously been shown to enhance the expression from genes of the *Zscan4* cluster and MERVL endogenous retroviruses (Dan et al., 2015; Walter et al., 2016). Expression of these elements is a distinguishing feature of ES cells that have entered a distinctive transcriptional state characterised by the activation of genes and retrotransposons normally restricted to pre-implantation 2-cell (2C)-stage embryos (Macfarlan et al., 2012). *Zscan4*/MERVL-expressing cells exist as a naturally-occuring rare subpopulation within mESC cell cultures. ES cells can fluctuate in an out of this 2C-state; typically 3-5% of cells express *Zscan4* at a time (Zalzman et al., 2010; Macfarlan et al., 2012). Similarly to preimplantation 2C-stage embryos, cells in the 2C-state display reduced expression of the pluripotency factors OCT4, SOX2 and NANOG and are characterised as having globally elevated levels of histone

acetylation (Macfarlan et al., 2012; Ishiuchi et al., 2015), reduced DNA methylation (Eckersley-Maslin et al., 2016) and increased chromatin mobility (as assessed by the Fluorescence Recovery After Photobleaching (FRAP) analyses of core histone proteins (Bošković et al., 2014; Ishiuchi et al., 2015). We hypotheise that TSA treatment in J1 and DNMT.TKO cells recapitulates a 2C-like transcriptional state. Indeed, upon TSA treatment in both cell lines, we observe an increase in the abundance of mRNAs originating from MERVL elements (Figures 3.16 and 3.17), a significant upregulation in the expression of *Zscan4* cluster genes (data not shown) and an increase in sensitivity to Tn5 transposition throughout the genome (Figure 4.12), potentially indicative of enhanced chromatin mobility. Systematic comparison of the transcriptome of our TSA-treated cells and that of isolated 2C-state cells (Eckersley-Maslin et al., 2016) is needed to validate this finding

Acute DNA demethylation in mES cells was previously reported to be insufficient to activate the 2C transcriptional network (Eckersley-Maslin et al., 2016). Congruently, DNMT.TKO cells do not show enhanced *Zscan4* or MERVL expression compared to wild type cells. However, the changes in gene expression and chromatin accessibility that are observed upon TSA treatment are more pronounced in DNMT.TKO cells compared to wild-type cells (Figures 3.8 and 4.8) suggesting that, in the absence of DNA methylation, ES cells are more prone to the activation of the 2C-state. Widespread DNA methylation gains occur during the developmental transition from the 2-cell embryonic stage to the post-implantation blastocyst. This is associated with a decrease in developmental potency, while 2C-cells are totipotent, cells of the inner cell mass of the blastocyst are pluripotent (ie. unable to contribute to extra-embryonic lineages) (Smith et al., 2012; Auclair et al., 2014). We propose that these changes in DNA methylation contribute to impeding the activation of totipotent transcriptional networks, together with changes to histone modifications and chromatin conformation

6.2 The role of DNA methylation on the occupancy of transcription factors genome-wide

Our second line of inquiry throughout this thesis was to determine the impact of DNA methylation on shaping transcription factor occupancy genome-wide. DNA methylation has been shown to alter the site selectivity of many transcription factors *in vitro* (Hu et al., 2013; Kribelbauer et al., 2017; Yin et al., 2017). Nevertheless, for most transcription factors there is a lack of *in vivo* evidence that the absence of DNA methylation results in differential localisation in the genome.

Our ATAC-seq data indicated that a subset of normally methylated loci significantly increase in accessibility in the absence of DNA methyltransferases likely representing novel transcription factor binding sites. The sequences underlying the loci that gained in accessibility in DNMT.TKO cells were enriched for the binding motifs of the transcription factors NRF1, NFY and YY1 for which evidence of methylation-sensitive binding exists (Satyamoorthy et al., 1993; Kumari and Usdin, 2001; Domcke et al., 2015; Hu et al., 2017). These results suggested that in the absence of DNA methylation lead to increases in the occupancy of these transcription factors.

In a parallel attempt to examine the role of DNA methylation in genome-wide transcription factor occupancy, we used ChIP-seq experiments in wild-type and DNMT.TKO cells to profile the binding landscape of seven candidate transcription factors (Chapter 5). Thus, we extended the number of transcription factors for which genome-wide occupancy has been profiled in the absence of DNA methylation (data has been published previously for the proteins CTCF, NRF1, Oct4 and N-Myc (Domcke et al., 2015; Stadler et al., 2011; Yin et al., 2017). We found that four of the transcription factors that we screened, GABPA, MAX, NRF1 and YY1, acquire novel binding sites in the absence of DNA methylation (Figures 5.12 and 5.13). Sites bound by these proteins specifically in DNMT.TKO cells harboured sequences that closely match their

respective consensus motifs (Figures 5.15 and 5.16). These results provide evidence that DNA methylation contributes to obstructing the binding of these methylation-sensitive transcription factors to a subset of their high-affinity binding sites. A subset of the novel transcription factor binding events that were observed in DNMT.TKO cells, occurred in the proximity to genes that were upregulated in the absence of DNA methylation (Table 5.6). Specifically, GABPA, MAX and NRF1 occupied the promoters of several germ-cell specific genes in DNMT.TKO but not wild-type cells. An attractive hypothesis is that these genes are maintained silent in somatic cells (Borgel et al., 2010) despite the ubiquitous expression of GABPA, MAX or NRF1 through the action of DNA methylation directly impeding their binding. Whether transcription factor binding to the promoters of these genes in the absence of DNA methylation is sufficient for their transcriptional activation requires elucidating. Reporter gene assays could be used to test this.

In recent years, a number of studies have reported that some transcription factors bind non-cognate sequences when these contain methylated CpGs (Rishi et al., 2010; Mann et al., 2013; Kribelbauer et al., 2017; Yin et al., 2017). We have profiled the genomic occupancy of two of these: CEBPB and KLF4. While KLF4 did not appear to show reduced occupancy in DNMT.TKO compared to wild-type cells, we detected significant loses in CEBPB binding in the absence of DNA methylation. However, whether or not reduced binding in DNMT.TKO cells was a consequence of DNA demethylation remains unclear. CEBPB was found to occupy both methylated and unmethylated versions of its consensus recognition motif (Figures 5.13 and 5.14). Moreover we cannot reject the possibility that decreased CEBPB binding is not due to a reduction of protein CEBPB in the nucleus.

Overall, there were few highly methylated ATAC-seq peaks which show reduced accessibility in DNMT.TKO compared to wild-type cells (Figure 4.7) suggesting that the role of DNA methylation in allowing the binding of transcription factors in ES cells is limited.

Although comparing DNMT.TKO mutant cells to their corresponding wild type cells has provided us with insights into the impact of DNA methylation loss on transcription factor binding, this approach does have its limitations. Indeed, in the absence of DNA methylation, it is possible that DNMT.TKO cells have adapted in a way to affect transcription factor binding. Redistribution of H3K27me3 is known to occur in acutely demethylated cells and this may limit the occupancy of transcription factors to retrotransposons or other genomic loci in the absence of DNA methylation (Brinkman et al., 2012; Reddington et al., 2013; King et al., 2016; Walter et al., 2016). In addition, the original genetic modification to the J1 cell line (knockout of the *Dnmt3a* gene) that preceded the generation of the DNMT.TKO cell line dates back to 1999 (Okano et al., 1999) meaning that these J1 and DNMT.TKO cells are likely to have experienced some genetic divergences that could contribute to the differences observed between the two. This is evidently the case as chromosome 15 is amplified in the J1 cells that we have used for many of our experiments. The consequences of this amplification remains largely uncharacterised. To validate some of our results, CRISPR/Cas9-based tools that have been designed to modify the DNA methylation state at specific CpG sites in the genome (Liu et al., 2016b) could be taken advantage of. Specifically, one could assess the occupancy of transcription factors *in vivo* upon experimental methylation or demethylation of specific CpGs within their binding site.

The abrogation of DNA methyltransferases in DNMT.TKO cells not only results in a depletion of DNA methylation, but also a loss of 5hmC and the further oxidised forms of cytosine due to the loss of a substrate for TET proteins (Ficz et al., 2011). 5-hydroxymethylcytosine has previously been found to affect the binding of some transcription factors to DNA (Spruijt et al., 2013) and it is possible that loss of 5hmC in DNMT.TKO cells could impact their genome-wide occupancy. We have not explored this possibility, which could be addressed through profiling chromatin accessibility or transcription factor binding in TET.TKO cells. However, as 5hmC is found at low per-allele frequency and is relatively rare compared to 5mC in wild-type

mES cells (Yu et al., 2012), its presence or absence is unlikely to contribute substantially to transcription factor occupancy.

The results presented in this thesis relate to experiments performed in mouse embryonic stem cells. Unlike any other mammalian cells known to date, these cells are able to persevere in the absence of DNA methylation implying that they employ other mechanisms to restrict transcription factor occupancy, chromatin accessibility and the transcription of deleterious elements. It is possible that in other cell lines where DNA methylation is essential, the role we have determined for DNA methylation in protecting from transcription factor occupancy and chromatin opening is more widespread.

Chapter 7:

Appendix

7.1 Supplementary Tables

Table S1: Top 100 significantly upregulated genes in TKO.DMSO versus J1.DMSO: Differential gene expression analysis of RNA-seq data was performed using the R package edgeR (Robinson et al., 2010); a FDR threshold of 10^{-5} was used to call significantly differentially expressed genes (marked in red). The average CpG methylation levels in the 620bp regions surrounding each gene's transcription start site is indicated. Grey shading indicates an average CpG methylation above 0.6. Chr. = Chromosome; Log2FC = Fold Change (log2); FDR = False Discovery Rate. NA = no CpG methylation data available.

Gene ID	Chr.	Mean FPKM				TKO.DMSO vs J1.DMSO		J1.TSA vs J1.DMSO		TKO.TSA vs TKO.DMSO		TKO.TSA vs J1.DMSO		Mean CpG methylation
		J1.DMSO	TKO.DMSO	J1.TSA	TKO.TSA	Log2FC	FDR	Log2FC	FDR	Log2FC	FDR	Log2FC	FDR	
1700080O16Rik	X	0.000	3.832	0.044	23.935	9.344	1.31E-32	3.090	1.73E-01	2.643	3.23E-33	11.987	7.08E-102	89.2
Gm773	X	0.000	2.578	0.304	17.125	8.158	9.18E-13	5.121	3.45E-03	2.734	1.15E-11	10.892	3.68E-36	86.9
Magea2	X	0.000	1.115	0.000	2.814	7.777	7.71E-10	0.000	1.00E+00	1.338	5.26E-03	9.115	5.29E-18	79.4
Magea5	X	0.023	6.486	0.020	14.558	7.774	2.47E-34	-0.184	9.34E-01	1.167	9.60E-05	8.941	3.51E-52	70.0
Pramel3	X	0.011	2.724	0.018	8.371	7.751	6.59E-44	0.655	7.27E-01	1.621	8.55E-13	9.372	2.96E-84	88.3
4930550L24Rik	X	0.023	5.984	0.645	45.322	7.667	1.48E-45	4.446	6.04E-07	2.919	3.88E-45	10.586	8.31E-138	81.3
1700019B21Rik	X	0.000	2.228	1.925	7.760	7.661	3.16E-11	7.448	1.17E-10	1.801	3.67E-07	9.463	1.74E-28	70.9
Magea6	X	0.000	1.360	0.000	2.128	7.575	1.83E-09	0.000	1.00E+00	0.652	1.80E-01	8.227	8.42E-14	NA
Magea8	X	0.025	4.618	0.060	14.959	7.353	8.19E-23	1.182	4.81E-01	1.697	7.34E-06	9.051	1.07E-39	83.1
Magea3	X	0.000	1.244	0.000	2.638	7.332	1.98E-09	0.000	1.00E+00	1.086	5.89E-03	8.418	5.89E-18	NA
Xlr4c	X	0.000	1.011	0.301	2.236	7.178	5.80E-07	5.422	1.69E-03	1.146	4.58E-02	8.325	4.16E-12	78.5
Gm41	X	0.000	2.328	1.238	15.607	7.177	1.01E-08	6.257	9.48E-06	2.741	4.96E-15	9.918	8.88E-36	76.9
Iqca	1	0.000	0.342	0.041	0.916	7.004	5.74E-08	4.005	4.93E-02	1.420	2.89E-04	8.424	1.26E-18	11.7
Gm595	X	0.017	1.507	0.045	5.543	6.113	1.16E-18	1.183	4.73E-01	1.875	2.35E-13	7.988	1.22E-51	70.2
Mgl2	11	0.127	6.726	0.257	5.860	5.720	3.33E-41	1.002	2.42E-01	-0.197	4.41E-01	5.522	3.68E-38	70.8
Xlr4b	X	0.031	1.480	0.349	3.072	5.363	1.46E-10	3.307	3.08E-03	1.056	3.78E-03	6.419	3.30E-20	75.8

Dnajc5g	5	0.052	1.999	0.166	0.769	5.209	1.67E-12	1.659	1.55E-01	-1.364	2.52E-03	3.845	2.47E-06	68.8
Gm6329	8	0.047	1.892	0.304	3.554	5.136	3.99E-09	2.523	4.91E-02	0.913	1.93E-02	6.048	1.47E-16	85.1
Clec4f	6	0.015	0.637	0.170	0.392	5.047	1.80E-07	3.197	6.86E-03	-0.684	2.17E-01	4.362	9.37E-06	84.0
Asz1	6	0.331	9.989	1.875	41.110	4.913	5.61E-43	2.488	5.59E-09	2.041	2.73E-18	6.954	1.72E-92	63.0
Rhox2a	X	0.052	1.770	0.335	4.831	4.900	6.86E-07	2.524	5.54E-02	1.452	2.11E-03	6.353	3.07E-15	67.4
Pet2	X	0.027	0.823	0.015	6.205	4.824	1.63E-15	-0.739	6.17E-01	2.911	2.80E-23	7.735	9.96E-60	83.7
Rhox2b	X	0.053	1.684	0.000	4.308	4.786	9.35E-07	-2.437	2.88E-01	1.360	2.56E-03	6.147	4.89E-15	86.2
Tuba3b	6	0.755	20.580	2.074	46.382	4.784	1.15E-16	1.469	1.08E-02	1.172	1.43E-02	5.956	1.20E-24	90.5
Tfec	6	0.020	0.690	0.092	0.131	4.740	4.46E-07	1.875	1.96E-01	-2.384	1.03E-04	2.356	6.79E-02	84.2
1700018B24Rik	3	0.080	2.161	0.546	10.250	4.737	1.06E-28	2.739	3.76E-07	2.245	1.36E-22	6.982	9.82E-82	75.2
Nxf2	X	0.130	3.440	0.888	3.276	4.713	2.67E-29	2.759	8.54E-08	-0.073	8.03E-01	4.640	5.46E-29	79.0
Aqp8	7	0.054	1.453	0.087	0.607	4.678	1.52E-09	0.704	6.22E-01	-1.236	9.77E-03	3.442	5.29E-05	78.8
Cyp4a12b	4	0.015	0.479	0.068	4.869	4.653	7.66E-07	1.874	1.94E-01	3.323	1.48E-21	7.976	7.35E-41	NA
Xlr4a	X	0.128	3.047	0.306	5.149	4.508	5.87E-12	1.193	2.16E-01	0.759	6.38E-02	5.266	1.98E-18	75.4
Usp9y	Y	0.114	2.505	0.793	10.081	4.462	3.35E-34	2.795	5.86E-13	2.008	1.26E-13	6.470	5.25E-71	89.3
Nostrin	2	0.064	1.387	0.090	1.074	4.362	9.79E-11	0.477	7.09E-01	-0.360	3.99E-01	4.002	2.13E-09	74.2
1700013H16Rik	X	0.978	19.820	10.126	80.219	4.354	1.22E-09	3.385	6.25E-07	2.017	8.84E-04	6.371	3.30E-18	78.2
Efhc2	X	1.119	21.469	2.737	26.415	4.275	5.78E-53	1.299	1.79E-05	0.299	2.17E-01	4.574	5.65E-61	42.3
Scml2	X	1.628	30.001	6.480	42.699	4.206	1.61E-52	1.993	9.28E-14	0.509	4.05E-02	4.715	1.78E-64	3.6
Fkbp6	5	2.224	40.071	1.663	18.853	4.190	1.57E-69	-0.394	2.41E-01	-1.087	2.17E-08	3.103	1.62E-38	75.7
Magea10	X	0.061	1.216	0.071	4.574	4.189	7.28E-11	0.174	9.01E-01	1.907	9.16E-11	6.096	2.46E-35	89.7
Dcdc2a	13	0.135	2.376	0.334	1.860	4.117	8.88E-35	1.287	1.44E-03	-0.355	1.90E-01	3.762	1.85E-29	5.7
Pkd2l1	19	0.140	2.243	0.531	3.276	4.003	8.06E-17	1.923	4.88E-04	0.549	1.20E-01	4.552	2.79E-23	75.0
Magea4	X	0.283	4.472	0.962	8.208	3.955	3.04E-16	1.740	6.68E-03	0.878	1.92E-03	4.833	5.30E-29	NA
Gm364	X	3.344	51.009	19.357	94.238	3.938	3.91E-42	2.539	2.52E-19	0.885	5.75E-04	4.823	3.63E-60	85.6
Xist	X	0.015	0.215	0.047	0.270	3.821	4.17E-10	1.635	2.47E-02	0.331	4.65E-01	4.152	4.34E-13	76.9

Xlr5a	X	0.049	0.710	0.083	4.650	3.798	9.47E-06	0.710	6.16E-01	2.711	1.47E-12	6.509	1.37E-27	NA
Dazl	17	3.512	47.785	14.317	89.572	3.768	9.83E-29	2.028	4.77E-10	0.906	3.38E-03	4.674	1.07E-41	74.5
Xlr3c	X	0.291	3.867	1.552	18.893	3.724	2.24E-17	2.401	7.90E-07	2.290	4.31E-16	6.014	1.56E-53	84.5
Cox7b2	5	1.622	19.918	9.729	60.235	3.618	8.73E-17	2.578	6.08E-09	1.595	1.77E-06	5.214	2.53E-35	82.5
AU022751	X	1.636	20.011	5.777	23.063	3.613	3.60E-80	1.819	7.90E-18	0.204	1.98E-01	3.817	2.17E-91	74.5
4933425L06Rik	13	0.052	0.681	0.505	3.549	3.604	5.33E-06	3.153	6.66E-05	2.372	1.45E-08	5.976	1.93E-22	67.6
Wfdc15a	2	0.544	6.999	1.441	16.367	3.593	2.93E-08	1.292	1.00E-01	1.223	5.80E-03	4.817	1.98E-16	85.7
Dpep3	8	0.160	1.945	0.592	6.729	3.584	4.81E-11	1.848	4.75E-03	1.788	2.89E-08	5.372	3.28E-31	93.1
Rhox1	X	0.665	7.726	1.598	7.900	3.557	1.03E-16	1.277	1.57E-02	0.032	9.28E-01	3.589	4.68E-18	54.2
Klb	5	0.124	1.417	0.151	1.079	3.518	1.29E-13	0.303	6.91E-01	-0.394	2.70E-01	3.123	3.41E-11	82.3
Ctcfl	2	0.428	4.862	1.123	5.381	3.508	2.90E-28	1.392	1.43E-04	0.148	5.90E-01	3.656	1.09E-31	60.6
Srgn	10	0.884	9.608	0.494	3.926	3.473	4.00E-16	-0.813	2.07E-01	-1.289	1.93E-04	2.184	7.19E-07	63.5
Hkdc1	10	0.090	0.985	0.284	0.306	3.462	7.61E-10	1.684	1.05E-02	-1.674	1.10E-04	1.789	4.76E-03	65.3
Olfr307	7	0.328	3.369	0.238	0.390	3.351	3.35E-07	-0.417	6.84E-01	-3.108	1.39E-07	0.244	7.88E-01	82.0
Gm6880	X	0.557	5.458	5.828	28.289	3.290	6.95E-09	3.378	5.26E-10	2.373	2.25E-10	5.663	3.55E-30	75.1
Gm5635	X	0.469	4.409	2.615	10.067	3.216	8.63E-12	2.456	1.17E-06	1.187	1.17E-05	4.403	1.42E-28	68.8
Nlrp4c	7	0.689	6.320	6.354	34.804	3.202	3.50E-21	3.208	1.23E-21	2.461	1.60E-17	5.663	3.63E-60	86.0
Trim43b	9	0.150	1.409	0.271	0.925	3.198	5.25E-08	0.833	2.67E-01	-0.599	1.85E-01	2.599	7.07E-06	66.3
Nxf3	X	1.653	14.958	1.327	16.170	3.187	1.49E-38	-0.314	3.12E-01	0.112	6.39E-01	3.299	4.90E-42	84.4
C1qtnf9	14	0.101	0.924	0.086	0.243	3.160	8.11E-07	-0.219	8.52E-01	-1.923	1.00E-04	1.236	1.27E-01	73.6
Ushbp1	8	0.595	5.244	0.246	0.635	3.148	8.82E-27	-1.254	3.13E-03	-3.041	3.45E-27	0.107	7.88E-01	82.2
Tex12	9	0.320	2.774	0.646	2.490	3.121	6.76E-06	1.004	2.66E-01	-0.162	7.50E-01	2.959	6.15E-06	57.2
Tktl2	8	0.339	2.935	2.646	21.583	3.121	1.28E-22	2.968	8.07E-21	2.877	2.37E-27	5.998	2.33E-77	71.3
Kdm5d	Y	0.458	3.936	0.315	4.617	3.119	8.58E-38	-0.527	1.37E-01	0.230	2.64E-01	3.349	4.86E-45	73.5
H2bfm	X	0.231	1.941	1.857	20.167	3.110	1.02E-06	3.029	7.01E-07	3.370	3.14E-20	6.480	2.43E-39	68.6
Zfp112	7	0.333	2.839	0.314	0.702	3.077	2.37E-19	-0.093	8.65E-01	-2.015	8.89E-12	1.061	7.28E-03	0.0

1700019E08Rik	2	0.733	6.112	1.828	9.778	3.067	6.27E-12	1.335	1.47E-02	0.680	2.06E-02	3.747	9.37E-21	73.1
Gpr146	5	0.072	0.612	0.460	0.477	3.055	7.17E-08	2.657	2.53E-06	-0.352	4.06E-01	2.703	9.88E-07	1.4
Apom	17	1.112	9.017	0.437	0.897	3.007	5.77E-13	-1.284	4.98E-02	-3.293	4.38E-16	-0.286	6.37E-01	80.1
Mlana	19	1.219	9.980	1.352	4.808	3.001	3.94E-11	0.118	8.71E-01	-1.048	3.28E-03	1.953	4.09E-05	71.6
Cubn	2	0.470	3.623	0.417	1.325	2.954	1.09E-40	-0.169	5.65E-01	-1.451	2.29E-13	1.503	4.07E-11	76.4
Gm715	X	0.180	1.449	0.679	2.155	2.911	3.58E-06	1.802	9.72E-03	0.568	1.57E-01	3.479	2.02E-10	0.5
Abo	2	0.259	1.946	0.758	1.182	2.902	8.76E-08	1.535	9.65E-03	-0.720	9.27E-02	2.182	5.30E-05	69.7
Slfn4	11	0.069	0.509	0.082	0.495	2.881	4.57E-07	0.274	7.66E-01	-0.037	9.31E-01	2.844	1.15E-07	74.4
Smtnl1	2	1.391	10.006	0.551	2.301	2.852	2.41E-19	-1.322	1.76E-03	-2.124	2.33E-13	0.728	3.71E-02	69.3
Kcnt1	2	0.072	0.538	0.425	0.490	2.834	1.98E-09	2.503	1.08E-07	-0.130	7.25E-01	2.705	2.46E-09	7.0
Xlr3b	X	0.501	3.521	0.766	3.973	2.820	3.56E-11	0.620	2.48E-01	0.178	6.19E-01	2.998	1.03E-13	79.0
Gpat2	2	2.235	15.671	6.633	30.503	2.814	3.76E-19	1.571	5.14E-07	0.961	7.41E-04	3.775	1.82E-33	44.0
Fermt3	19	0.820	5.776	0.701	1.894	2.813	1.73E-21	-0.227	5.85E-01	-1.612	5.20E-10	1.202	1.63E-04	77.2
Gm9	X	0.213	1.474	1.241	9.153	2.802	5.83E-06	2.551	1.96E-05	2.636	9.34E-11	5.438	1.43E-25	75.8
Sox17	1	0.157	1.086	1.055	4.365	2.771	2.75E-09	2.718	1.65E-09	2.002	6.30E-12	4.773	4.80E-35	10.4
Hus1b	13	1.323	8.697	1.796	6.874	2.722	6.17E-16	0.450	2.86E-01	-0.342	2.43E-01	2.380	6.46E-13	39.4
Trim43a	9	0.455	3.004	0.751	4.135	2.713	8.01E-07	0.711	2.49E-01	0.464	3.23E-01	3.177	5.43E-10	75.3
H19	7	20.575	132.168	120.979	1299.316	2.684	2.01E-09	2.556	3.27E-09	3.297	3.91E-14	5.981	1.73E-34	86.3
Hrh3	2	0.139	0.899	0.274	0.269	2.652	9.57E-07	0.971	1.49E-01	-1.729	8.11E-05	0.923	1.63E-01	0.2
Slc1a1	19	1.255	7.757	6.583	13.391	2.637	1.39E-15	2.400	1.44E-13	0.788	8.23E-03	3.425	3.71E-26	1.5
Ccdc169	3	0.154	0.946	0.801	1.631	2.624	8.32E-07	2.379	5.22E-06	0.781	1.94E-02	3.404	5.99E-14	72.7
Ace	11	0.167	1.017	0.775	2.443	2.614	2.71E-07	2.221	7.35E-06	1.267	1.43E-03	3.881	9.89E-17	3.0
Tph2	10	0.129	0.806	0.485	1.282	2.605	1.20E-06	1.865	9.81E-04	0.668	5.22E-02	3.273	1.43E-12	15.8
Xlr3a	X	1.161	6.954	1.296	9.665	2.598	1.53E-10	0.178	7.28E-01	0.476	1.85E-01	3.074	1.78E-15	74.6
1700048O20Rik	9	0.249	1.519	1.450	5.053	2.597	5.53E-06	2.529	3.33E-06	1.736	1.09E-05	4.334	1.28E-18	76.6
Tdrd9	12	0.226	1.309	0.468	1.547	2.549	1.42E-10	1.062	1.64E-02	0.242	4.72E-01	2.791	8.89E-14	81.6

Spink10	18	0.477	2.755	0.312	1.015	2.539	1.10E-07	-0.572	4.35E-01	-1.444	2.17E-04	1.095	4.26E-02	0.1
Nxf7	X	0.836	4.809	3.033	8.665	2.533	1.75E-06	1.866	2.72E-04	0.851	6.41E-02	3.384	1.13E-11	82.6
Lims2	18	0.222	1.313	0.514	0.635	2.526	8.59E-07	1.185	4.96E-02	-1.038	8.01E-03	1.488	7.64E-03	11.5
Pdgfra	5	0.294	1.670	1.176	2.447	2.508	1.84E-16	2.002	7.80E-11	0.552	2.73E-02	3.061	3.84E-26	13.7
4933406J08Rik	2	0.740	4.158	1.782	8.509	2.506	5.50E-09	1.281	3.73E-03	1.034	3.78E-03	3.540	2.18E-18	54.8
H2-Eb1	17	0.777	4.500	5.402	8.861	2.502	7.61E-16	2.765	1.43E-20	0.977	1.25E-05	3.479	3.38E-35	80.3

Table S2: All 65 significantly upregulated LTR retrotransposon element types in TKO.DMSO versus J1.DMSO: Differential analysis of RNA-seq data was performed using the R package DESeq2 (Love et al., 2014). An adjusted p value threshold of 0.05 was used to call significantly differences (significant gains and losses are marked in red and blue respectively). Log2FC = Fold Change (log2).

Repeat ID	Repeat family	Total genomic size (bp)	Number of copies	Est. Repeat age (million years)	Mean CpG percent	Base Mean	TKO.DMSO vs J1.DMSO		J1.TSA vs J1.DMSO		TKO.TSA vs TKO.DMSO		TKO.TSA vs J1.DMSO	
							p.adj	Log2FC	p.adj	Log2FC	p.adj	Log2FC	p.adj	Log2FC
IAPEy-int	ERVK	2341002	755	32.70	1.75	95397.1	2.683	3.39E-17	1.273	3.11E-04	2.444	6.10E-15	5.127	1.01E-64
MMERGLN-int	ERV1	1265344	856	115.35	1.16	663487.9	2.467	3.41E-24	1.921	1.13E-14	3.119	9.53E-40	5.585	1.58E-127
LTR86A2	ERVL	12303	79	142.05	0.51	1710.9	2.057	4.69E-05	1.062	5.83E-02	1.193	1.63E-02	3.251	1.29E-13
IAPA_MM-int	ERVK	37463	126	40.13	5.20	65351.2	2.035	7.51E-22	1.001	1.51E-05	3.175	1.46E-54	5.209	8.85E-147
IAPEz-int	ERVK	11883078	7319	27.08	2.36	7833824.7	1.778	4.83E-32	0.703	2.46E-05	2.759	7.05E-79	4.537	4.47E-214
IAPLTR1_Mm	ERVK	551227	1492	16.00	5.38	517643.2	1.757	4.83E-26	0.759	3.37E-05	2.845	3.61E-70	4.602	8.35E-184
RLTR19A	ERVK	44831	153	77.02	0.62	2538.5	1.404	3.38E-06	0.797	1.61E-02	1.250	1.11E-05	2.654	2.35E-23
RLTR47_MM	ERV1	13906	45	77.40	0.99	514.3	1.371	8.48E-11	0.606	1.41E-02	0.777	1.53E-04	2.148	2.18E-28
MMERGLN_LTR	ERV1	167469	435	22.53	2.65	76438.8	1.317	1.05E-06	0.828	4.08E-03	1.945	1.94E-15	3.262	5.93E-43
RLTR45-int	ERVK	918744	1317	50.89	1.11	82322.1	1.264	1.61E-05	0.053	9.37E-01	-0.226	5.15E-01	1.038	1.59E-04
RLTR44E	ERVK	41391	122	48.36	3.06	27247.0	1.247	7.63E-04	-0.504	2.07E-01	-1.533	2.24E-06	-0.286	4.78E-01

MMTV-int	ERVK	34354	35	118.57	0.48	29083.2	1.227	1.63E-03	-0.551	1.84E-01	-1.591	2.40E-06	-0.364	3.66E-01
RLTR9C	ERVK	175420	477	68.89	0.67	5067.1	1.157	3.84E-10	0.962	2.82E-07	0.393	5.26E-02	1.550	2.81E-19
RLTR26	ERVK	139769	408	90.88	1.19	29296.1	1.147	2.73E-03	-0.437	3.01E-01	-1.508	5.08E-06	-0.361	3.60E-01
MuLV-int	ERV1	306109	392	134.11	0.57	23855.8	1.141	9.65E-04	-0.555	1.35E-01	-1.387	4.13E-06	-0.246	5.16E-01
RLTR20D	ERVK	206997	777	88.93	0.81	33202.1	1.131	1.80E-03	-0.441	2.66E-01	-1.443	4.12E-06	-0.312	4.11E-01
RLTR16B_MM	ERVK	239283	635	72.53	1.15	33554.0	1.101	1.59E-03	-0.338	4.05E-01	-1.366	6.35E-06	-0.265	4.82E-01
MLTR18A_MM	ERVK	109687	299	106.16	0.52	54024.4	1.086	1.53E-03	-0.208	6.53E-01	-0.867	5.96E-03	0.218	5.63E-01
RLTR10F	ERVK	136494	482	34.07	3.23	30446.5	1.075	6.45E-04	-0.581	8.27E-02	-1.367	5.64E-07	-0.293	3.74E-01
ORR1D2-int	ERVL-MaLR	575296	1361	114.25	0.48	36319.7	1.044	5.84E-04	-0.215	5.94E-01	-0.990	2.81E-04	0.053	8.83E-01
RLTR13D3A1	ERVK	294068	613	47.35	0.78	42350.2	1.043	1.94E-03	0.325	4.07E-01	-0.188	6.28E-01	0.855	4.83E-03
RLTR13C1	ERVK	220441	309	57.79	0.73	23035.6	1.039	1.39E-03	-0.555	1.07E-01	-1.244	1.11E-05	-0.205	5.68E-01
IAPLTR4_I	ERVK	102274	183	43.16	3.00	5996.8	1.002	2.63E-03	0.566	1.07E-01	-0.143	7.21E-01	0.859	4.07E-03
LTRIS4A	ERV1	79353	228	41.15	1.50	9584.6	0.993	1.35E-09	0.461	1.40E-02	-0.124	5.56E-01	0.869	3.68E-08
RMER6B	ERVK	405871	1360	72.60	0.48	47681.9	0.991	2.61E-04	-0.267	4.09E-01	-1.095	5.52E-06	-0.104	7.49E-01
RLTR6C_Mm	ERV1	42097	102	33.91	2.99	7187.3	0.975	1.89E-03	1.705	2.87E-10	1.793	5.80E-12	2.768	6.64E-28
RMER21B	ERV1	289464	1252	104.72	0.61	47745.4	0.973	7.30E-03	-0.355	3.78E-01	-1.189	1.30E-04	-0.216	5.77E-01
RMER6BA	ERVK	225220	957	79.99	0.45	49629.1	0.960	6.69E-04	-0.320	3.13E-01	-1.054	2.39E-05	-0.094	7.82E-01
RLTR1E_MM	ERV1	19320	86	71.41	2.90	3438.0	0.954	1.79E-04	0.577	3.29E-02	0.040	9.06E-01	0.994	1.27E-05
ERVB4_2-LTR_MM	ERVK	67645	144	36.11	2.76	5281.5	0.952	1.07E-13	0.514	2.34E-04	0.090	5.92E-01	1.042	2.04E-17
IAPEY_LTR	ERVK	476732	1410	42.55	5.11	19051.1	0.940	1.21E-03	1.020	1.50E-04	2.085	3.02E-18	3.025	2.64E-38
RLTR13B2	ERVK	225139	314	21.10	2.22	34495.1	0.933	6.59E-04	-0.092	8.34E-01	-0.134	6.81E-01	0.799	1.23E-03
LTRIS6	ERV1	42103	149	34.78	1.95	30805.5	0.933	4.84E-02	-0.485	2.77E-01	-1.220	8.55E-04	-0.287	5.16E-01
MLT1F1-int	ERVL-MaLR	23751	89	142.35	0.74	279.7	0.924	2.67E-04	1.165	3.16E-07	0.295	2.55E-01	1.219	2.04E-08
LTRIS5	ERV1	27054	85	51.50	1.43	28211.2	0.913	4.91E-02	-0.509	2.37E-01	-1.254	4.45E-04	-0.341	4.15E-01

IAPLTR2_Mm	ERVK	626514	1388	35.36	5.12	87009.2	0.902	6.69E-04	0.361	2.08E-01	1.196	2.11E-07	2.098	6.52E-22
RLTR12D	ERVK	203443	647	101.71	0.60	26056.8	0.853	1.04E-02	-0.350	3.20E-01	-1.212	9.98E-06	-0.358	2.67E-01
MMERVK10D3_I-int	ERVK	493747	458	96.75	1.36	9646.3	0.849	3.01E-03	1.383	3.04E-08	1.491	4.06E-10	2.340	2.67E-24
RLTR1A2_MM	ERV1	171796	406	67.32	1.86	11427.6	0.820	1.03E-07	0.233	2.04E-01	-0.263	1.22E-01	0.557	2.23E-04
RMER17A	ERVK	873643	1819	65.36	0.71	54954.1	0.783	1.57E-02	-0.161	7.05E-01	-0.439	1.45E-01	0.345	2.66E-01
RLTR44-int	ERVK	526075	653	67.45	1.32	2638.6	0.782	1.03E-07	0.725	5.55E-07	0.930	4.05E-12	1.712	2.57E-39
RMER6D	ERVK	804327	2411	83.77	0.40	15781.2	0.779	2.77E-12	0.241	7.91E-02	-0.438	1.48E-04	0.342	3.42E-03
RLTR41A2	ERV1	155787	440	67.79	1.06	9547.3	0.774	1.99E-02	1.184	1.88E-05	1.270	1.53E-06	2.044	3.11E-16
MURVY-int	ERV1	1950428	544	39.91	1.19	3149.7	0.684	1.74E-02	1.362	4.62E-09	1.120	1.00E-06	1.805	1.15E-16
RLTR44B	ERVK	84952	275	72.27	2.21	451.0	0.682	1.82E-03	0.478	3.46E-02	0.456	2.13E-02	1.138	2.89E-10
RLTR10D2	ERVK	71666	225	33.58	2.90	1312.5	0.677	6.45E-04	0.730	6.18E-05	0.929	2.57E-08	1.606	1.62E-23
LTRIS_Mm	ERV1	251936	555	13.97	1.71	31218.7	0.675	1.80E-03	0.303	1.87E-01	0.129	5.96E-01	0.804	1.75E-05
IAPEY3C_LTR	ERVK	137710	475	23.49	3.18	4704.6	0.636	1.39E-03	0.498	1.23E-02	0.294	1.35E-01	0.930	3.14E-08
MMERVK10D3_LTR	ERVK	55961	174	30.88	3.11	727.3	0.632	4.84E-02	0.322	2.89E-01	0.299	3.02E-01	0.931	1.04E-04
MamGypLTR3a	Gypsy	27627	161	144.27	0.65	313.0	0.629	1.65E-02	0.597	1.44E-02	0.224	3.87E-01	0.853	5.32E-05
RMER6C	ERVK	1942752	4887	80.87	0.39	32597.9	0.618	2.63E-03	0.025	9.65E-01	-0.504	7.59E-03	0.114	6.24E-01
RLTR9E	ERVK	467242	1318	59.41	1.23	42535.3	0.613	5.18E-03	0.335	1.39E-01	0.104	6.78E-01	0.718	1.29E-04
MamGypLTR1c	Gypsy	24041	145	140.12	0.58	375.4	0.581	1.07E-04	0.424	5.71E-03	-0.196	1.97E-01	0.385	6.82E-03
IAPLTR2a	ERVK	243139	627	22.57	4.98	23619.4	0.562	4.42E-03	0.296	1.46E-01	1.659	5.63E-26	2.221	1.87E-46
RLTR18-int	ERVK	163514	506	98.06	0.45	12897.2	0.562	2.27E-03	0.176	4.13E-01	-0.257	1.64E-01	0.305	8.79E-02
RMER17B	ERVK	1686444	2898	75.28	1.17	39445.3	0.506	1.94E-03	-0.167	3.75E-01	-0.624	1.11E-05	-0.118	5.09E-01
MERV1_I-int	ERV1	958198	1365	94.55	0.56	14551.0	0.477	4.10E-03	1.474	3.26E-28	2.420	4.65E-78	2.897	8.83E-112
RLTR13D3A	ERVK	89842	162	64.97	0.71	1696.1	0.451	2.62E-03	0.144	4.11E-01	-0.079	6.45E-01	0.372	5.85E-03
RMER13A1	ERVK	523855	1082	98.67	0.94	9858.8	0.437	1.56E-02	0.584	1.71E-04	0.752	1.81E-07	1.189	5.05E-18
RMER19A	ERVK	1017676	2095	71.23	0.76	22727.9	0.411	8.91E-04	0.513	4.13E-06	1.223	1.97E-34	1.634	2.84E-61

RLTR9A2	ERVK	71406	227	61.11	0.76	1259.6	0.401	3.42E-03	0.038	8.69E-01	-0.448	1.98E-04	-0.047	7.69E-01
MTB-int	ERVL-MaLR	414280	592	51.87	0.37	17621.0	0.339	3.33E-02	0.082	6.74E-01	-0.167	2.57E-01	0.172	2.34E-01
MamGyp-int	Gypsy	11669	28	150.11	1.11	3059.3	0.331	1.13E-03	0.465	2.08E-07	0.628	2.42E-14	0.959	2.50E-32
RLTR30	ERV1	101242	390	86.29	0.85	2027.3	0.285	1.27E-02	-0.275	1.12E-02	-0.437	2.60E-06	-0.152	1.59E-01

Table S1: All 27 LINE1 elements that significantly gain in ATAC-seq coverage in TKO.TSA relative to TKO.DMSO cells. Differential analysis of ATAC-seq and RNA-seq data was performed using the R package DESeq2 (Love et al., 2014). An adjusted p value threshold of 0.05 was used to call significantly differences (significant gains and losses are marked in red and blue respectively).

LINE1 repeat ID	Total genomic size (bp)	Number of copies	Mean CpG percent	Est. Repeat age (million years)	Base Mean	ATAC-seq Fold Change (log2)				RNA-seq Fold Change (log2)			
						J1.TSA vs J1.DMSO	TKO.DMSO vs J1.DMSO	TKO.TSA vs TKO.DMSO	TKO.TSA vs J1.DMSO	J1.TSA vs J1.DMSO	TKO.DMSO vs J1.DMSO	TKO.TSA vs TKO.DMSO	TKO.TSA vs J1.DMSO
L1Md_F	5919327	4016	1.19	55.2	180931.1	0.123	-0.030	0.234	0.204	0.547	0.204	0.733	0.530
L1Md_A	32690648	16844	0.98	17.6	3879065	0.093	-0.115	0.229	0.114	0.659	0.411	0.882	0.471
L1Md_T	48698338	23688	0.95	17.1	4796072	0.047	-0.209	0.219	0.010	0.473	0.067	0.521	0.454
Lx3_Mus	8796836	12651	0.43	50.8	87951.44	0.123	-0.145	0.210	0.065	0.330	0.175	0.627	0.452
Lx3A	6103541	9223	0.41	62.8	201870.4	0.142	-0.129	0.210	0.080	-0.663	0.197	-0.492	-0.689
Lx3B	3425540	5993	0.44	67.5	279397.5	0.133	-0.117	0.196	0.079	-0.628	0.197	-0.448	-0.645
L1_Mus4	8782822	11880	0.38	49.5	103247.9	0.108	-0.098	0.188	0.091	0.381	0.106	0.840	0.734
Lx2A	821827	1364	0.54	59.9	27167.28	0.101	-0.098	0.186	0.087	0.201	0.372	-0.147	-0.519
L1Md_F2	74093504	64898	0.76	30.8	2450219	0.105	-0.047	0.184	0.137	0.297	0.135	0.472	0.337
Lx3C	3903581	6475	0.43	70.8	67132	0.115	-0.121	0.179	0.058	0.490	0.361	0.746	0.385
Lx4A	6018144	10564	0.44	71.1	106726.2	0.111	-0.099	0.176	0.077	-0.016	0.437	0.430	-0.007
L1_Mur1	4541516	7793	0.33	67.6	215981	0.105	-0.071	0.175	0.105	-0.583	0.161	-0.459	-0.620
Lx2B2	6354359	11012	0.42	77.0	141446.8	0.110	-0.110	0.171	0.061	0.084	0.608	0.302	-0.306

L1_Mus2	15963233	21079	0.50	39.4	527421.2	0.091	-0.082	0.170	0.088	-0.232	0.011	-0.086	-0.096
L1_Mus1	22026666	27252	0.62	34.0	805247.8	0.100	-0.079	0.169	0.090	-0.090	0.068	-0.001	-0.069
Lx	16377948	24834	0.46	59.6	310533.6	0.106	-0.094	0.168	0.074	-0.125	0.104	-0.186	-0.290
Lx2B	3846006	6191	0.38	73.4	266416.2	0.113	-0.105	0.168	0.062	-0.634	0.009	-0.637	-0.646
Lx2	10435637	18521	0.42	63.8	253143.2	0.102	-0.119	0.165	0.045	-0.172	0.211	-0.139	-0.350
L1_Mus3	36756625	48489	0.41	48.2	762727.1	0.103	-0.076	0.164	0.088	-0.079	0.127	0.145	0.017
Lx2A1	1319632	2739	0.44	71.3	45532.98	0.107	-0.082	0.163	0.081	0.685	0.380	1.091	0.711
Lx5	9588832	16339	0.40	89.6	209658.2	0.085	-0.099	0.147	0.048	-0.475	-0.083	-0.253	-0.170
Lx5b	5149060	7975	0.38	87.5	75322.06	0.084	-0.104	0.137	0.033	0.311	0.654	0.590	-0.063
Lx5c	7571398	16422	0.42	94.4	239372.3	0.079	-0.071	0.127	0.056	-0.544	0.102	-0.296	-0.398
Lx4B	3033439	5680	0.44	81.2	138756.2	0.086	-0.081	0.121	0.039	-0.019	0.010	-0.744	-0.754
Lx8b	8425451	24166	0.46	114.6	233873.7	0.061	-0.051	0.117	0.066	-0.525	-0.206	-0.300	-0.094
L1M3e	183095	652	0.43	131.0	3299.103	0.095	-0.013	0.103	0.090	0.045	-0.105	-0.070	0.035
Lx8	16469159	49141	0.57	128.8	876629.6	0.045	-0.039	0.086	0.047	-0.223	0.056	-0.206	-0.262

Table 1: Panel of potential transcription factors for ChIP-seq experiments. Transcription factors highlighted in red are those that were selected for the first series of ChIP_seq experiments. Prelimaniry ChIP-qPCR was performed for the transcription factors highlighted in blue. bHLH: Basic Helix-loop-helix factor; CRE element: 5'-TGA**CG**TCA-3'; Ebox element: 5'-CA**CG**TG-3'.

Transcription Factor	Family or DNA binding domain type	CpG present in consensus recognition motif	Evidence for DNA methylation sensitive binding
ARNT	bHLH	**YES** (ARE Element)	**NO**
BHLHE40	bHLH	**YES** (Ebox element)	**NO**
MAX	bHLH	**YES** (Ebox element)	**NO**
MYC (c-Myc)	bHLH	**YES** (Ebox element)	*In vitro* (Prendergast & Ziff 1991)
TCF3 (E2-A)	bHLH	**YES** (Ebox element)	**NO**
USF1	bHLH	**YES** (Ebox element)	**NO**
HIF1A	bHLH	**YES** (5'-A**CG**TG-3')	**YES** (Wenger *et al*. 1998)
ATF2	Basic leucine zipper	**YES** (CRE element)	*In vitro* (Jones *et al.* 2006)
ATF3	Basic leucine zipper	**YES** (CRE element)	**NO**
CEBPB	Basic leucine zipper	**NO** (CCAAT element) **YES** (CRE element)	*In vivo* and *in vitro* (Mann *et al.* 2013)
CEBPA	Basic leucine zipper	**NO** (CCAAT element) **YES** (CRE element)	*In vitro* (Rishi *et al.* 2010)
CREB1	Basic leucine zipper	**YES** (CRE element)	*In vitro* (Spruijt *et al.* 2013)
JUN (c-Jun)	Basic leucine zipper	**YES** (if dimerised to CREB)	*In vitro* (Mann *et al.* 2013)
JUND	Basic leucine zipper	**YES** (if dimerised to CREB)	*In vitro* (Mann *et al.* 2013)
E2F4	E2F winged helix	**YES** (5'-TTTCC**CG**C-3')	**NO**
E2F1	E2F winged helix	**YES** (5'-TTTCC**CG**C-3')	*In vitro* (Campanero *et al.* 2003)
EGR1	C2H2-type zinc finger	**YES** (5'-G**CG**(G/T)GGG**CG**-3')	**NO**
KAISO (ZBTB33)	ZBTB zinc finger	**YES** (5'-**CGCG**-3')	*In vitro* (Yoon *et al.* 2001)
NRF1	NRF1/Ewg family	**YES** (5'-G**CG**CATG**CG**C-3')	**NO**
SP1	SP/KLF zinc finger protein	**YES** (5'-C**CG**CCC-3')	Mixed (Clark *et al.* 1997, Guo *et al.* 2012, Harrington *et al.* 1988)
YY1	Gli-Kruppel zinc finger	**YES** (5'-C**CG**CCATNTT-3')	*In vitro* (Kim *et al.* 2003)
CTCF	Zinc Finger factor	**NO**	Mixed
CUX1 (CDP)	Homeodomain protein	**NO**	**NO**
ELF1	ETS family	**NO**	**NO**
ELK1	ETS family	**NO**	**NO**
ELK4 (SAP1A)	ETS family	**NO**	**NO**
GABPA (NRF2)	ETS family	Adjacent to core motif (5'-c**G**GAAT-3')	*In vitro* (Lucas *et al.* 2009)
IRF1	Winged helix-turn-helix	**NO**	**NO**
KLF4	Kruppel-like zinc finger	**NO**	*In vivo* and *vitro* (Hu *et al.* 2013)
KLF5	Kruppel-like zinc finger	**NO**	**NO**
REST	Kruppel-like zinc finger	**NO**	**NO**
NANOG	Homeodomain protein	**NO**	**NO**

OCT4 (POU5F1)	Homeodomain protein	**NO**	**NO**
PATZ1 (MAZR)	AT-hook zinc finger	**NO**	**NO**
RBP-J	Su(H) family	**NO**	*In vitro* (Bartel *et al* 2011)
SMAD4	SMAD MH1 family	**NO** (SMAD binding element)	**NO**
SOX2	HMG-domain protein	**NO**	**NO**
STAT3	STAT family	**NO**	**NO**
TBX3	T-box family	**NO**	**NO**

7.2 Supplementary Figures

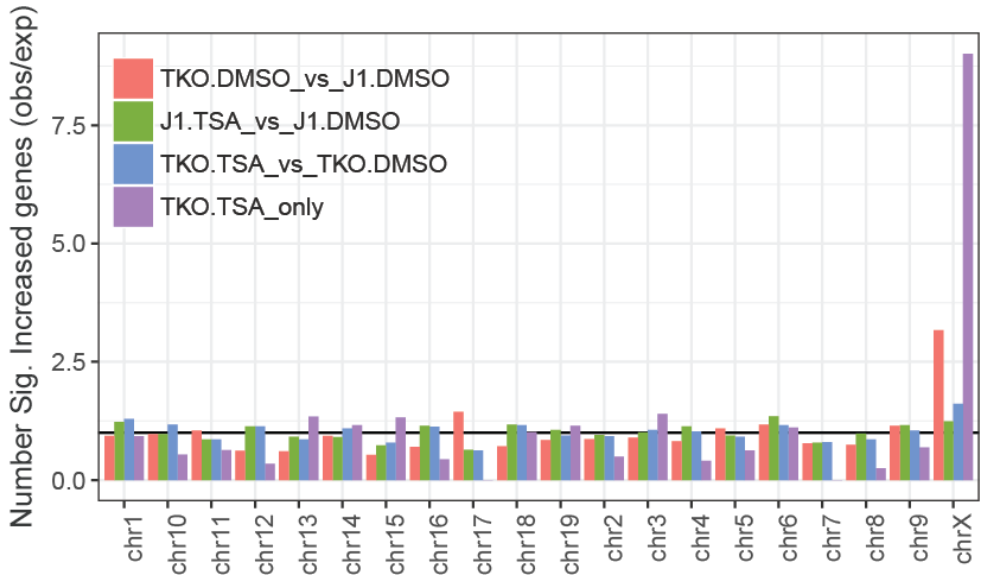


Figure S1: Bar graph representing the number of genes significantly upregulated per chromosome (Observed) relative to the expected number (Expected) based on the total number of genes per chromosome. Significantly upregulated genes were determined for DNMT.TKO compared to wild type J1 mESCs (red), TSA- compared to DMSO- treated J1 cells (green) and TSA-treated DNMT.TKO cells compared to DMSO-treated DNMT.TKO cells (blue). In purple are genes identified as significantly upregulated in TKO.TSA versus TKO.DMSO but not through other comparisons. The Observed/Expected ratios for each chromosome were calculated as follows: Observed = Number of genes on chromosome that are significantly upregulated in DNMT.TKO cells / Total number of genes significantly upregulated in DNMT.TKO cells. Expected = Total number of genes on chromosome / Total number of genes.

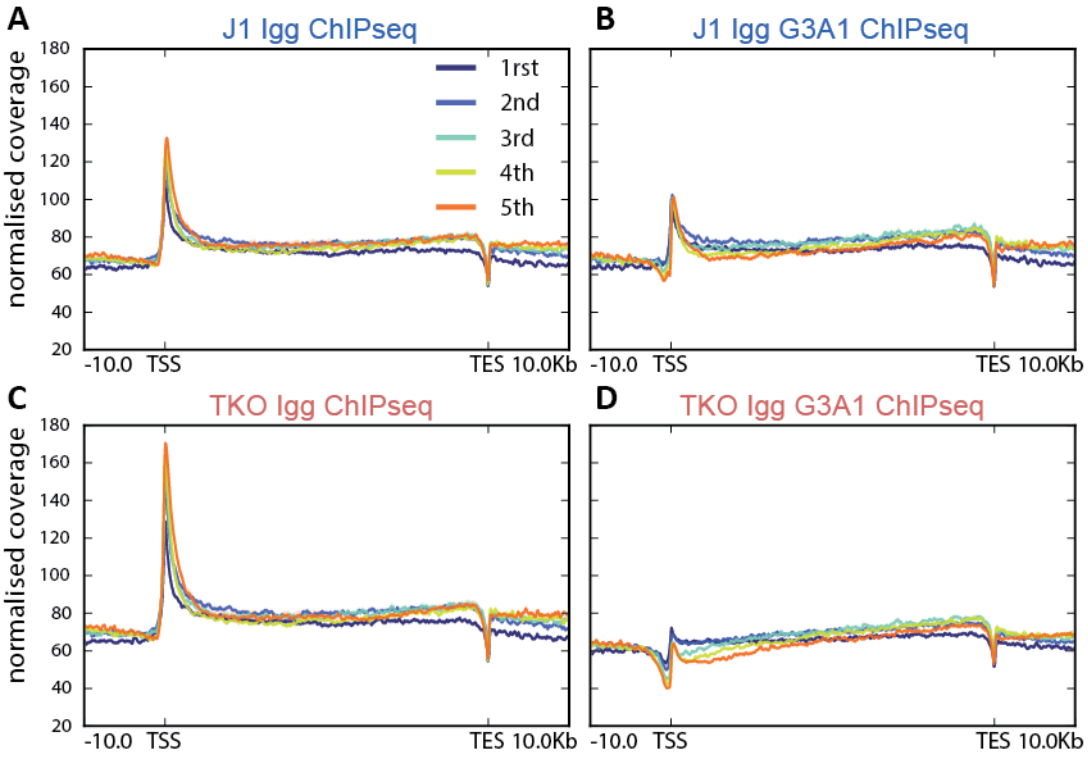


Figure S1: Metaplots showing Igg ChIP-seq signal at genic regions grouped by transcript expression level. Each line represents an equal number of transcripts ranked according to their expression level as measured by the mean RNA-seq FPKM values obtained across both wild-type (J1, figures **A-B**) or DNMT.TKO (TKO, figures **C-D**) biological replicates. RNA-seq FPKM values were calculated using fragments mapping to exons only. rAb_Igg ChIP-seq signal is plotted in figures **A** and **C**, mAb_Igg ChIP-seq signal is plotted in figures **B** and **D**. 5^{th} : Top 20% transcribed genes (5^{th} quintile), 4^{th}: 20%-40%, 3^{rd}: 40%-60%, 2^{nd} : 60%-80%, 1rst : Bottom 20% transcribed genes (1rst quintile). Each category includes 3044 genes.

7.3 Supplied scripts for bioinformatics analyses

R script to provide scaling factors to DeSeq2 for the differential analysis of H3ac ChIP-seq data based on the reads mapping to an exogenous reference genome:

```
library(DESeq2)
library(ggplot2)

##################calculate Chicken size factors#####################
#input is annotated galGal4 GeneBody file, using galGal4 bam files pre-downsampled to the relative amount of
galGal4    #reads in Input

galGal4.ReadCount.Table <- read.table("galGal4_geneBodies.featureCount.txt",header=T,sep="\t")
colnames(galGal4.ReadCount.Table)[7:14] <- c("J1.DMSO_REP1", "J1.DMSO_REP2", "J1.TSA_REP1",
"J1.TSA_REP2", "TKO.DMSO_REP1", "TKO.DMSO_REP2", "TKO.TSA_REP1", "TKO.TSA_REP2")
colnames(galGal4.ReadCount.Table)[1] <- "id"

galGal4.DESeq.Table<- galGal4.ReadCount.Table[,c("id","J1.DMSO_REP1", "J1.DMSO_REP2", "J1.TSA_REP1",
"J1.TSA_REP2", "TKO.DMSO_REP1", "TKO.DMSO_REP2", "TKO.TSA_REP1", "TKO.TSA_REP2")]
#copies only the relevant columns to a new table

rownames(galGal4.DESeq.Table) = make.names(galGal4.DESeq.Table$id, unique=TRUE)

galGal4.DESeq.Table <- galGal4.DESeq.Table[,2:9]
#now edits the table to remove the row name column (now we've successfully put this as the row name instead)

str(galGal4.DESeq.Table)
colnames(galGal4.DESeq.Table)

galGal4.DESeq.Samples.col <- colnames(galGal4.DESeq.Table)
#saves list of column names

galGal4.DESeq.Condition <- c(rep("J1.DMSO",2), rep("J1.TSA", 2), rep("TKO.DMSO",2), rep("TKO.TSA",2))
galGal4.DESeq.SampleInfo.Table <-data.frame(galGal4.DESeq.Condition)
#saves table of the above condition info:

colnames(galGal4.DESeq.SampleInfo.Table)<-c("condition")
#labels the above column

rownames(galGal4.DESeq.SampleInfo.Table)<-colnames(galGal4.DESeq.Table)
#rows are labelled UNT_r1_ReadCount etc

galGal4.DESeq.SampleInfo.Table
#views sample info table

galGal4.DESeq.Analysis <-DESeqDataSetFromMatrix(countData=galGal4.DESeq.Table,
colData=galGal4.DESeq.SampleInfo.Table, design = ~ condition)

cds=estimateSizeFactors(galGal4.DESeq.Analysis)

sizeFactors(cds)
galGal4_sizeFactors <- sizeFactors(cds)
#outputs size factors

####################now use these size factors in mm10 DESeq2 analysis###################
#input is mm10 annotated MACS2.mm10.p10 Peak file, using raw (non-normalised) Bam files
####################
setwd("Z:/genomic_data/Martin_H3ac_ChIP/DATA/CountingReads/")
mm10.ReadCount.Table <-read.table("H3ac_Nat.ChIP_MACS2.mm10.p10.featureCount.txt",header=T,sep="\t")
```

```
colnames(mm10.ReadCount.Table)[7:14] <- c("J1.DMSO_REP1", "J1.DMSO_REP2", "J1.TSA_REP1",
"J1.TSA_REP2", "TKO.DMSO_REP1", "TKO.DMSO_REP2", "TKO.TSA_REP1", "TKO.TSA_REP2")
colnames(mm10.ReadCount.Table)[1] <- "id"
coordinates <- mm10.ReadCount.Table[,c("id", "Chr", "Start","End")]

mm10.DESeq.Table<- mm10.ReadCount.Table[,c("id","J1.DMSO_REP1", "J1.DMSO_REP2", "J1.TSA_REP1",
"J1.TSA_REP2", "TKO.DMSO_REP1", "TKO.DMSO_REP2", "TKO.TSA_REP1", "TKO.TSA_REP2")]

rownames(mm10.DESeq.Table)<-mm10.DESeq.Table$id
mm10.DESeq.Table<-mm10.DESeq.Table[,2:9]
mm10.DESeq.Samples.col <-colnames(mm10.DESeq.Table)
mm10.DESeq.Condition <- c(rep("J1.DMSO",2), rep("J1.TSA", 2), rep("TKO.DMSO",2), rep("TKO.TSA",2))
mm10.DESeq.SampleInfo.Table <-data.frame(mm10.DESeq.Condition)
colnames(mm10.DESeq.SampleInfo.Table)<-c("condition")
rownames(mm10.DESeq.SampleInfo.Table)<-colnames(mm10.DESeq.Table)
mm10.DESeq.SampleInfo.Table

mm10.DESeq.Analysis<-
DESeqDataSetFromMatrix(countData=mm10.DESeq.Table,colData=mm10.DESeq.SampleInfo.Table, design
=~condition)

sizeFactors(mm10.DESeq.Analysis)<-galGal4_sizeFactors
#uses galGal4 size factors

mm10.DESeq.Analysis.results=DESeq(mm10.DESeq.Analysis)
#NB prints 'using pre-existing size factors' so you know this has been successful

res_J1.TSA.vs.J1.DMSO <- results(mm10.DESeq.Analysis.results, alpha = 0.05, contrast = c("condition", "J1.TSA",
"J1.DMSO"), independentFiltering = F)
res_TKO.DMSO.vs.J1.DMSO <- results(mm10.DESeq.Analysis.results, alpha = 0.05, contrast = c("condition",
"TKO.DMSO", "J1.DMSO"), independentFiltering = F)
res_TKO.TSA.vs.J1.DMSO <- results(mm10.DESeq.Analysis.results, alpha = 0.05, contrast = c("condition",
"TKO.TSA", "J1.DMSO"), independentFiltering = F)
res_TKO.TSA.vs.TKO.DMSO <- results(mm10.DESeq.Analysis.results, alpha = 0.05, contrast = c("condition",
"TKO.TSA", "TKO.DMSO"), independentFiltering = F)
res_TKO.TSA.vs.J1.TSA <- results(mm10.DESeq.Analysis.results, alpha = 0.05, contrast = c("condition", "TKO.TSA",
"J1.TSA"), independentFiltering = F)

summary(res_J1.TSA.vs.J1.DMSO)
summary(res_TKO.DMSO.vs.J1.DMSO)
summary(res_TKO.TSA.vs.J1.DMSO)
summary(res_TKO.TSA.vs.TKO.DMSO)
summary(res_TKO.TSA.vs.J1.TSA)

df_TKO.DMSO.vs.J1.DMSO <- data.frame(res_TKO.DMSO.vs.J1.DMSO)

########################
```

Perl script used to count the number of reads mapping to one or more copies of a repetitive element. Original script was written by Hamish King and edited for our use.

```
#!/usr/bin/perl
use Getopt::Long;
use Pod::Usage;
use Data::Dumper;
use Parallel::ForkManager;
use List::Util qw/max/;
use strict;

pod2usage("\nUsage: -rmsk <rmsk file from UCSC> -genome <mm9, hg19> -list <bamfile list> \n") if (($#ARGV<0)
&& (-t STDIN));

&GetOptions ("rmsk=s"=> \my $rmskfile,"genome=s"=> \my $genome,"list=s"=> \my $bamfilelist);

my $outfilerepnamerpk = "$genome.rmsk.RepName.RPK.txt";
my $outfilerepnamerpkm = "$genome.rmsk.RepName.RPKM.txt";
my $outfilerepnamereadcount = "$genome.rmsk.RepName.ReadCount.txt";
my $dollar12 = '$12';my $dollar13 = '$13';my $dollar11 = '$11';my $dollar2 = '$2';my $dollar3 = '$3';
my $fork= new Parallel::ForkManager(25);
my $max_processors = 25;
`mkdir TMP/`;
`mkdir TMP.OUTPUT/`;

#Split repnames into separate files for parallelization
   `cut -f 11,12,13 $rmskfile | sort | uniq > $genome.rmsk.RepName.list.txt`;

   my $number_of_entries_in_your_dataset = `cut -f 1 $genome.rmsk.RepName.list.txt | wc -l`;
   my $number_split = int($number_of_entries_in_your_dataset/$max_processors);
   my $split_Files = `split -d -l $number_split $genome.rmsk.RepName.list.txt TMP/tmp`;


##Extract full list of bam files and compile filenames and counts into arrays

   open(BAMFILELIST, "$bamfilelist") or die "Could not open $bamfilelist";

         my @bamnamearray = ();
         my @bamfilearray = ();
         my @bamfilesizearray = ();

         while(my $line = <BAMFILELIST>){
                  chomp $line;
                  my ($bamfile,$bamname,$bamfilecount)=split (/\t+/,$line);

push(@bamnamearray,$bamname);push(@bamfilearray,$bamfile);push(@bamfilesizearray,$bamfilecount);
                  print "$bamname\t$bamfilecount\n"
         }
   close BAMFILELIST;

##

my $arraySize = @bamfilearray;

open(REPNAMEOUTRPK, ">$outfilerepnamerpk") or die "Could not open $outfilerepnamerpk";
open(REPNAMEOUTRPKM, ">$outfilerepnamerpkm") or die "Could not open $outfilerepnamerpkm";
open(REPNAMEOUTREADCOUNT, ">$outfilerepnamereadcount") or die "Could not open
$outfilerepnamereadcount";

print REPNAMEOUTRPK
"RepName\tRepClass\tRepType\tSize\tRepeats\tMedianCpGDensity\tMeanCpGDensity\tStDevCpGDensity\tStEr
```

```
rCpGDensity\tmilliDiv.Median\tmilliDiv.Mean\tmilliDiv.StDev\tmilliDiv.StErr\t";print REPNAMEOUTRPK join("\t",
@bamnamearray);print REPNAMEOUTRPK "\n";
print REPNAMEOUTRPKM
"RepName\tRepClass\tRepType\tSize\tRepeats\tMedianCpGDensity\tMeanCpGDensity\tStDevCpGDensity\tStEr
rCpGDensity\tmilliDiv.Median\tmilliDiv.Mean\tmilliDiv.StDev\tmilliDiv.StErr\t";print REPNAMEOUTRPKM
join("\t", @bamnamearray);print REPNAMEOUTRPKM "\n";
print REPNAMEOUTREADCOUNT
"RepName\tRepClass\tRepType\tSize\tRepeats\tMedianCpGDensity\tMeanCpGDensity\tStDevCpGDensity\tStEr
rCpGDensity\tmilliDiv.Median\tmilliDiv.Mean\tmilliDiv.StDev\tmilliDiv.StErr\t";print REPNAMEOUTREADCOUNT
join("\t", @bamnamearray);print REPNAMEOUTREADCOUNT "\n";

opendir(DIRECTORY, "TMP/") or die $!; ### open the directory

my $superfork= new Parallel::ForkManager($max_processors);

while (my $file = readdir(DIRECTORY)) { ### read the directory
   if($file=~/^\./){next;
   }

########## Start fork ##########
my $pid= $superfork->start and next;

##Fork process

   #Open repname lists and output bed file
   open(REPNAME, "TMP/$file") or die "Could not open TMP/$file";
   open(TEMPREPNAMEOUTRPK, ">TMP.OUTPUT/$outfilerepnamerpk.$file") or die "Could not open
TMP.OUTPUT/$outfilerepnamerpk.$file";
   open(TEMPREPNAMEOUTRPKM, ">TMP.OUTPUT/$outfilerepnamerpkm.$file") or die "Could not open
TMP.OUTPUT/$outfilerepnamerpkm.$file";
   open(TEMPREPNAMEOUTREADCOUNT, ">TMP.OUTPUT/$outfilerepnamereadcount.$file") or die "Could not
open TMP.OUTPUT/$outfilerepnamereadcount.$file";

      while(my $repnameline = <REPNAME>){
         chomp $repnameline;
         my ($name,$class,$type)=split (/\t+/,$repnameline);
         `awk '$dollar11 == "$name"' $rmskfile | cut -f 6,7,8 > TMP/$name.bed`;

         ##Count number of elements
            my $numberofrepeatscount = `wc -l TMP/$name.bed`;chomp $numberofrepeatscount;
               my ($numberofrepeats,$bed)=split (/\s/,$numberofrepeatscount);

         #Calculate total size of repclass intervals
            my $size =`cat TMP/$name.bed | awk -F'\t' 'BEGIN{SUM=0}{ SUM+=$dollar3-$dollar2 }END{print
SUM}'`;
                  chomp $size;
                  my $sizekb = $size/1000;

            #Generate Average CpG Density counts
               my $commandinterval1 = `bedtools nuc -pattern CG -C -fi
/databank/igenomes/Mus_musculus/UCSC/$genome/Sequence/WholeGenomeFasta/genome.fa -bed
TMP/$name.bed | grep -v seq_len | awk '{density = $dollar13 / $dollar12; print density}' | st -complete | sed -n
2p`;
               chomp $commandinterval1;
               my ($CGnumber, $CGmin, $CGq1, $CGmedian, $CGq3, $CGmax, $CGsum, $CGmean, $CGstddev,
$CGstderr)=split (/\t+/,$commandinterval1);

            #Calculate Substitution rates
               my $command = `awk '$dollar11 == "$name"' $rmskfile | cut -f 3 | st --complete | sed -n 2p`;
               chomp $command;
               my ($Subnumber, $Submin, $Subq1, $Submedian, $Subq3, $Submax, $Subsum, $Submean,
$Substddev, $Substderr)=split (/\t+/,$command);
```

```
        #Compile info into array and print out to temp output files
            my @dataline_rpk =
("$name\t$class\t$type\t$size\t$numberofrepeats\t$CGmedian\t$CGmean\t$CGstddev\t$CGstderr\t$Submedi
an\t$Submean\t$Substddev\t$Substderr");
            my @dataline_rpkm =
("$name\t$class\t$type\t$size\t$numberofrepeats\t$CGmedian\t$CGmean\t$CGstddev\t$CGstderr\t$Submedi
an\t$Submean\t$Substddev\t$Substderr");
            my @dataline_readcount =
("$name\t$class\t$type\t$size\t$numberofrepeats\t$CGmedian\t$CGmean\t$CGstddev\t$CGstderr\t$Submedi
an\t$Submean\t$Substddev\t$Substderr");

        ##Generate Bam counts
            foreach(my $i = 0; $i <= $arraySize - 1; $i += 1){
               my $bamfilecount = `samtools view -c $bamfilearray[$i] -L TMP/$name.bed`;
               chomp $bamfilecount;
               my $bamfilesize =  $bamfilesizearray[$i]/1000000;

               my $bamfilerpk = $bamfilecount/$sizekb;
               my $bamfilerpkm = $bamfilerpk/$bamfilesize;

               push(@dataline_rpk,"$bamfilerpk");
               push(@dataline_rpkm,"$bamfilerpkm");
               push(@dataline_readcount,"$bamfilecount");
            }

            print TEMPREPNAMEOUTRPK join("\t", @dataline_rpk), "\n";
            print TEMPREPNAMEOUTRPKM join("\t", @dataline_rpkm), "\n";
            print TEMPREPNAMEOUTREADCOUNT join("\t", @dataline_readcount), "\n";
    }

######### end fork ###############
$superfork->finish;
}
$superfork->wait_all_children;

`cat TMP.OUTPUT/*ReadCount* | sort -k 1 >> $outfilerepnamereadcount`;
`cat TMP.OUTPUT/*RPKM* | sort -k 1 >> $outfilerepnamerpkm`;
`cat TMP.OUTPUT/*RPK.txt* | sort -k 1 >> $outfilerepnamerpk`;

#Delete folder with temp files
`rm -r TMP`;
`rm -r TMP.OUTPUT`;

############################
```

Chapter 8:

Bibliography

Aapola, U., Shibuya, K., Scott, H.S., Ollila, J., Vihinen, M., Heino, M., Shintani, A., Kawasaki, K., Minoshima, S., Krohn, K., et al. (2000). Isolation and Initial Characterization of a Novel Zinc Finger Gene, DNMT3L, on 21q22.3, Related to the Cytosine-5- Methyltransferase 3 Gene Family. Genomics *65*, 293–298.

Allevato, M., Bolotin, E., Grossman, M., Mane-Padros, D., Sladek, F.M., and Martinez, E. (2017). Sequence-specific DNA binding by MYC/MAX to low-affinity non-E-box motifs. PloS One *12*, e0180147.

Allis, C.D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nat. Rev. Genet. *17*, 487–500.

Ambrosi, C., Manzo, M., and Baubec, T. (2017). Dynamics and Context-Dependent Roles of DNA Methylation. J. Mol. Biol. *429*, 1459–1475.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat. Genet. *23*, 185–188.

Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Fejes Toth, K., Bestor, T., and Hannon, G.J. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol. Cell *31*, 785–799.

Arber, W. (1974). DNA Modification and Restriction. In Progress in Nucleic Acid Research and Molecular Biology, W.E. Cohn, ed. (Academic Press), pp. 1–37.

Auclair, G., Guibert, S., Bender, A., and Weber, M. (2014). Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. Genome Biol. *15*, 545.

Baets, J., Duan, X., Wu, Y., Smith, G., Seeley, W.W., Mademan, I., McGrath, N.M., Beadell, N.C., Khoury, J., Botuyan, M.-V., et al. (2015). Defects of mutant DNMT1 are linked to a spectrum of neurological disorders. Brain J. Neurol. *138*, 845–861.

Barau, J., Teissandier, A., Zamudio, N., Roy, S., Nalesso, V., Hérault, Y., Guillou, F., and Bourc'his, D. (2016). The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. Science *354*, 909–912.

Barlow, D.P., and Bartolomei, M.S. (2014). Genomic Imprinting in Mammals. Cold Spring Harb. Perspect. Biol. *6*, a018382.

Bartels, S.J.J., Spruijt, C.G., Brinkman, A.B., Jansen, P.W.T.C., Vermeulen, M., and Stunnenberg, H.G. (2011). A SILAC-based screen for Methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein. PloS One *6*, e25884.

Batchelor, A.H., Piper, D.E., Brousse, F.C. de la, McKnight, S.L., and Wolberger, C. (1998). The Structure of GABPα/β: An ETS Domain- Ankyrin Repeat Heterodimer Bound to DNA. Science *279*, 1037–1041.

Baubec, T., Ivánek, R., Lienert, F., and Schübeler, D. (2013). Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family. Cell *153*, 480–492.

Baubec, T., Colombo, D.F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A.R., Akalin, A., and Schübeler, D. (2015). Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature *advance online publication*.

Beard, C., Li, E., and Jaenisch, R. (1995). Loss of methylation activates Xist in somatic but not in embryonic cells. Genes Dev. *9*, 2325–2334.

Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482–485.

Bestor, T.H., and Ingram, V.M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. Proc. Natl. Acad. Sci. U. S. A. *80*, 5559–5563.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells: The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. J. Mol. Biol. *203*, 971–983.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. *8*, 1499–1504.

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91–99.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., 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., Farcas, A.M., Kondo, T., King, H.W., McGouran, J.F., Hanssen, L.L.P., Ito, S., Cooper, S., Kondo, K., Koseki, Y., et al. (2014). Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. Cell *157*, 1445–1459.

Blackwood, E.M., Lüscher, B., and Eisenman, R.N. (1992). Myc and Max associate in vivo. Genes Dev. *6*, 71–80.

Blattler, A., Yao, L., Wang, Y., Ye, Z., Jin, V.X., and Farnham, P.J. (2013). ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes. Epigenetics Chromatin *6*, 13.

Bogdanović, O., and Veenstra, G.J.C. (2009). DNA methylation and methyl-CpG binding proteins: developmental requirements and function. Chromosoma *118*, 549–565.

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., Delorenzi, M., et al. (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. *24*, 1157–1168.

Borgel, J., Guibert, S., Li, Y., Chiba, H., Schübeler, D., Sasaki, H., Forné, T., and Weber, M. (2010). Targets and dynamics of promoter DNA methylation during early mouse development. Nat. Genet. *42*, 1093–1100.

Bošković, A., Eid, A., Pontabry, J., Ishiuchi, T., Spiegelhalter, C., Raghu Ram, E.V.S., Meshorer, E., and Torres-Padilla, M.-E. (2014). Higher chromatin mobility supports totipotency and precedes pluripotency in vivo. Genes Dev. *28*, 1042–1047.

Bostick, M., Kim, J.K., Estève, P.-O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. Science *317*, 1760–1764.

Bottardi, S., Aumont, A., Grosveld, F., and Milot, E. (2003). Developmental stage-specific epigenetic control of human beta-globin gene expression is potentiated in hematopoietic progenitor cells prior to their transcriptional activation. Blood *102*, 3989–3997.

Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature *431*, 96–99.

Bourc'his, D., Xu, G.-L., Lin, C.-S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the Establishment of Maternal Genomic Imprints. Science *294*, 2536–2539.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311–322.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Names, A., Temper, V., Razin, A., and Cedar, H. (1994). Spl elements protect a CpG island from de novo methylation. Nature *371*, 435–438.

Brinkman, A.B., Gu, H., Bartels, S.J.J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., et al. (2012). Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res. *22*, 1128–1138.

Brockdorff, N. (2011). Chromosome silencing mechanisms in X-chromosome inactivation: unknown unknowns. Development *138*, 5057–5065.

Brocks, D., Schmidt, C.R., Daskalakis, M., Jang, H.S., Shah, N.M., Li, D., Li, J., Zhang, B., Hou, Y., Laudato, S., et al. (2017). DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. Nat. Genet. *49*, 1052–1060.

Bröske, A.-M., Vockentanz, L., Kharazi, S., Huska, M.R., Mancini, E., Scheller, M., Kuhl, C., Enns, A., Prinz, M., Jaenisch, R., et al. (2009). DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. Nat. Genet. *41*, 1207–1215.

Brunmeir, R., Lagger, S., Simboeck, E., Sawicka, A., Egger, G., Hagelkruys, A., Zhang, Y., Matthias, P., Miller, W.J., and Seiser, C. (2010). Epigenetic Regulation of a Murine Retrotransposon by a Dual Histone Modification Mark. PLOS Genet. *6*, e1000927.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Busslinger, M., Hurst, J., and Flavell, R.A. (1983). DNA methylation and the regulation of globin gene expression. Cell *34*, 197–206.

Cameron, E.E., Bachman, K.E., Myöhänen, S., Herman, J.G., and Baylin, S.B. (1999). Synergy of demethylation and histone deacetylase inhibition in the re-expression of genes silenced in cancer. Nat. Genet. *21*, 103–107.

Campanero, M.R., Armstrong, M.I., and Flemington, E.K. (2000). CpG methylation as a mechanism for the regulation of E2F activity. Proc. Natl. Acad. Sci. *97*, 6481–6486.

Challen, G.A., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J.S., Bock, C., Vasanthakumar, A., Gu, H., Xi, Y., et al. (2012). Dnmt3a is essential for hematopoietic stem cell differentiation. Nat. Genet. *44*, 23–31.

Chen, Z., and Riggs, A.D. (2011). DNA Methylation and Demethylation in Mammals. J. Biol. Chem. *286*, 18347–18353.

Chen, K., Xi, Y., Pan, X., Li, Z., Kaestner, K., Tyler, J., Dent, S., He, X., and Li, W. (2013). DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res. *23*, 341–351.

Chen, T., Ueda, Y., Dodge, J.E., Wang, Z., and Li, E. (2003). Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b. Mol. Cell. Biol. *23*, 5594–5605.

Chiurazzi, P., Pomponi, M.G., Willemsen, R., Oostra, B.A., and Neri, G. (1998). In Vitro Reactivation of the FMR1 Gene Involved in Fragile X Syndrome. Hum. Mol. Genet. *7*, 109–113.

Clark, S.J., Harrison, J., and Molloy, P.L. (1997). Sp1 binding is inhibited by mCpmCpG methylation. Gene *195*, 67–71.

Clark, S.J., Lee, H.J., Smallwood, S.A., Kelsey, G., and Reik, W. (2016). Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. Genome Biol. *17*.

Coffee, B., Zhang, F., Warren, S.T., and Reines, D. (1999). Acetylated histones are associated with FMR1 in normal but not fragile X-syndrome cells. Nat. Genet. *22*, 98–101.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215–219.

Collings, C.K., Waddell, P.J., and Anderson, J.N. (2013). Effects of DNA methylation on nucleosome stability. Nucleic Acids Res. *41*, 2918–2931.

Consortium, M.G.S. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520.

Cooper, D.N., Taggart, M.H., and Bird, A.P. (1983). Unmethylated domains in vertebrate DNA. Nucleic Acids Res. *11*, 647–658.

Cotton, A.M., Price, E.M., Jones, M.J., Balaton, B.P., Kobor, M.S., and Brown, C.J. (2015). Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. Hum. Mol. Genet. *24*, 1528–1539.

Dan, J., Yang, J., Liu, Y., Xiao, A., and Liu, L. (2015). Roles for Histone Acetylation in Regulation of Telomere Elongation and Two-cell State in Mouse ES Cells. J. Cell. Physiol. *230*, 2337–2344.

Daskalos, A., Nikolaidis, G., Xinarianos, G., Savvari, P., Cassidy, A., Zakopoulou, R., Kotsinas, A., Gorgoulis, V., Field, J.K., and Liloglou, T. (2009). Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. Int. J. Cancer *124*, 81–87.

Davis, C.M., Constantinides, P.G., van der Riet, F., van Schalkwyk, L., Gevers, W., and Parker, M.I. (1989). Activation and demethylation of the intracisternal A particle genes by 5-azacytidine. Cell Differ. Dev. Off. J. Int. Soc. Dev. Biol. *27*, 83–93.

De Smet, C., Lurquin, C., Lethé, B., Martelange, V., and Boon, T. (1999). DNA methylation is the primary silencing mechanism for a set of germ line- and tumor-specific genes with a CpG-rich promoter. Mol. Cell. Biol. *19*, 7327–7335.

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

Delhommeau, F., Dupont, S., Valle, V.D., James, C., Trannoy, S., Massé, A., Kosmider, O., Le Couedic, J.-P., Robert, F., Alberdi, A., et al. (2009). Mutation in TET2 in Myeloid Cancers. N. Engl. J. Med. *360*, 2289–2301.

Dhalluin, C., Carlson, J.E., Zeng, L., He, C., Aggarwal, A.K., and Zhou, M.M. (1999). Structure and ligand of a histone acetyltransferase bromodomain. Nature *399*, 491–496.

Dodge, J.E., Okano, M., Dick, F., Tsujimoto, N., Chen, T., Wang, S., Ueda, Y., Dyson, N., and Li, E. (2005). Inactivation of Dnmt3b in Mouse Embryonic Fibroblasts Results in DNA Hypomethylation, Chromosomal Instability, and Spontaneous Immortalization. J. Biol. Chem. *280*, 17986–17991.

Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *advance online publication*.

Doskočil, J., and Šormová, Z. (1965). The occurrence of 5-methylcytosine in bacterial deoxyribonucleic acids. Biochim. Biophys. Acta BBA - Nucleic Acids Protein Synth. *95*, 513–515.

Duncan, B.K., and Miller, J.H. (1980). Mutagenic deamination of cytosine residues in DNA. Nature *287*, 560–561.

Dunn, D.B., and Smith, J.D. (1958). The occurrence of 6-methylaminopurine in deoxyribonucleic acids. Biochem. J. *68*, 627–636.

Eckersley-Maslin, M.A., Svensson, V., Krueger, C., Stubbs, T.M., Giehr, P., Krueger, F., Miragaia, R.J., Kyriakopoulos, C., Berrens, R.V., Milagre, I., et al. (2016). MERVL/Zscan4

Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. Cell Rep. *17*, 179–192.

Eden, A., Gaudet, F., Waghmare, A., and Jaenisch, R. (2003). Chromosomal Instability and Tumors Promoted by DNA Hypomethylation. Science *300*, 455–455.

Edwards, J.R., Yarychkivska, O., Boulard, M., and Bestor, T.H. (2017). DNA methylation and DNA methyltransferases. Epigenetics Chromatin *10*, 23.

Egan, B., Yuan, C.-C., Craske, M.L., Labhart, P., Guler, G.D., Arnott, D., Maile, T.M., Busby, J., Henry, C., Kelly, T.K., et al. (2016). An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon EZH2 Inhibition. PloS One *11*, e0166438.

Ehrlich, M., Gama-Sosa, M.A., Carreira, L.H., Ljungdahl, L.G., Kuo, K.C., and Gehrke, C.W. (1985). DNA methylation in thermophilic bacteria: N4-methylcytosine, 5-methylcytosine, and N6-methyladenine. Nucleic Acids Res. *13*, 1399–1412.

Endoh, M., Endo, T.A., Shinga, J., Hayashi, K., Farcas, A., Ma, K.-W., Ito, S., Sharif, J., Endoh, T., Onaga, N., et al. (2017). PCGF6-PRC1 suppresses premature differentiation of mouse embryonic stem cells by regulating germ cell-related genes. ELife *6*, e21064.

Fang, M., Ou, J., Hutchinson, L., and Green, M.R. (2014). The BRAF oncoprotein functions through the transcriptional repressor MAFG to mediate the CpG Island Methylator phenotype. Mol. Cell *55*, 904–915.

Farthing, C.R., Ficz, G., Ng, R.K., Chan, C.-F., Andrews, S., Dean, W., Hemberger, M., and Reik, W. (2008). Global Mapping of DNA Methylation in Mouse Promoters Reveals Epigenetic Reprogramming of Pluripotency Genes. PLoS Genet. *4*.

Feldmann, A., Ivánek, R., Murr, R., Gaidatzis, D., Burger, L., and Schübeler, D. (2013). Transcription Factor Occupancy Can Mediate Active Turnover of DNA Methylation at Regulatory Regions. PLoS Genet *9*, e1003994.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature *473*, 398–402.

Figueroa, M.E., Wahab, O.A., Lu, C., Ward, P.S., Patel, J., Shih, A., Li, Y., Bhagwat, N., Vasanthakumar, A., Fernandez, H.F., et al. (2010). Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell *18*, 553–567.

Filion, G.J.P., Zhenilo, S., Salozhin, S., Yamada, D., Prokhortchouk, E., and Defossez, P.-A. (2006). A family of human zinc finger proteins that bind methylated DNA and repress transcription. Mol. Cell. Biol. *26*, 169–181.

Fouse, S.D., Shen, Y., Pellegrini, M., Cole, S., Meissner, A., Van Neste, L., Jaenisch, R., and Fan, G. (2008). Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, Polycomb binding and histone H3 lys4/lys27 trimethylation. Cell Stem Cell *2*, 160–169.

Franco, P.J., Farooqui, M., Seto, E., and Wei, L.N. (2001). The orphan nuclear receptor TR2 interacts directly with both class I and class II histone deacetylases. Mol. Endocrinol. Baltim. Md *15*, 1318–1328.

Frank, C.L., Manandhar, D., Gordân, R., and Crawford, G.E. (2016). HDAC inhibitors cause site-specific chromatin remodeling at PU.1-bound enhancers in K562 cells. Epigenetics Chromatin *9*, 15.

Friedli, M., and Trono, D. (2015). The Developmental Control of Transposable Elements and the Evolution of Higher Species. Annu. Rev. Cell Dev. Biol. *31*, 429–451.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. U. S. A. *89*, 1827–1831.

Fujita, N., Watanabe, S., Ichimura, T., Tsuruzoe, S., Shinkai, Y., Tachibana, M., Chiba, T., and Nakao, M. (2003). Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. J. Biol. Chem. *278*, 24132–24138.

Gallinari, P., Di Marco, S., Jones, P., Pallaoro, M., and Steinkühler, C. (2007). HDACs, histone deacetylation and gene transcription: from molecular biology to cancer therapeutics. Cell Res. *17*, 195–211.

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

Gaston, K., and Fried, M. (1995). CpG methylation has differential effects on the binding of YY1 and ETS proteins to the bi-directional promoter of the Surf-1 and Surf-2 genes. Nucleic Acids Res. *23*, 901–909.

Gaudet, F., Hodgson, J.G., Eden, A., Jackson-Grusby, L., Dausman, J., Gray, J.W., Leonhardt, H., and Jaenisch, R. (2003). Induction of Tumors in Mice by Genomic Hypomethylation. Science *300*, 489–492.

Gaudet, F., Rideout, W.M., Meissner, A., Dausman, J., Leonhardt, H., and Jaenisch, R. (2004). Dnmt1 Expression in Pre- and Postimplantation Embryogenesis and the Maintenance of IAP Silencing. Mol. Cell. Biol. *24*, 1640–1648.

Glozak, M.A., Sengupta, N., Zhang, X., and Seto, E. (2005). Acetylation and deacetylation of non-histone proteins. Gene *363*, 15–23.

Goll, M.G., and Bestor, T.H. (2005). Eukaryotic Cytosine Methyltransferases. Annu. Rev. Biochem. *74*, 481–514.

Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.-L., Zhang, X., Golic, K.G., Jacobsen, S.E., and Bestor, T.H. (2006). Methylation of tRNAAsp by the DNA Methyltransferase Homolog Dnmt2. Science *311*, 395–398.

Goodier, J.L. (2016). Restricting retrotransposons: a review. Mob. DNA *7*.

Gordon, S., Akopyan, G., Garban, H., and Bonavida, B. (2005). Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. Oncogene *25*, 1125–1142.

Gross, D.S., and Garrard, W.T. (1988). Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. *57*, 159–197.

Guallar, D., Pérez-Palacios, R., Climent, M., Martínez-Abadía, I., Larraga, A., Fernández-Juan, M., Vallejo, C., Muniesa, P., and Schoorlemmer, J. (2012). Expression of endogenous retroviruses is negatively regulated by the pluripotency marker Rex1/Zfp42. Nucleic Acids Res. *40*, 8993–9007.

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.

Gugneja, S., and Scarpulla, R.C. (1997). Serine Phosphorylation within a Concise Amino-terminal Domain in Nuclear Respiratory Factor 1 Enhances DNA Binding. J. Biol. Chem. *272*, 18732–18739.

Guibert, S., Forné, T., and Weber, M. (2012). Global profiling of DNA methylation erasure in mouse primordial germ cells. Genome Res. *22*, 633–641.

Guo, D., Wu, B., Yan, J., Li, X., Sun, H., and Zhou, D. (2012). A possible gene silencing mechanism: Hypermethylation of the Keap1 promoter abrogates binding of the transcription factor Sp1 in lung cancer cells. Biochem. Biophys. Res. Commun. *428*, 80–85.

Guo, J., Li, T., Schipper, J., Nilson, K.A., Fordjour, F.K., Cooper, J.J., Gordân, R., and Price, D.H. (2014). Sequence specificity incompletely defines the genome-wide occupancy of Myc. Genome Biol. *15*.

Habibi, E., Brinkman, A.B., Arand, J., Kroeze, L.I., Kerstens, H.H.D., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N.C., et al. (2013). Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. Cell Stem Cell *13*, 360–369.

Hainer, S.J., McCannell, K.N., Yu, J., Ee, L.-S., Zhu, L.J., Rando, O.J., and Fazzio, T.G. (2016). DNA methylation directs genomic localization of Mbd2 and Mbd3 in embryonic stem cells. ELife *5*, e21964.

Hansen, R.S., Canfield, T.K., Fjeld, A.D., and Gartler, S.M. (1996). Role of late replication timing in the silencing of X-linked genes. Hum. Mol. Genet. *5*, 1345–1353.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., and Tilghman, S.M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature *405*, 486–489.

Harrington, M.A., Jones, P.A., Imagawa, M., and Karin, M. (1988). Cytosine methylation does not affect binding of transcription factor Sp1. Proc. Natl. Acad. Sci. U. S. A. *85*, 2066–2070.

Hartman, H.B., Yu, J., Alenghat, T., Ishizuka, T., and Lazar, M.A. (2005). The histone-binding code of nuclear receptor co-repressors matches the substrate specificity of histone deacetylase 3. EMBO Rep. *6*, 445–451.

Hashimoto, H., Liu, Y., Upadhyay, A.K., Chang, Y., Howerton, S.B., Vertino, P.M., Zhang, X., and Cheng, X. (2012). Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic Acids Res. *40*, 4841–4849.

Hassig, C.A., Fleischer, T.C., Billin, A.N., Schreiber, S.L., and Ayer, D.E. (1997). Histone deacetylase activity is required for full transcriptional repression by mSin3A. Cell *89*, 341–347.

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

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.

Hendrich, B., and Bird, A. (1998). Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. Mol. Cell. Biol. *18*, 6538–6547.

Hendrich, B., Hardeland, U., Ng, H.-H., Jiricny, J., and Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature *401*, 301–304.

Hezroni, H., Sailaja, B.S., and Meshorer, E. (2011). Pluripotency-related, Valproic Acid (VPA)-induced Genome-wide Histone H3 Lysine 9 (H3K9) Acetylation Patterns in Embryonic Stem Cells. J. Biol. Chem. *286*, 35977–35988.

Hill, P.W.S., Amouroux, R., and Hajkova, P. (2014). DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: An emerging complex story. Genomics *104*, 324–333.

Hill, P.W.S., Leitch, H.G., Requena, C.E., Sun, Z., Amouroux, R., Roman-Trufero, M., Borkowska, M., Terragni, J., Vaisvila, R., Linnett, S., et al. (2018). Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. Nature.

Hinoue, T., Weisenberger, D.J., Lange, C.P.E., Shen, H., Byun, H.-M., Berg, D.V.D., Malik, S., Pan, F., Noushmehr, H., Dijk, C.M. van, et al. (2012). Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res. *22*, 271–282.

Hirasawa, R., Chiba, H., Kaneda, M., Tajima, S., Li, E., Jaenisch, R., and Sasaki, H. (2008). Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. Genes Dev. *22*, 1607–1616.

Hollenhorst, P.C., Ferris, M.W., Hull, M.A., Chae, H., Kim, S., and Graves, B.J. (2011). Oncogenic ETS proteins mimic activated RAS/MAPK signaling in prostate cells. Genes Dev. *25*, 2147–2157.

Höller, M., Westin, G., Jiricny, J., and Schaffner, W. (1988). Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated. Genes Dev. *2*, 1127–1135.

Holliday, R., and Pugh, J.E. (1975). DNA modification mechanisms and gene activity during development. Science *187*, 226–232.

Hon, G.C., Rajagopal, N., Shen, Y., McCleary, D.F., Yue, F., Dang, M.D., and Ren, B. (2013). Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat. Genet. *45*, 1198–1206.

Hotchkiss, R.D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. J. Biol. Chem. *175*, 315–332.

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H.N., Shin, J., Cox, E., Rho, H.S., Woodard, C., et al. (2013). DNA methylation presents distinct binding sites for human transcription factors. ELife *2*, e00726.

Hu, T., Zhu, X., Pi, W., Yu, M., Shi, H., and Tuan, D. (2017). Hypermethylated LTR retrotransposon exhibits enhancer activity. Epigenetics *12*, 226–237.

Huang, E.Y., Zhang, J., Miska, E.A., Guenther, M.G., Kouzarides, T., and Lazar, M.A. (2000). Nuclear receptor corepressors partner with class II histone deacetylases in a Sin3-independent repression pathway. Genes Dev. *14*, 45–54.

Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science *339*, 957–959.

Huff, J.T., and Zilberman, D. (2014). Dnmt1-Independent CG Methylation Contributes to Nucleosome Positioning in Diverse Eukaryotes. Cell *156*, 1286–1297.

Hughes, L.A.E., Melotte, V., Schrijver, J. de, Maat, M. de, Smit, V.T.H.B.M., Bovée, J.V.M.G., French, P.J., Brandt, P.A. van den, Schouten, L.J., Meyer, T. de, et al. (2013). The CpG Island Methylator Phenotype: What's in a Name? Cancer Res. *73*, 5858–5868.

Iguchi-Ariga, S.M., and Schaffner, W. (1989). CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes Dev. *3*, 612–619.

Illingworth, R.S., and Bird, A.P. (2009). CpG islands--'a rough guide'. FEBS Lett. *583*, 1713–1720.

Ishiuchi, T., Enriquez-Gasca, R., Mizutani, E., Bošković, A., Ziegler-Birling, C., Rodriguez-Terrones, D., Wakayama, T., Vaquerizas, J.M., and Torres-Padilla, M.-E. (2015). Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. Nat. Struct. Mol. Biol. *22*, 662–671.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. Science *333*, 1300–1303.

Iurlaro, M., von Meyenn, F., and Reik, W. (2017). DNA methylation homeostasis in human and mouse development. Curr. Opin. Genet. Dev. *43*, 101–109.

Iyer, L.M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. Cell Cycle Georget. Tex *8*, 1698–1710.

Jackson, M., Krassowska, A., Gilbert, N., Chevassut, T., Forrester, L., Ansell, J., and Ramsahoye, B. (2004). Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. Mol. Cell. Biol. *24*, 8862–8871.

Jackson, V., Shires, A., Tanphaichitr, N., and Chalkley, R. (1976). Modifications to histones immediately after synthesis. J. Mol. Biol. *104*, 471–483.

Jackson-Grusby, L., Beard, C., Possemato, R., Tudor, M., Fambrough, D., Csankovszki, G., Dausman, J., Lee, P., Wilson, C., Lander, E., et al. (2001). Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. Nat. Genet. *27*, 31–39.

Jähner, D., and Jaenisch, R. (1985). Retrovirus-induced de novo methylation of flanking host sequences correlates with gene inactivity. Nature *315*, 594–597.

Jones, B., and Chen, J. (2006). Inhibition of IFN-? transcription by site-specific methylation during T helper cell development. EMBO J. *25*, 2443–2452.

Jones, P.L., Veenstra, G.J., Wade, P.A., Vermaak, D., Kass, S.U., Landsberger, N., Strouboulis, J., and Wolffe, A.P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat. Genet. *19*, 187–191.

Jørgensen, H.F., Ben-Porath, I., and Bird, A.P. (2004). Mbd1 Is Recruited to both Methylated and Nonmethylated CpGs via Distinct DNA Binding Domains. Mol. Cell. Biol. *24*, 3387–3395.

Josse, J., Kaiser, A.D., and Kornberg, A. (1961). Enzymatic Synthesis of Deoxyribonucleic Acid VIII. FREQUENCIES OF NEAREST NEIGHBOR BASE SEQUENCES IN DEOXYRIBONUCLEIC ACID. J. Biol. Chem. *236*, 864–875.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

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

Karimi, M.M., Goyal, P., Maksakova, I.A., Bilenky, M., Leung, D., Tang, J.X., Shinkai, Y., Mager, D.L., Jones, S., Hirst, M., et al. (2011). DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements and chimaeric transcripts in mouse ES cells. Cell Stem Cell *8*.

Kato, Y., Kaneda, M., Hata, K., Kumaki, K., Hisano, M., Kohara, Y., Okano, M., Li, E., Nozaki, M., and Sasaki, H. (2007). Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. Hum. Mol. Genet. *16*, 2272–2280.

Keshet, I., Lieman-Hurwitz, J., and Cedar, H. (1986). DNA methylation affects the formation of active chromatin. Cell *44*, 535–543.

Kim, J., Kollhoff, A., Bergmann, A., and Stubbs, L. (2003). Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. Hum. Mol. Genet. *12*, 233–245.

Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

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

King, A.D., Huang, K., Rubbi, L., Liu, S., Wang, C.-Y., Wang, Y., Pellegrini, M., and Fan, G. (2016). Reversible Regulation of Promoter and Enhancer Histone Landscape by DNA Methylation in Mouse Embryonic Stem Cells. Cell Rep. *17*, 289–302.

Klose, R.J., Sarraf, S.A., Schmiedeberg, L., McDermott, S.M., Stancheva, I., and Bird, A.P. (2005). DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. Mol. Cell *19*, 667–678.

Klose, R.J., Cooper, S., Farcas, A.M., Blackledge, N.P., and Brockdorff, N. (2013). Chromatin sampling--an emerging perspective on targeting polycomb repressor proteins. PLoS Genet. *9*, e1003717.

Kokura, K., Kaul, S.C., Wadhwa, R., Nomura, T., Khan, M.M., Shinagawa, T., Yasukawa, T., Colmenares, C., and Ishii, S. (2001). The Ski protein family is required for MeCP2-mediated transcriptional repression. J. Biol. Chem. *276*, 34115–34121.

Koziol, M.J., Bradshaw, C.R., Allen, G.E., Costa, A.S.H., Frezza, C., and Gurdon, J.B. (2016). Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. Nat. Struct. Mol. Biol. *23*, 24–30.

Kriaucionis, S., and Heintz, N. (2009). The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. Science *324*, 929–930.

Kribelbauer, J.F., Laptenko, O., Chen, S., Martini, G.D., Freed-Pastor, W.A., Prives, C., Mann, R.S., and Bussemaker, H.J. (2017). Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. Cell Rep. *19*, 2383–2395.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., et al. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet. *4*, e1000242.

Kumari, D., and Usdin, K. (2001). Interaction of the Transcription Factors USF1, USF2, and α-Pal/Nrf-1 with the FMR1 Promoter IMPLICATIONS FOR FRAGILE X MENTAL RETARDATION SYNDROME. J. Biol. Chem. *276*, 4357–4364.

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T.W., et al. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev. *22*, 908–917.

Lalande, M. (1996). Parental imprinting and human disease. Annu. Rev. Genet. *30*, 173–195.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Landolin, J.M., Johnson, D.S., Trinklein, N.D., Aldred, S.F., Medina, C., Shulha, H., Weng, Z., and Myers, R.M. (2010). Sequence features that drive human promoter function and tissue specificity. Genome Res. *20*, 890–898.

Lane, N., Dean, W., Erhardt, S., Hajkova, P., Surani, A., Walter, J., and Reik, W. (2003). Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. Genesis *35*, 88–93.

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

Leeb, M., Pasini, D., Novatchkova, M., Jaritz, M., Helin, K., and Wutz, A. (2010). Polycomb complexes act redundantly to repress genomic repeats and genes. Genes Dev. *24*, 265–276.

Leitch, H.G., McEwen, K.R., Turp, A., Encheva, V., Carroll, T., Grabole, N., Mansfield, W., Nashun, B., Knezovich, J.G., Smith, A., et al. (2013). Naive pluripotency is associated with global DNA hypomethylation. Nat. Struct. Mol. Biol. *20*, 311–316.

Leung, D., Du, T., Wagner, U., Xie, W., Lee, A.Y., Goyal, P., Li, Y., Szulwach, K.E., Jin, P., Lorincz, M.C., et al. (2014). Regulation of DNA methylation turnover at LTR retrotransposons and imprinted loci by the histone methyltransferase Setdb1. Proc. Natl. Acad. Sci. 201322273.

Ley, T.J., Ding, L., Walter, M.J., McLellan, M.D., Lamprecht, T., Larson, D.E., Kandoth, C., Payton, J.E., Baty, J., Welch, J., et al. (2010). DNMT3A Mutations in Acute Myeloid Leukemia. N. Engl. J. Med. *363*, 2424–2433.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

Li, H., Liefke, R., Jiang, J., Kurland, J.V., Tian, W., Deng, P., Zhang, W., He, Q., Patel, D.J., Bulyk, M.L., et al. (2017). Polycomb-like proteins link the PRC2 complex to CpG islands. Nature *549*, 287–291.

Liang, G., Chan, M.F., Tomigahara, Y., Tsai, Y.C., Gonzales, F.A., Li, E., Laird, P.W., and Jones, P.A. (2002). Cooperativity between DNA Methyltransferases in the Maintenance Methylation of Repetitive Elements. Mol. Cell. Biol. *22*, 480–491.

Liao, J., Karnik, R., Gu, H., Ziller, M.J., Clement, K., Tsankov, A.M., Akopian, V., Gifford, C.A., Donaghey, J., Galonska, C., et al. (2015). Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nat. Genet. *advance online publication*.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma. Oxf. Engl. *30*, 923–930.

Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E., Lee, T.I., and Young, R.A. (2012). Transcriptional Amplification in Tumor Cells with Elevated c-Myc. Cell *151*, 56–67.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science *341*, 1237905.

Liu, J., Zhu, Y., Luo, G.-Z., Wang, X., Yue, Y., Wang, X., Zong, X., Chen, K., Yin, H., Fu, Y., et al. (2016a). Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. Nat. Commun. *7*, 13052.

Liu, X.S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R.A., and Jaenisch, R. (2016b). Editing DNA Methylation in the Mammalian Genome. Cell *167*, 233-247.e17.

Long, H.K., Blackledge, N.P., and Klose, R.J. (2013). ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. Biochem. Soc. Trans. *41*, 727–740.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lu, F., Liu, Y., Jiang, L., Yamaguchi, S., and Zhang, Y. (2014). Role of Tet proteins in enhancer activity and telomere elongation. Genes Dev. *28*, 2103–2119.

Lucas, M.E., Crider, K.S., Powell, D.R., Kapoor-Vazirani, P., and Vertino, P.M. (2009). Methylation-sensitive regulation of TMS1/ASC by the Ets factor, GA-binding protein-alpha. J. Biol. Chem. *284*, 14698–14709.

Lynch, M.D., Smith, A.J.H., Gobbi, M.D., 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.

Lyon, M.F. (1961). Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature *190*, 372–373.

Lyst, M.J., Ekiert, R., Ebert, D.H., Merusi, C., Nowak, J., Selfridge, J., Guy, J., Kastan, N.R., Robinson, N.D., de Lima Alves, F., et al. (2013). Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. Nat. Neurosci. *16*, 898–902.

Maatouk, D.M., Kellam, L.D., Mann, M.R.W., Lei, H., Li, E., Bartolomei, M.S., and Resnick, J.L. (2006). DNA methylation is a primary mechanism for silencing postmigratory primordial germ cell genes in both germ cell and somatic cell lineages. Development *133*, 3411–3418.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). ES cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57–63.

Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev. *8*, 2282–2292.

Maiti, A., and Drohat, A.C. (2011). Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES. J. Biol. Chem. *286*, 35334–35338.

Maksakova, I.A., Romanish, M.T., Gagnier, L., Dunn, C.A., van de Lagemaat, L.N., and Mager, D.L. (2006). Retroviral Elements and Their Hosts: Insertional Mutagenesis in the Mouse Germ Line. PLoS Genet. *2*.

Maksakova, I.A., Thompson, P.J., Goyal, P., Jones, S.J., Singh, P.B., Karimi, M.M., and Lorincz, M.C. (2013). Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. Epigenetics Chromatin *6*, 15.

Mann, I.K., Chatterjee, R., Zhao, J., He, X., Weirauch, M.T., Hughes, T.R., and Vinson, C. (2013). CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. Genome Res. *23*, 988–997.

Martín Caballero, I., Hansen, J., Leaford, D., Pollard, S., and Hendrich, B.D. (2009). The methyl-CpG binding proteins Mecp2, Mbd2 and Kaiso are dispensable for mouse embryogenesis, but play a redundant function in neural differentiation. PloS One *4*, e4315.

Matsui, T., Leung, D., Miyashita, H., Maksakova, I.A., Miyachi, H., Kimura, H., Tachibana, M., Lorincz, M.C., and Shinkai, Y. (2010). Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature *464*, 927–931.

Mattson, G., Conklin, E., Desai, S., Nielander, G., Savage, M.D., and Morgensen, S. (1993). A practical approach to crosslinking. Mol. Biol. Rep. *17*, 167–183.

Maurano, M.T., Wang, H., John, S., Shafer, A., Canfield, T., Lee, K., and Stamatoyannopoulos, J.A. (2015). Role of DNA Methylation in Modulating Transcription Factor Occupancy. Cell Rep.

McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. *40*, 4288–4297.

McMorrow, T., van den Wijngaard, A., Wollenschlaeger, A., van de Corput, M., Monkhorst, K., Trimborn, T., Fraser, P., van Lohuizen, M., Jenuwein, T., Djabali, M., et al. (2000). Activation of the beta globin locus by transcription factors and chromatin modifiers. EMBO J. *19*, 4986–4996.

Meehan, R.R., Lewis, J.D., McKay, S., Kleiner, E.L., and Bird, A.P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell *58*, 499–507.

Messerschmidt, D.M., Knowles, B.B., and Solter, D. (2014). DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes Dev. *28*, 812–828.

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

Millar, C.B., Guy, J., Sansom, O.J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P.D., Bishop, S.M., Clarke, A.R., and Bird, A. (2002). Enhanced CpG Mutability and Tumorigenesis in MBD4-Deficient Mice. Science *297*, 403–405.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., 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.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development *99*, 371–382.

Mutskov, V., Gerber, D., Angelov, D., Ausio, J., Workman, J., and Dimitrov, S. (1998). Persistent interactions of core histone tails with nucleosomal DNA following acetylation and transcription factor binding. Mol. Cell. Biol. *18*, 6293–6304.

Nan, X., Ng, H.-H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386–389.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. Nature *543*, 72–77.

Ng, H.H., Zhang, Y., Hendrich, B., Johnson, C.A., Turner, B.M., Erdjument-Bromage, H., Tempst, P., Reinberg, D., and Bird, A. (1999). MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. Nat. Genet. *23*, 58–61.

Nickel, J., Short, M.L., Schmitz, A., Eggert, M., and Henkawitz, R. (1995). Methylation of the mouse M-lysozyme downstream enhancer inhibits heterotetrameric GABP binding. Nucleic Acids Res. *23*, 4785–4792.

Oda, M., Yamagiwa, A., Yamamoto, S., Nakayama, T., Tsumura, A., Sasaki, H., Nakao, K., Li, E., and Okano, M. (2006). DNA methylation regulates long-range gene silencing of an X-linked homeobox gene cluster in a lineage-specific manner. Genes Dev. *20*, 3382–3394.

Okano, M., Xie, S., and Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat. Genet. *19*, 219.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell *99*, 247–257.

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., and Guenther, M.G. (2014). Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome. Cell Rep. *9*, 1163–1170.

Otani, J., Nankumo, T., Arita, K., Inamoto, S., Ariyoshi, M., and Shirakawa, M. (2009). Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX–DNMT3–DNMT3L domain. EMBO Rep. *10*, 1235–1241.

Palacios, D., Summerbell, D., Rigby, P.W.J., and Boyes, J. (2010). Interplay between DNA Methylation and Transcription Factor Availability: Implications for Developmental Activation of the Mouse Myogenin Gene. Mol. Cell. Biol. *30*, 3805–3815.

Panning, B., and Jaenisch, R. (1996). DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes Dev. *10*, 1991–2002.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Kazadi, D., Sun, J., Federation, A., Chao, J., Elliott, O., Liu, Z.-P., et al. (2015). RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. Cell *161*, 774–789.

Perdomo-Sabogal, A., Nowick, K., Piccini, I., Sudbrak, R., Lehrach, H., Yaspo, M.-L., Warnatz, H.-J., and Querfurth, R. (2016). Human Lineage-Specific Transcriptional Regulation through GA-Binding Protein Transcription Factor Alpha (GABPa). Mol. Biol. Evol. *33*, 1231–1244.

Perini, G., Diolaiti, D., Porro, A., and Valle, G.D. (2005). In vivo transcriptional regulation of N-Myc target genes is controlled by E-box methylation. Proc. Natl. Acad. Sci. U. S. A. *102*, 12117–12122.

Philipsen, S., and Suske, G. (1999). A tale of three fingers: the family of mammalian Sp/XKLF transcription factors. Nucleic Acids Res. *27*, 2991–3000.

Picelli, S., Björklund, Å.K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively-scaled sequencing projects. Genome Res. gr.177881.114.

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. Nat. Biotechnol. *28*, 1057–1068.

Prendergast, G.C., and Ziff, E.B. (1991). Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. Science *251*, 186–189.

Prokhortchouk, A., Hendrich, B., Jørgensen, H., Ruzov, A., Wilm, M., Georgiev, G., Bird, A., and Prokhortchouk, E. (2001). The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes Dev. *15*, 1613–1618.

Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Baglivo, I., Pedone, P.V., Grimaldi, G., Riccio, A., et al. (2011). In Embryonic Stem Cells, ZFP57/KAP1 Recognize a Methylated Hexanucleotide to Affect Chromatin and DNA Methylation of Imprinting Control Regions. Mol. Cell *44*, 361–372.

Rafehi, H., Balcerczyk, A., Lunke, S., Kaspi, A., Ziemann, M., Kn, H., Okabe, J., Khurana, I., Ooi, J., Khan, A.W., et al. (2014). Vascular histone deacetylation by pharmacological HDAC inhibition. Genome Res. *24*, 1271–1284.

Raiber, E.-A., Beraldi, D., Ficz, G., Burgess, H.E., Branco, M.R., Murat, P., Oxley, D., Booth, M.J., Reik, W., and Balasubramanian, S. (2012). Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. Genome Biol. *13*, R69.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A.P., and Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. *97*, 5237–5242.

Ratel, D., Ravanat, J.-L., Berger, F., and Wion, D. (2006). N6-methyladenine: the other methylated base of DNA. BioEssays News Rev. Mol. Cell. Dev. Biol. *28*, 309–315.

Reddington, J.P., Perricone, S.M., Nestor, C.E., Reichmann, J., Youngson, N.A., Suzuki, M., Reinhardt, D., Dunican, D.S., Prendergast, J.G., Mjoseng, H., et al. (2013). Redistribution of H3K27me3 upon DNA hypomethylation results in de-repression of Polycomb target genes. Genome Biol. *14*, R25.

Reichmann, J., Crichton, J.H., Madej, M.J., Taggart, M., Gautier, P., Garcia-Perez, J.L., Meehan, R.R., and Adams, I.R. (2012). Microarray analysis of LTR retrotransposon silencing identifies Hdac1 as a regulator of retrotransposon expression in mouse embryonic stem cells. PLoS Comput. Biol. *8*, e1002486.

Rhee, I., Bachman, K.E., Park, B.H., Jair, K.-W., Yen, R.-W.C., Schuebel, K.E., Cui, H., Feinberg, A.P., Lengauer, C., Kinzler, K.W., et al. (2002). DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. Nature *416*, 552–556.

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. Cytogenet. Genome Res. *14*, 9–25.

Rishi, V., Bhattacharya, P., Chatterjee, R., Rozenberg, J., Zhao, J., Glass, K., Fitzgerald, P., and Vinson, C. (2010). CpG methylation of half-CRE sequences creates C/EBPα binding sites that activate some tissue-specific genes. Proc. Natl. Acad. Sci. *107*, 20311–20316.

Ristevski, S., O'Leary, D.A., Thornell, A.P., Owen, M.J., Kola, I., and Hertzog, P.J. (2004). The ETS Transcription Factor GABPα Is Essential for Early Embryogenesis. Mol. Cell. Biol. *24*, 5844–5849.

Robert, M.-F., Morin, S., Beaulieu, N., Gauthier, F., Chute, I.C., Barsalou, A., and MacLeod, A.R. (2003). DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. Nat. Genet. *33*, 61–65.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rosmarin, A.G., Resendes, K.K., Yang, Z., McMillan, J.N., and Fleming, S.L. (2004). GA-binding protein transcription factor: a review of GABP as an integrator of intracellular signaling and protein–protein interactions. Blood Cells. Mol. Dis. *32*, 143–154.

Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P.V., Layard-Liesching, H., Verp, S., Marquis, J., et al. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. Nature *463*, 237–240.

Rowe, H.M., Friedli, M., Offner, S., Verp, S., Mesnard, D., Marquis, J., Aktas, T., and Trono, D. (2013). De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. Development *140*, 519–529.

Rozenberg, J.M., Shlyakhtenko, A., Glass, K., Rishi, V., Myakishev, M.V., FitzGerald, P.C., and Vinson, C. (2008). All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. BMC Genomics *9*, 67.

de Ruijter, A.J.M., van Gennip, A.H., Caron, H.N., Kemp, S., and van Kuilenburg, A.B.P. (2003). Histone deacetylases (HDACs): characterization of the classical HDAC family. Biochem. J. *370*, 737–749.

Ruiz-Carrillo, A., Wangh, L.J., and Allfrey, V.G. (1975). Processing of newly synthesized histone molecules. Science *190*, 117–128.

Sabò, A., and Amati, B. (2014). Genome Recognition by MYC. Cold Spring Harb. Perspect. Med. *4*, a014191.

Sabò, A., Kress, T.R., Pelizzola, M., de Pretis, S., Gorski, M.M., Tesi, A., Morelli, M.J., Bora, P., Doni, M., Verrecchia, A., et al. (2014). Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis. Nature *511*, 488–492.

Sado, T., Okano, M., Li, E., and Sasaki, H. (2004). De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. Development *131*, 975–982.

Saito, M., and Ishikawa, F. (2002). The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. J. Biol. Chem. *277*, 35434–35439.

Sarraf, S.A., and Stancheva, I. (2004). Methyl-CpG Binding Protein MBD1 Couples Histone H3 Methylation at Lysine 9 by SETDB1 to DNA Replication and Chromatin Assembly. Mol. Cell *15*, 595–605.

Satoh, J., Kawana, N., and Yamamoto, Y. (2013). Pathway Analysis of ChIP-Seq-Based NRF1 Target Genes Suggests a Logical Hypothesis of their Involvement in the Pathogenesis of Neurodegenerative Diseases. Gene Regul. Syst. Biol. *7*, 139–152.

Satyamoorthy, K., Park, K., Atchison, M.L., and Howe, C.C. (1993). The intracisternal A-particle upstream element interacts with transcription factor YY1 to activate transcription: pleiotropic effects of YY1 on distinct DNA promoter elements. Mol. Cell. Biol. *13*, 6621–6628.

Schübeler, D., Lorincz, M.C., Cimbora, D.M., Telling, A., Feng, Y.-Q., Bouhassira, E.E., and Groudine, M. (2000). Genomic Targeting of Methylated DNA: Influence of Methylation on Transcription, Replication, Chromatin Structure, and Histone Acetylation. Mol. Cell. Biol. *20*, 9103–9112.

Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., Popp, C., Thienpont, B., Dean, W., and Reik, W. (2012). The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. Mol. Cell *48*, 849–862.

Sharif, J., Muto, M., Takebayashi, S., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., et al. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature *450*, 908–912.

Sharif, J., Endo, T.A., Nakayama, M., Karimi, M.M., Shimada, M., Katsuyama, K., Goyal, P., Brind'Amour, J., Sun, M.-A., Sun, Z., et al. (2016). Activation of Endogenous Retroviruses in Dnmt1−/− ESCs Involves Disruption of SETDB1-Mediated Repression by NP95 Binding to Hemimethylated DNA. Cell Stem Cell *19*, 81–94.

Sharp, A.J., Stathaki, E., Migliavacca, E., Brahmachary, M., Montgomery, S.B., Dupre, Y., and Antonarakis, S.E. (2011). DNA methylation profiles of human active and inactive X chromosomes. Genome Res. *21*, 1592–1600.

Shen, J.C., Rideout, W.M., and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. *22*, 972–976.

Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R.A., and Issa, J.-P.J. (2007). Genome-Wide Profiling of DNA Methylation Reveals a Class of Normally Methylated CpG Island Promoters. PLoS Genet. *3*.

Sherwood, R.I., Hashimoto, T., O'Donnell, C.W., Lewis, S., Barkal, A.A., van Hoff, J.P., Karun, V., Jaakkola, T., and Gifford, D.K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat. Biotechnol. *32*, 171–178.

Shi, Y., Seto, E., Chang, L.S., and Shenk, T. (1991). Transcriptional repression by YY1, a human GLI-Krüppel-related protein, and relief of repression by adenovirus E1A protein. Cell *67*, 377–388.

Shoji, M., Tanaka, T., Hosokawa, M., Reuter, M., Stark, A., Kato, Y., Kondoh, G., Okawa, K., Chujo, T., Suzuki, T., et al. (2009). The TDRD9-MIWI2 complex is essential for piRNA-mediated retrotransposon silencing in the mouse male germline. Dev. Cell *17*, 775–787.

Siegfried, Z., Eden, S., Mendelsohn, M., Feng, X., Tsuberi, B.-Z., and Cedar, H. (1999). DNA methylation represses transcription in vivo. Nat. Genet. *22*, 203–206.

Sigova, A.A., Abraham, B.J., Ji, X., Molinie, B., Hannett, N.M., Guo, Y.E., Jangi, M., Giallourakis, C.C., Sharp, P.A., and Young, R.A. (2015). Transcription factor trapping by RNA in gene regulatory elements. Science aad3346.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci. *39*, 381–399.

Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods *advance online publication*.

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

Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature *484*, 339–344.

Song, J., Rechkoblit, O., Bestor, T.H., and Patel, D.J. (2011). Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation. Science *331*, 1036–1040.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. Cell *161*, 555–568.

Spencer, D.H., Russler-Germain, D.A., Ketkar, S., Helton, N.M., Lamprecht, T.L., Fulton, R.S., Fronick, C.C., O'Laughlin, M., Heath, S.E., Shinawi, M., et al. (2017). CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression. Cell *168*, 801-816.e13.

Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W.T.C., Bauer, C., Münzel, M., Wagner, M., Müller, M., Khan, F., et al. (2013). Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. Cell *152*, 1146–1159.

Stadler, M.B., Murr, R., Burger, L., Ivánek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Stalder, J., Larsen, A., Engel, J.D., Dolan, M., Groudine, M., and Weintraub, H. (1980). Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase I. Cell *20*, 451–460.

Startk, R., and Brown, G. (2011). Diffbind:differential binding analysis of ChIP-Seq peak data. http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf.

Stelzer, Y., Shivalila, C.S., Soldner, F., Markoulaki, S., and Jaenisch, R. (2015). Tracing Dynamic Changes of DNA Methylation at Single-Cell Resolution. Cell *163*, 218–229.

Stelzer, Y., Wu, H., Song, Y., Shivalila, C.S., Markoulaki, S., and Jaenisch, R. (2016). Parent-of-Origin DNA Methylation Dynamics during Mouse Development. Cell Rep. *16*, 3167–3180.

Stirzaker, C., Taberlay, P.C., Statham, A.L., and Clark, S.J. (2014). Mining cancer methylomes: prospects and challenges. Trends Genet. TIG *30*, 75–84.

Stocking, C., and Kozak, C.A. (2008). Endogenous retroviruses. Cell. Mol. Life Sci. *65*, 3383–3398.

Suetake, I., Hayata, D., and Tajima, S. (2006). The Amino-Terminus of Mouse DNA Methyltransferase 1 Forms an Independent Domain and Binds to DNA with the Sequence Involving PCNA Binding Motif. J. Biochem. (Tokyo) *140*, 763–776.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet. *9*, 465–476.

Swartz, M.N., Trautner, T.A., and Kornberg, A. (1962). Enzymatic Synthesis of Deoxyribonucleic Acid XI. FURTHER STUDIES ON NEAREST NEIGHBOR BASE SEQUENCES IN DEOXYRIBONUCLEIC ACIDS. J. Biol. Chem. *237*, 1961–1967.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. Science *324*, 930–935.

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.

Tang, W.W.C., Dietmann, S., Irie, N., Leitch, H.G., Floros, V.I., Bradshaw, C.R., Hackett, J.A., Chinnery, P.F., and Surani, M.A. (2015). A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. Cell *161*, 1453–1467.

Tate, P.H., and Bird, A.P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. Curr. Opin. Genet. Dev. *3*, 226–231.

Tatton-Brown, K., Seal, S., Ruark, E., Harmer, J., Ramsay, E., del Vecchio Duarte, S., Zachariou, A., Hanks, S., O'Brien, E., Aksglaede, L., et al. (2014). Mutations in the DNA methyltransferase gene DNMT3A cause an overgrowth syndrome with intellectual disability. Nat. Genet. *advance online publication*.

Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. Genome Res. *21*, 1800–1812.

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.

Tian, H.-P., Lun, S.-M., Huang, H.-J., He, R., Kong, P.-Z., Wang, Q.-S., Li, X.-Q., and Feng, Y.-M. (2015). DNA Methylation Affects the SP1-regulated Transcription of FOXF2 in Breast Cancer Cells. J. Biol. Chem. *290*, 19173–19183.

Tomala, M., Lavrentieva, A., Moretti, P., Rinas, U., Kasper, C., Stahl, F., Schambach, A., Warlich, E., Martin, U., Cantz, T., et al. (2010). Preparation of bioactive soluble human leukemia inhibitory factor from recombinant Escherichia coli using thioredoxin as fusion partner. Protein Expr. Purif. *73*, 51–57.

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J.G., Baylin, S.B., and Issa, J.P. (1999). CpG island methylator phenotype in colorectal cancer. Proc. Natl. Acad. Sci. U. S. A. *96*, 8681–8686.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trowbridge, J.J., Snow, J.W., Kim, J., and Orkin, S.H. (2009). DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. Cell Stem Cell *5*, 442–449.

Tse, C., Sera, T., Wolffe, A.P., and Hansen, J.C. (1998). Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. Mol. Cell. Biol. *18*, 4629–4638.

Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H.R., et al. (2006). Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes Cells *11*, 805–814.

Tucker, K.L., Talbot, D., Lee, M.A., Leonhardt, H., and Jaenisch, R. (1996). Complementation of methylation deficiency in embryonic stem cells by a DNA methyltransferase minigene. Proc. Natl. Acad. Sci. *93*, 12920–12925.

Turcan, S., Rohle, D., Goenka, A., Walsh, L.A., Fang, F., Yilmaz, E., Campos, C., Fabius, A.W.M., Lu, C., Ward, P.S., et al. (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature *483*, 479–483.

Ueda, A., Akagi, T., and Yokota, T. (2017). GA-Binding Protein Alpha Is Involved in the Survival of Mouse Embryonic Stem Cells. Stem Cells Dayt. Ohio.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat. Methods *5*, 829–834.

Virbasius, C.A., Virbasius, J.V., and Scarpulla, R.C. (1993). NRF-1, an activator involved in nuclear-mitochondrial interactions, utilizes a new DNA-binding domain conserved in a family of developmental regulators. Genes Dev. *7*, 2431–2445.

Vogelauer, M., Wu, J., Suka, N., and Grunstein, M. (2000). Global histone acetylation and deacetylation in yeast. Nature *408*, 495.

Voo, K.S., Carlone, D.L., Jacobsen, B.M., Flodin, A., and Skalnik, D.G. (2000). Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA Methyltransferase, Human Trithorax, and Methyl-CpG Binding Domain Protein 1. Mol. Cell. Biol. *20*, 2108–2121.

Waalwijk, C., and Flavell, R.A. (1978). DNA methylation at a CCGG sequence in the large intron of the rabbit β-globin gene: tissue-specific variations. Nucleic Acids Res. *5*, 4631–4641.

Wade, P.A., Gegonne, A., Jones, P.L., Ballestar, E., Aubry, F., and Wolffe, A.P. (1999). Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. Nat. Genet. *23*, 62–66.

Walsh, C.P., Chaillet, J.R., and Bestor, T.H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat. Genet. *20*, 116–117.

Walter, M., Teissandier, A., Pérez-Palacios, R., and Bourc'his, D. (2016). An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. ELife *5*, e11418.

Walter, M.J., Ding, L., Shen, D., Shao, J., Grillot, M., McLellan, M., Fulton, R., Schmidt, H., Kalicki-Veizer, J., O'Laughlin, M., et al. (2011). Recurrent DNMT3A mutations in patients with myelodysplastic syndromes. Leukemia *25*, 1153–1158.

Wan, J., Su, Y., Song, Q., Tung, B., Oyinlade, O., Liu, S., Ying, M., Ming, G., Song, H., Qian, J., et al. (2017). Methylated cis-regulatory elements mediate KLF4-dependent gene transactivation and cell migration. ELife *6*, e20068.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897.

Wang, Z., Zang, C., Cui, K., Schones, D.E., Barski, A., Peng, W., and Zhao, K. (2009). Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes. Cell *138*, 1019–1031.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. *39*, 457–466.

Weill, L., Shestakova, E., and Bonnefoy, E. (2003). Transcription factor YY1 binds to the murine beta interferon promoter and regulates its transcriptional capacity with a dual activator/repressor role. J. Virol. *77*, 2903–2914.

Wenger, R.H., Kvietikova, I., Rolfs, A., Camenisch, G., and Gassmann, M. (1998). Oxygen-regulated erythropoietin gene expression is dependent on a CpG methylation-free hypoxia-inducible factor-1 DNA-binding site. Eur. J. Biochem. *253*, 771–777.

Wiench, M., John, S., Baek, S., Johnson, T.A., Sung, M.-H., Escobar, T., Simmons, C.A., Pearce, K.H., Biddie, S.C., Sabo, P.J., et al. (2011). DNA methylation status predicts cell type-specific enhancer activity. EMBO J. *30*, 3028–3039.

Wierstra, I. (2008). Sp1: Emerging roles—Beyond constitutive activation of TATA-less housekeeping genes. Biochem. Biophys. Res. Commun. *372*, 1–13.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A.C., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature *473*, 343–348.

Wu, T.P., Wang, T., Seetin, M.G., Lai, Y., Zhu, S., Lin, K., Liu, Y., Byrum, S.D., Mackintosh, S.G., Zhong, M., et al. (2016). DNA methylation on N(6)-adenine in mammalian embryonic stem cells. Nature *532*, 329–333.

Xie, W., Barr, C.L., Kim, A., Yue, F., Lee, A.Y., Eubanks, J., Dempster, E.L., and Ren, B. (2012). Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. Cell *148*, 816–831.

Xiol, J., Cora, E., Koglgruber, R., Chuma, S., Subramanian, S., Hosokawa, M., Reuter, M., Yang, Z., Berninger, P., Palencia, A., et al. (2012). A Role for Fkbp6 and the Chaperone Machinery in piRNA Amplification and Transposon Silencing. Mol. Cell *47*, 970–979.

Xu, G.-L., Bestor, T.H., Bourc'his, D., Hsieh, C.-L., Tommerup, N., Bugge, M., Hulten, M., Qu, X., Russo, J.J., and Viegas-Péquignot, E. (1999). Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature *402*, 187–191.

Yang, W.M., Inouye, C., Zeng, Y., Bearss, D., and Seto, E. (1996). Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD3. Proc. Natl. Acad. Sci. U. S. A. *93*, 12845–12850.

Yang, X., Han, H., De Carvalho, D.D., Lay, F.D., Jones, P.A., and Liang, G. (2014). Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. Cancer Cell *26*, 577–590.

Yildirim, O., Li, R., Hung, J.-H., Chen, P.B., Dong, X., Ee, L., Weng, Z., Rando, O.J., and Fazzio, T.G. (2011). Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. Cell *147*, 1498–1510.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

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.

Yisraeli, J., Frank, D., Razin, A., and Cedar, H. (1988). Effect of in vitro DNA methylation on beta-globin gene expression. Proc. Natl. Acad. Sci. U. S. A. *85*, 4638–4642.

Yoder, J.A., Soman, N.S., Verdine, G.L., and Bestor, T.H. (1997). DNA (cytosine-5)-methyltransferases in mouse cells and tissues. studies with a mechanism-based probe11Edited by K. Yamamoto. J. Mol. Biol. *270*, 385–395.

Yokomori, N., Kobayashi, R., Moore, R., Sueyoshi, T., and Negishi, M. (1995). A DNA methylation site in the male-specific P450 (Cyp 2d-9) promoter and binding of the heteromeric transcription factor GABP. Mol. Cell. Biol. *15*, 5355–5362.

Yoon, H.-G., Chan, D.W., Reynolds, A.B., Qin, J., and Wong, J. (2003). N-CoR Mediates DNA Methylation-Dependent Repression through a Methyl CpG Binding Protein Kaiso. Mol. Cell *12*, 723–734.

Yoshida, M., Kijima, M., Akita, M., and Beppu, T. (1990). Potent and specific inhibition of mammalian histone deacetylase both in vivo and in vitro by trichostatin A. J. Biol. Chem. *265*, 17174–17179.

Yu, J., Li, Y., Ishizuka, T., Guenther, M.G., and Lazar, M.A. (2003). A SANT motif in the SMRT corepressor interprets the histone code and promotes histone deacetylation. EMBO J. *22*, 3403–3410.

Yu, M., Yang, X.-Y., Schmidt, T., Chinenov, Y., Wang, R., and Martin, M.E. (1997). GA-binding Protein-dependent Transcription Initiator Elements EFFECT OF HELICAL SPACING BETWEEN POLYOMAVIRUS ENHANCER A FACTOR 3(PEA3)/Ets-BINDING SITES ON INITIATOR ACTIVITY. J. Biol. Chem. *272*, 29060–29067.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.-X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., et al. (2012). Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. Cell *149*, 1368–1380.

Zalzman, M., Falco, G., Sharova, L.V., Nishiyama, A., Thomas, M., Lee, S.-L., Stagg, C.A., Hoang, H.G., Yang, H.-T., Indig, F.E., et al. (2010). Zscan4 regulates telomere elongation and genomic stability in ES cells. Nature *464*, 858–863.

Zhang, L., Lu, X., Lu, J., Liang, H., Dai, Q., Xu, G.-L., Luo, C., Jiang, H., and He, C. (2012). Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. Nat. Chem. Biol. *8*, 328–330.

Zhang, T., Xu, M., Makowski, M.M., Lee, C., Kovacs, M., Fang, J., Willems, E., Trent, J.M., Hayward, N.K., Vermeulen, M., et al. (2017). SDHD Promoter Mutations Ablate GABP Transcription Factor Binding in Melanoma. Cancer Res. *77*, 1649–1661.

Zhang, X., Su, J., Jeong, M., Ko, M., Huang, Y., Park, H.J., Guzman, A., Lei, Y., Huang, Y.-H., Rao, A., et al. (2016). DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. Nat. Genet. *48*, 1014–1023.

Zhang, Y., Ng, H.-H., Erdjument-Bromage, H., Tempst, P., Bird, A., and Reinberg, D. (1999). Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. Genes Dev. *13*, 1924–1935.

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.

Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.-Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., et al. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477–481.