Influence of DNA methylation on transcription factor binding

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Silvia Domcke

aus
München, Deutschland

Basel, 2017

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Prof. Dr. Dirk Schübeler und Dr. François Spitz.

Basel, den 18.04.2017

Prof. Dr. Martin Spiess (Dekan)

"While Occam's razor is a useful tool in the physical sciences, it can be a very dangerous implement in biology. It is thus very rash to use simplicity and elegance as a guide in biological research. While DNA could be claimed to be both simple and elegant, it must be remembered that DNA almost certainly originated fairly close to the origin of life when things were necessarily simple or they would not have got going. Biologists must constantly keep in mind that what they see was not designed, but rather evolved."

---- Francis Crick in 'What mad Pursuit' (1988) ----

Acknowledgements

This thesis would not have been possible without the support and contribution of many people.

I would like to thank my PhD advisor Dirk Schübeler for giving me the opportunity to work in his group and for many excellent as well as thought-provoking scientific discussions. I especially appreciate that it was always possible to get critical input and advice while still having a lot of freedom and flexibility.

For direct contributions to this study I would like to thank the co-authors of the published manuscript, in particular Anaïs Bardet. In addition, many thanks go to Christiane Wirbelauer, who generated the inducible-NGN2 TKO ES cells and performed their neuronal differentiation. I am grateful to Michael Stadler from the Computational Biology group for insightful discussions on how to deal with repetitive elements in the bioinformatic analysis and to Tuncay Baubec for initial advice on the CTCF project.

Thank you to all current members of the Schübeler lab for a great and rigorous scientific environment and pleasant working atmosphere. Special thanks go to Arnaud Krebs and all my fellow PhD students for their friendship and fun experiences and travels outside the lab.

I would like to thank the members of my thesis committee, François Spitz and Nico Thomä, for taking the time and interest to discuss and offer valuable input on my project.

For funding and providing me with the opportunity to visit courses and conferences as well as meet great people, I am grateful to the Boehringer Ingelheim Fonds.

I am indebted to my parents for always supporting me. Finally I would like to thank Christian, my friends, my physiotherapists and the Swiss mountains for keeping me sane in difficult times!

Table of contents

List of figures and tables

List of abbreviations

AP-1	activator protein 1
ARNT	aryl hydrocarbon receptor nuclear translocator
ATAC	assay for transposase-accessible chromatin
ATF	activating transcription factor
ATP	adenosine triphosphate
bp	base pair
bZIP	basic leucine zipper
cAMP	cyclic adenosine monophosphate
CGI	CpG island
ChIP	chromatin immunoprecipitation
CpG	cytosine nucleotide followed by a guanine nucleotide
CRE	cAMP-responsive element
CREB	cAMP response element-binding protein
CRISPR	clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
DHS	DNase hypersensitive site
DNA	deoxyribonucleic acid
DNMT	DNA methyltransferase
dox	doxycycline
E2F	E2 factor
EHMT	euchromatin histone methyltransferase
ERV	endogenous retrovirus
ERV-K	endogenous retrovirus group K
ES cell	embryonic stem cell
ETS	E twenty-six
FOS	Fos proto-oncogene
GAL4	galactose-responsive transcription factor 4
H19	imprinted maternally expressed transcript (non-coding)
H3KXme3	histone 3 lysine X (e.g. 9, 27) trimethylation
HDAC	histone deacetylase
HIF1A	hypoxia-inducible factor 1-alpha
HIV	human immunodeficiency virus
HNF6	hepatocyte nuclear factor 6
IAP	intracisternal A-particle
ICR	imprinting control region
IGF-2	insulin-like growth factor 2
JUN	Jun proto-oncogene
kb	kilobase

KZFP	KRAB (Krüppel-associated box) zinc-finger protein
LTR	long terminal repeat
MBD	methyl-CpG binding domain
MEF	mouse embryonic fibroblast
MPRA	massively parallel reporter assay
MYC	v-myc avian myelocytomatosis viral oncogene homolog
NF-κB	nuclear factor 'kappa-light-chain-enhancer' of activated B-cells
NFY	nuclear transcription factor Y
NGN2	neurogenin2
NOMe-seq	nucleosome occupancy and methylome sequencing
NP	neuronal progenitor
NPC	neural precursor cell
NRF1	nuclear respiratory factor 1
PCR	polymerase chain reaction
PDGFRA	platelet derived growth factor receptor alpha
PGC	primordial germ cell
PHO5	phosphatase encoded by gene 5 of the yeast Pho regulon
POU5F1	POU class 5 homeobox 1
PWM	position weight matrix
qAMP	quantitative analysis of DNA methylation using real-time PCR
qPCR	quantitative polymerase chain reaction
REST	RE1-silencing transcription factor
RMCE	recombinase-mediated cassette exchange
RNA	ribonucleic acid
RPKM	reads per kilobase per million mapped reads
SETDB1	SET domain bifurcated 1
SP1	Specificity protein 1
TAT	tyrosine aminotransferase
TBP	TATA-binding protein
TE	transposable element
TET	ten-eleven translocation
TF	transcription factor
TKO	triple knockout; *here:* of *DNMT1/3a/3b* genes
TN	terminal neuron
TSS	transcription start site
UHRF1	ubiquitin-like with PHD and ring finger domains 1
USF	upstream transcription factor
WT	wildtype

Protein names are in capital letters. Gene and transcript names are in italics.

1. Summary

Eukaryotic transcription factors (TFs) are key determinants of gene activity, yet they bind only a fraction of their corresponding DNA sequence motifs in any given cell type. Chromatin has the potential to restrict accessibility of binding sites; however, in which context chromatin states are instructive for TF binding remains mainly unknown. This thesis explores the contribution of DNA methylation to constrained TF binding by studying CTCF as a known methylation-sensitive TF and applying a genome-wide approach to identify further sensitive factors in mouse stem and differentiated cells.

CTCF is perhaps the most prominent example for a TF that can be prevented from binding by DNA methylation *in vivo*. However, it is restricted by methylation only at a subset of its genomic binding sites, such as the *H19/Igf2* imprinting control region (ICR). In order to understand this context-dependency of CTCF methylation sensitivity, we compared CTCF binding in isogenic mouse stem cells with and without DNA methylation. Two features distinguish the fraction of sites that are bound only in the absence of DNA methylation: CpG-containing variants of the canonical CTCF motif as well as higher CpG density in the flanking regions. The *H19/Igf2* ICR indeed fulfils these criteria and we show that CTCF methylation sensitivity there is independent of the complete ICR sequence, the chromosomal context and H3K9me3 marks.

In order to go beyond CTCF and identify more methylation-sensitive TFs *a priori*, we mapped DNase I hypersensitive sites, as an indicator of TF binding, in mouse stem cells with and without DNA methylation. Methylation-restricted sites are enriched for TF motifs containing CpGs, especially for those of NRF1. In fact, NRF1 occupies several thousand additional sites in the unmethylated genome, resulting in increased genic and non-genic transcription. Restoring *de novo* methyltransferase activity initiates remethylation at these sites and outcompetes NRF1 binding. Even strong overexpression of NRF1 is unable to prompt binding at methylated regions.

This suggests that binding of methylation-sensitive TFs relies on additional determinants to induce local hypomethylation. In support of this model, deletion of neighbouring motifs in *cis* or of a TF in *trans* causes local hypermethylation and subsequent loss of NRF1 binding. This competition between DNA methylation and TFs *in vivo* reveals a case of cooperativity between TFs that acts indirectly via DNA methylation.

Nevertheless, the vast majority of TF binding events do not change upon removal of DNA methylation in stem cells. To investigate whether more TFs are affected in differentiated cells, for which DNA methylation is essential, we generated methylation-deficient neuronal cells that survive for several days in culture. Changes in genic transcription and chromatin accessibility are surprisingly limited in the absence of DNA methylation, although again a subset of TF motifs are enriched in methylation-restricted sites, such as NRF1 and HNF6. While this closely resembles the situation in stem cells, we observe a striking activation of specific classes of endogenous retroviruses (ERV) only in the differentiated methylation mutant. Several lines of evidence indicate that methylation-sensitive TF binding at the cAMP-responsive element (CRE motif) is responsible for ERV activation in differentiated methylation mutants including mouse cortex, which might provide a link to the ensuing cell death.

Taken together, only a low percentage of TF binding events are restricted by DNA methylation in stem or differentiated cells. However, a subset of factors is methylation-sensitive at CpG-containing motifs. These factors rely on other TFs to keep their motif in an unmethylated state and their aberrant binding can have devastating consequences by repeat activation.

Understanding the influence of DNA methylation on TF binding constitutes one step towards better interpretation of the rapidly growing number of epigenetic and TF binding maps. The success of the approach taken here suggests that it can be applied to other chromatin components and modifications, which should enable comprehensive prediction of TF binding and ultimately gene expression in development and disease.

2. General introduction

2.1 Eukaryotic transcription factors bind a fraction of their target sites

Dynamic regulation of gene expression enables prokaryotes to adapt to external conditions and multicellular eukaryotes to form diverse cell types in spite of a largely invariant DNA blueprint. The ability to turn genes on and off is central to every life form and all biological processes. How genes are regulated has thus been a fundamental question in biology ever since their discovery. Early work in prokaryotes identified a new type of gene product, the 'regulator', which interacts with the DNA immediately upstream of genes and controls their expression (Jacob and Monod; 1961). In eukaryotes such 'regulators', or transcription factors (TFs), bind to the DNA in a sequence-specific manner not only at gene-proximal promoter regions as in prokaryotes, but also at distal enhancer elements (Banerji et al., 1981; Maniatis et al., 1987; Moreau et al., 1981). In the 1980s several eukaryotic TFs were cloned and biochemically characterised, leading Johnson and McKnight to declare that 'a major effort is now under way to identify sequence-specific DNA-binding proteins, to match them to their cognate sites within or around eukaryotic genes, and to elucidate how the binding of such proteins results in increased or decreased transcription of the associated gene' (Johnson and McKnight, 1989).

Nearly three decades later, extensive progress has indeed been made in the identification and cataloguing of various eukaryotic TF classes (Weirauch and Hughes, 2011); however, matching TFs to their genomic binding sites remains a challenge that has been surprisingly difficult to tackle. In contrast to prokaryotic TFs that bind highly defined sequence motifs in a predictable manner, TFs in higher eukaryotes recognise short highly degenerate DNA sequences (Fig. 2-1a) (Wunderlich and Mirny, 2009). As a result, the consensus sequences for each factor are extremely common in the genome,

and in fact occur frequently in and around most genes (Biggin, 2011; Wunderlich and Mirny, 2009). Only a miniscule fraction of these target sites is actually occupied by the TF in any given cell type (Fig. 2-1b) (Biggin, 2011). Even if the *in vitro* binding specificity of the factor is known, predicting which of the seemingly identical sites are bound in a cell thus remains an unsolved problem and presents a substantial barrier in our path towards understanding eukaryotic gene regulation (Biggin, 2011; Slattery et al., 2014; Todeschini et al., 2014). Yet within the crowded nucleoplasm, TFs somehow manage to bind to defined DNA sites and regulate gene expression in a highly reproducible and cell type-specific manner.

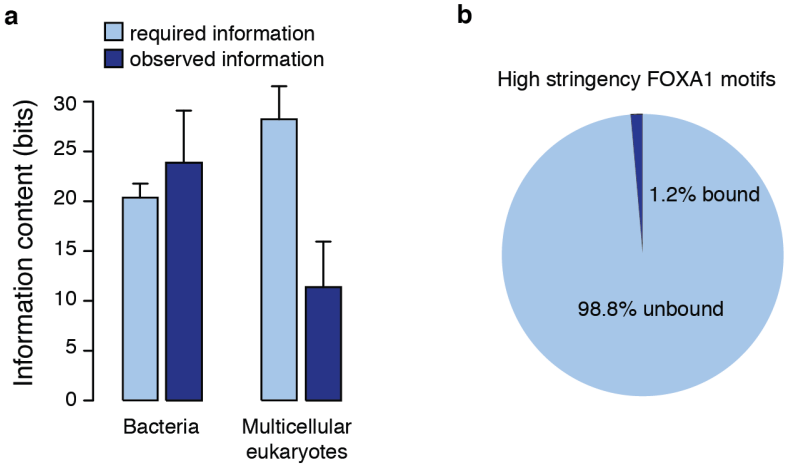


Figure 2-1. Information content of eukaryotic TF motifs is not sufficient to specify their binding sites in a large genome.
a) Comparison of required and actual information content of TF binding motifs in bacteria and multicellular eukaryotes. Shown is the minimum required information content $I_{min} = \log_2(N)$ needed to specify a unique address in a genome of size N (light blue), and the mean information content of actual TF binding motifs for roughly 100 bacterial and multicellular eukaryotic motifs (dark blue). The error bars represent the standard deviation, which for the required information content is due to the range of genome sizes. Graph adapted from Wunderlich et al., 2009. **b)** Example of the fraction of high-confidence motif sequences bound in a given cell type for the pioneer TF FOXA1. High-stringency FOXA1 motifs were called with MotifLocator. Of these sites, FOXA1 only occupies 1.2% in MCF7 cells as measured by ChIP-seq (~ 12,000 peaks, 1% FDR). Adapted from Lupien et al., 2008.

2.2 Role of chromatin in binding site restriction

The organisation of eukaryotic genomes into complex nucleoprotein structures that are absent in their smaller prokaryotic counterparts was first attributed

only to the need for compaction. However, chromatin soon emerged as the most likely candidate for restricting the access of TFs to specific regulatory sites (Voss and Hager, 2014). Accessibility of the DNA within chromatin, as measured by susceptibility to DNase digestion, was recognised as a unifying feature of active regulatory regions in eukaryotes (Elgin, 1981; Weintraub and Groudine, 1976; Wu et al., 1979) that is highly cell type-specific (Thurman et al., 2012). While TFs tend to show a similar principal motif preference on both naked and chromatinised genomic DNA, binding locations differ considerably between the two templates (Liu et al., 2006). The occupancy levels of many different classes of TFs *in vivo* correlate well with the degree of accessibility of those regions (Biggin, 2011). Accordingly, predictions of TF binding based on accessibility data are a vast improvement over pure sequence-based models (Pique-Regi et al., 2011). These observations raise the question how differential chromatin accessibility and TF binding are connected and which aspects of chromatin are involved in the binding site restriction of TFs.

The existence of at least two different chromatin states was described nearly a century ago in moss (Heitz, 1928). In recent years an ever more fine-grained distinction of chromatin states, which differ in the transcriptional activity of the contained genes, has been proposed based on location of chromatin proteins or post-translational modifications of histones (Ernst and Kellis, 2012; Filion et al., 2010). Indeed many eukaryote-specific chromatin components correlate or anticorrelate with TF occupancy *in vivo* (Fig. 2-2). Nevertheless, whether a specific chromatin state is simply permissive to TF binding, actively directs TF binding, or is a result of TF binding is often unclear, and with it the sequence of events that connect chromatin states and gene activity (Slattery et al., 2014). The setting and removing of chromatin features in the context of transcription, as well as their interplay with each other and with TFs or chromatin-modifying enzymes is a dauntingly complex system to disentangle. In the following I will briefly present three of the most promising candidates for chromatin-mediated TF binding site restriction: nucleosomes themselves, post-translational modifications of their histone tails, as well as methylation of the underlying DNA.

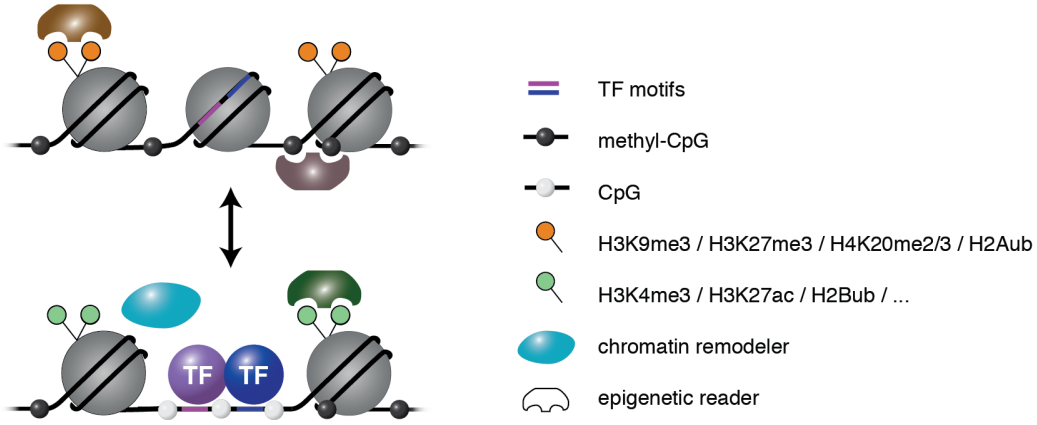


Figure 2-2. Chromatin states differ at bound and unbound TF motifs.
Nucleosomes, repressive histone modifications and DNA methylation have all been associated with binding site restriction of TFs in vertebrate genomes. The transition from an unbound inactive (top) to a TF-bound active (bottom) regulatory region involves changes in nucleosome occupancy and positioning, chromatin remodeling activity, changes in histone modifications and DNA methylation as well as differential recruitment of epigenetic readers. However, it remains unclear which of these chromatin features have an instructive role in shaping cell type-specific TF binding patterns and thus gene regulation and which are adopted downstream of TF binding. Sensitivity to chromatin states likely varies across TFs, but these differences can be masked at co-bound sites.

2.2.1 Nucleosomes and transcription factor binding

The basic unit of chromatin is the nucleosome, in which a DNA stretch of 147 bp length is tightly wrapped around an octamer of histone proteins (Richmond and Davey, 2003). Early *in vitro* reconstitution experiments and observation of glucose-mediated nucleosome loss at the yeast *Pho5* promoter implied that transcription initiation is impeded in the presence of nucleosomes (Han and Grunstein, 1988; Knezetic and Luse, 1986). On the other hand, the yeast TF GAL4 was shown to be capable of displacing nucleosomes over its binding site *in vitro* (Workman and Kingston, 1992). *In vivo*, occupied TF binding sites are indeed devoid of nucleosomes (Yuan et al., 2005), but the sensitivity of different TFs to nucleosomes covering their motifs varies broadly. Today the accepted view is that some TFs, termed pioneer TFs, are capable of engaging their target sites in closed chromatin (Iwafuchi-Doi and Zaret, 2014). This has been suggested to occur through binding of partial motifs displayed on the nucleosome surface, ultimately leading to nucleosome displacement (Soufi et al., 2015). While pioneering activity has been attributed

to roughly a dozen TFs (Iwafuchi-Doi and Zaret, 2014), it remains unclear for most factors to which extent they are restricted in their binding by nucleosome occupancy. The majority of TFs are likely unable to initially breach the nucleosome barrier on their own and require exposure of their binding sites through other means (John et al., 2011; Svaren and Hörz, 1997). These could involve a combination of spontaneous unwrapping and rebinding of the histone octamer (Bucceri et al., 2006; Li et al., 2005; Polach and Widom, 1995), the action of ATP-dependent chromatin remodelers (Lorch et al., 2010) and cooperative TF binding competing with nucleosomes for access to DNA (Adams and Workman, 1995; Miller and Widom, 2003; Spitz and Furlong, 2012). The presence of nucleosomes has thus been suggested to substantially contribute to the binding site selectivity of most TFs, with pioneer TFs being an important exception (Slattery et al., 2014). Nonetheless, even known pioneer factors only bind a fraction of their sequence motifs in a cell type-specific manner (Fig. 2-1b) (Iwafuchi-Doi and Zaret, 2014; Lupien et al., 2008), so other layers besides nucleosome occupancy must contribute to binding site restriction.

2.2.2 Histone modifications and transcription factor binding

Beyond the mere absence or presence of nucleosomes, certain post-translational modifications of the contained core histone proteins are positively or negatively associated with TF occupancy (Fig. 2-2) (Ernst and Kellis, 2013). An estimated 60% of nucleosomes are substantially modified on their histone tails in mammals (Ho et al., 2014). The facultative or constitutive silent heterochromatic state is characterised by low levels of acetylation and high levels of specific methylated (H3K9, H3K27, and H4K20) and ubiquitinylated (H2A) sites (Kouzarides, 2007; Li et al., 2007).

H3K9me3 is the hallmark of constitutive heterochromatin as found for example in pericentric regions of the chromosome. When a transcriptionally active gene is brought near pericentric heterochromatin, the gene can become silenced. This phenomenon was first discovered in the fruit fly *Drosophila*

melanogaster when studying position-effect variegation of X-ray induced chromosomal rearrangements and has been attributed to spreading of the H3K9me3 mark into active chromatin (Girton and Johansen, 2008; Tschiersch et al., 1994). *In vitro* studies demonstrated that the interaction of Heterochromatin Protein 1 (HP1) with H3K9 methylated histones mediates dose-dependent repression of transcription (Loyola et al., 2001). Facultative heterochromatin is mainly characterised by H3K27me3 and H2A119ub1 marks set by the Polycomb-group of proteins, which are critical for repression of key transcriptional regulators during development (Shilatifard, 2006).

For both types of heterochromatin, it is currently unclear how gene silencing is actually brought about *in vivo* and to which extent histone modifications are set upstream or downstream of changes in TF binding and transcription (Shilatifard, 2006; Zhang et al., 2015). While it has been suggested that even pioneer TFs are blocked from binding by the presence of repressive histone marks (Iwafuchi-Doi and Zaret, 2014), experimental evidence for this model is still lacking. For example, access of specific TFs and the transcription machinery does not seem to be blocked by H3K27me3, yet transcription initiation is inhibited (Dellino et al., 2004). The sensitivity of different TFs to various histone modifications thus remains unclear to date and with it the mechanisms underlying gene repression in heterochromatin.

2.2.3 DNA methylation and transcription factor binding

Apart from nucleosomes and the posttranslational modification of their histone tails, modifications of the DNA itself could affect TF binding. In particular, methylation of cytosines in the context of CpG dinucleotides has long been associated with gene repression (Cedar, 1988). Since TF binding site restriction by DNA methylation is the main focus of this thesis, this mark will be discussed in more detail in the following paragraphs in terms of evolution, genomic distribution and interplay with TF binding.

2.2.3.1 Evolution of cytosine methylation and repeat silencing

Methylation of the fifth carbon on cytosine is an ancient DNA modification that is catalysed by the same enzymatic superfamily in bacteria, archaea, and eukaryotes (Goll and Bestor, 2005). DNA methylation likely arose as a sort of 'genomic immune system', to defend the host against the invasion of virus DNA and transposable elements (TEs) (Bestor, 1990). TEs threaten the host genome not only through potentially deleterious insertional mutagenesis, but can also induce rearrangements through homologous recombination of non-allelic repeats, produce neomorphic chimeric transcripts with host genes and overload the host with the sheer amount of their transcripts (Bestor, 2003). Recognizing and methylating these foreign DNA sequences enables their transcriptional repression and prevents their further replication within the host genome. Inactivated TEs are riddled with mutations over time, further depriving them of transcriptional competence, with C to T transitions by deamination of methylated cytosines being a substantial contributor (Cooper and Youssoufian, 1988; Lander et al., 2001).

The development of such an effective silencing mechanism of TEs allowed for their accumulation in the host genome. This is thought to account for the strong correlation between genome size, repeat content and DNA methylation observed across organisms (Bestor, 1990; Bird, 1995; Lechner et al., 2013). In fact an astounding 50 to 70% of the human genome is made up of such repetitive elements (Lander et al., 2001; Padeken et al., 2015). It has been proposed that the presence of TEs in our genomes is in fact a 'penalty' of sexual reproduction (Bestor, 2003). In asexual organisms, a harmful transposon is dependent on the survival of the host genome and reduces the fitness of the host and itself in a similar manner, preventing it from spreading through the population. In sexual organisms on the other hand, TEs can spread quite rapidly due to their ability to colonise new genomes during zygote formation. Even harmful TEs become fixed in a population if they reduce host fitness by anything less than one half (Hickey, 1982). Indeed there is a good agreement between transposon aggressiveness and the

extent of sexual out-crossing that occurred during the evolution of closely-related species (Bestor, 2003).

Intriguingly, cytosine methylation has been lost several times in the course of animal evolution, such as in the invertebrate lineages leading to *Drosophila* and the nematode *Caenorhabditis elegans* (Zemach and Zilberman, 2010). It is also uncommon in fungi such as saccharomycetes and most species of green algae (Suzuki and Bird, 2008; Zemach and Zilberman, 2010). This loss could be due to the fact that their unicellular ancestors primarily reproduced asexually and thus could dispense with the ability to silence TEs by DNA methylation. Today's invertebrate lineages likely similarly evolved from a primarily asexual state that had lost the ability to use methylation to silence TEs. Instead they came to rely on alternative repressive pathways upon sexual reproduction, such as histone modifications or piRNAs (Aravin et al., 2007; Korf et al., 1998). While some invertebrate genomes thus contain DNA methylation, it is not necessarily targeted towards TEs as observed in the sea squirt *Ciona intestinalis*, and there is no evidence it is involved in silencing amongst these lineages (Feng et al., 2010; Zemach et al., 2010). At the same time, the loss of the ancestral methylation-dependent TE silencing pathway in early animal evolution implies that the vertebrate lineage independently 're-evolved' the use of methylation for TE defence but could in addition build on the existing methylation-independent silencing mechanisms from invertebrate ancestors. This makes vertebrates less dependent on strictly maintaining high methylation levels at all times. Land plants on the other hand, whose use of methylation for TE silencing goes back in an uninterrupted line to ancestral eukaryotes (Zemach et al., 2010), do not show any major fluctuations in methylation during their life cycle (Zemach and Zilberman, 2010).

The presence of other repeat silencing pathways could explain why vertebrates can undergo periods of global low methylation in the germline. This allows them to reset their (epi)genome to a basic, totipotent state before establishing sex-specific and germ cell-specific epigenetic signatures and

transcription profiles (Messerschmidt et al., 2014). At the same time, however, transcription and transposition in the germline is the way to evolutionary success for TEs, since activity in somatic cells would harm the host fitness without increasing the copy number of the TE in the host's descendants. Indeed there is measurable transcriptional activity of ERVs, the evolutionarily youngest endogenous retroviruses, in both the mouse and human germline (Brûlet et al., 1983; Dupressoir and Heidmann, 1996; Göke et al., 2015; Grow et al., 2015; Seisenberger et al., 2012; Tang et al., 2015).

The host is faced here with the challenge of not only silencing existing copies of these TEs, but also recognizing new transposition events, while at the same time not impeding transcription at older insertions that have been co-opted to have regulatory functions. The piRNA pathway, the primary repeat silencing strategy in *Drosophila*, seems to function as an immediate *de novo* silencing response in the vertebrate germline, using the transcripts generated by TEs as a targeting mechanism (Molaro and Malik, 2016). In an alternative and evolutionary slower response, KRAB zinc-finger proteins (KZFPs) can recognise defined sequence elements through a unique combination of zinc fingers and globally repress elements of the same family without need for their expression (Molaro and Malik, 2016). These proteins make up the largest single family of transcriptional regulators in mammals and are abundantly expressed in the germline (Ecco et al., 2016). Long an understudied group of proteins, very recently hundreds of KZFPs could be assigned to their targets within specific TE families in humans (Imbeault et al., 2017; Schmitges et al., 2016). In an 'arms race' between host and TEs, retroelements have been suggested to change their sequence to evade KZFP binding, whereas KZFPs counteract this development by gene duplication and diversification (Molaro and Malik, 2016). Indeed the speed of KZFP gene duplication mirrors that of retroelement family diversification (Thomas and Schneider, 2011). Both the piRNA and the KZFP pathway are thought to ultimately lead to deposition of the repressive H3K9me3 mark at TEs (Padeken et al., 2015), which is essential for silencing in the hypomethylated vertebrate germline (Liu et al., 2014).

2.2.3.2 Distribution of CpGs and methylation in vertebrate genomes

In vertebrates, DNA methylation is set by the *de novo* methylating enzymes DNMT3a and 3b in the context of CpG dinucleotides and maintained upon cell division by DNMT1 (Hermann et al., 2004). Methylation levels can be reduced either passively through cell divisions (Chen et al., 2003), or actively by the TET family of enzymes (Tahiliani et al., 2009). Apart from the brief phases of global demethylation in the germline however, vertebrate genomes are unique in that they are characterised by almost blanket methylation, suggesting this is the default state. Most (~90%) 5-methylcytosine residues in human DNA lie within TE repeats (Yoder et al., 1997). Deamination of methylated CpGs leads to their progressive loss over time, and this cost of the genome defence is not limited to repetitive regions (Cooper and Youssoufian, 1988). Accordingly, the CpG dinucleotide occurs at only 20% of the expected frequency in vertebrate genomes. Exceptions to this rule are CpG islands (CGIs) that overlap frequently with promoter regions (Bird, 1986). These regions are able to maintain their expected CpG content, since they tend to be unmethylated in the germline apart from some exceptions (Smallwood et al., 2011).

Methylation of CGI promoters has been shown to cause robust transcriptional repression (Busslinger et al., 1983; Schubeler et al., 2000) and is at the basis of the two established incidents of long-term mono-allelic silencing (Illingworth and Bird, 2009): X chromosome inactivation (Jaenisch and Bird, 2003; Panning and Jaenisch, 1996) and genomic imprinting (Bourc'his et al., 2001; Li et al., 1993). Although methylation of CpGs was thus primarily evolved for repeat defence, this silencing mechanism has likely been co-opted by vertebrates for other means. Of note, the retrotransposons that are still transcriptionally competent and rely on methylation for their silencing in differentiated cells (Jähner et al., 1982; Walsh et al., 1998) similarly have high CpG content (Fig. 2-3a). Another feature these three prime examples of DNA-methylation mediated silencing have in common is that silencing is essentially irreversible over the life span of the organism. Thus DNA methylation has been suggested to 'lock down' inactive sequences and commit them to long-term silencing even in the presence of all factors needed

for their activation (Bestor et al., 2015; Jones, 2012). DNA methylation and silencing of CpG-rich regions often go hand in hand with accumulation of the H3K9me3 mark, not only at repeats (Dong et al., 2008).

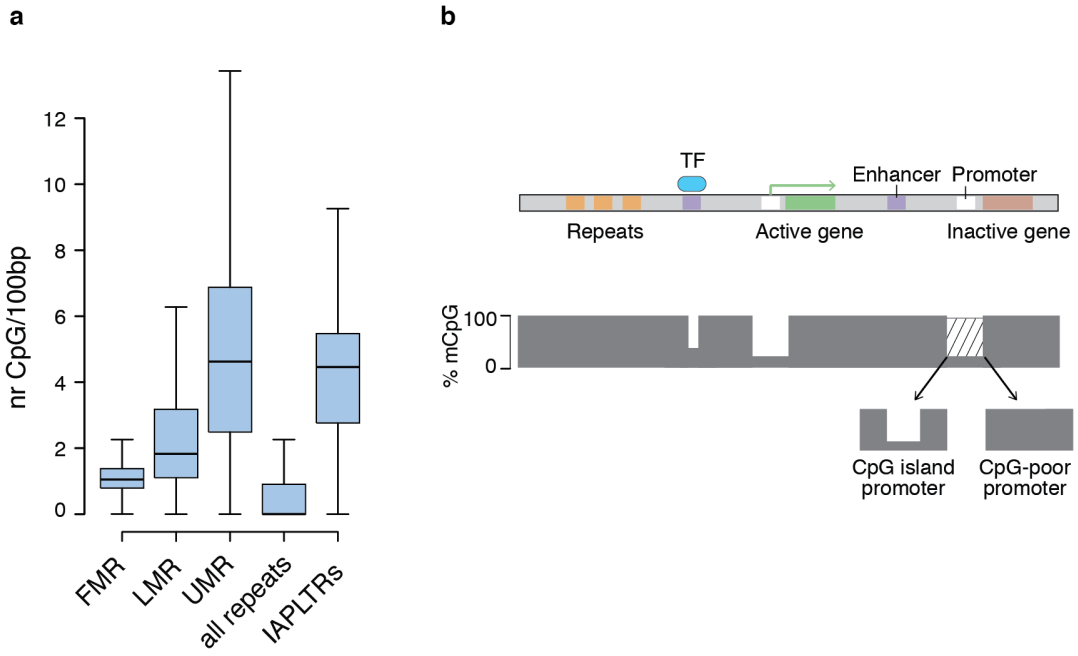


Figure 2-3. Unequal distribution of CpGs and DNA methylation in the vertebrate genome.
a) High CpG content can be found at unmethylated regions and evolutionarily young repeats. Segmenting the genome of mouse embryonic stem cells into fully methylated regions (FMR), lowly methylated regions (average 30%, LMR) and unmethylated regions (UMR) reveals a clear correlation of methylation and CpG content. Most of the genome, including repeat regions, consists of FMRs. The vast majority of CGIs are UMRs, which frequently lie close to gene transcription start sites. LMRs mostly reside distal to gene transcription start sites and overlap with distal regulatory regions/ enhancers. While repeats/ TEs generally have a low CpG content in line with other FMRs, promoter regions of the evolutionarily youngest and most active group in rodents (IAPLTRs) still retain a high CpG content in spite of being fully methylated. For details on UMRs, LMRs and FMRs see Stadler et al., 2011. Boxplots show median (black line), 25th and 75th percentiles (boundaries), minimum and maximum (whiskers). **b)** Vertebrate genomes are characterised by local dips in otherwise blanket methylation (mCpG) over active enhancers, CGIs and active CpG-poor promoters. Schematic representation adapted from Schübeler, 2015.

In spite of the clear link between cytosine methylation and gene silencing at CGIs, surprisingly few of them actually change their methylation state during development. Apart from germline-specific genes that require DNA methylation for their silencing in somatic cells, most CGI promoters remain unmethylated in all tissue types regardless of their activity (Borgel et al., 2010). Instead they acquire the Polycomb-group-mediated H3K27me3 mark

when silent (Fig. 2-3b) (Lynch et al., 2012; Tanay et al., 2007). However, perturbations of their methylation state are frequently observed in diseases, especially in cancer, where methylation of CGI promoters for tumour suppressor genes has been reported (Jones, 2012).

The advent of bisulfite and next-generation sequencing enabled genome-wide mapping of DNA methylation at nucleotide resolution across different cell types (Lister et al., 2009; Stadler et al., 2011; Ziller et al., 2013). This revealed that not only CGIs but also other active regulatory regions are characterised by low methylation levels (Fig. 2-3b). These CpG-poor regulatory regions include roughly one quarter of promoters, regulating for the most part tissue-specific genes, and the vast majority of enhancers (Bestor et al., 2015). In contrast to CGI promoters, CpG-poor regulatory regions tend to have higher levels of methylation when inactive (Fig. 2-3b) (Schübeler, 2015). The majority of dynamic methylation changes observed between cell types, tissues or individuals occurs at distal enhancers and is matched by both differential TF occupancy and gene activity (Zhang et al., 2013a). It is tempting to attribute these observations to an instructive role of DNA methylation in tissue-specific gene silencing, by regulating binding of TFs.

However, recently it was shown that certain TFs such as REST, CTCF and several other factors are able to bind methylated CpG-poor regulatory regions and induce their local demethylation (Boller et al., 2016; Boulard et al., 2015; Han et al., 2001; Stadler et al., 2011; Wang et al., 2015). It is currently unclear how this local reduction in methylation levels is brought about, but it likely involves a component of active demethylation (Feldmann et al., 2013). Thus dynamic changes in methylation patterns at CpG-poor regulatory regions across cell types could also be a mere consequence of differential TF binding. Indeed there is currently no experimental evidence for methylation-dependent silencing occurring at CpG-poor regions (Bestor et al., 2015; Schübeler, 2015). Interestingly, TF binding has been implicated in maintaining the unmethylated state even at CGIs (Brandeis et al., 1994; Krebs et al., 2014; Macleod et al., 1994). In the following I will review the existing evidence

for an instructive role of DNA methylation in regulating TF binding both *in vitro* and *in vivo*.

2.2.3.3 DNA methylation and transcription factor binding *in vitro*

TF binding is associated with absence of DNA methylation at regulatory regions genome-wide (Baubec and Schübeler, 2014; Gal-Yam et al., 2008; Naveh-Many and Cedar, 1981). This raises the question if methylation patterns are the cause or the consequence of differential TF binding (Fig. 2-4a). Given the ability of methylation to silence inserted DNA (Jähner et al., 1982; Stein et al., 1982) the historic view was that methylation patterns directly determine the activity of genes (Cedar et al., 1983). However, even for the long-established examples of methylation-dependent silencing such as TE repression, it remains unclear how this effect is actually achieved by the cytosine modification. There are two popular explanations (Fig. 2-4b): On the one hand, methylation could block TF binding in an indirect manner through methyl-CpG-binding domain proteins (MBDs) recognizing dense arrays of methylated CpGs and recruiting histone deacetylases (Nan et al., 1998; 1996). This would lead to chromatin compaction and thus exclusion of TFs independent of their sequence motifs. However, *in vivo* evidence for this model on a genome-wide level is still lacking. Deletion of individual MBDs does not affect gene expression (Hendrich et al., 2001; Tudor et al., 2002), although it cannot be excluded that different MBD family members can compensate for each other. On the other hand, methylation of cytosines within a sequence motif could directly obstruct binding by affecting the shape and base readout of the matching TF (Dantas Machado et al., 2015). Such sensitivity of TFs to methylation of their binding site was indeed observed *in vitro* for USF, c-MYC, NF-κB, E2F and CTCF as well as for an undefined factor at the cAMP-responsive element (CRE), which all preferentially bound to an unmethylated stretch of DNA in gel shift assays (Bednarik et al., 1991; Campanero et al., 2000; Iguchi-Ariga and Schaffner, 1989; Prendergast and Ziff, 1991; Watt and Molloy, 1988). Crystal structures to support a direct disruptive effect of the methyl-group on these protein-DNA interactions are still

missing. For other factors there has been conflicting evidence. For example, SP1 binding has been reported to be indifferent to methylation (Höller et al., 1988), to be blocked by methylation (Clark et al., 1997) or to prevent methylation from accumulating at CGIs (Brandeis et al., 1994).

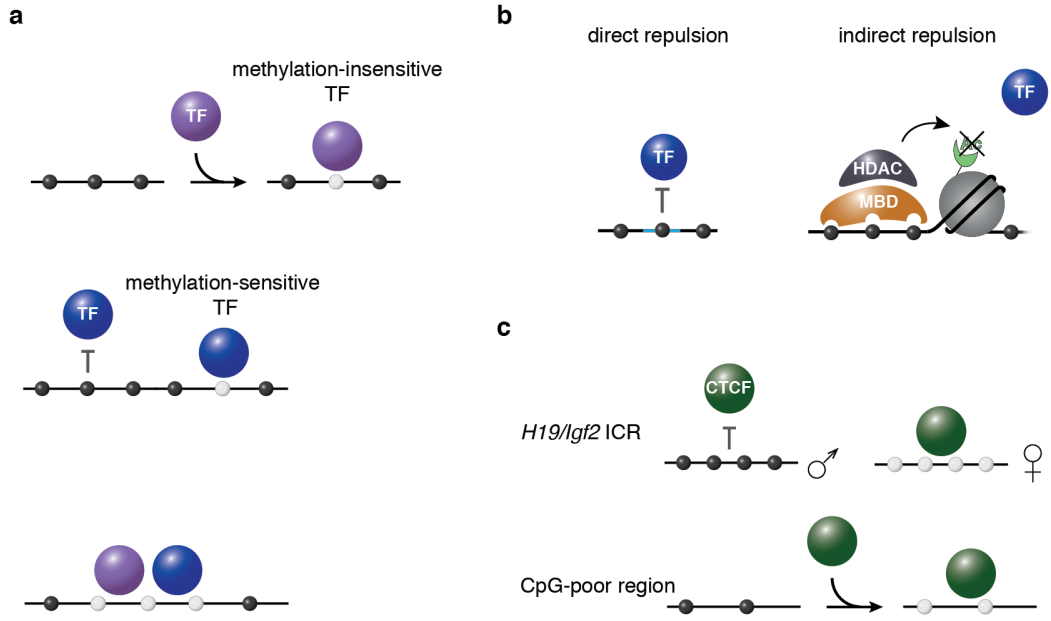


Figure 2-4. Absence of a simple rule for the relationship between DNA methylation and TF binding.
a) Possible scenarios to explain the genome-wide anticorrelation between DNA methylation and TF binding outside of CpG islands. TFs could be methylation-insensitive, capable of binding methylated sites and inducing local demethylation, thus shaping cell type-specific methylation patterns (top). Alternatively, TFs might be methylation-sensitive and require an unmethylated state to enable their binding, giving DNA methylation an instructive role (middle). These two extreme scenarios could apply differently across factors and even sequence contexts. Insensitive factors might induce demethylation and thus enable other sensitive factors to bind the same region (bottom), but this differential behaviour would not be apparent from measuring steady-state methylation and TF binding profiles. **b)** Suggested mechanisms of DNA methylation-based repression. DNA methylation could directly impede TF binding by steric influence of the methyl-CpG group in the DNA sequence motif on the protein-DNA interaction (left). Alternatively, methyl-CpG binding domain proteins (MBDs) have been proposed to bind arrays of methylated CpGs and induce chromatin compaction by recruiting histone deacetylases (HDACs), thus indirectly blocking TF binding independent of specific sequence motifs (right). **c)** Methylation sensitivity of CTCF is context dependent. CTCF is unable to bind the methylated paternal allele of the *H19/Igf2* ICR (top). On the other hand, binding of CTCF can occur in the presence of methylation at a reporter construct and induce local hypomethylation, as shown in Stadler et al., 2011 (bottom).

Technology development has enabled large-scale studies of the effect of DNA methylation on *in vitro* TF binding in recent years. Spruijt et al. used

mass spectrometry to identify proteins in nuclear extracts from mouse embryonic stem (ES) cells that bind an immobilised methylated or unmethylated DNA template (Spruijt et al., 2013). This study identified some TFs that preferentially bound the unmethylated template, including two of the TFs listed above, NF-κB and ATF/CREB factors that normally bind the CRE motif. However, in fact only one DNA sequence template consisting mostly of ACG repeats was used in this study, which conflicts with the sequence-specific binding nature of TFs. In an alternative approach Hu et al. spotted 1,321 human TFs on a protein microarray and measured binding to synthesised *in vitro* methylated templates in competition with unmethylated templates (Hu et al., 2013). While this study interrogated 154 CpG-containing TF motifs, it only reports factors that preferentially bind methylated sites and thus the findings cannot be systematically compared with Spruijt et al. or the gel shift experiments. Of note, Hu et al. describe several factors that alter their motif preference in the presence of DNA methylation. Mann et al. performed the inverse experiment and used a double-stranded *in vitro* methylated or unmethylated DNA microarray with 65,536 octamers which they incubated with eight mouse bZIP TF family members (Mann et al., 2013). While they observed preferential binding to methylated sequences for some factors, others were blocked by DNA methylation, e.g. CREB. A recent study in the flowering plant *Arabidopsis thaliana* avoided synthesis of DNA oligomers and instead used fragments of genomic DNA, thus obviating the need to methylate *in vitro* (O'Malley et al., 2016): 1,812 *in vitro*-expressed TFs were bound to beads and incubated with naked genomic DNA fragments. Comparison with largely unmethylated DNA fragments generated by PCR nominated roughly 180,000 TF binding sites occluded by DNA methylation. Fewer sites were gained and cytosine content of TF motifs correlated with binding sensitivity to 5-methylcytosine. Of note, *Arabidopsis* genomes are methylated at cytosines also outside of the CpG context and thus contain more than twice as much methylation as vertebrate genomes (Cokus et al., 2008; Schmitz et al., 2013).

Additional studies to investigate *in vitro* binding preferences of TFs in the presence of DNA methylation are underway. Perhaps most prominent is the

adaption of the systematic evolution of ligands by exponential enrichment (SELEX) method for methylated DNA templates, which relies on affinity-tagged DNA-binding domains, barcoded selection of bound oligonucleotides, and multiplexed sequencing (Jolma et al., 2010).

Interestingly, the main focus of these large-scale studies in vertebrates in recent years has been to identify TFs that preferentially bind methylated DNA, since a negative effect of methylation on TF binding is largely taken for granted. However, apart from the single-factor/ single-locus examples mentioned above, there is currently no evidence for widespread binding-site restriction by DNA methylation in vertebrates.

2.2.3.4 DNA methylation and transcription factor binding *in vivo*

Large-scale *in vitro* studies of TF methylation sensitivity are valuable starting points and will hopefully be expanded in coming years. However, it is becoming increasingly clear that TF binding depends to a large part on the sequence-, chromatin- and cellular context and these factors will need to be considered if we want to reach the ultimate goal of predicting genome-wide TF binding and gene activity. For example, DNA methylation seems to have disparate effects at CpG-dense versus CpG-poor regions. Many additional factors come into play within a cell that are not captured in *in vitro* binding experiments, such as the presence of DNA methylation readers and writers, e.g. MBDs, co-factors and other TFs as well as various other chromatin components mentioned above. In addition, binding affinities measured *in vitro* can be in a realm that is not naturally relevant within cells in terms of DNA binding site and protein concentration. Of note, just because a factor is sensitive to methylation in a certain sequence context does not mean the TF is actually restricted in its binding by this mark in a cell, e.g. if high affinity sites are all unmethylated in the first place or inaccessible for other reasons. Accordingly, transferring observations of TF binding behaviour and methylation sensitivity from *in vitro* to *in vivo* binding site predictions has been difficult. For example, *in vitro* blocking of CREB binding at the *Tat* promoter sequence suggested that methylation would be responsible for regulating

binding at this site. However, removal of methylation was unable to induce binding in a cell type that is normally inactive for this gene and where CREB is highly expressed (Weih et al., 1991).

Currently there are no studies that systematically investigate the influence of DNA methylation on binding site restriction *in vivo*. This is likely in part due to the fact that DNA methylation is essential for cell survival in most tested mammalian cells (Chen et al., 2007; Liao et al., 2015), making loss of function effects hard to study. At the same time this means that in spite of the largely correlative nature of the DNA methylation and gene expression relationship, this mark has a crucial role in cell survival. Cell death has been attributed in turn to misregulation of critical genes (Jackson-Grusby et al., 2001) or activation of repeats (Walsh et al., 1998; Yoder et al., 1997) and was linked to DNA damage response (Shaknovich et al., 2011) and mitotic catastrophe (Chen et al., 2007). To date it remains unclear to which extent it is driven by differential TF binding or other global responses.

2.2.3.5 Example CTCF

The CCCTC-binding factor CTCF is likely the most prominent example for a methylation-sensitive TF and is one of the few cases that has indeed been shown to bind in a methylation-sensitive manner not only *in vitro* but also *in vivo*. This factor nicely illustrates the complex relationship between DNA methylation and TF binding and how far we are from fully understanding it in spite of a wealth of experiments.

Over the past fifteen years, the relationship between CTCF binding and DNA methylation has been studied in detail at the imprinting control region (ICR) of the *H19/Igf2* locus, resulting in more than a hundred publications on this topic and region. CpG methylation in the core motifs was shown to prevent CTCF binding *in vitro* (Bell and Felsenfeld, 2000; Hark et al., 2000; Renda et al., 2007). *In vivo*, CTCF binds and acts as an insulator only at the unmethylated maternal allele but not at the methylated paternal allele, giving rise to allele-specific gene expression (Fig. 2-4c) (Szabó et al., 2000). Demethylation of the *H19/Igf2* ICR by treatment with 5-Aza-deoxycytidine

leads to biallelic binding and expression (Ito et al., 2013), whereas mutation of CTCF binding sites in the ICR results in gain of methylation at the maternal allele (Schoenherr et al., 2003; Szabó et al., 2004).

While the exact mechanism that repels CTCF from the methylated allele remains to be elucidated, these observations have led to the belief that methylation within the core motif is generally instructive for CTCF binding (Filippova, 2008). Indeed, on a genome-wide level, an inverse relationship between CTCF binding and methylation was found in many cell types (Mukhopadhyay et al., 2004; Wang et al., 2012a). However, these studies do not address whether DNA methylation itself prevents binding *in vivo* or whether bound sites become hypomethylated as a secondary effect. Indeed it has been demonstrated that CTCF binding itself can create reduced methylation states, by binding to a methylated CpG-poor region and leading to local demethylation (Fig. 2-4c) (Stadler et al., 2011). In stem cells without DNA methylation (DNA methyltransferase triple knockout cells, TKOs) (Tsumura et al., 2006), CTCF binding was not drastically altered on a genome-wide level, with the notable exception of several imprinted regions including the *H19/Igf2* ICR (Stadler et al., 2011). These findings argue against a general role for DNA methylation in preventing CTCF binding *in vivo* and stand in stark contrast to the methylation sensitivity observed at the *H19/Igf2* ICR (Fig. 2-4c). It is currently unclear how to reconcile these findings and which factors influence CTCF methylation sensitivity in the cellular context.

2.3 Studying binding site restriction of transcription factors *in vivo*

The difficulty to find a unifying rule for how different chromatin components impact TFs has made it increasingly clear that their influence on binding is likely both factor and context specific (Slattery et al., 2014): For each chromatin component a whole spectrum of sensitivities could exist among the various TFs and sequence context can contribute to further distinguish

otherwise identical binding sites for an individual TF, e.g. by impacting DNA shape or co-factor binding (Levo and Segal, 2014; Slattery et al., 2014).

Removing individual chromatin components and assessing the impact on TF binding across the whole genome *in vivo* would thus be invaluable for gaining insights into the complex role these features have in TF binding site restriction and to tease apart causation and correlation (Fig. 2-5). Indeed methods to measure genome-wide binding of specific TFs, such as ChIP-seq, or chromatin accessibility as an indicator of the entire cellular TF binding landscape, such DNase- or more recently ATAC-seq, are well developed (Levo and Segal, 2014). The removal of individual repressive chromatin components is less straightforward, since they are generally essential for cell survival. DNA methylation has two advantages that make it a prime candidate for a proof-of-concept study and thus the focus of this thesis: First, it can be mapped to nucleotide resolution by bisulfite sequencing. Second, mouse ES cells have been shown to survive in its absence (Tsumura et al., 2006), providing us with an ideal model system to study the influence of DNA methylation on TF binding.

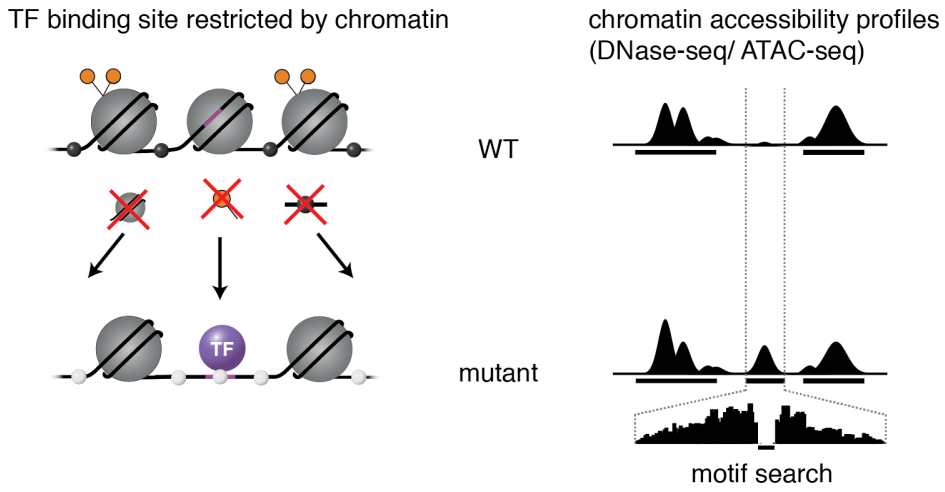


Figure 2-5. Studying binding site restriction by chromatin.
Hypothetical experimental approach for studying the influence of individual chromatin features on TF binding site restriction. Individual chromatin features such as nucleosomes, repressive histone marks or DNA methylation could be genetically removed or depleted. Determining genome-wide chromatin accessibility with DNase-seq or ATAC-seq as indicator for TF binding in each of these mutants would allow identification of sites only bound in the absence of a given modification. Sequence analysis of these sites should nominate candidate TFs that are otherwise blocked from binding by this mark. Importantly, gene expression changes need to be limited in mutants to allow assignment of differentially accessible sites to loss of binding site restriction rather than secondary effects.

2.4 Open questions and scope of this thesis

Taken together, we currently lack an understanding of how and to which extent different chromatin features influence TF binding. DNA methylation is a mark that is comparatively easy to manipulate and measure at base-pair resolution. Nonetheless, its influence on TF binding remains unclear in spite of a vast array of literature that has been amassed on this subject over the past more than three decades.

Some of the key open questions in the field are: Can DNA methylation have an instructive role in TF binding site restriction in vertebrate cells or are methylation patterns only generated downstream of TF binding? If yes, for which factors among the large TF family does DNA methylation block binding? Does this occur at all motifs for candidate factors or only in certain chromatin or sequence contexts? How is binding site restriction actually brought about: Is it due to indirect changes in chromatin environment or by direct steric alterations in the sequence-specific DNA-protein interaction? Finally, which role does TF binding site restriction play in the essential nature of DNA methylation?

In view of the exploding number of epigenetic and TF binding maps being collected across species, tissues, developmental and disease stages, answering these questions would bring us one step closer towards predicting dynamic TF binding and ultimately gene regulation during development and disease.

For this thesis, I addressed these open questions by investigating the influence of DNA methylation on TF binding in the cellular context. First, we focused on understanding genome-wide CTCF methylation sensitivity. This TF is the most prominent example for a methylation-sensitive TF, yet it remains unclear in which sequence contexts DNA methylation restricts its binding. Second, we aimed to identify further methylation-sensitive TFs by comparing genome-wide TF binding in mouse ES cells in the presence and absence of DNA methylation. For the identified factors, we investigated possible mechanisms of binding site restriction and studied the interplay

between TFs and DNA methylation. Third, since DNA methylation is essential for cell survival only in differentiated cells and is thus expected to have a larger impact there, we expanded this approach to a differentiated cell state in the form of methylation-deficient neurons. Apart from analysing differential TF binding in this context, we also explored the impact of DNA methylation loss on expression and cell survival.

3. Results

3.1 Methylation sensitivity of the transcription factor CTCF

3.1.1 Abstract

CTCF plays a key role in the three-dimensional organisation and transcriptional regulation of vertebrate genomes. Binding of this TF has been shown to be sensitive to DNA methylation at the *H19/Igf2* imprinting control region (ICR), yet it is not restricted by DNA methylation at the vast majority of genomic sites. In order to understand this apparent context-dependent influence of DNA methylation, we compared CTCF binding in isogenic mouse stem cells with and without DNA methylation. We find that the couple hundred CTCF sites only bound in the absence of DNA methylation are characterised by CpGs at certain positions in the motif as well as a higher CpG density in the flanking regions. Of note, these features also hold true at the *H19/Igf2* ICR. In addition, we show that methylation sensitivity at this well-studied region is indeed encoded in the sequence and not dependent on the chromosomal location or allele-specific enrichment of H3K9me3. Comparing CTCF binding at ectopically inserted methylated and unmethylated sequence libraries is a means to test the impact of these and other sequence features on CTCF methylation sensitivity. Clonal variability in methylation states was observed for ectopic genomic insertions of the *H19/Igf2* ICR fragments. However, we suggest several strategies to overcome this issue and to comprehensively decode the context-dependent influence of DNA methylation on CTCF binding, thus facilitating the interpretation of epigenomic and topological maps.

3.1.2 Introduction

CTCF is one of the best-studied and most frequently cited examples for a methylation-sensitive TF, making it the natural starting point for elucidating the influence of DNA methylation on TF binding. This essential and highly conserved zinc-finger protein has been implicated in a myriad of biological processes (Ohlsson et al., 2001). By mediating long-range intrachromosomal interactions, CTCF is thought to demarcate the boundaries of topologically associated domains (Ghirlando and Felsenfeld, 2016), i.e. chromosome neighbourhoods which frequently interact within but not between each other. Thus, this 'Master Weaver of the genome' (Phillips and Corces, 2009) limits interactions across its binding site and acts as an insulator (Bell et al., 1999; Hark et al., 2000).

The influence of DNA methylation on CTCF binding has been studied in detail at the imprinting control region (ICR) of the *H19/Igf2* locus, which contains four CTCF binding sites in mice. CTCF only binds at the unmethylated maternal allele, but not at the methylated paternal allele (Szabó et al., 2000). The vast majority of CTCF sites however are not restricted by DNA methylation in ES cells (Stadler et al., 2011). In contrast, CTCF was shown to be able to bind a methylated CpG-poor region and induce its demethylation. It is currently unclear how to reconcile these observations and in which genomic contexts methylation indeed affects CTCF binding.

The impact of altered CTCF binding can be substantial, even if it occurs only at few genomic sites. Aberrations in methylation states and CTCF binding and thus enhancer looping have been linked to misexpression of *H19/Igf2* in Beckwith-Wiedemann and Russell-Silver growth defect syndromes (Herold et al., 2012) as well as more recently to oncogene activation in glioma (Flavahan et al., 2016). Being able to predict which CTCF sites are impacted by methylation changes is therefore crucial in order to interpret the growing number of genome-wide epigenetic and topological maps. To this end, we aimed to investigate the context-dependent influence of DNA methylation on CTCF binding in a systematic manner and to determine the sequence features involved in its methylation sensitivity.

3.1.3 Results

In order to investigate which sequence parameters influence methylation sensitivity of CTCF in the cellular context, we planned to compare binding at differentially methylated sequences in an otherwise controlled environment. This is achieved by inserting sequence variants into the same ectopic genomic site via Cre recombinase-mediated cassette exchange (RMCE) (Feng et al., 1999; Lienert et al., 2011). The sequences can be methylated *in vitro* prior to insertion and integrated in the methylated as well as in the unmethylated state (Fig. 3-1, right). This strategy enables measurement of CTCF binding at an identical sequence and chromosomal location that only differs in the methylation state. Comparison of sequences that are bound by CTCF regardless of methylation levels and those that are only bound in the unmethylated state should elucidate which parameters are required for methylation-sensitive binding. Hundreds of sequences can be inserted in parallel in a library approach and methylation status and CTCF binding at these different inserts can be read out in the same experiment (Krebs et al., 2014; Barisic et al. unpublished). This makes it feasible to systematically test the influence of many sequence parameters in a high-throughput fashion.

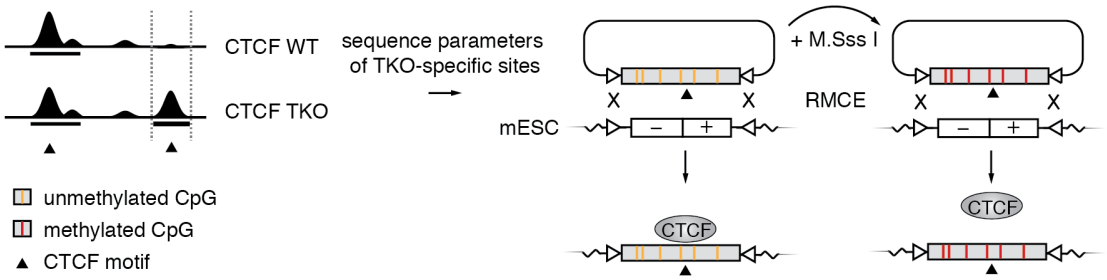


Figure 3-1. Overview of experimental approach for studying CTCF methylation sensitivity.
Comparison of CTCF binding measured by ChIP-seq in cells with (WT) and without (*Dnmt* TKO) DNA methylation can identify putative methylation-restricted CTCF binding sites only occupied in the TKO (left). Characteristic sequence features of these sites are tested for their contribution to CTCF methylation sensitivity using a genomic editing approach (right): The same sequences are inserted unmethylated and premethylated (using M.SssI) into an ectopic genomic site in mouse ES cells (mESC) by recombinase-mediated cassette exchange (RMCE), using rounds of positive and negative selection. Maintenance of methylation state and CTCF binding at the inserts are monitored by bisulfite sequencing and ChIP-qPCR. Insertion of a complex plasmid library of sequence variants in a methylated and unmethylated state would allow measurement of relative CTCF binding at differentially methylated but otherwise identical sequences for hundreds of variants in parallel in a single experiment.

3.1.3.1 A subset of CTCF binding events occur only in the absence of DNA methylation

In order to reduce the sequence search space and have a starting point for library design, we first sought to identify potential sequence features that might be involved in methylation-sensitive CTCF binding (Fig. 3-1, left). We compared CTCF binding in cells with and without DNA methylation to identify binding sites that are restricted by DNA methylation. This was already previously performed by our lab in an ES *Dnmt* TKO cell line generated by traditional mouse genetics and a wildtype ES cell line of a different strain background (Stadler et al., 2011). However, overall low fold-changes made it difficult to distinguish sites that were differentially bound due to methylation loss or due to biological and experimental variability.

To minimise clonal and strain influence and enable confident detection of subtle differences, we compared CTCF binding by high-coverage ChIP-seq in WT and isogenic *Dnmt* TKO ES cell lines generated by CRISPR genome editing (Domcke et al., 2015) (Fig. 3-2a,b). Variation was slightly larger between cell lines than between replicates (Pearson correlation coefficient of 0.89 vs. 0.94, Fig. 3-2a,b,c), implying that some sites are indeed bound in a methylation-dependent manner. However, overall there were few TKO-specific CTCF sites and a similar amount of WT-specific sites. To validate these cell line-specific sites by independent means, we analysed high-coverage DNase-seq data at all low-confidence CTCF motifs in both cell lines (Domcke et al., 2015). When comparing changes between TKO and WT, we observe correlation for ChIP-seq and DNase-seq signal (Fig. 3-2d). Importantly, TKO-specific sites show higher correlation in these measures than WT-specific sites, which are largely not shared between the methods. Sites with at least two-fold change in both ChIP-seq and DNase-seq were selected for further analysis (Fig. 3-2e). The 202 TKO-specific sites called with this cut-off are fully methylated in WT, in contrast to the 69 WT-specific sites, implying that they might indeed be methylation-sensitive binding events (Fig. 3-2f).

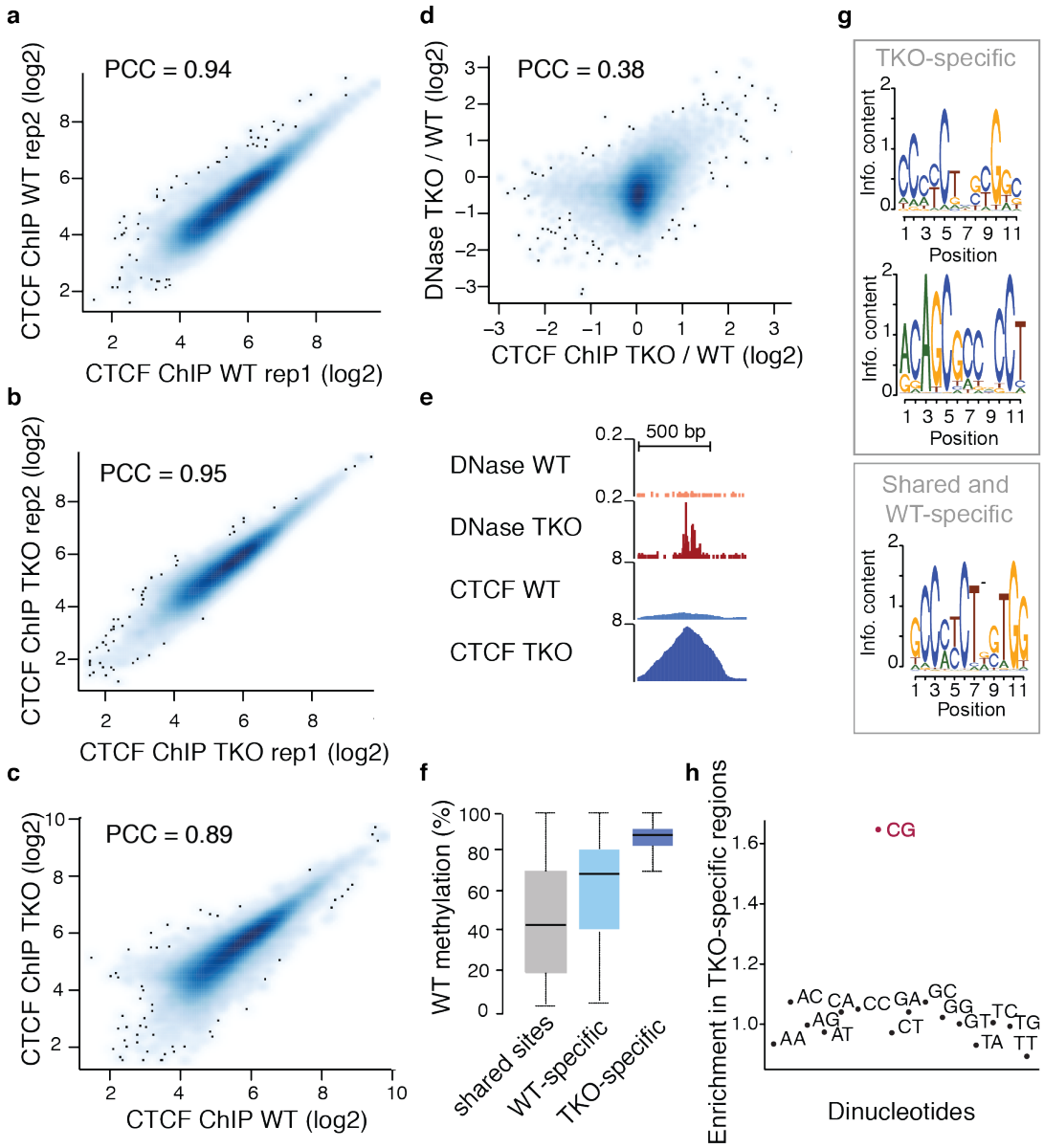


Figure 3-2. Identification of putative methylation-sensitive CTCF sites and their sequence characteristics.

a, b) Comparison of CTCF binding in WT (a) or TKO (b) ES cells for two biological ChIP-seq replicates. PCC = Pearson correlation coefficient. **c)** Comparison of CTCF binding in WT and TKO ES cells as measured by ChIP-seq (mean of replicates). **d)** Comparison of differential CTCF binding and chromatin accessibility between TKO and WT ES cells in a 300 bp window around CTCF motifs. **e)** Example for chromatin accessibility measured by DNase-seq and CTCF binding measured by ChIP-seq at a 'TKO-specific' CTCF site (chr6: 29,042,850-29,043,500). **f)** WT methylation levels measured by whole-genome bisulfite sequencing in 300 bp CTCF peak regions bound either in both WT and TKO ES cells or only in one of the two cell lines. Boxplots show median (black line), 25th and 75th percentiles (boundaries), minimum and maximum (whiskers). **g)** Top enriched motifs identified by *de novo* motif enrichment in TKO-specific (top; found in 48% and 20% of regions; p-value 1e-94 and 1e-38) or shared and WT-specific peaks (bottom; found in 68% of shared and 59% of WT-specific regions; p-value 1e-25356 and 1e-21), compared to remaining sites overlapping CTCF motifs. **h)** Enrichment of dinucleotides in flanking regions (excluding the CTCF motif) of TKO-specific over shared peaks (CG observed/expected = 0.3 in TKO-specific, 0.2 in shared sites).

3.1.3.2 TKO-specific sites contain more CpGs in the motif and the flanking regions

Next we asked whether there are some sequence features that set the TKO-specific sites apart from the majority of constitutive CTCF sites. *De novo* motif enrichment found CTCF motifs to be highly enriched in the called peaks, in line with how the original set of regions was selected. Interestingly, the two most strongly enriched motifs found in nearly three quarters of the TKO-specific sites are two non-canonical CTCF motifs (Fig. 3-2g). These motif variants both contain a central CpG, in contrast to the canonical CTCF motif found in more than two thirds of the non-changing sites, as well as in most of the WT-specific sites (Fig. 3-2g). Two recent studies reporting methylation sensitivity of CTCF at single loci also focus on motifs containing CpGs at just these positions: The motif in the *Pou5f1* locus that was artificially methylated by targeting a dCas9-DNMT3a fusion protein by Liu et al. closely matches the first variant (Liu et al., 2016), whereas the motif in the insulator of the *Pdgfra* oncogene studied by Flavahan et al. corresponds to the second variant (Flavahan et al., 2016). In contrast, the motif inserted into a methylated CpG-poor sequence that was bound by CTCF and demethylated is the best match to the canonical CTCF motif (Stadler et al., 2011).

In addition to the motif itself, flanking regions could also impact methylation sensitivity (Levo and Segal, 2014). Analysis of the sequence composition of flanking regions revealed that CpG dinucleotides are enriched in the surroundings of TKO-specific sites (Fig. 3-2h). Of note, motifs containing CpGs are more likely to occur in regions with overall higher CpG density, so these two features might be linked.

The *H19/Igf2* ICR can serve as a positive control for sequence features involved in methylation-sensitive CTCF binding. As expected, we observe a two-fold increase in CTCF ChIP-seq and DNase-seq signal at this region (Fig. 3-3a) and the binding sites fall within our TKO-specific peak set. The ICR indeed contains the CpG-variant forms of the CTCF motif (Fig. 3-3b). In addition, it has a higher CpG density than the rest of the genome (CpG content observed/expected = 0.4 versus 0.2 for regions with constitutive

CTCF binding). Thus, CpG density of the flanking region and CpGs at certain positions in the motif are likely features contributing to context-dependent CTCF methylation sensitivity.

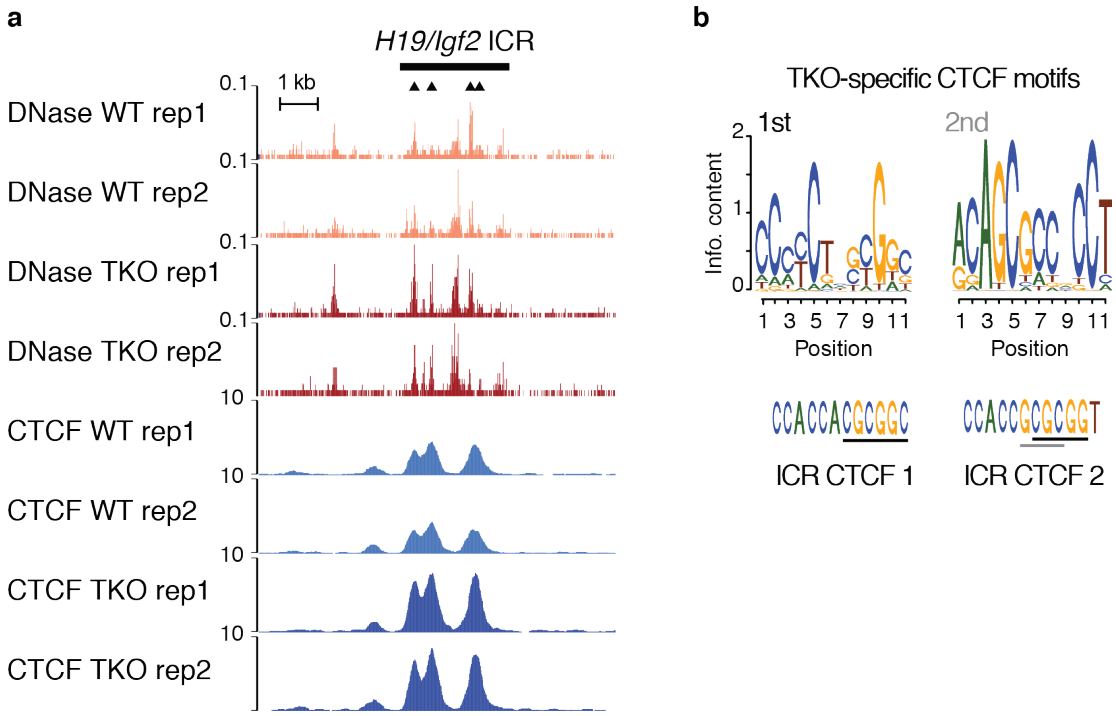


Figure 3-3. CTCF binding sites in the *H19/Igf2* ICR.
a) DNase-seq and ChIP-seq signal at the *H19/Igf2* ICR in WT and TKO ES cells for two biological replicates each. Triangles mark the position of the CTCF motifs in the ICR. **b)** Comparison of motifs identified in the majority of TKO-specific CTCF binding sites (PWM, top) and the first two CTCF sites in the *H19/Igf2* ICR (bottom). CpGs are found in similar positions of the motif as in the most (black) or second most (grey) enriched TKO-specific motifs.

3.1.3.3 Methylation sensitivity of CTCF can be recapitulated at an ectopic site

Having identified possible methylation-sensitive CTCF sites and some of their key sequence features, we next wanted to validate these by inserting unmethylated and premethylated sequence variants into an ectopic genomic site and measuring relative CTCF binding by ChIP-qPCR. As a proof of principle, we decided to first insert fragments of the *H19/Igf2* ICR to probe whether its known methylation sensitivity is indeed purely sequence and not location dependent and can be recapitulated at an ectopic genomic site. First we tested seven differently sized fragments for their ability to maintain both

the methylated and unmethylated state by profiling methylation post-insertion with targeted bisulfite sequencing (Fig. 3-4). Long (~ 2 kb) ICR fragments containing all four CTCF sites can indeed maintain both the methylated and the unmethylated state after insertion in the ectopic site (Fig. 3-4, i). This is also the case for one short fragment containing the first CTCF binding site (Fig. 3-4, i), but not for the other tested fragments containing one or two CTCF motifs (Fig. 3-4, ii and iii).

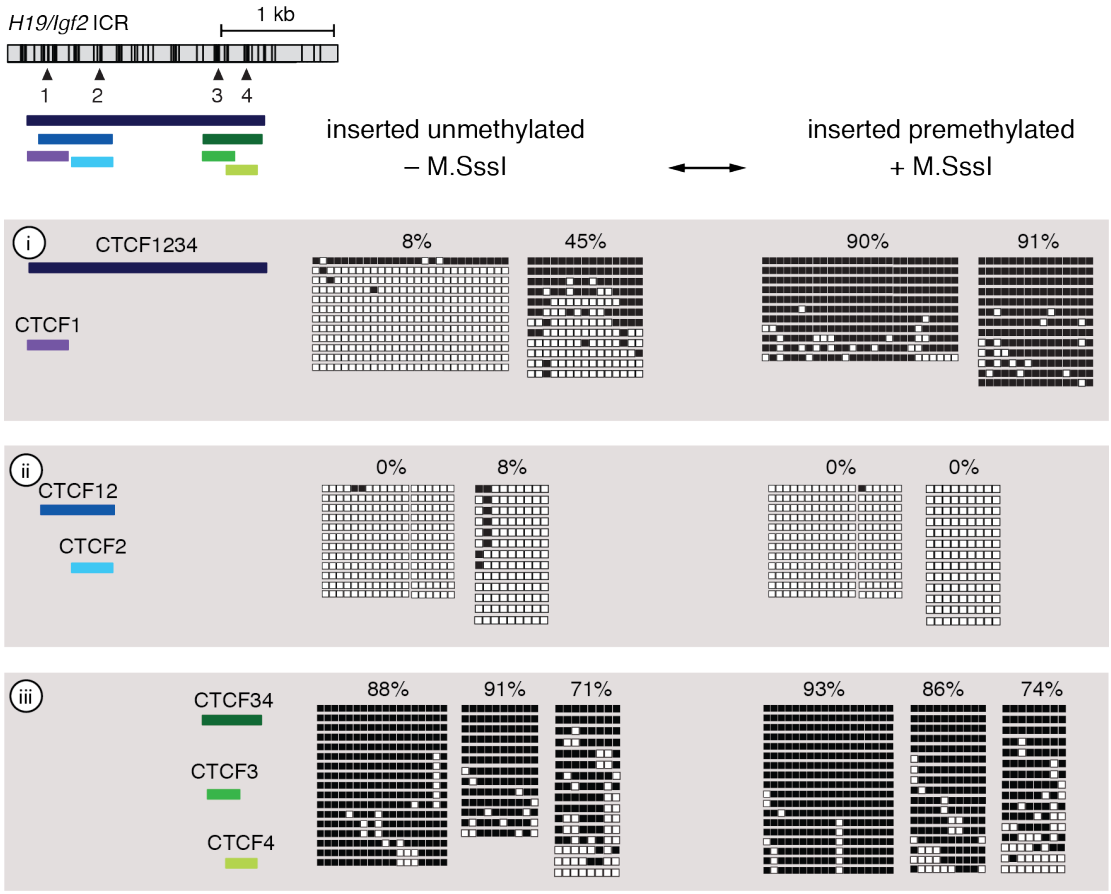


Figure 3-4. Methylation state of fragments of the *H19/Igf2* ICR after insertion into an ectopic genomic site.
Fragments of the *H19/Igf2* ICR (top; vertical bars = position of CpGs, numbered triangles = CTCF motifs) of different length containing different CTCF sites (coloured horizontal bars) were inserted by RMCE into an ectopic genomic site. The same fragments were inserted both unmethylated (– M.SssI) and fully premethylated *in vitro* (+ M.SssI). For the individual fragments shown on the left, methylation levels were then measured *after* insertion. Results are shown in the same order as the indicated fragments once for the unmethylated and once for the premethylated version. Every line of boxes corresponds to a sequenced bisulfite PCR amplicon (black box = methylated CpG, white box = unmethylated CpG). Average methylation of these amplicons is summarised in percent. While some sequences can maintain a differential methylation state upon insertion (i), others lose (ii, + M.SssI) or gain (iii, – M.SssI) methylation.

Nonetheless, in all cases CTCF binding measured by ChIP-qPCR anticorrelates strongly with the methylation state, as expected for methylation-sensitive binding events (Fig. 3-5a). Thus, methylation-sensitive CTCF binding can be recapitulated at an ectopic genomic site.

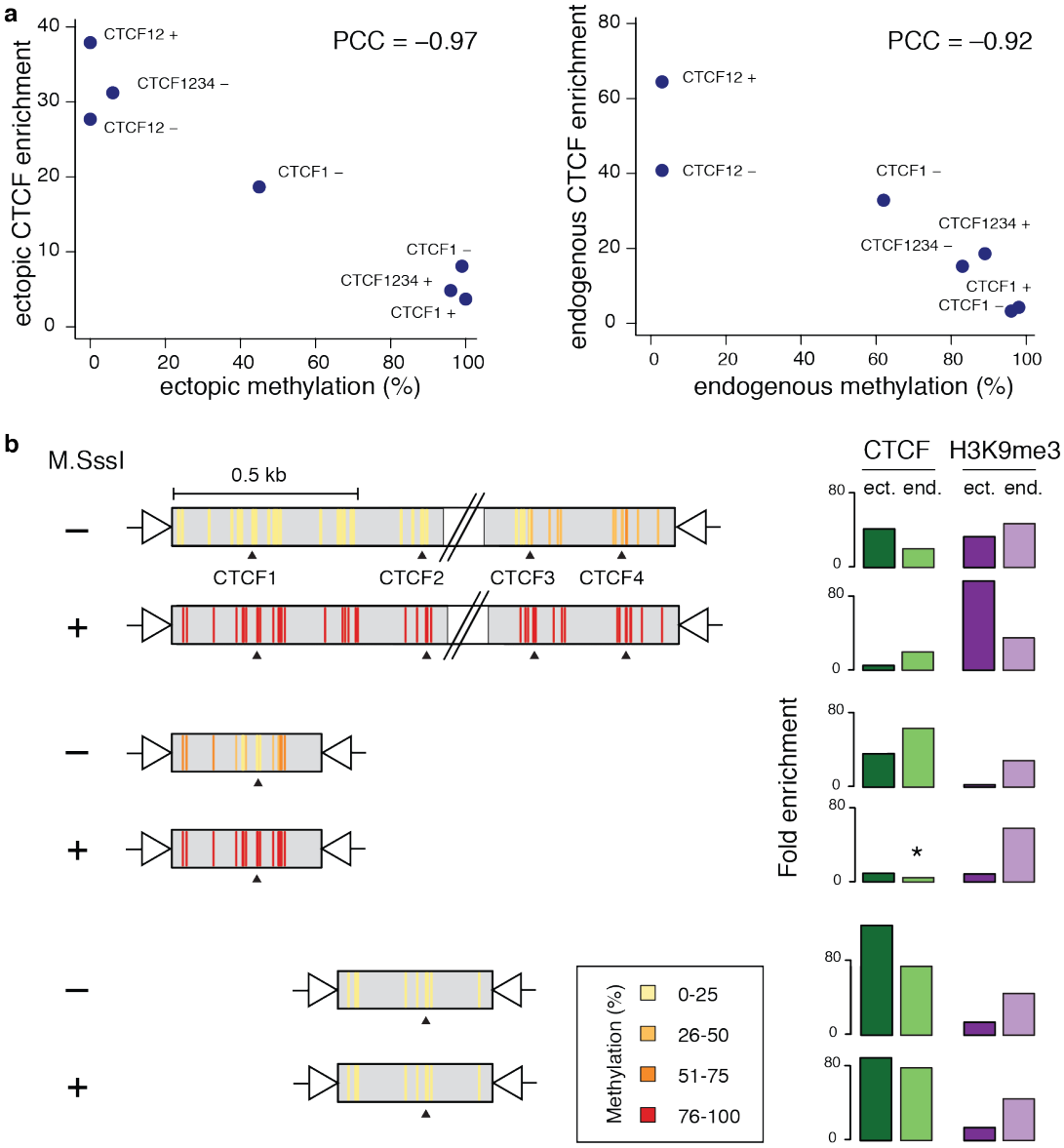


Figure 3-5. Methylation-sensitive CTCF binding is recapitulated at an ectopic genomic site.

a) Anticorrelation between CTCF binding and DNA methylation at both ectopic ICR fragments (left) and the endogenous ICR (right) across clones with different inserts. Clones named after inserted CTCF sites and pre-insertion methylation state (+/–). PCC = Pearson correlation coefficient. **c)** Methylation levels (left) and CTCF and H3K9me3 enrichment (right) at three exemplary fragments of the *H19/Igf2* ICR inserted unmethylated (– M.SssI) or premethylated (+ M.SssI) into the ectopic site. Methylation after insertion was measured by bisulfite sequencing (bars represent individual CpGs coloured according to average methylation level). CTCF (green) and H3K9me3 (purple) enrichment was measured by ChIP-qPCR (dark, ect.) and compared to the levels at the endogenous *H19/Igf2* ICR (light, end.). For the methylated short fragment, CTCF binding could not be detected at the endogenous site (middle, *).

3.1.3.4 CTCF binding is independent of H3K9me3 in the ectopic site

H3K9me3 is present in an allele-specific manner at the endogenous *H19/Igf2* locus (Singh et al., 2011) and was also removed in experiments where methylation was reduced by 5-Aza-deoxycytidine, leading to CTCF binding at the paternal allele (Ito et al., 2013). Therefore, this heterochromatic mark could be responsible for preventing CTCF binding rather than DNA methylation itself. To test this, we measured H3K9me3 enrichment at the inserted fragments and found that only the longest methylated ICR fragment could recruit this mark (Fig. 3-5b). Since we also observe loss of CTCF binding at the short methylated fragments in the absence of H3K9me3 (Fig. 3-5b), CTCF repulsion at these sequences is likely truly DNA methylation dependent.

3.1.3.5 Methylation levels of inserted sequences vary between clones

When measuring CTCF binding at the fragments inserted in the ectopic site, enrichments at the endogenous *H19/Igf2* ICR on chromosome 7, as well as at other unrelated CTCF binding sites in the genome, were used as positive controls. To our surprise, we were unable to measure enrichment at the endogenous ICR for the short methylated fragment that is unable to bind CTCF in several attempts, although ChIP enrichments at other unrelated genomic regions were strong (Fig. 3-5b). When we profiled methylation at the endogenous ICR we surprisingly observed an increase in methylation to almost 100%, rather than the expected 50%, in line with the absence of CTCF binding (Fig. 3-6a). Upon measuring methylation and CTCF binding at the endogenous ICR for all clones, we observed that it always matched the ectopic site, both in terms of DNA methylation and CTCF ChIP enrichment (Fig. 3-6a,b,c).

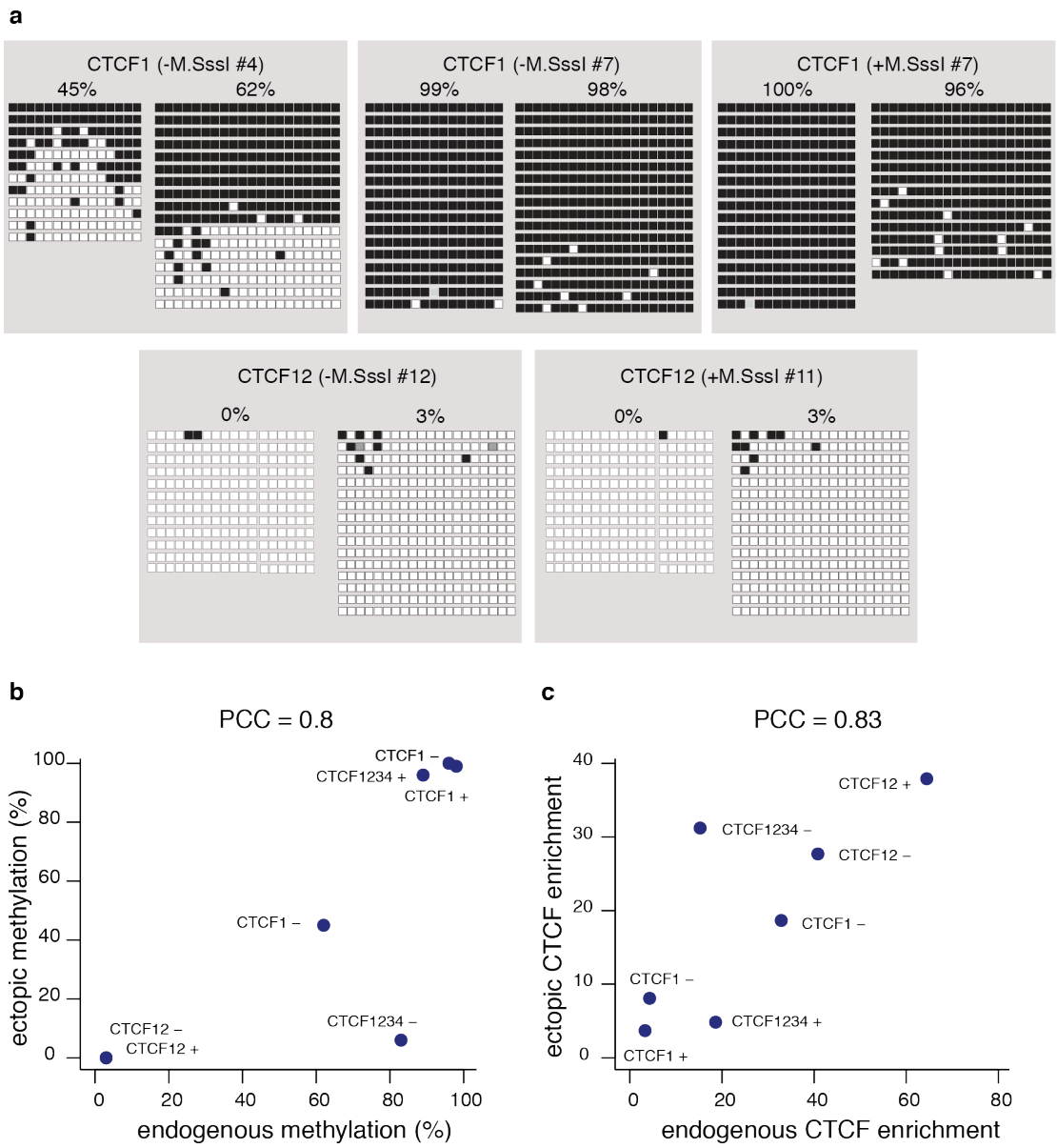


Figure 3-6. Comparison of DNA methylation and CTCF enrichment for ectopic fragments of the *H19/Igf2* ICR and their endogenous counterpart.
a) Bisulfite sequencing results of ectopic and matching endogenous ICR sequences within the same ES cell clone. Fragments containing one or two CTCF sites were inserted into the ectopic genomic site either unmethylated (– M.SssI) or premethylated (+ M.SssI); ES cell clones (large grey boxes) are named accordingly. For each ES cell clone, methylation levels at the ectopic (left) and endogenous (right) site are shown: Every line of boxes corresponds to a sequenced bisulfite PCR amplicon (black box = methylated CpG, white box = unmethylated CpG, grey box = NA). Average methylation of these amplicons is summarised in percent. **b,c)** Strong correlation of DNA methylation levels (b) and CTCF enrichment (c) at the ectopic site and the endogenous counterpart across clones with different inserts. Methylation was measured by bisulfite sequencing, CTCF enrichment by ChIP-qPCR. ES cell clones are named after CTCF binding sites inserted in the ectopic site and pre-insertion methylation state (+/–). PCC = Pearson correlation coefficient.

We hypothesised that the observed changes in DNA methylation at the endogenous ICR might be an artefact of the bisulfite sequencing method, since conversion of cytosines in mixed populations can lead to preferential amplification of one of the alleles in the subsequent PCR step. Accordingly we performed methylation-sensitive restriction digest coupled to qPCR on unconverted material to determine the methylation status of central CpGs in the constructs, thereby avoiding this amplification bias (Fig. 3-7a). The results were identical to those of the bisulfite sequencing, implying that the change of methylation at the endogenous imprinted region is indeed present and not a methodological artefact (Fig. 3-7b).

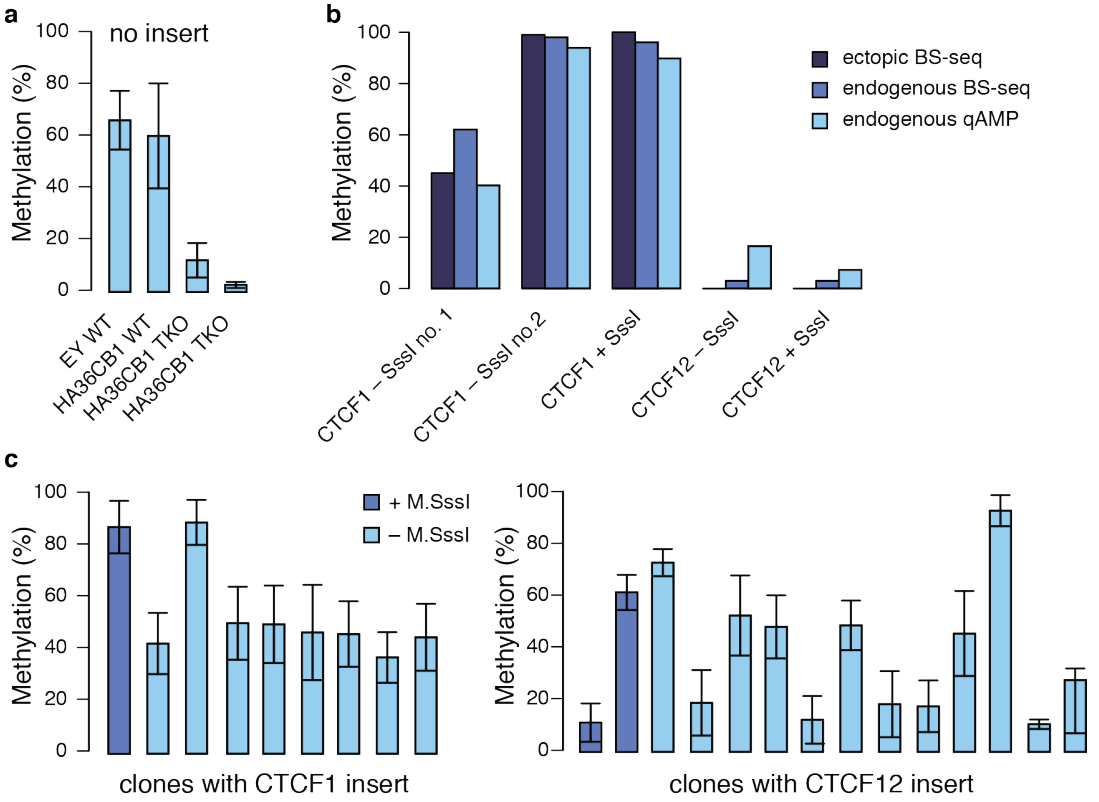


Figure 3-7. Methylation state of the *H19/Igf2* ICR varies between ES cell clones.
a) Methylation-dependent restriction digest allows quantitative analysis of DNA methylation using real-time PCR (qAMP), without the need for bisulfite conversion. qAMP delivers the expected results for the methylation level of the *H19/Igf2* ICR in ES cell clones that did not undergo RMCE. **b)** Similar methylation levels are measured for *H19/Igf2* ICR fragments containing one or two CTCF sites at the ectopic and endogenous site using qAMP and bisulfite sequencing (BS-seq), regardless of previous treatment of inserts with methylase M.SssI (+/–). **c)** Methylation levels at the endogenous *H19/Igf2* ICR measured by qAMP for different ES cell clones after insertion of a fragment containing the first (left) or the first two (right) CTCF sites of the H19 ICR. Methylation levels are independent of insert direction or methylation levels prior to insertion (+/–).
Error bars: standard deviation for levels measured with four different restriction enzymes.

To test how frequent this change in methylation state is, we redid insertions of some of the ICR fragments and analysed a larger number of ES cell clones (Fig. 3-7c). We observed clear clonal differences for the methylation state at the endogenous ICR. These always matched the ectopic site, but were not linked to the insert direction or the methylation level of the fragments prior to insertion. Since the ectopic site is actually on the same chromosome as the endogenous *H19/Igf2* ICR, we wondered whether there could be cross-talk between the two locations that could explain the agreement observed in all tested cases for methylation levels and CTCF binding between the ectopic and endogenous site. Therefore, we repeated the insertions in a different cell line, where the RMCE site is in a different genomic location. Again we observed strong variation in the methylation level of the endogenous ICR (Fig. 3-8a). In addition, we removed the inserts in a new round of RMCE and measured methylation of the endogenous ICR before and after insert removal (Fig. 3-8b). Methylation levels were not influenced by removal of the insert, but rather stably maintained.

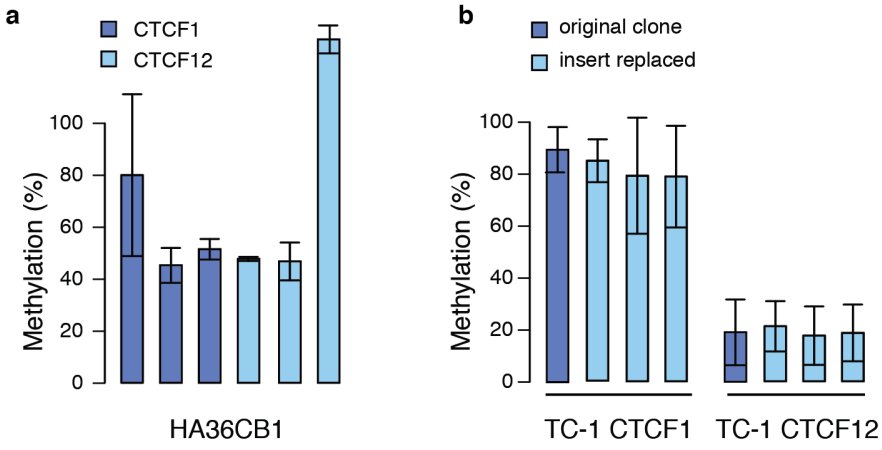


Figure 3-8. Variable methylation states of the endogenous *H19/Igf2* ICR sequence do not depend on the genomic location or presence of an ectopic insert.
a) Variability in methylation state of the endogenous ICR between ES cell clones with ectopic inserts does not depend on the location of the ectopic site. In the HA36CB1 cell line, the RMCE site is in a different genomic location than in the cell line used for previous targeting experiments (TC-1). Fragments containing the first or first two CTCF sites of the ICR were inserted there. After insertion, methylation levels were measured by qAMP at the endogenous *H19/Igf2* ICR. Levels above 100% are due to differences in Ct values from undigested control (see Methods). **b)** Aberrant methylation levels are maintained after insert removal. Methylation levels at the endogenous *H19/Igf2* ICR measured by qAMP before and after removal of the inserted ICR fragments (CTCF1 or CTCF12) from the ectopic site.
Error bars: standard deviation for levels measured with four different restriction enzymes.

These two findings argue against cross-talk between the endogenous and ectopic site. Rather, they imply that sequences from imprinted regions can adapt fully methylated or unmethylated states in our ES cell culture and clonal selection conditions, which vary from clone to clone but are identical for the same sequence within the same cell, regardless of the genomic location. Once established, these levels are then stably maintained over many generations within a clonal population (Fig. 3-8b). In agreement with these observations, it has been previously reported for human ES cells that *in vitro* culture can lead to alterations in methylation levels at ICRs (Frost et al., 2011). These findings require adaptations to the experimental approach since variability in methylation states for identical sequences is not compatible with use of the planned library scheme. As a consequence, this line of experiments was not explored further in the scope of this thesis.

3.1.4 Discussion

We identified several hundred putative methylation-sensitive CTCF binding sites in ES cells that are only occupied in the absence of DNA methylation. These contain motifs with CpGs at certain positions and lie in CpG-denser regions of the genome, setting them apart from the vast majority of CTCF binding sites.

Of note, a recent study comparing CTCF binding in WT and *Dnmt1/3b* hypomorph HCT116 cell lines observed increased CTCF binding at a subset of sites in the cell line with reduced methylation levels (Maurano et al., 2015). Many of these sites contained a CpG in critical positions of the motif and were located in CpG-dense regions, in line with our findings. Overall we observe fewer cell line-specific sites compared to said study, although DNA methylation is completely removed, not only reduced, in our system. This might in part be due to our isogenic approach, which uses cell lines that are very similar in terms of cell culture age etc., thus minimizing variability not directly linked to DNA methylation loss.

CTCF methylation sensitivity can be recapitulated at an ectopic site and is independent of the H3K9me3 mark, implying that this behaviour is indeed dependent on the local sequence context and DNA methylation state. However, the methylation level of inserted ICR fragments (and their endogenous counterparts) showed strong clonal variability. This complicates the planned use of the library approach to further test and validate sequence parameters involved in methylation-sensitive CTCF binding. Differences in methylation state for cells with the same inserted sequence would make it difficult to link them to the correct ChIP enrichment. To circumvent this issue, it is feasible to instead perform NOMe-seq on inserted libraries (Jessen et al., 2004; Kelly et al., 2012). This method uses GC methyltransferase footprinting and bisulfite sequencing to measure both DNA methylation and TF occupancy on the same single molecule and can thus also be applied to heterogeneous cell populations.

It is possible that the clonal variability we observe is a special feature of imprinted regions. Given their unique allele-specific methylation *in vivo*, these

sequences might be 'metastable' and more likely to adapt binary methylation states in contrast to other genomic regions, which show less clonal variability in our experience. At the same time, other sequences are even less likely to maintain an 'imposed' methylation state after insertion (data not shown). Inserting libraries into WT and TKO cells instead of premethylating the inserts could circumvent the difficulty of maintaining the methylation state after insertion. The sequences of interest are mostly methylated in the WT cell and would likely acquire this state after insertion, especially since inserted sequences are more likely to gain methylation than to lose it compared to their endogenous state (Krebs et al., 2014). Binding could thus be compared to their unmethylated counterpart in TKO cells.

The sequences profiled in a library approach could include the 202 putative methylation-sensitive sites identified here, as well as variations in motif and flanking region CpG density, or other regions which are not methylated in ES cells but are of interest due to their sequence composition or methylation state in other cell types. Further possible sequence parameters to vary would be motif strength, neighbouring TF motifs or scanning point mutations. It will be particularly interesting to dissect the *H19/Igf2* ICR by replacing the CpGs in the motifs or reducing CpG density in the surroundings, in order to test which of these manipulations affect methylation sensitivity at this well-studied sequence.

Of note, many regions with the described characteristics remain unbound also in the TKO cell line. This is in line with the general observation that only a very small fraction of TF motifs are actually bound in any condition (Wang et al., 2012b) and makes it clear that other factors besides DNA methylation and CpG content regulate CTCF binding. In agreement with this notion, 'reactivated' CTCF sites identified in methylation-deficient HCT116 cells were found to coincide with sites that are bound in one of the 40 other profiled cell types (Maurano et al., 2015), rather than constituting a novel binding site repertoire. Since more than 90% of CTCF binding sites are shared across cell types (Chen et al., 2012), it is not likely that many more sites are restricted by DNA methylation in other cell types. It should be noted however, that while we

do not observe a major role for DNA methylation in CTCF binding site restriction in ES cells, we cannot exclude that CTCF is methylation-sensitive at sites which are already bound and unmethylated in WT ES cells.

Taken together, CTCF binding is restricted by DNA methylation only at a subset of sites in ES cells. These sites tend to contain more CpGs both in the motif and the flanking regions. The targeted genomic editing approach suggested here – modified to account for heterogeneity in methylation levels – would provide clear experimental evidence for the contribution of these sequence features to CTCF methylation sensitivity. This in turn would facilitate the prediction of effects of epigenetic alterations on genome topology and gene expression.

3.2 Binding site restriction by DNA methylation in embryonic stem cells

3.2.1 Abstract

It is currently unclear to which extent DNA methylation influences TF binding in the cellular context. In order to identify methylation-sensitive TFs *a priori*, we compared DNase I hypersensitivity, as an indicator of TF binding, in mouse embryonic stem cells with and without DNA methylation. While most sites remain unchanged, a subset of sites is indeed only accessible in the absence of DNA methylation and is enriched for CpG-containing TF motifs, most prominently of NRF1. This TF occupies several thousand additional sites in the unmethylated genome, resulting in increased transcription. Restoring *de novo* methyltransferase activity initiates remethylation at these sites and outcompetes NRF1 binding. Even strong overexpression of NRF1 is unable to prompt binding at methylated regions. Together, this suggests that binding of methylation-sensitive TFs relies on additional determinants to induce local hypomethylation. In support of this model, removal of neighbouring motifs in *cis* or of a TF in *trans* causes local hypermethylation and subsequent loss of NRF1 binding. This competition between DNA methylation and TFs reveals a case of cooperativity between TFs that acts indirectly via DNA methylation. Methylation removal by methylation-insensitive factors enables occupancy of methylation-sensitive factors, a principle that rationalises hypomethylation of regulatory regions.

3.2.2 Published manuscript

LETTER

doi:10.1038/nature16462

Competition between DNA methylation and transcription factors determines binding of NRF1

Silvia Domcke[1,2]*, Anaïs Flore Bardet[1]*, Paul Adrian Ginno[1], Dominik Hartl[1,2], Lukas Burger[1,3] & Dirk Schübeler[1,2]

Eukaryotic transcription factors (TFs) are key determinants of gene activity, yet they bind only a fraction of their corresponding DNA sequence motifs in any given cell type[1]. Chromatin has the potential to restrict accessibility of binding sites; however, in which context chromatin states are instructive for TF binding remains mainly unknown[1,2]. To explore the contribution of DNA methylation to constrained TF binding, we mapped DNase-I-hypersensitive sites in murine stem cells in the presence and absence of DNA methylation. Methylation-restricted sites are enriched for TF motifs containing CpGs, especially for those of NRF1. In fact, the TF NRF1 occupies several thousand additional sites in the unmethylated genome, resulting in increased transcription. Restoring *de novo* methyltransferase activity initiates remethylation at these sites and outcompetes NRF1 binding. This suggests that binding of DNA-methylation-sensitive TFs relies on additional determinants to induce local hypomethylation. In support of this model, removal of neighbouring motifs in *cis* or of a TF in *trans* causes local hypermethylation and subsequent loss of NRF1 binding. This competition between DNA methylation and TFs *in vivo* reveals a case of cooperativity between TFs that acts indirectly via DNA methylation. Methylation removal by methylation-insensitive factors enables occupancy of methylation-sensitive factors, a principle that rationalizes hypomethylation of regulatory regions.

Methylation of DNA at cytosines within CpG dinucleotides has the potential to block TF binding either directly through interference with base recognition or indirectly through recruitment of methylation-specific binding proteins[3]. DNA methylation has been reported to block binding of some TFs *in vitro*[3]. However, this does not necessarily translate to a similar effect *in vivo*[4,5]. In addition, sensitivity *in vivo* can be highly locus-specific as observed for the TF CTCF, which only responds to methylation at a very limited set of chromosomal loci[6–9]. Intriguingly, some TFs such as REST and CTCF have been shown to bind methylated regions and trigger their demethylation[8,10,11]. Thus, although it is established that active regulatory regions are bound by TFs and generally display low levels of DNA methylation[8,12], it remains contentious whether this relationship reflects the cause or consequence of altered TF binding[13,14]. Determining factor-specific sensitivity of binding events in a cellular context is therefore imperative for understanding how DNA methylation affects gene expression and to functionally interpret epigenomic maps. To identify TFs that are restricted in their binding by DNA methylation *in vivo*, we mapped DNase-I-hypersensitive sites (DHSs), an indicator of TF binding, in wild-type murine embryonic stem (ES) cells and upon global removal of DNA methylation (Fig. 1a).

DNA methylation is essential for mouse development and survival of most tested mammalian cell types, with the exception of murine ES cells[15]. Therefore, these cells provide an opportunity to compare TF binding in the presence and absence of DNA methylation. To reduce genetic or clonal variability we used CRISPR/Cas9 to generate genetic deletions of both *de novo* DNA methyltransferases *Dnmt3a* and *Dnmt3b* and the maintenance enzyme *Dnmt1* in the ES cell line 159 (see Methods) for which we previously performed base-pair-resolution methylation profiling[8] (Extended Data Fig. 1a).

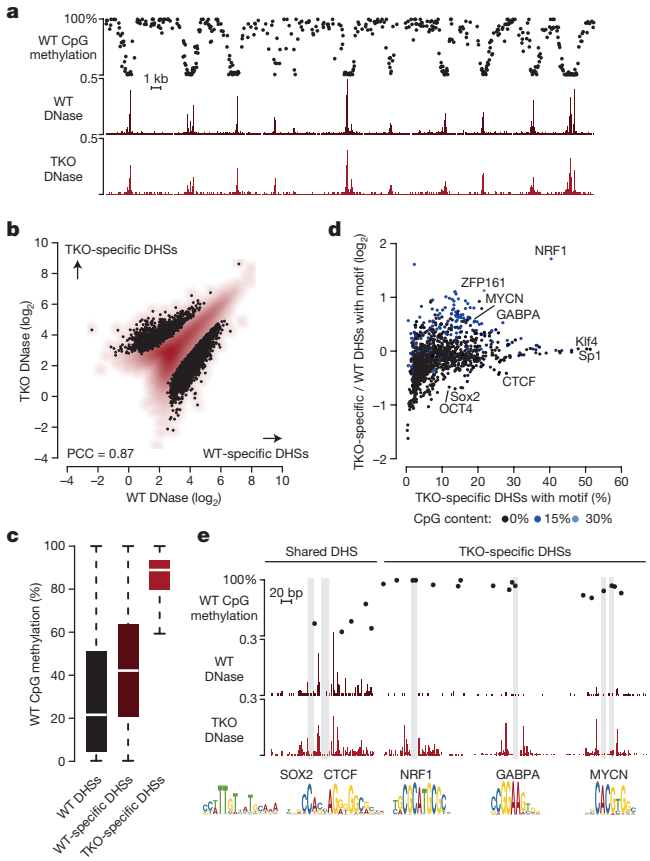


Figure 1 | DHSs that form upon removal of DNA methylation are enriched for specific TF motifs. a, Wild-type (WT) methylation, and wild-type and TKO DNase-seq signal at a representative genomic region (chr17: 25,920,000–25,972,499). **b**, DNase-seq signal at all DHSs in wild type and TKO. Black dots mark DHSs significantly enriched in wild type ($n = 2{,}837$) or TKO ($n = 1{,}543$). PCC, Pearson correlation coefficient. **c**, Average wild-type methylation of CpGs within all wild-type, wild-type-specific (the subset of wild-type DHSs that are not present in TKO DHSs) or TKO-specific DHSs. Boxplots show median (white line), 25th and 75th percentiles (boundaries), minimum and maximum (whiskers). **d**, Motif occurrences in TKO-specific DHSs compared to all wild-type DHSs. Blue colouring illustrates motif CpG content. **e**, Representative genomic regions showing shared (left, chr6: 31,189,871–31,190,470) and TKO-specific (right, chr1: 51,483,272–51,483,871, chr6: 48,413,300–48,413,899 and chr10: 62,623,300–62,623,899) DHS footprints. Motif locations are highlighted in grey.

[1]Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, CH 4058 Basel, Switzerland. [2]University of Basel, Faculty of Sciences, Petersplatz 1, CH 4003 Basel, Switzerland. [3]Swiss Institute of Bioinformatics, Maulbeerstrasse 66, CH 4058 Basel, Switzerland.
*These authors contributed equally to this work.

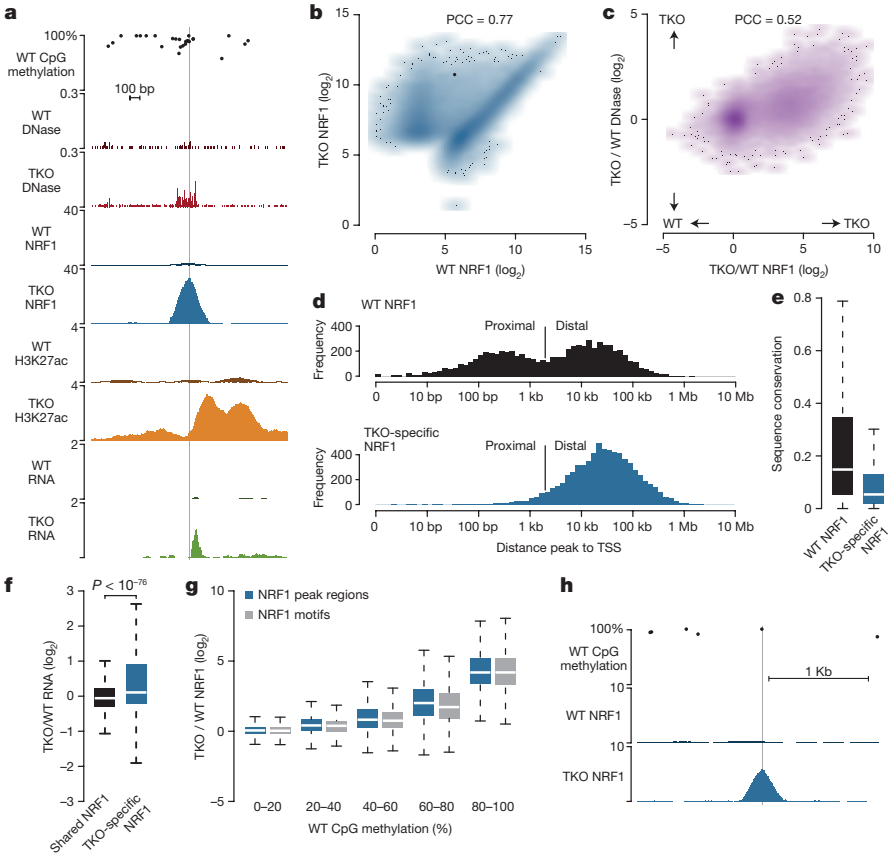


Figure 2 | NRF1 binds several thousand new sites in the unmethylated genome. a, Wild-type methylation, and wild-type and TKO DNase-seq, NRF1 ChIP-seq, H3K27ac ChIP-seq and RNA-seq signal at a TKO-specific distal genomic region (chr4: 99,235,857–99,236,456). The NRF1 motif location is highlighted in grey. **b**, Wild-type and TKO NRF1 ChIP-seq signal at all peak regions. The thick black dot represents the region in **a**. **c**, Changes in NRF1 binding and DNase-seq signal between wild type and TKO at all NRF1 peak regions. **d**, Distance of all wild-type (top; $n = 8{,}835$) or TKO-specific (bottom; $n = 7{,}205$) NRF1 peaks to the nearest transcriptional start site (TSS). Cutoff between proximal and distal sites is 2 kb. **e**, Average sequence conservation (PhastCons score) of all wild-type or TKO-specific NRF1 peak regions. Boxplots show median (white line), 25th and 75th percentiles (boundaries), minimum and maximum (whiskers). **f**, Expression change (in reads per kilobase per million (RPKM)) of genes closest to shared and TKO-specific NRF1 peaks. *P* value from a Wilcoxon test. **g**, Change in NRF1 binding between TKO and wild-type at all peak regions grouped according to their average methylation. Blue boxes represent changes within entire peak regions, grey boxes only those within NRF1 motifs. $n > 800$ in all groups. **h**, Wild-type methylation, and wild-type and TKO NRF1 ChIP-seq signal at a genomic region with no additional CpGs within 1.8 kb around the motif (grey line).

The resulting triple knockout (TKO) cells showed no detectable DNA methylation by several measures (Extended Data Fig. 1b, c) and limited changes in global expression patterns, as previously reported for a TKO cell line generated by classical mouse genetics[15,16] (Extended Data Fig. 1d, e).

Hypersensitivity to digestion by DNase I is an indicator of TF binding that does not require a priori knowledge of the TFs involved[17]. We mapped DHSs with high coverage in both wild-type cells and the isogenic TKO cells and observed that the vast majority of DHSs remain unchanged (Fig. 1a, b, Extended Data Fig. 2a–d and Extended Data Table 1). This suggests that the binding patterns of most TFs expressed in murine ES cells are not altered upon global removal of DNA methylation. In addition, we observed a fraction of DHSs that are specific to each cell state in a reproducible manner (Fig. 1b and Extended Data Fig. 2e). These DHSs are preferentially located distal to transcriptional start sites (TSS) and within CpG-poor regions (Extended Data Fig. 2f, g). In contrast to wild-type-specific DHSs (the subset of wild-type DHSs that are not present in TKO DHSs), newly formed sites in the TKO cell line lie within regions that were methylated in the wild-type cells, indicating that they could be methylation-dependent (Fig. 1c and Extended Data Fig. 2h).

Searching for known TF motifs and hexamer sequences enriched in TKO-specific DHSs resulted in a small number of candidate methylation-sensitive TFs including NRF1, GABPA and MYCN (Fig. 1d, Extended Data Fig. 3a and Supplementary Table 1). These factors are expressed at similar levels in both cell lines and probably form TKO-specific DNase I footprints (Fig. 1d, e and Extended Data Fig. 3b, c). In contrast, motifs enriched in wild-type-specific DHSs do not reveal footprints limited to this cell state (Fig. 1b and Extended Data Fig. 3c, d). Notably, TKO-specific DHSs are enriched for motifs containing CpG dinucleotides, even though they reside within regions that are generally CpG-poor (Fig. 1d, Extended Data Fig. 2g and Supplementary Table 1). The most prominently enriched motif in TKO-specific DHSs contains two CpGs, consistent with a direct inhibition by DNA methylation, and belongs to the highly conserved TF nuclear respiratory factor 1 (NRF1)[18] (Fig. 1d, e). Previous *in vitro* experiments with NRF1 suggested that DNA methylation blocks binding[19,20], but also that it preferentially binds to methylated sequences[21]. Given its strong signal and because only one factor has been reported to bind this motif, we focused on further analysis of NRF1.

Chromatin immunoprecipitation of NRF1 followed by sequencing (ChIP-seq) revealed that more than 7,000 sites, in addition to those already occupied in wild-type cells, show reproducible increased NRF1 binding in the absence of DNA methylation (Fig. 2a, b, Extended Data Fig. 4a–d and Extended Data Table 1). Newly bound NRF1 sites correlate with TKO-specific DHSs, validating the comparative DHS

approach (Fig. 2a, c). They occur distal to genes (Fig. 2d) in regions of low CpG content (Extended Data Fig. 4e) and poor sequence conservation (Fig. 2e), suggesting that a large fraction could represent non-functional sites otherwise blocked by DNA methylation. Nevertheless, increase of NRF1 binding is matched by a significant increase in expression of the nearest genes, indicative of an impact on transcription (Fig. 2f). Additionally, for some TKO-specific sites, lysine 27 acetylation of histone H3, a mark of active regulatory regions, appears and aberrant NRF1-dependent transcripts are initiated directly at the binding sites (Fig. 2a and Extended Data Fig. 4a, f–j).

TKO-specific NRF1 sites mostly contain a high confidence motif with at least one but usually two CpGs (Extended Data Fig. 4k and Supplementary Table 2). These motifs display intermediate to full methylation in wild-type cells, yet increased NRF1 binding in TKO is strongest at highly methylated motifs, suggesting that methylation of the core motif directly prevents binding in wild-type cells (Fig. 2g and Extended Data Fig. 4l). TKO-specific binding of NRF1 is independent of the density of methylated CpGs in the surrounding region, strongly arguing against an involvement of indirect repression through methyl-CpG binding-domain proteins[22] (Extended Data Fig. 4m–o). This is exemplified at a locus that harbours no CpG within 1.8 kb around the motif (Fig. 2h); despite this absence of additional CpGs, NRF1 binds in a strictly methylation-dependent manner.

In the experiments described so far, ES cells were cultured in the presence of serum and LIF, which recapitulates the genome-wide methylation observed in the postimplantation epiblast[23]. Culturing in the presence of two kinase inhibitors (2i) is an alternative regime that mimics the inner cell mass of the blastocyst and coincides with downregulation of the *de novo Dnmts*[24,25]. Here it provides the opportunity to measure NRF1 binding at physiological levels of low methylation and without genetic alteration of the *Dnmt* genes. Transferring wild-type cells cultured originally in serum to 2i conditions leads to increased NRF1 binding at the vast majority of previously identified TKO-specific sites (Fig. 3a, b and Extended Data Fig. 5a–c). Similarly, this coincides with hypomethylation of these sites in 2i conditions as revealed by whole-genome, as well as high-coverage amplicon, methylation profiling (Fig. 3a and Extended Data Fig. 5d–f). Small differences in NRF1 binding between 2i and TKO conditions are readily explained by remaining levels of methylation at a subset of sites in 2i (Extended Data Fig. 5c, g). These include examples where the motif remains methylated and unbound even though the surrounding region is demethylated (Extended Data Fig. 5h), providing additional support for our observation that methylation of the core motif alone is the critical determinant of NRF1 binding *in vivo*.

To test if NRF1 binding to these new sites inhibits their *de novo* methylation, we transferred ES cells cultured in 2i back to medium with serum. This leads to transcriptional upregulation of the *de novo Dnmt* genes and genomic remethylation over time[25]. Profiling of NRF1 binding, as well as whole-genome and amplicon methylation, revealed that the majority of methylation-dependent sites become remethylated and that NRF1 binding can no longer be detected (Fig. 3a, c and Extended Data Fig. 5h–m). This shows that *de novo* methylation can outcompete binding of NRF1, implying that binding and creation of a DHS is not sufficient to protect against *de novo* methylation for this TF.

Although levels of *Nrf1* expression remained mostly unchanged between the tested conditions (Extended Data Fig. 3b and Extended Data Fig. 5a), we assessed if variations in NRF1 protein abundance could account for differential occupancy. Therefore we overexpressed NRF1 at least tenfold and profiled its genomic binding in wild-type cells (Extended Data Fig. 6a). This revealed an increase in binding at previously occupied sites but also novel sites (Fig. 3a and Extended Data Fig. 6b, c). The latter, however, do not overlap with methylation-dependent sites and contain weak NRF1 motifs, reflecting less specific binding to regions of open chromatin (Fig. 3a and Extended Data Fig. 6d, e). This shows that methylation of individual core motifs, and not NRF1 protein levels, determines genomic occupancy.

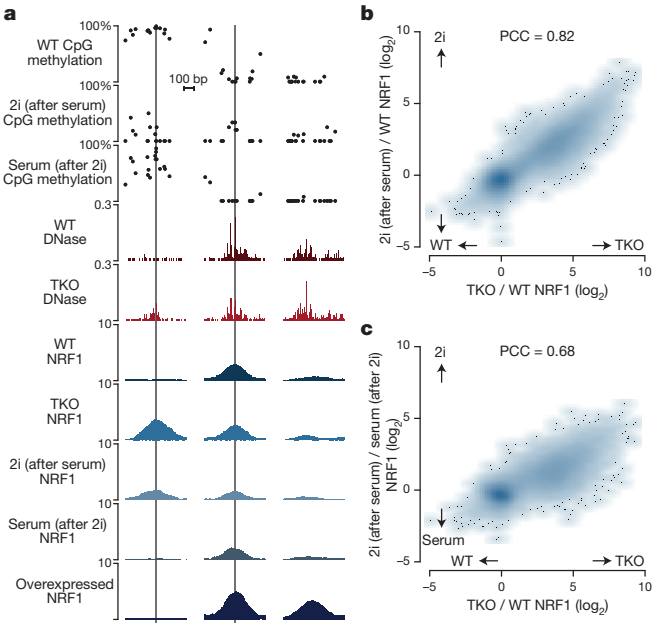


Figure 3 | *De novo* methylation outcompetes NRF1 binding. a, Wild-type methylation, wild-type and TKO DNase-seq, and NRF1 ChIP-seq signal in wild-type under different culture regimes, in TKO and in wild-type cells overexpressing NRF1 at representative genomic regions. Left, TKO-specific site (chr12: 82,788,342–82,794,341). Middle, shared site in wild-type and TKO, with increased binding upon overexpression (chr5: 148,104,611–148,105,210). Right, site only bound upon overexpression (chr18: 36,030,688–36,036,687). Grey lines indicate the location of the NRF1 motif. **b**, Change of NRF1 ChIP-seq signal between wild-type and TKO or 2i culture (after culture with serum) at all NRF1 peaks regions. **c**, Change of NRF1 ChIP-seq signal between TKO and wild-type versus between culture with 2i (after culture with serum) and culture with serum (after culture with 2i) at all NRF1 peak regions.

To test if cell-type-specific methylation patterns could similarly explain differential binding of NRF1, we differentiated ES cells into neuronal progenitors and investigated NRF1 binding. We found that the gain of methylation at NRF1 motifs in neuronal progenitors coincides with loss of NRF1 binding (Extended Data Fig. 7a–c) and matching lower expression of neighbouring genes (Extended Data Fig. 7d, e). This tight link between DNA methylation, NRF1 binding and transcription holds true beyond the murine system, as seen by genomic profiling of NRF1 in human normal breast cells (HMEC) and a breast cancer cell line (HCC1954)[26] (Extended Data Fig. 7f–i), as well as in other cell type comparisons (H1hESC and GM12878)[27] (Extended Data Fig. 7j–m). Thus, data from different organisms and cellular states including cancer may indicate that methylation-dependent binding of NRF1 is a general phenomenon that affects gene regulation.

We next sought to test the methylation sensitivity of NRF1 without global reduction of DNA methylation, by using reporter constructs inserted into a defined chromosomal locus of ES cells by Cre recombinase[28]. NRF1 sites with 400 bp of their surrounding genomic sequence were inserted either unmethylated or premethylated at CpGs *in vitro* (Extended Data Fig. 8a–c). As expected, this revealed reduced binding to the premethylated compared to the untreated template (Extended Data Fig. 8b). Thus, sensitivity of NRF1 to methylation of the underlying motif can be recapitulated in an ectopic site. We previously showed that CTCF can bind a motif added to a premethylated reporter and cause local reduction of methylation[8]. When we exchanged the CTCF motif with that of NRF1, we did not observe NRF1 binding or loss of methylation. Only upon forced demethylation is NRF1 capable of binding this minimal sequence context (Fig. 4a). Therefore NRF1 can bind its motif autonomously, but only if unmethylated. Genome-wide binding and single-locus reporter experiments indicate that NRF1 is

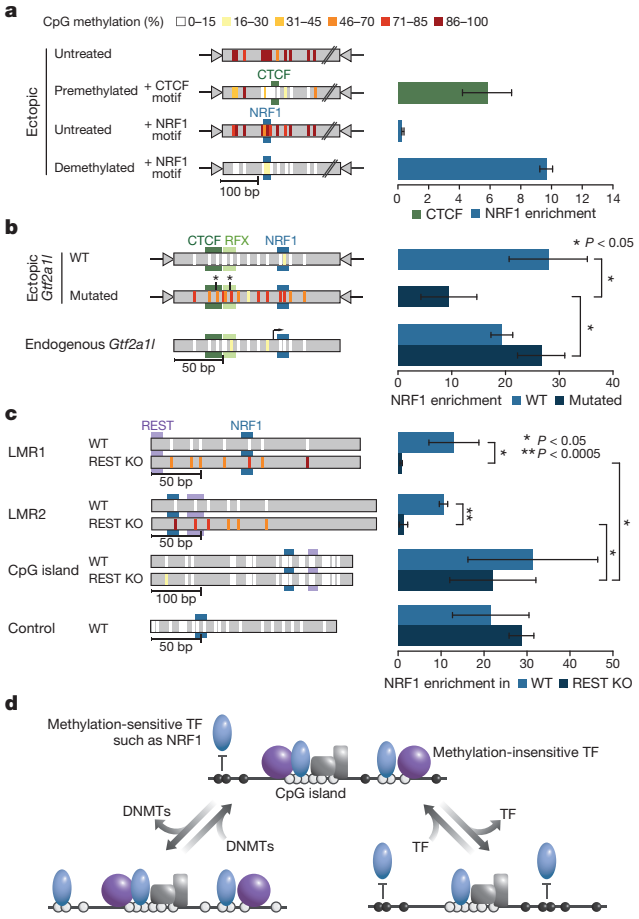


Figure 4 | NRF1 binds to unmethylated core motifs via TF-mediated local hypomethylation. a–c, Methylation levels of individual CpGs (left, amplicon Bis-seq) and TF occupancy (right, ChIP–qPCR) for reporters inserted into a defined ectopic genomic locus or for endogenous regions. TF motif locations are marked as coloured boxes. ChIP–qPCR enrichments are the mean of three biological replicates; error bars represent standard deviation; *P* values from two-sided *t*-tests. **a**, CTCF or NRF1 motifs were added to the same sequence inserted as either premethylated (CTCF)[8], untreated or chemically demethylated (NRF1). **b**, The *Gtf2a1l* promoter was inserted with intact or mutated (asterisks) CTCF and RFX motifs[28]. In both cases the corresponding endogenous locus serves as control. **c**, Endogenous regions bound by REST in wild-type and containing adjacent NRF1 motifs in low-methylated regions (LMR) and an unmethylated CpG island, profiled in wild-type and REST knockout (KO) cells. The control region is REST independent. **d**, Image of the model. In wild-type cells, NRF1 binding is blocked by DNA methylation and only occurs at unmethylated motifs (top). Motif methylation requires the activity of the DNMTs (bottom left), while motif demethylation can be mediated upon adjacent binding of methylation-insensitive TFs (bottom right). Circles represent unmethylated (white) or methylated (black) CpGs.

sensitive to DNA methylation of its motif and that it cannot protect it from *de novo* methylation. This leads to the prediction that NRF1 relies on other features that keep its motif in an unmethylated state.

As some TFs, such as CTCF, can locally mediate low methylation levels[8,28,29], we hypothesized that such factors could direct NRF1 binding in wild-type cells. Consistent with this model, constitutive NRF1 binding sites reside in regions that are co-bound by many TFs, as reflected by broad DHSs and overlap with existing TF localization maps (Extended Data Fig. 9a, b). To experimentally test this hypothesis we inserted reporter constructs harbouring an endogenous promoter sequence including a NRF1 motif (Extended Data Fig. 9c). Deletion of the CTCF and RFX motifs within this construct leads to its hypermethylation[28] but notably also to decreased NRF1 binding (Fig. 4b). This establishes a dependence of NRF1 in *cis* on motifs of TFs that mediate local hypomethylation. To further explore this hierarchical model, we assessed whether removal of a demethylating TF affects NRF1 binding. We previously showed that REST (also known as NSRF) creates regions of low methylation at its binding sites in CpG-poor regions, which become remethylated when REST is genetically removed[8,10]. Even though REST and NRF1 have not been functionally linked, we identified a few sites where NRF1 binds adjacent to REST (Extended Data Fig. 9d), enabling us to monitor NRF1 occupancy as a function of REST. At sites that occur within CpG-poor low-methylated regions, we observe *de novo* methylation upon deletion of REST that extends well into the NRF1 motif and coincides with loss of NRF1 binding in both cases tested (Fig. 4c). Of note, the absence of REST does not affect proximal NRF1 binding within a CpG island, as it remains hypomethylated regardless of REST occupancy, possibly because CpG islands are bound by additional factors that confer hypomethylation[29] (Fig. 4c). Thus, NRF1 binding *in vivo* critically relies on the local DNA sequence context in *cis* and TFs in *trans* to ensure a hypomethylated binding site (Fig. 4d).

This study proposes several TFs that might be restricted by DNA methylation but also suggests that the majority of factors expressed in mouse ES cells do not respond to global loss of DNA methylation. A critical question remains whether differentiated cells, for which DNA methylation has been shown to be essential, express a larger set of methylation-sensitive factors.

Our study of NRF1 binding in different and dynamic methylomes establishes an example of genome-wide, methylation-sensitive TF binding *in vivo*. Combined with site-specific genetic and epigenetic perturbation, it provides a proof of principle for a model whereby DNA methylation can guide TF binding in a highly factor- and context-specific manner (Fig. 4d).

NRF1 has previously been proposed to be a pioneer factor based on its ability to form a DHS *de novo*[30]. We show that NRF1 only bears canonical hallmarks of a pioneer factor[2] in the absence of DNA methylation, where it indeed can bind autonomously and form a DHS. In the presence of DNA methylation, it behaves as a 'settler' TF, as it requires the assistance of superordinate TFs to ensure hypomethylation of its motif. This suggests that the ability to mediate a hypomethylated state upon binding could be an additional relevant characteristic for a pioneer TF in vertebrates. Notably, we show that NRF1 binding to an unmethylated site does not protect against *de novo* methylation. This provides clear evidence for competition between TFs and DNMTs, and argues that active demethylation and/or efficient obstruction of *de novo* methylation is required not only for the establishment of NRF1 binding, but also for its maintenance. This exemplifies the idea that TF hierarchies can be mediated via a local epigenetic mark—DNA methylation removal by methylation-insensitive factors enables occupancy of methylation-sensitive factors in a form of indirect cooperativity that does not require physical interaction between both TFs[1]. It illustrates that TF binding patterns at enhancers and promoters are both guided by and actively shape the balance between active demethylation and *de novo* methylation (Fig. 4d). This supports a model in which the role of DNA methylation in restricting genomic binding of TFs is dependent on the specific factor, the local activity of methylating and demethylating enzymes, and the genomic context of individual motif occurrences.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 April; accepted 16 November 2015.
Published online 16 December 2015.

1. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39,** 381–399 (2014).
2. Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes Dev.* **28,** 2679–2692 (2014).

3. Tate, P. H. & Bird, A. P. Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr. Opin. Genet. Dev.* **3,** 226–231 (1993).
4. Becker, P. B., Ruppert, S. & Schütz, G. Genomic footprinting reveals cell type-specific DNA binding of ubiquitous factors. *Cell* **51,** 435–443 (1987).
5. Weih, F., Nitsch, D., Reik, A., Schütz, G. & Becker, P. B. Analysis of CpG methylation and genomic footprinting at the tyrosine aminotransferase gene: DNA methylation alone is not sufficient to prevent protein binding *in vivo*. *EMBO J.* **10,** 2559–2567 (1991).
6. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405,** 482–485 (2000).
7. Hark, A. T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19/Igf2* locus. *Nature* **405,** 486–489 (2000).
8. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480,** 490–495 (2011).
9. Maurano, M. T. *et al.* Role of DNA methylation in modulating transcription factor occupancy. *Cell Rep.* **12,** 1184–1195 (2015).
10. Feldmann, A. *et al.* Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.* **9,** e1003994 (2013).
11. Wu, H. & Zhang, Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell* **156,** 45–68 (2014).
12. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500,** 477–481 (2013).
13. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Rev. Genet.* **13,** 484–492 (2012).
14. Schübeler, D. Function and information content of DNA methylation. *Nature* **517,** 321–326 (2015).
15. Tsumura, A. *et al.* Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes Cells* **11,** 805–814 (2006).
16. Karimi, M. M. *et al.* DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell* **8,** 676–687 (2011).
17. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (2012).
18. Virbasius, C. A., Virbasius, J. V. & Scarpulla, R. C. 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 (1993).
19. Kumari, D. & Usdin, K. 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 (2001).
20. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152,** 1146–1159 (2013).
21. Hu, S.*et al.* DNA methylation presents distinct binding sites for human transcription factors. *eLife* **2,** e00726 (2013).
22. Baubec, T., Ivanek, R., Lienert, F. & Schübeler, D. Methylation-dependent and -independent genomic targeting principles of the MBD protein Family. *Cell* **153,** 480–492 (2013).
23. Borgel, J. *et al.* Targets and dynamics of promoter DNA methylation during early mouse development. *Nature Genet.* **42,** 1093–1100 (2010).
24. Ficz, G. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13,** 351–359 (2013).
25. Habibi, E. *et al.* Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. *Cell Stem Cell* **13,** 360–369 (2013).
26. Hon, G. C. *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* **22,** 246–258 (2012).
27. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
28. Lienert, F. *et al.* Identification of genetic elements that autonomously determine DNA methylation states. *Nature Genet.* **43,** 1091–1097 (2011).
29. Krebs, A. R., Dessus-Babus, S., Burger, L. & Schübeler, D. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *eLife* **3,** e04094 (2014).
30. Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnol.* **32,** 171–178 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to S. Dessus-Babus, K. Jacobeit and T. Roloff (FMI) for processing deep-sequencing samples, to C. Wirbelauer for technical assistance and to A. Arnold for technical advice. We thank M. Stadler and D. Gaidatzis for bioinformatic advice and members of our laboratory, N. Thomae (FMI) and M. Lorincz (UBC Vancouver) for comments on the manuscript. We apologize to colleagues whose work we could not cite owing to space limitations. Research in the laboratory of D.S. is supported by the Novartis Research Foundation, the European Union (NoE 'EpiGeneSys' FP7-HEALTH-2010-257082 and the 'Blueprint' consortium FP7-282510), the European Research Council (EpiGePlas) and the Swiss initiative in Systems Biology (RTD Cell Plasticity). A.F.B. and P.A.G. are supported by EMBO postdoctoral long-term fellowships and S.D. and D.H. by predoctoral fellowships from the Boehringer Ingelheim Fonds.

Author Contributions A.F.B., L.B., S.D. and D.S. initiated and designed the study; S.D. performed the experiments; A.F.B. performed the data analysis; S.D. contributed to data analysis; S.D. and P.A.G. generated the TKO cell line; D.H. generated the overexpression construct; L.B. advised on data analysis; D.S. supervised all aspects of the project; the manuscript was prepared by S.D., A.F.B. and D.S. All authors discussed results and commented on the manuscript.

Author Information Genome-wide datasets generated for this study are deposited at GEO under the accession number GSE67867. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.S. (dirk@fmi.ch).

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture. Mouse ES cells HA36CB1/159-2 (denoted hereafter as 159) derived from mixed 129-C57Bl/6 background blastocysts[22], TC-1 cells and REST knockout and corresponding wild-type cells[31,32] were cultivated without feeders on 0.2% gelatine-coated dishes in DMEM, supplemented with 15% fetal calf serum, 1× non-essential amino acids, 2 mM L-glutamine, LIF and 0.001% β-mercaptoethanol (37 °C, 7% CO_2). Serum-free cultivation was performed in N2B27 medium, supplemented with 1× non-essential amino acids, 2 mM L-glutamine, LIF and 0.001% β-mercaptoethanol, as well as MEK inhibitor PD0325901 (1 μM) and GSK3 inhibitor CHIR99021 (3 μM), together known as 2i. For switching between culturing conditions, cells were cultured for at least three weeks under the new conditions before performing downstream experiments. Mouse 159 ES cells were differentiated to neuronal progenitors as previously described[33]. HMECs were purchased from Lonza (CC-2551), cultivated according to the supplier's instructions and collected after two passages. HCC1954 cells were cultured in RPMI 1640 medium supplemented with 10% fetal calf serum, 1× nonessential amino acids and 1× L-glutamine (37 °C, 5% CO_2).

Generation of isogenic DNMT TKO cell lines. Mouse 159 ES cells were co-nucleofected with three plasmids expressing mammalian-codon optimized Cas9 and sgRNAs targeting the region coding for the active PCQ/N loop in *Dnmt1*, *Dnmt3a*, and *Dnmt3b* (parental vector pX330, guide oligo sequences: *Dnmt1* (CACCTGTGGTGGGCCACCCTGCCA, AAACTGGCAGGGTGG CCCACCACA), *Dnmt3a* (CACCGACAATGGAGAGGTCATTGC, AAACGCAATGACCTCTCCATTGTC), *Dnmt3b* (CACCCGTTAGAGAG ATCATTGCAT, AAACATGCAATGATCTCTCTAACG)). A plasmid conveying resistance against puromycin was co-transfected. Puromycin selection (2 μg ml^{-1}) was carried out one day after transfection for 48 h. After five days of recovery, individual colonies were picked and genotyped by methylation-sensitive HpaII digest, using methylation-insensitive MspI digest as control. For clones in which loss in methylation was observed, *Dnmt* genes were sequenced to confirm successful targeting of all six alleles. Global 5-methylcytosine and 5-hydroxy-methylcytosine levels in positive TKO clones were measured by Zymo Research (http://www.zymoresearch.com), using high-pressure liquid chromatography coupled to mass spectrometry.

RNA isolation. RNA was isolated with the RNeasy mini kit (Qiagen) with on-column DNA digestion. For RNA-seq, two micrograms of total RNA from three independent cultures were depleted from ribosomal RNA using the Ribo-Zero rRNA removal kit (Epicentre).

DNase footprinting. DNase treatment of wild-type and TKO cells was performed essentially as previously described, with some modifications[34]. Briefly, intact nuclei were extracted using 0.03% NP-40 in an isotonic buffer. After NP-40 removal, batches of 5 million nuclei were incubated for 4 min at 37 °C with a range of DNase I (DPRF, Worthington) concentrations in the presence of Ca^{2+}. The digestion was stopped by addition of EDTA and SDS and the samples were treated with proteinase K and RNase A. Phenol-chloroform extracted DNA was separated on a 5–30% sucrose gradient by ultracentrifugation for 24 h and fractionated with a Gilson fraction collector FC 203B. Fractions were precipitated with ethanol and resuspended in TE buffer. Both successful digestion and size separation were verified by agarose gel electrophoresis. In addition, qPCR for amplicons within or outside known DHSs was used to confirm enrichment of DHSs in DNase-treated versus untreated and size-selected versus total DNA (primer sequences available upon request). Low-coverage sequencing of a barcoded pool of samples derived from different fractions of the sucrose gradient and treated with different DNase concentrations was used to select the sample with the highest information content. Based on this, the fraction of the gradient containing the shortest fragments (1–100 bp) was chosen for high-coverage sequencing.

Chromatin immunoprecipitation. Chromatin immunoprecipitation (ChIP) was carried out essentially as previously described[35], using a monoclonal antibody against NRF1 (Abcam, ab55744) and a polyclonal one against H3K27ac (Abcam, ab4729). ChIP–qPCRs were performed on at least three independent ChIP replicates according to standard protocols. Primer sequences are available upon request.

Knockdown by siRNA. TKO cells were reverse transfected with four preselected siRNAs targeting *Nrf1* (Qiagen, FlexiTube GeneSolution, GS18181) and Lipofectamine RNAiMax (Life Technologies) in three biological replicates, using the supplier's positive and negative controls (Qiagen, AllStars Mm Cell Death Control siRNA, SI04939025, AllStars Negative Control siRNA, SI03650318). To test knockdown efficiency, RNA was isolated after 72 h, reverse transcribed (PrimeScript, Takara) and *Nrf1* and *Gapdh* levels were determined according to standard protocols using predesigned TaqMan probes (Applied Biosystems, 4331182 and 4448489). Protein levels were measured by western blot on nuclear extracts. The most efficient siRNA targeting *Nrf1* (Mm_Nrf1_7 FlexiTube siRNA, SI05183738) and the negative control siRNA were used for RNA-seq experiments.

Transient overexpression. For transient overexpression, NRF1 was placed under the control of the CAG promoter. *Nrf1* cDNA was amplified from a random hexamer reverse transcription cDNA library (Superscript III, Invitrogen) generated from total RNA extracts and cloned into pL1-CAGGS-bio-MCS-polyA-1L[22]. Primer sequences are available upon request. This plasmid was reverse transfected into mouse 159 ES cells using Lipofectamine 2000 (Invitrogen). ChIP was performed 12 h after transfection. Overexpression was verified by western blot on nuclear extracts.

Recombinase-mediated cassette exchange. DNA fragments to be inserted into the ectopic genomic site in TC-1 cells were amplified from genomic DNA and cloned into a plasmid containing a multiple cloning site flanked by two inverted L1 *Lox* sites. We inserted two endogenous NRF1 binding sites (chr8: 113,271,870–113,272,282 and chr8: 123,020,293–123,020,670 for Extended Data Fig. 8) as well as part of the *Mrap* promoter (chr16: 90,738,245–90,738,944 for Fig. 4a), into which we integrated an NRF1 motif with Quickchange PCR mutagenesis by replacing the T at position chr16: 90,738,825 with CATG. Primer sequences are available upon request. Both unmethylated plasmids and plasmids that were *in vitro* methylated with M.SssI (NEB) were used for the recombinase-mediated cassette exchange reaction[36]. Complete *in vitro* methylation of the plasmids was confirmed by digestion with HpaII/MspI. Recombinase-mediated cassette exchange was performed in TC-1 ES cells as previously described[28,35]. Single clones were picked 12 days after nucleofection and tested for successful insertion events by PCR. To remove methylation after insertion, clones were treated with 25 nM 5-Aza-2′-deoxycytidine (Sigma) for 4 days. For analysis of wild-type and mutated fragments of the *Gtf2a1l* promoter, we used previously described clones that were generated in the same way[28].

Targeted amplicon bisulfite sequencing. For high coverage amplicon bisulfite sequencing of NRF1 binding sites target regions containing the highest confidence NRF1 motif (CGCATGCG) were selected based on high NRF1 ChIP enrichments in the TKO cell line, absence of enrichment in the wild-type and wild-type methylation levels of at least 80%. Primers for 200–400 bp amplicons were designed using our AmpliconBiSeq R package (https://github.com/BIMSBbioinfo/AmpliconBiSeq) and 56 pairs were randomly selected from this set. In addition, primers for 6 NRF1 motifs that were unbound in the TKO cell line, 9 unmethylated regions (UMRs), 9 fully methylated regions (FMRs), 9 constitutive REST/CTCF LMRs and T7/lambda were included as controls, resulting in 96 primer pairs in total (Supplementary Table 3). Primers were commercially synthesized in a 96-well plate format (Microsynth). Genomic DNA was isolated at the same time point as collection for ChIP. Bisulfite conversion was performed on 2 μg of the RNaseA-treated DNA mixed with 3.2 pM M.SssI methylated T7 and unmethylated lambda DNA as conversion controls (EpiTect Bisulfite kit, Qiagen). Bisulfite-converted DNA was amplified in a 96-well format with the designed specific primers using the following cycling conditions: 20 touch-down cycles from 55 to 50 °C with 30 s at 95 °C, 30 s at 55/50 °C and 30 s at 72 °C, followed by 36 cycles of 30 s at 95 °C, 30 s at 50 °C and 30 s at 72 °C and a final 5 min extension step at 72 °C. Then 5 μl of each individual PCR reaction were combined and the pool was size-selected using Agencourt AMPure XP beads (Beckman Coulter) before library preparation. Methylation profiling for insertions as well as REST motif-containing LMRs/UMR was performed with the same settings (genomic coordinates and primers in Supplementary Table 3).

Library preparation and next-generation sequencing. DNase-seq libraries were prepared essentially according to standard Illumina protocols, using 40 ng of the precipitated fractions of the sucrose gradient as starting material. To reduce amplification bias, end-repaired, A-tailed and adaptor ligated DNA was amplified in 6 cycles of PCR with KAPA HiFi Hot Start polymerase. Adaptor dimers were subsequently removed with Agencourt AMPure XP beads (Beckman Coulter). For sequencing of total RNA, strand-specific RNA-seq libraries were prepared from rRNA depleted samples using the ScriptSeq v2 protocol (Epicentre). Libraries for ChIP-seq were prepared according to standard Illumina library preparation protocols, with matching input sequenced for each IP. Twelve cycles of PCR (NEB Q5 Hot Start HiFi PCR) were performed on end-repaired, A-tailed and adaptor-ligated DNA before gel size-selection. Libraries for whole genome bisulfite sequencing were prepared essentially as previously described[8]. Briefly, 5 μg of sonicated genomic DNA were end repaired and 3′-end adenylated using the Illumina TruSeq DNA LT Sample Preparation kit (Illumina 15025064). Paired-end adapters were ligated to the DNA fragments and adaptor-ligated DNA was purified by 2% agarose gel electrophoresis. The gel-purified DNA was converted with the EpiTect bisulfite kit (Qiagen). Converted libraries were enriched by 10 cycles of PCR using PfuTurbo Cx Hotstart DNA Polymerase (Agilent) and purified using AMPure XP beads. For amplicon bisulfite sequencing, libraries of purified PCR pools were prepared according to standard Illumina library preparation protocols using 12 cycles of

PCR (NEB Q5 Hot Start HiFi PCR). Quality of the libraries and size distribution was assessed on an Agilent 2100 Bioanalyzer (Agilent Technologies). For RNA-seq, DNase-seq benchmarking, ChIP-seq and amplicon bisulfite sequencing, three to six samples with different barcodes were mixed at equimolar ratios per pool. Sequencing was performed on an Illumina HiSeq 2500 machine (DNase-seq, RNA-seq, ChIP-seq: 50 bp read length, single-end; whole-genome bisulfite sequencing: 100 bp read length, paired end) or a MiSeq machine (DNase-seq benchmarking: 25 bp read length, paired end; amplicon bisulfite sequencing: 250 bp read length, paired end) according to Illumina standards.

Sequencing data processing. RNA-seq reads were mapped to the mouse reference transcriptome (NCBIM37.67) using TopHat[37] version 1.3.1 with parameter–no-novel-juncs. DNase-seq reads were trimmed for Illumina adaptors. DNase-seq and ChIP-seq reads were mapped to the mouse reference genome (mm9 only chromosomes 1 to 19, X, Y and M) or human reference genome (hg19 only chromosomes 1 to 22, X, Y and M) using Bowtie[38] version 1.0.0 with parameters -v 3 -m 1–best–strata. Whole-genome Bis-seq reads were processed with QuasR[39] and positions covered by at least 10 reads were used. Amplicon bisulfite sequencing samples were analysed with the AmpliconBiSeq R package (https://github.com/BIMSBbioinfo/AmpliconBiSeq). Amplicons with at least 100× (TKO-specific NRF1 sites) or 30× (insertions) coverage were selected for downstream analysis.

Visualization of read densities. We used the first bp (5′-end) of the DNase-seq reads (DNase I cut site), the ChIP-seq reads extended to 200 bp (average estimated fragment length) and split RNA-seq reads to calculate the read density normalized to one million reads in the library for each genomic position (BigWig files). Screenshots of genomic regions were taken using the UCSC genome browser[40].

Identification of enriched regions. DHSs were identified as regions with enriched DNase I cuts using a sliding window approach. The mean read density for each region of 51 bp was calculated by steps of 10 bp within mappable regions and outside ENCODE blacklisted regions[27]. Regions with a mean density of 0.001 (about 10 DNase I cuts) and at least 10 bp covered were merged and kept if their length was at least 100 bp. Enriched ChIP-seq regions over corresponding input were identified using the peak calling software Peakzilla[41] with default parameters.

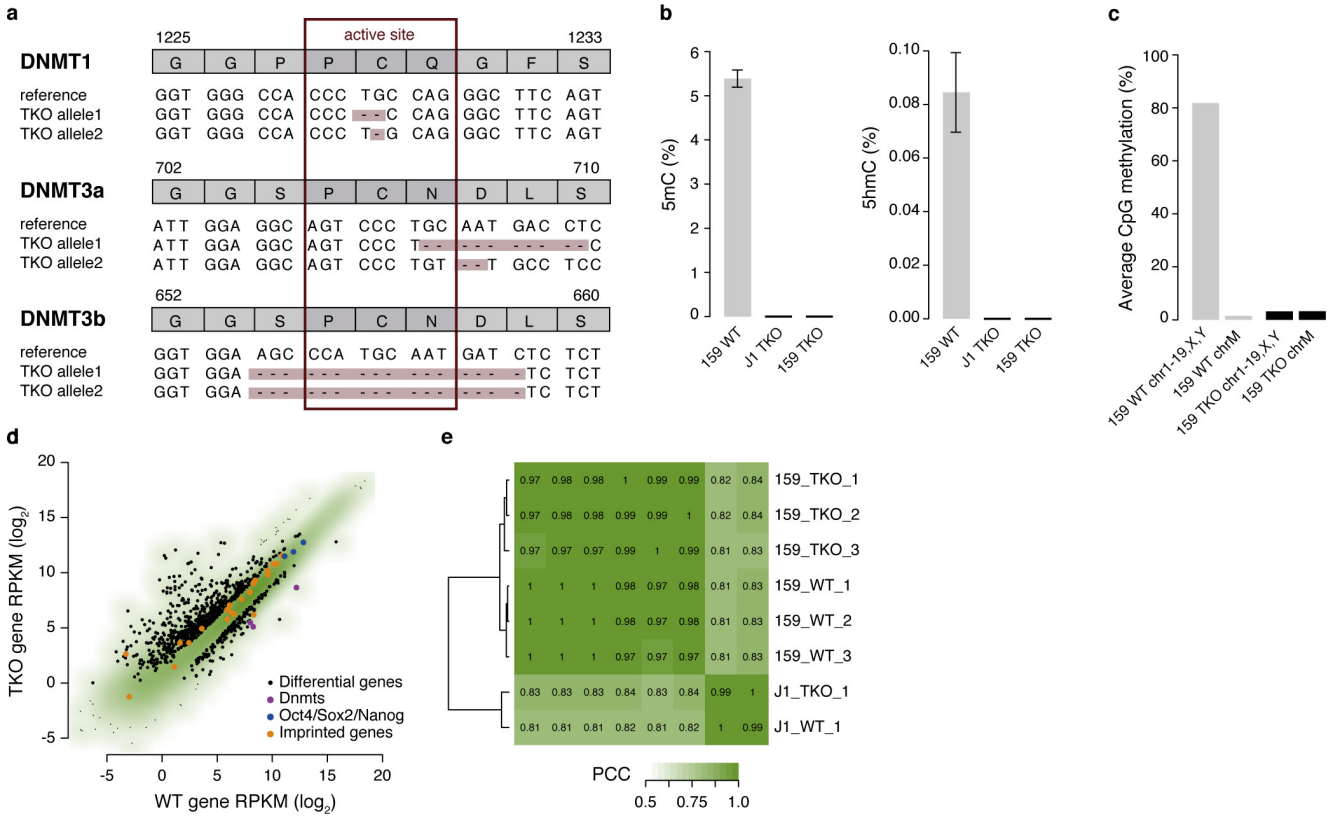
Correlation of read counts. We used the first bp (5′-end) of the DNase-seq reads (DNase I cut site), the ChIP-seq reads extended to 200 bp (average estimated fragment length) and split RNA-seq reads to calculate raw read counts for regions of interest (merged DNase-seq or ChIP-seq enriched regions or genes). The R package DESeq[42] was used to normalize the raw read counts and identify differential regions using a fold change threshold of 2 and an adjusted P value threshold of 10^{-3} for DNase-seq and ChIP-seq regions and 10^{-5} for RNA-seq data sets. We generated scatterplots and calculated Pearson correlation coefficients (PCC) from the normalized read counts using R.

Functional analyses. Germline-specific imprinted regions were used from ref. 43. Peaks were assigned to their closest gene transcriptional start site (TSS) using the mouse reference transcriptome (NCBIM37.67) and human reference transcriptome (GRCh37.71). The conservation rate of regions was calculated using the PhastCons 11 way placental mammals[44].

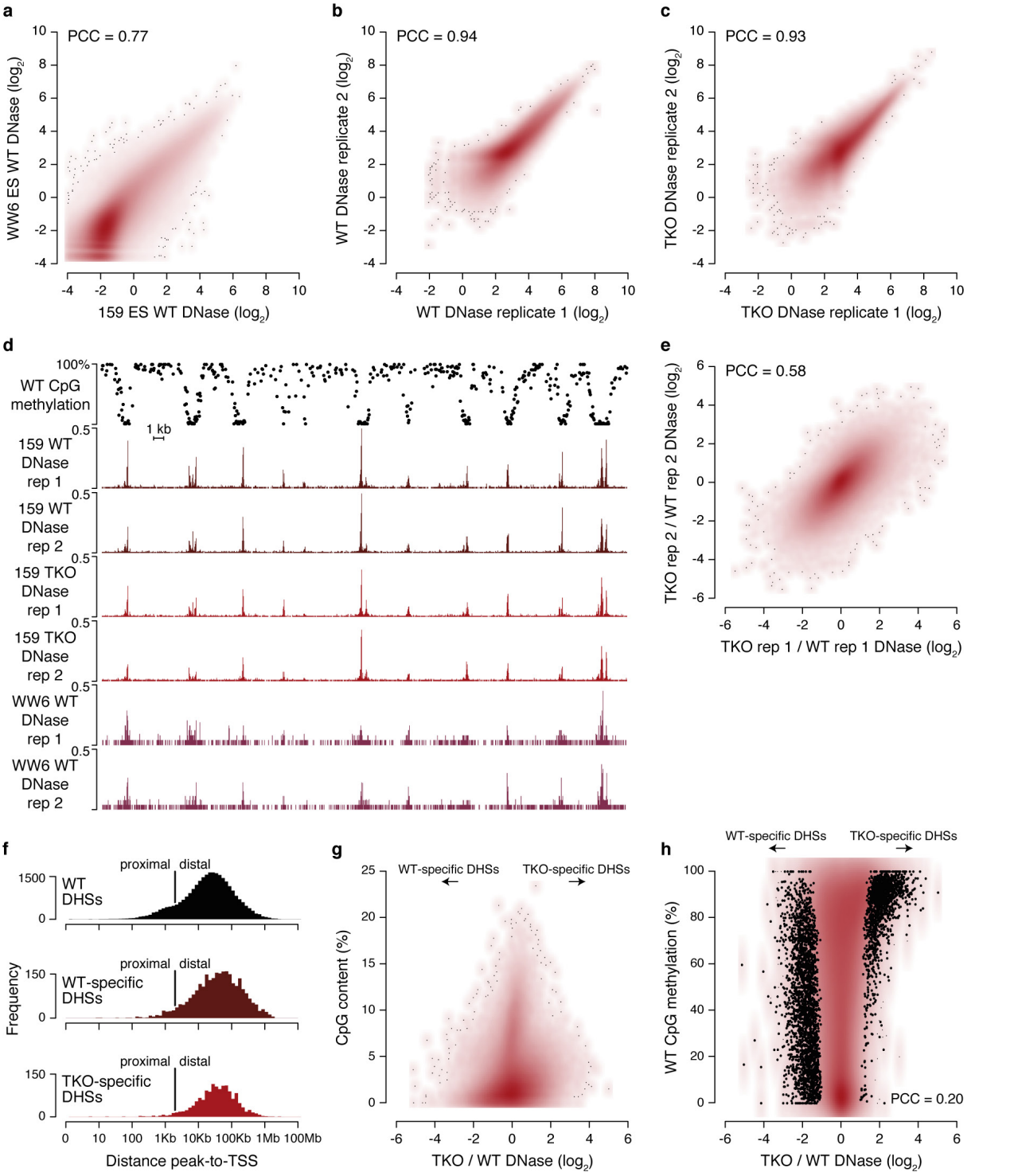
Motif-enrichment analysis. We searched DHS regions for known motifs from JASPAR[45], ref. 46 and UniPROBE[47] using MAST[48] (from the MEME suite programs version 4.1.1) with a P value threshold of 2.44×10^{-4} ($(0.25)^6$) (see Supplementary Table 1). The statistical significance of the differential motif enrichment was assessed by a hypergeometric P value.

Published data sets. RNA-seq data sets in J1 mouse ES cells were obtained from GEO with the accession numbers GSM727427 and GSM727428 (ref. 16), in mouse ES cells cultured in serum from GSM590126, GSM758167 and GSM758168 (ref. 49), and 2i from GSM758168, GSM590128 and GSM590129 (ref. 49), in neuronal progenitors from GSM778489 and GSM778490 (ref. 50), in HMEC cells from GSM721141 (ref. 26), in HCC1954 cells from GSM721140 (ref. 26), in h1hESC from GSM758566 (ref. 51) and in GM12878 from GSM758559 (ref. 51). DNase-seq data sets in mouse ES cells were obtained from GSM1014159 (ref. 51). Bis-seq data sets in mouse ES cells were obtained from GSM748786 (ref. 8), in neuronal progenitors from GSM748788 (ref. 8), in HMEC cells from GSM721195 (ref. 26), in HCC1954 cells from GSM721194 (ref. 26), in H1-hESC from GSM1002649 (ref. 27) and in GM12878 from GSM1002650 (ref. 27). ChIP-seq data sets were obtained for NRF1 in H1-hESC from GSM935308 (ref. 27) and in GM12878 from GSM935309 (ref. 27), in mouse ES cells for MeCP2 from GSM972976 (ref. 22), for CTCF from GSM747534 (ref. 8), for REST from GSM671094 (ref. 52), for ZFX from GSM288352 (ref. 53), for KLF4 from GSM288354 (ref. 53), for ESRRB from GSM288355 (ref. 53), for cMYC from GSM288356 (ref. 53), for nMYC from GSM288357 (ref. 53), for OCT4 from GSM307137 (ref. 54), for SOX2 from GSM307138 (ref. 54) and for NANOG from GSM307141 (ref. 54).

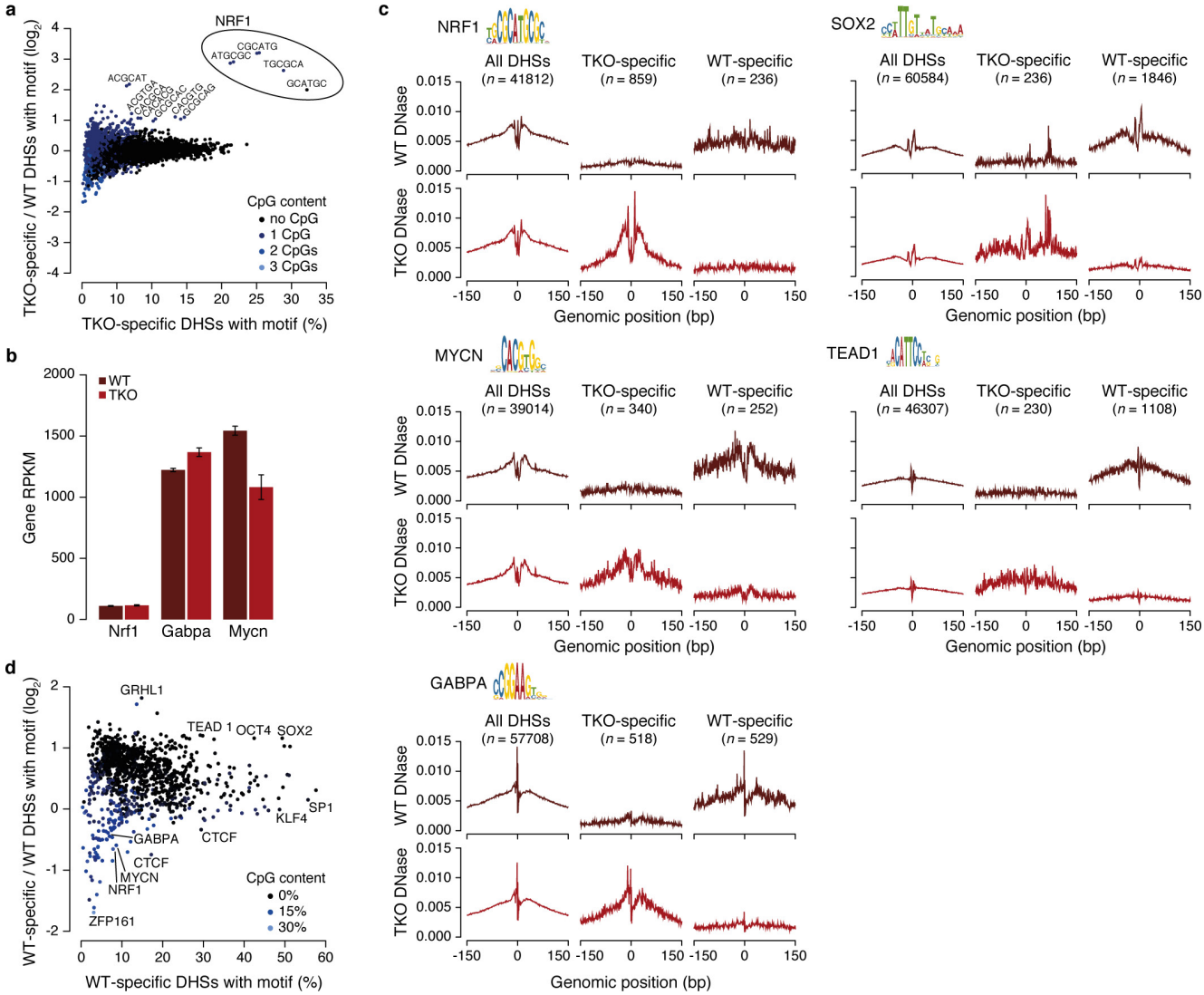
31. Jørgensen, H. F., Chen, Z. -F., Merkenschlager, M. & Fisher A. G. Is REST is required for ESC pluripotency? *Nature* **457,** E4–E5, E7 (2009).
32. Chen, Z. F., Paquette, A. J. & Anderson, D. J. NRSF/REST is required *in vivo* for repression of multiple neuronal target genes during embryogenesis. *Nature Genet.* **20,** 136–142 (1998).
33. Bibel, M., Richter, J., Lacroix, E. & Barde, Y.-A. Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. *Nature Protocols* **2,** 1034–1043 (2007).
34. John, S.*et al.* Genome-scale mapping of DNase I hypersensitivity. *Curr. Protoc. Mol. Biol.* **Chapter 27,** Unit 21.27–21.27.20 (2013).
35. Jermann, P., Hoerner, L., Burger, L. & Schübeler, D. Short sequences can efficiently recruit histone H3 lysine 27 trimethylation in the absence of enhancer activity and DNA methylation. *Proc. Natl Acad. Sci. USA* **111,** E3415–E3421 (2014).
36. Schübeler, D. *et al.* Genomic targeting of methylated DNA: influence of methylation on transcription, replication, chromatin structure, and histone acetylation. *Mol. Cell. Biol.* **20,** 9103–9112 (2000).
37. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25,** 1105–1111 (2009).
38. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
39. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: quantification and annotation of short reads in R. *Bioinformatics* **31,** 1130–1132 (2015).
40. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
41. Bardet, A. F. *et al.* Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics* **29,** 2705–2713 (2013).
42. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
43. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148,** 816–831 (2012).
44. Siepel, A. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
45. Sandelin, A. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32,** D91–D94 (2004).
46. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152,** 327–339 (2013).
47. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Res.* **37,** D77–D82 (2009).
48. Bailey, T. L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14,** 48–54 (1998).
49. Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149,** 590–604 (2012).
50. Tippmann, S. C. *et al.* Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. *Mol. Syst. Biol.* **8,** 593 (2012).
51. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515,** 355–364 (2014).
52. Arnold, P. *et al.* Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res.* **23,** 60–73 (2013).
53. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133,** 1106–1117 (2008).
54. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134,** 521–533 (2008).



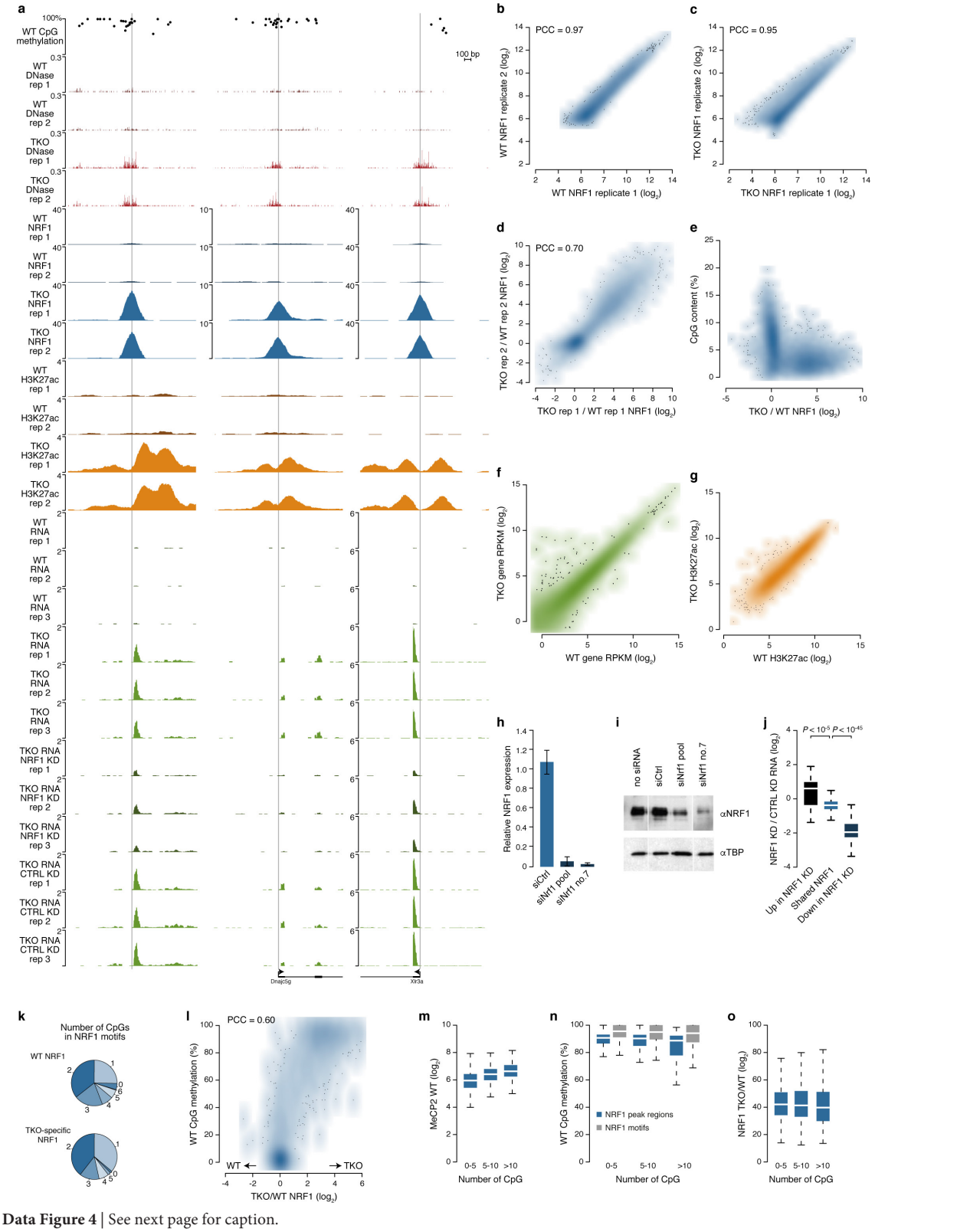
Extended Data Figure 1 | Characterization of an isogenic DNMT TKO cell line created with CRISPR/Cas9. a, Frameshift deletions (brown) introduced at the active PCQ/N loops of the three DNA methyltransferases by CRISPR/Cas9 genome editing. **b**, Levels of 5-methyl-C and 5-hydroxymethyl-C in the wild-type, isogenic (mouse ES cell line 159) and traditional (J1) TKO cell lines as determined by mass spectrometry. **c**, Average CpG methylation in wild-type and TKO cell lines determined by whole-genome bisulfite sequencing. Methylation in the TKO cell line is comparable to background levels represented by the methylation in chromosome M. **d**, Gene expression levels (RPKM) in isogenic wild type and TKO (159). Black dots represent significantly differentially expressed genes in wild type or TKO, with expected unpregulation of germline genes[16]. The *Dnmt* genes are among the most downregulated genes (purple), while the majority of genes that reside within imprinted domains are upregulated roughly twofold (orange). Prominent marker genes of ES cells (*Oct4, Sox2* and *Nanog*, blue) remain unaltered. **e**, Hierarchical clustering of gene expression correlations for three independent 159 ES cell line wild-type and TKO replicates, and published J1 wild-type and TKO RNA-seq samples[16]. Overall, gene expression clusters by strain rather than presence of DNA methylation. This reflects the strong influence of genetic background on the global gene expression program and supports our approach of focusing further analysis on the isogenic TKO.



Extended Data Figure 2 | Characteristics of DNase-hypersensitive sites. **a**, DNase-seq signal in our 159 ES cell line (wild-type) and an ENCODE WW6 ES cell (wild-type) DNase-seq sample[27] using a tiling window (500 bp) over the whole genome in mappable regions not blacklisted by ENCODE, illustrating that our protocol for genome-wide detection of DHSs matches available data sets in mouse ES cells. PCC was calculated on all DHSs. **b**, **c**, DNase-seq signal and PCC at all DHSs for independent biological replicates of wild type (**b**) and TKO (**c**). **d**, Wild-type methylation and replicates for DNase-seq signal in the 159 ES cell line (wild-type and TKO) and ENCODE WW6 (wild-type) at the genomic region from Fig. 1a (chr17: 25,920,000–25,972,499), illustrating that most DHSs remain unchanged upon removal of DNA methylation, in agreement with the overall similarity in gene expression. **e**, Change in DNase-seq signal and PCC between wild type and TKO using different replicate samples, illustrating a high reproducibility of quantitative DHS changes between wild type and TKO. **f**, Distance of all wild-type, wild-type-specific or TKO-specific DHSs from closest gene transcriptional start site (TSS). Proximal and distal separation is at 2 kb. **g**, Change in DNase-seq signal between TKO and wild-type as a function of CpG content for all wild-type and TKO DHSs, illustrating that most changes occur in CpG-poor regions. **h**, Change in DNase-seq signal between TKO and wild-type versus average CpG methylation of all wild-type and TKO DHSs matching Fig. 1c, showing that TKO-specific DHSs (right) lie in regions with high methylation in wild type. Black dots represent significantly enriched DHSs (see Methods) in wild type ($n = 2{,}837$) or TKO ($n = 1{,}543$) from Fig. 1b.

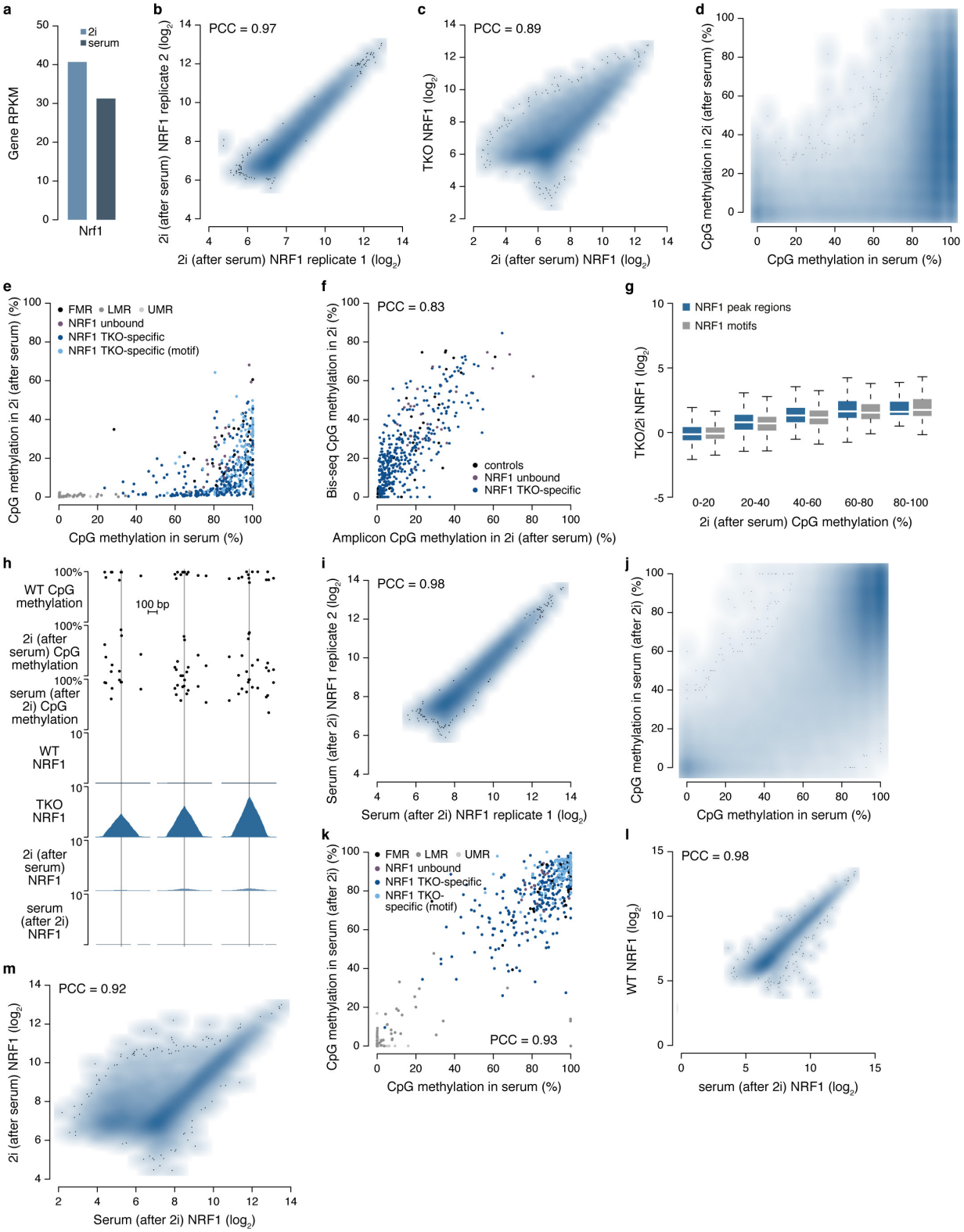


Extended Data Figure 3 | Motif enrichment in cell-line-specific DNase-hypersensitive sites. a, Occurrence of all possible hexamers in TKO-specific DHSs compared to all wild-type DHSs. Blue colouring illustrates hexamer CpG content. Hexamers representing the NRF1 motif are highlighted by a circle. Most strongly enriched hexamers are labelled (only one of two reverse complements). **b**, Gene expression levels (RPKM) of candidate methylation-sensitive TFs in wild type and TKO indicating that differential abundance does not account for DHS formation upon loss of DNA methylation. Error bars are standard deviation from three biological replicates. **c**, Footprints of candidate TF motifs enriched in TKO-specific (NRF1, MYCN, GABPA) or wild-type-specific (SOX2, TEAD1) DHSs shown as metaplot of wild-type (brown) or TKO (red) DNase-seq signal for all motifs in all wild-type and TKO (left), TKO-specific (middle) and wild-type-specific (right) DHSs. Number of regions is indicated above each metaplot. A DNase footprint is apparent at the NRF1 motif and, to a lesser extent, at MYCN and GABPA motifs specifically in TKO-specific sites in the TKO sample, whereas footprints at SOX2 and TEAD1 motifs in wild-type-specific sites are less unique to that cell state. **d**, Motif occurrences in wild-type-specific DHSs compared to all wild-type DHSs. Blue colouring illustrates motif CpG content.



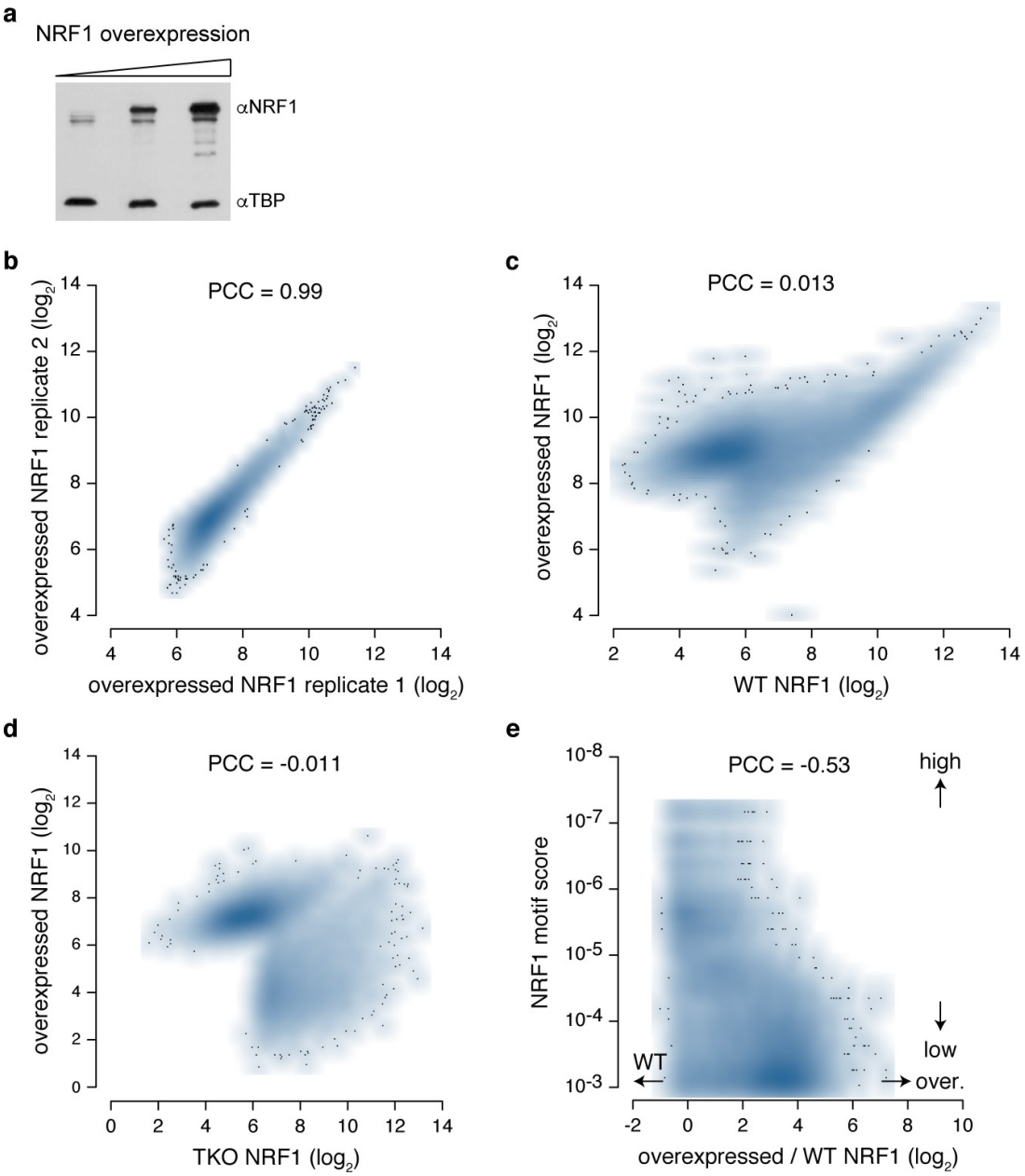
Extended Data Figure 4 | See next page for caption.

Extended Data Figure 4 | Characteristics of NRF1 binding sites. **a**, Wild-type methylation, and wild-type and TKO DNase-seq, NRF1 ChIP-seq, H3K27ac ChIP-seq and RNA-seq signal also upon *Nrf1* and mock knockdown in TKO at TKO-specific distal (left, chr4: 99,235,170–99,237,170; from Fig. 2a) and proximal (middle, chr5: 31,409,700–31,411,700; right, chrX: 70,341,500–70,343,500) genomic regions. The transcripts initiated directly at the NRF1 binding sites in TKO cells are specifically reduced upon knockdown of *Nrf1*, implying that they are indeed NRF1-dependent. **b**, **c**, NRF1 ChIP-seq signal at all NRF1 peak regions for independent biological replicates of wild type (**b**) and TKO (**c**). **d**, Change in NRF1 ChIP-seq signal and PCC between wild type and TKO using different replicate samples, illustrating a high reproducibility of quantitative NRF1 changes between wild type and TKO. **e**, Change in NRF1 ChIP-seq signal between TKO and wild type versus CpG content of all wild-type and TKO NRF1 peak regions, illustrating that most changes occur in CpG-poor regions. **f**, RNA expression levels (RPKM) in wild type and TKO at all wild-type and TKO NRF1 peak regions, illustrating the appearance of a few aberrant TKO-specific transcripts directly at NRF1 binding sites. **g**, H3K27ac ChIP-seq signal in wild type and TKO at all wild-type and TKO NRF1 peak regions, illustrating appearance of TKO-specific acetylation at a few NRF1 binding sites. **h**, Knockdown efficiency for the pool of three siRNAs and most efficient single siRNA targeting *Nrf1* in TKO cells. Mean of three independent biological replicates normalized to GAPDH; error bars reflect standard deviation. Genetic deletion of *Nrf1* with CRISPR/Cas9 was lethal (data not shown). **i**, Reduction in nuclear NRF1 levels upon siRNA knockdown with pool of three siRNAs and most efficient single siRNA targeting *Nrf1* as measured by western blot. Blot was cropped for clarity, all samples were loaded on the same gel (for uncropped gels see Supplementary Fig. 1). **j**, Expression change (in RPKM) of genes closest to shared and TKO-specific NRF1 peaks between TKO cells treated either with negative control siRNA or the most efficient single siRNA targeting *Nrf1*, showing highly significant loss in expression after knockdown. *P* values from Wilcoxon tests. **k**, Number of CpGs in NRF1 motifs closest to peak summit in all wild-type (top) or TKO-specific (bottom) NRF1 peaks, illustrating that motifs in TKO-specific NRF1 peaks contain at least one CpG. **l**, Change in NRF1 ChIP-seq signal between TKO and wild type versus average methylation in wild type at all NRF1 sites corresponding to Fig. 2g, illustrating that increased NRF1 binding in TKO occurs at regions that were methylated in wild type. **m**–**o**, Average wild-type MeCP2 ChIP-seq signal[22] (**m**), wild-type methylation in NRF1 peak regions or in NRF1 motifs closest to peak summits (**n**) and change of NRF1 signal between wild type and TKO (**o**) within 500 bp regions around TKO-specific NRF1 peak summits grouped according to CpG density (0–5 CpGs, $n = 3{,}680$; 5–10 CpGs, $n = 2{,}477$; >10 CpGs, $n = 680$). If indirect repression could contribute to differential NRF1 binding, we would expect a more pronounced increase of NRF1 binding at sites with higher CpG density upon demethylation of the genome, as methyl-CpG binding domain proteins (MBDs) such as MeCP2 bind preferentially to regions with a high density of methylated CpGs rather than fully methylated regions with low CpG density. TKO-specific binding of NRF1 is independent of CpG density and MeCP2 enrichment in the methylated genome, strongly arguing against an involvement of indirect repression in NRF1 binding site restriction.

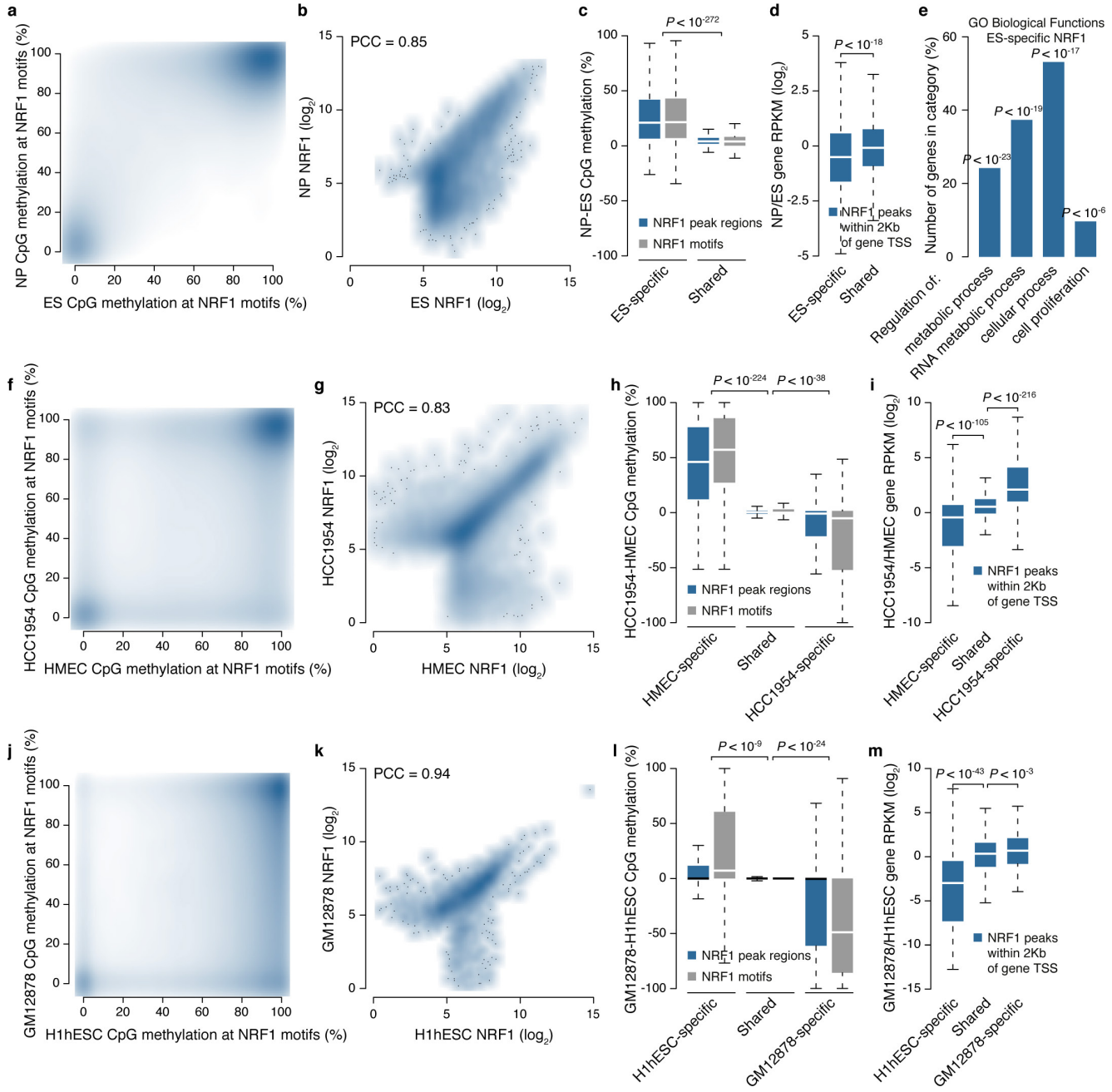


Extended Data Figure 5 | See next page for caption.

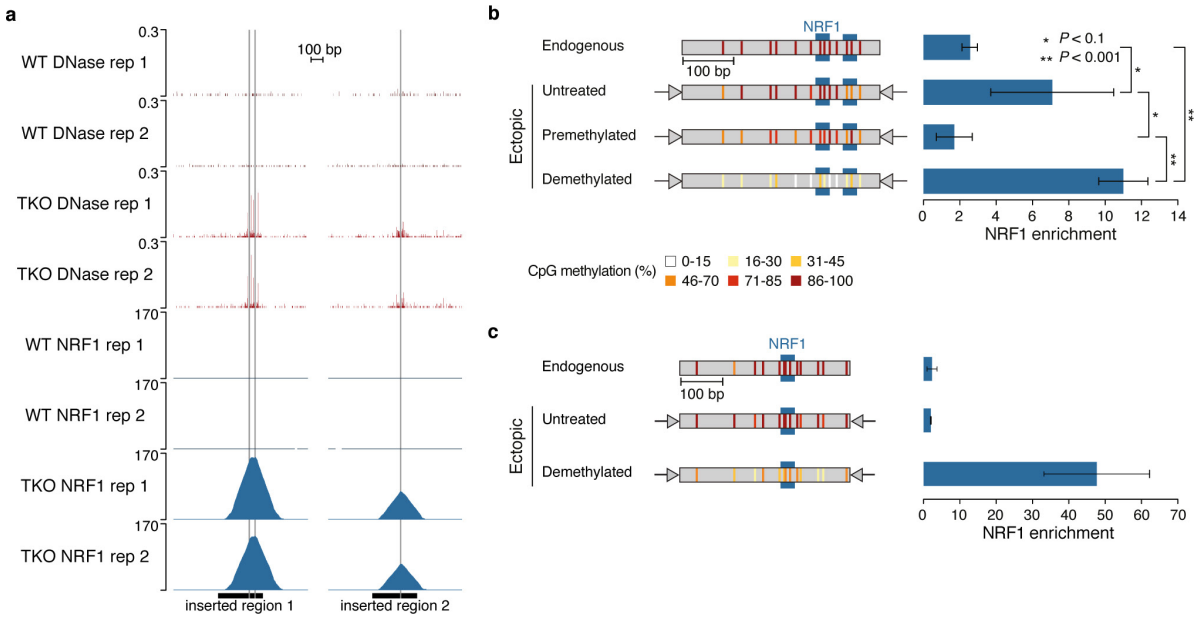
Extended Data Figure 5 | NRF1 binding in different culture conditions. **a**, *Nrf1* gene expression levels (RPKM) in 2i and serum culture conditions[49]. **b**, NRF1 ChIP-seq signal in wild-type cells adapted to 2i culture conditions (after culture with serum) for two biological replicates. **c**, NRF1 ChIP-seq signal in wild-type cells adapted to 2i (after culture with serum) and TKO. **d**, Methylation in wild-type cells cultured in serum and 2i (after culture with serum) at all NRF1 motifs. **e**, Methylation in serum and 2i (after culture with serum) measured by amplicon Bis-seq for fully methylated (FMR), low methylated (LMR), unmethylated (UMR) controls, 6 unbound NRF1 sites and 56 TKO-specific NRF1 sites. **f**, Comparison and PCC of DNA methylation levels by amplicon Bis-seq and whole-genome Bis-seq upon culture in 2i (after culture with serum). **g**, Average 2i (after culture with serum) methylation in NRF1 peak regions or NRF1 motifs within peaks versus change in NRF1 signal between TKO and 2i (after culture with serum) at all NRF1 peaks, illustrating that reduced NRF1 binding in 2i compared to TKO can be explained by residual methylation. **h**, Methylation in wild-type cells cultured in serum, cultured in 2i (after culture with serum) and cultured in serum (after culture in 2i) and NRF1 ChIP-seq signal in wild type, TKO, cultured in 2i (after culture with serum) and cultured in serum (after culture with 2i) at TKO-specific regions with higher 2i methylation in NRF1 motifs (grey lines) than surrounding region (left, chr10: 66,251,100–66,251,700; middle, chr4: 15,976,050–15,976,650; right, chr19: 55,833,420–55,834,020). NRF1 is unable to bind if CpGs in the motif remain methylated in 2i, even if the surrounding region is unmethylated. **i**, NRF1 ChIP-seq signal in wild-type cells adapted back to serum (after culture with 2i) for two biological replicates. **j**, Methylation in wild-type cells cultured in serum and adapted back to serum (after culture with 2i) at all NRF1 motifs. **k**, Methylation in wild-type cells cultured in serum and adapted back to serum (after culture with 2i) measured by amplicon Bis-seq for FMR, LMR and UMR controls, 6 unbound NRF1 sites and 56 TKO-specific NRF1 sites. **l**, NRF1 ChIP-seq signal in wild-type cells adapted back to serum (after culture with 2i) and original serum conditions. **m**, NRF1 ChIP-seq signal in wild-type cells adapted back to serum (after culture with 2i) and adapted to 2i (after culture with serum).



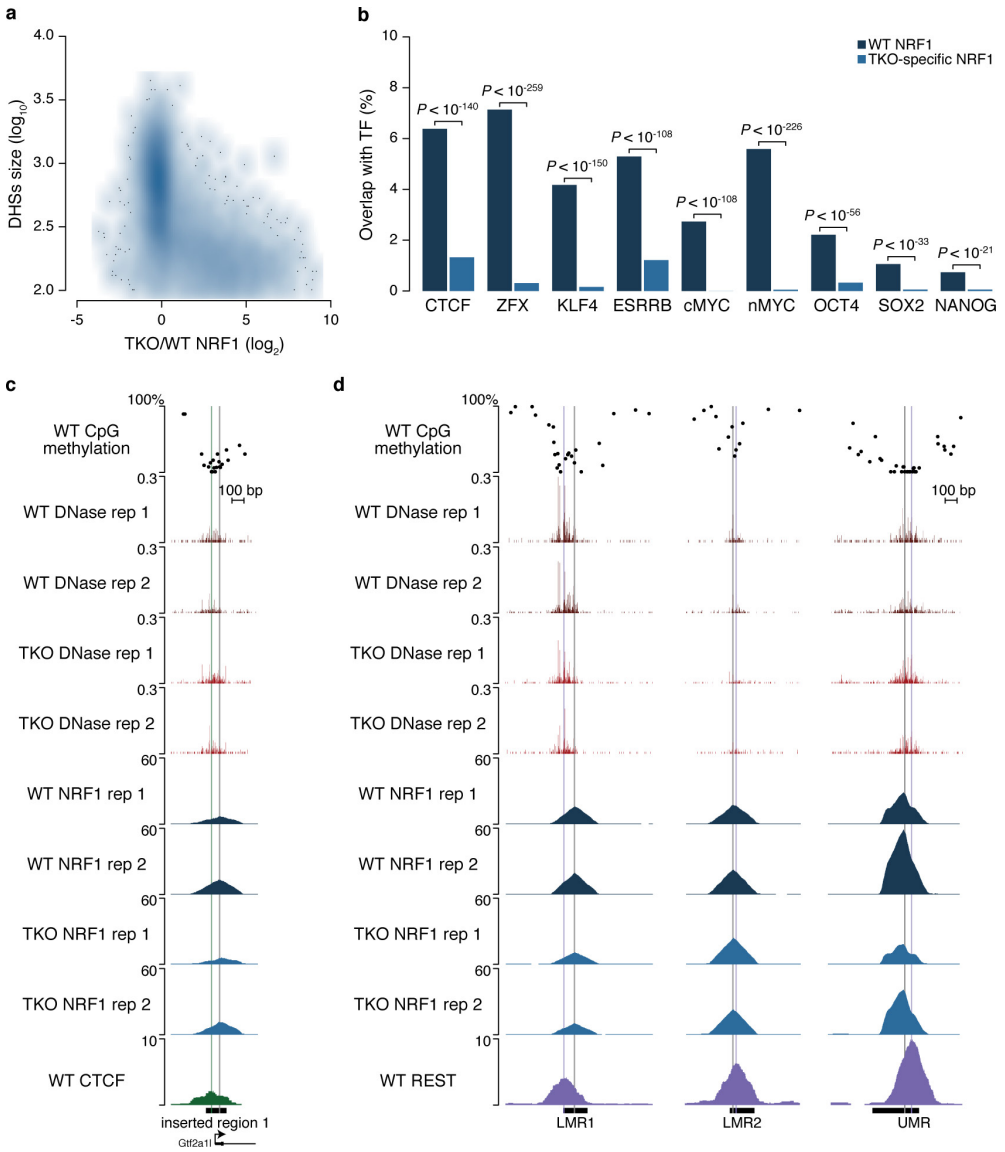
Extended Data Figure 6 | Overexpression of NRF1 is unable to induce binding to TKO-specific sites. a, Transient overexpression of NRF1 under control of the CMV (middle) or CAG promoter (right, used for ChIP experiments) leads to strong increase in nuclear NRF1 protein levels compared to endogenous levels (left) as measured by western blot (for uncropped gel data see Supplementary Fig. 1). The overexpressed protein contains a protein tag accounting for the higher molecular weight. **b**, NRF1 ChIP-seq signal upon transient NRF1 overexpression for two biological replicates. **c**, NRF1 ChIP-seq signal in wild type and upon overexpression. **d**, NRF1 ChIP-seq signal in TKO and overexpression conditions only at TKO- and overexpression-specific NRF1 peak regions, illustrating that TKO-specific NRF1 sites are distinct from overexpression-specific sites. **e**, Change in NRF1 ChIP-seq signal between overexpression and wild type versus the score (MAST position *P* value) of NRF1 motifs closest to the summit, illustrating that sites gaining most NRF1 upon overexpression do not contain high-confidence motifs.



Extended Data Figure 7 | Cell-type-specific binding of NRF1 correlates with methylation and expression changes. a–e, Comparison of NRF1 binding in ES and neuronal progenitor cells. Methylation in ES and neural progenitors[8] at all NRF1 motifs (**a**), NRF1 ChIP-seq signal in ES and neuronal progenitors at all NRF1 peaks (**b**), neuronal progenitor minus ES methylation of peak regions or NRF1 motifs in ES-specific ($n = 4,934$) and shared ($n = 4,951$) NRF1 peaks (negligible number of neuronal-progenitor-specific peaks) (**c**), expression of the genes[50] closest to ES-specific and shared NRF1 peaks (**d**), selection of gene ontology (GO) biological functions enriched in genes closest to ES-specific and shared NRF1 peaks (**e**). *P* values from Wilcoxon tests. **f–i**, Comparison of NRF1 binding in HMEC and HCC1954 cells. Methylation in HMEC and HCC1954[26] at all NRF1 motifs (**f**), NRF1 ChIP-seq signal in HMEC and HCC1954 at all NRF1 peaks (**g**), HCC1954 minus HMEC methylation of peak regions or NRF1 motifs in HMEC-specific ($n = 2,726$), HCC1954-specific ($n = 2,685$) and shared ($n = 12,180$) NRF1 peaks (**h**), expression of the genes[26] closest to HMEC-specific, HCC1954-specific and shared NRF1 peaks (**i**). **j–m**, Comparison of NRF1 binding in H1-hESC and GM12878 cells. Methylation in H1-hESC and GM12878[27] at all NRF1 motifs (**j**), NRF1 ChIP-seq signal in H1-hESC and GM12878[27] at all NRF1 peaks (**k**), GM12878 minus H1-hESC methylation of peak regions or NRF1 motifs in H1-hESC- ($n = 618$), GM12878-specific ($n = 561$) and shared ($n = 3,198$) NRF1 peaks (**l**), expression of the genes[27] closest to H1-hESC-specific, GM12878-specific and shared NRF1 peaks (**m**).



Extended Data Figure 8 | NRF1 binding to the unmethylated motif can be recapitulated at an ectopic site. a, Wild-type and TKO DNase-seq and NRF1 ChIP-seq signal for two biological replicates at the endogenous counterparts of the inserted regions profiled in Extended Data Fig. 8b (left, chr8: 123,019,920–123,021,030) and Extended Data Fig. 8c (right, chr8: 113,271,460–113,272,690). **b**, Methylation (amplicon Bis-seq, left, coloured lines indicate position and methylation status of CpGs) and NRF1 binding (ChIP-qPCR, right) for an endogenous methylation-dependent NRF1 site (chr8: 123,020,293–123,020,670) and upon insertion of this region into a defined ectopic genomic locus. The position of the two NRF1 motifs containing two CpGs each is indicated in blue. The reporter construct was inserted either unmethylated or *in vitro* premethylated with M.SssI. In the untreated construct one motif becomes completely methylated upon insertion, whereas the other only gains roughly 50% methylation, and NRF1 binding is detected. The pre-methylated construct maintains at least one CpG with almost complete methylation in both core motifs present and shows strongly reduced NRF1 binding by comparison. Thus, the methylation sensitivity of NRF1 can be recapitulated in an ectopic site even in the absence of global changes in DNA methylation. As expected, forcing complete demethylation of both core motifs in the premethylated insert by treatment of the cells with 5-aza-2′-deoxycytidine leads to further increased NRF1 binding compared to the untreated template. ChIP–qPCR enrichments are the mean of three independent biological replicates; error bars reflect standard deviation. See Supplementary Table 3 for methylation source data. **c**, Methylation (amplicon Bis-seq, left, coloured lines indicate position and methylation status of CpGs) and NRF1 binding (ChIP–qPCR, right) for an endogenous methylation-dependent NRF1 site (chr8: 113,271,870–113,272,282) and upon insertion of this region into a defined ectopic genomic locus. The untreated template gains full methylation in the core motif (blue) and does not show detectable NRF1 binding. Forcing complete demethylation by treatment with 5-aza-2′-deoxycytidine enables NRF1 to bind the site in the ectopic locus. ChIP–qPCR enrichments are mean of three independent biological replicates; error bars reflect standard deviation. See Supplementary Table 3 for methylation source data.



Extended Data Figure 9 | Constitutive NRF1 sites are co-bound by other TFs. a, Change in NRF1 ChIP-seq signal between TKO and wild type versus size of DHSs overlapping NRF1 peak regions, illustrating that wild-type NRF1 sites tend to overlap with larger DHSs. **b**, Overlap of wild-type and TKO-specific NRF1 peak regions with published ChIP-seq peak regions from other TFs expressed in ES cells[8,53,54], illustrating that wild-type NRF1 sites coincide with other TF binding events. *P* values from hypergeometric tests. **c**, Wild-type methylation, wild-type and TKO DNase-seq, and NRF1 and CTCF[8] ChIP-seq signal for two biological replicates at the endogenous *Gtf2a1l* promoter (chr17: 89,067,600–89,068,350). The region used for the insertion experiments in Fig. 4b is indicated below. **d**, Wild-type methylation, wild-type and TKO DNase-seq for two biological replicates and NRF1 and REST[52] ChIP-seq signal at adjacent NRF1 and REST binding sites (left, chr15: 100,703,260–100,704,500; middle, chr2: 180,152,200–180,153,150; right, chr2: 118,604,800–118,605,900). Regions profiled with amplicon Bis-seq in REST wild-type and REST KO cells in Fig. 4c and the position of the TF motifs are indicated below.

Extended Data Table 1 | Number of raw and mapped reads and enriched regions for all high-throughput sequencing samples

Type	Sample	Number of raw reads	Number of mapped reads	Percent of mapped reads	Number of enriched regions
RNA-seq	RNA_WT_1	81902703	64369711	79	NA
	RNA_WT_2	73623473	58374618	79	NA
	RNA_WT_3	69703377	55284618	79	NA
	RNA_TKO_1	70027452	55087030	79	NA
	RNA_TKO_2	82516808	64299099	78	NA
	RNA_TKO_3	73579964	58274831	79	NA
	RNA_TKO_CTRL_KD_1	64533537	51328388	80	NA
	RNA_TKO_CTRL_KD_2	79559568	63285076	80	NA
	RNA_TKO_CTRL_KD_3	79175188	62605040	79	NA
	RNA_TKO_NRF1_KD_1	78404841	56795481	72	NA
	RNA_TKO_NRF1_KD_2	74983199	57017015	76	NA
	RNA_TKO_NRF1_KD_3	77279139	57332826	74	NA
DNase-seq	DNASE_WT_1	131089973	96126970	73	125477
	DNASE_WT_2	238165244	170325464	71	222894
	DNASE_TKO_1	210534561	152434201	72	198796
	DNASE_TKO_2	170287886	117016188	69	132369
ChIP-seq	NRF1_CHIP_WT_1	40570927	31414178	77	6835
	NRF1_CHIP_WT_2	40365286	31447763	78	9847
	NRF1_INPUT_WT	22773779	18247061	80	NA
	NRF1_CHIP_TKO_1	32306980	25333581	78	11965
	NRF1_CHIP_TKO_2	45342909	35349643	78	13264
	NRF1_INPUT_TKO	24937026	19810205	79	NA
	NRF1_CHIP_to2i_1	51059626	40940267	80	7088
	NRF1_CHIP_to2i_2	50939344	36209344	71	9470
	NRF1_INPUT_to2i	30416060	23460617	77	NA
	NRF1_CHIP_toSerum_1	42310254	33037271	78	4941
	NRF1_CHIP_toSerum_2	42928737	33018296	77	5562
	NRF1_INPUT_toSerum	25103067	19493583	78	NA
	NRF1_CHIP_Over_1	77223442	56769391	73	18021
	NRF1_CHIP_Over_2	73340571	54380146	74	10479
	NRF1_INPUT_Over	70242507	52149318	74	NA
	NRF1_CHIP_NP_1	117333886	47952332	41	4564
	NRF1_INPUT_NP_1	35321797	15075613	43	NA
	NRF1_CHIP_NP_2	115065799	56350753	49	4906
	NRF1_INPUT_NP_2	25142679	11305749	45	NA
	H3K27ac_CHIP_WT_1	41972346	34720037	83	30616
	H3K27ac_CHIP_WT_2	40822025	34615432	85	29224
	H3K27ac_CHIP_TKO_1	50829570	41308561	81	29455
	H3K27ac_CHIP_TKO_2	45485455	38417647	84	30927
	NRF1_CHIP_HMEC_1	35943107	26822872	75	11585
	NRF1_CHIP_HMEC_2	40718156	28412905	70	13395
	NRF1_INPUT_HMEC	37667963	30523846	81	NA
	NRF1_CHIP_HCC1954_1	41562818	30632702	74	13896
	NRF1_CHIP_HCC1954_2	31412483	22848551	73	12594
	NRF1_INPUT_HCC1954	36664966	29527716	81	NA
Bis-seq	BISSEQ_TKO	257428499	174929585	68	NA
	BISSEQ_to2i	191890338	100546767	52	NA
	BISSEQ_toSerum	215409126	146458573	68	NA

KD = knockdown; CTRL = negative control siRNA; to2i = adapted to 2i (after serum); toSerum = adapted to serum (after 2i); Over = overexpression of NRF1

3.2.3 Addendum

To assess NRF1-dependence of transcripts arising at TKO-specific NRF1 sites, we performed CRISPR/Cas9-based targeted mutagenesis of the *Nrf1* gene in ES TKO cells. We obtained 20% cutting efficiency for the best guide RNA, when measured on the pool of cells, and genotyped 86 cell clones. However from the 21 detected unique insertions/ deletions most were in-frame (3 to 36 bp). Thus in spite of isolating 13 lines mutated on both alleles, all of these clones harboured at least one apparently functional copy of the gene. This led to substantial residual NRF1 levels in TKO clones homozygous for CRISPR-induced mutations in the *Nrf1* gene (Fig. 3-9). The exclusive recovery of mutants that still express a functional protein implies a strong negative selection against knockouts and in turn argues that *Nrf1* is an essential gene in ES cells. Therefore we proceeded with partial siRNA-mediated knockdown of *Nrf1* (see Chapter 3.2.2).

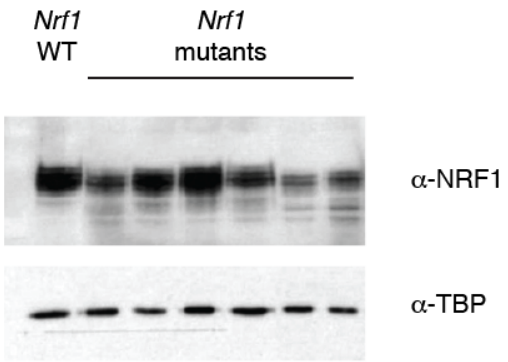


Figure 3-9. Residual NRF1 levels in six ES TKO cell lines homozygous for CRISPR-induced *Nrf1* mutations.
Western Blot was performed on nuclear extracts, using TBP as loading control. NRF1 levels in cells containing the wildtype *Nrf1* gene are shown as comparison.

3.3 Binding site restriction by DNA methylation in differentiated cells

3.3.1 Abstract

DNA methylation is not required for large-scale restriction of TF binding or cell survival in mouse ES cells. In contrast, it is essential in differentiated cells. This raises the possibility that more TF binding events are affected in these cells upon removal of DNA methylation. To test this assumption, we generated a differentiated cell type without DNA methylation by inducible expression of the neuronal TF Neurogenin2 (NGN2) in *Dnmt* TKO ES cells. The resulting TKO neurons survive for ten days in culture and closely resemble their WT counterparts in terms of morphology and gene expression, with the exception of upregulated germline-specific genes. Chromatin accessibility measured by ATAC-seq is remarkably similar between neurons with and without DNA methylation, although a subset of sites is only accessible in the TKO. These sites are enriched for different CpG-containing TF motifs, including those of HNF6 and NRF1, a methylation-sensitive TF we already described in ES cells. In contrast to the limited changes observed for chromatin accessibility and gene expression, specific long terminal repeat (LTR) retrotransposons, the intracisternal A particles (IAPs), are derepressed by several orders of magnitude in TKO neurons. Sequence comparison of activated and silent elements of the same IAP subtype reveals that the presence and strength of the cAMP-responsive element (CRE motif) within the LTR is highly predictive of the level of activation in TKO neurons. We suggest that DNA methylation is required to block binding of TFs at the CRE motif in neurons and thus prevents the potentially lethal derepression of transposable elements. Importantly, the same transposon family is activated in methylation-deficient fibroblasts and postnatal mouse cortex. This raises the possibility that the mechanisms we describe here are general responses to loss of DNA methylation in differentiated cells both in culture and *in vivo*.

3.3.2 Introduction

Only a small subset of sites (~ 3%) show differential chromatin accessibility, an indicator of TF binding, in mouse ES cells upon removal of DNA methylation (see Chapter 3.2.2). This finding is not entirely surprising given the limited gene expression changes observed in stem cells lacking DNA methylation and the fact that these *Dnmt* TKO ES cells have no proliferation or morphology defects (Domcke et al., 2015; Tsumura et al., 2006). Indeed since mouse ES cells are isolated from preimplantation blastocysts, whose genomes are globally demethylated (Auclair and Weber, 2012), mechanisms need to be in place at this developmental stage to ensure cell survival in spite of low DNA methylation levels. That said, TKO ES cells are unable to differentiate (Jackson et al., 2004; Tsumura et al., 2006) and are the only mammalian cell type known to survive without DNA methylation. Deletion of *Dnmt1* in differentiated cells or even human ES cells, which are thought to represent a later stage of development than their murine counterparts (Nichols and Smith, 2009), leads to rapid cell death (Chen et al., 2007; Liao et al., 2015) (Fig. 3-10).

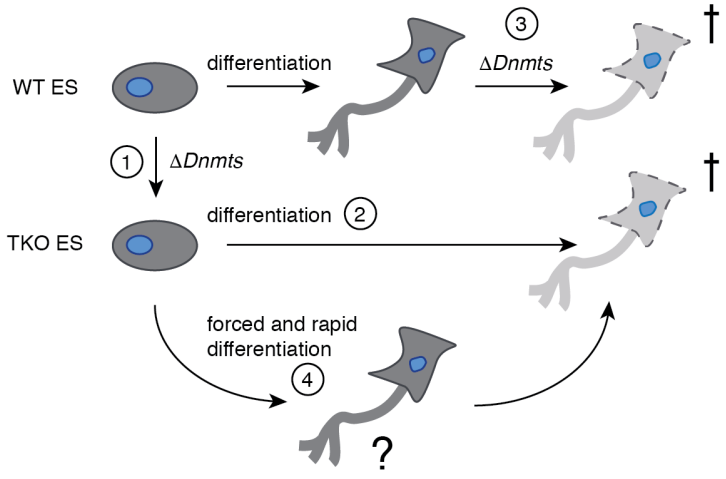


Figure 3-10. DNA methylation is essential in differentiated cells.
Mouse ES cells are able to survive complete loss of DNA methylation by deletion of the three *Dnmts* (1). However, these TKO ES cells are unable to differentiate (2). Deletion of *Dnmts* in differentiated cells leads to cell death (3). To nonetheless study the impact of DNA methylation on TF binding in differentiated cells, we attempted to generate a committed cell state without DNA methylation by forced and fast neuronal differentiation of TKO ES cells (4). This would enable us to compare chromatin accessibility and thus TF binding with WT cells of the same developmental stage, before the ensuing cell death.

This cell death has been attributed in turn to misregulation of critical genes (Jackson-Grusby et al., 2001) or activation of repeats (Walsh et al., 1998; Yoder et al., 1997) and was linked to the induction of DNA damage (Shaknovich et al., 2011) and mitotic catastrophe (Chen et al., 2007). The disparate effect of DNA methylation loss in stem cells and differentiated cells raises key questions regarding the role of DNA methylation in influencing TF binding and development: First, are more TF binding events affected by DNA methylation in differentiated cells compared to ES cells? Indeed a different TF repertoire is expressed and gene regulatory regions vary in methylation states in a tissue-specific manner (Ziller et al., 2013). Second, is differential TF binding responsible for the cellular lethality observed in methylation mutants or does methylation loss impair cell survival independent of TF methylation sensitivity? It has been inherently challenging to study these critical questions, for the exact reason that DNMTs are essential in differentiated cells. Here we attempt to address these questions by generating a differentiated cell state without DNA methylation and comparing expression and genome-wide TF binding to the isogenic WT counterpart (Fig. 3-10).

3.3.3 Results

3.3.3.1 Differentiated cells lacking DNA methylation

To study the role of DNA methylation in regulating TF binding in differentiated cells, we first sought to generate a differentiated cell state without DNA methylation. Although TKO ES cells are reportedly unable to differentiate (Jackson et al., 2004; Tsumura et al., 2006), we nevertheless tested the ability of our TKO ES cells to form glutamatergic pyramidal neurons. In addition to yielding reproducible homogeneous populations, neurons have the advantage that they are morphologically distinct and we have previously characterised their transcriptome and epigenome in great detail (Mohn et al., 2008; Stadler et al., 2011; Tippmann et al., 2012). First we tried an established protocol based on retinoic acid treatment of non-adherent cell aggregates that forms first neuronal progenitors (NPs) and then terminal neurons (TNs) in the course of three weeks (Bibel et al., 2007). TKO ES cells can indeed produce embryoid bodies and NPs with this protocol, but they die shortly after dissociation of the cell aggregates around day nine and before reaching the TN stage (Fig. 3-11a). We reasoned that a protocol that generates neurons faster might allow us to test if this lethality is not only dependent on the neuronal cell fate but also reflects time in culture following loss of pluripotency. Thus we tested if TKO ES cells might be more amenable to the rapid and efficient neuronal differentiation induced by ectopic expression of the neural TF Neurogenin2 (NGN2), which generates functional glutamatergic neurons already five days after induction (Thoma et al., 2012; Zhang et al., 2013b). Using CRISPR-Cas9 gene editing we made several *Dnmt* TKO clones in an ES cell line containing a stable insertion of *Ngn2* under the control of pTRE-tight, thus allowing dox-inducible NGN2 expression (Fig. 3-11b). The TKO clone with the highest differentiation potential indeed adopted neuronal morphology, formed axonal networks similar to the WT and survived around nine to ten days before cell death ('TKO neurons') (Fig. 3-11c). Absence of DNA methylation in this clone was confirmed by

mass spectrometry (Fig. 3-11d). Of note, absolute cell survival time is similar for the two tested differentiation methods. However, only the faster NGN2-expression protocol reaches the neuron stage already within this time span. This implies that the time spent in culture rather than the actual cell fate might be relevant for cell lethality in differentiated cells lacking DNA methylation.

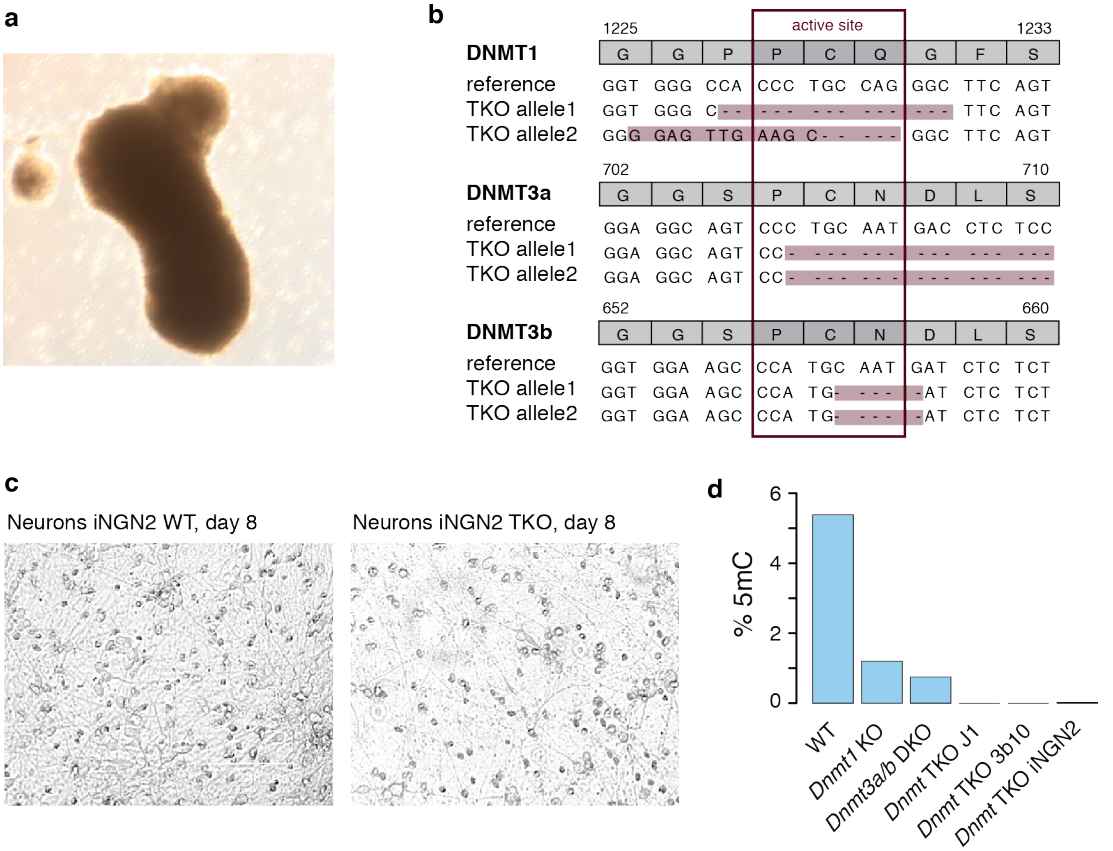


Figure 3-11. Neuronal morphology, genotype and methylation levels of TKO cells.
a) Non-adherent differentiated TKO cell aggregates generated by LIF withdrawal for four days followed by four days of retinoic acid treatment (”neuronal progenitors”, NPs) according to Bibel et al., 2007. TKO cells do not show an obvious morphological phenotype compared to WT at this stage, although aggregates tend to be somewhat smaller. TKO cells die shortly after aggregate dissociation and plating and do not form terminal neurons using this differentiation protocol. **b)** Frameshift deletions (red) introduced by CRISPR/Cas9 genome editing at the active PCQ/N loops of the three *Dnmt* genes in the ES TKO clone with inducible Ngn2 expression that showed the highest differentiation potential. For *Dnmt3a* and *Dnmt3b* only one mutated allele and no WT allele could be detected when sequencing the genotyping PCR product, even using high-coverage (> 10,000x) Illumina sequencing. **c)** Morphology of WT and TKO neurons. Both WT and TKO ES cells were differentiated by dox-inducible NGN2 expression and are shown on day 8 after induction. **d)** Comparison of levels of methylated cytosine (5mC) between WT ES cells and different *Dnmt* mutants as measured by mass spectrometry. *Dnmt* mutants were generated by traditional mouse genetics (*Dnmt1* KO and *Dnmt3a/b* DKO values from Le et al., 2011, *Dnmt* TKO J1) or CRISPR-Cas9 gene editing (*Dnmt* TKO 3b10 used in Chapter 3.2.2, *Dnmt* TKO iNgn2 clone used for neuronal differentiation with inducible NGN2 expression).

3.3.3.2 Limited changes in gene expression in TKO neurons

Apart from the neuronal morphology, we asked whether the differentiated TKOs also resemble neurons in terms of gene expression. We collected RNA from several differentiation batches and performed total RNA-seq in four replicates. The genes that are most upregulated in TKO neurons compared to the ES stage are almost exclusively involved in neuronal functions (Fig. 3-12a), confirming that the methylation-deficient differentiated cells we generated have indeed key neuronal features.

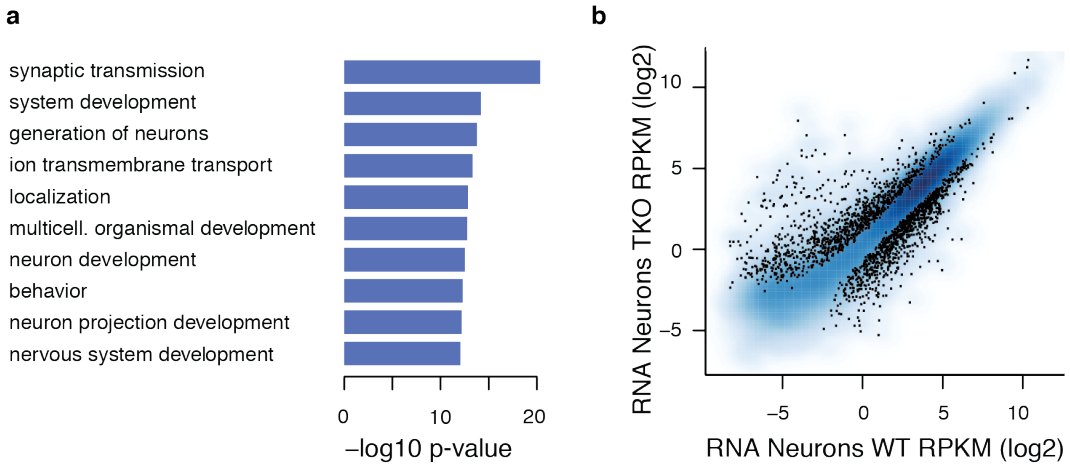


Figure 3-12. Gene expression changes in TKO neurons compared to ES stage and WT neurons.
a) Gene ontology (GO) categories overrepresented among 200 most differentially expressed genes between TKO ES and TKO neurons generated by inducible expression of NGN2 (day 8-9), as measured by RNA-seq. The ten GO categories with the lowest p-value are shown (hypergeometric test). **b)** Exonic gene expression levels (RPKM) in isogenic WT and TKO neurons generated by inducible NGN2 expression. Black: differentially expressed genes (at least 2-fold change, adjusted p-value < 1e-5).

In order to identify methylation-sensitive TF binding events by comparing chromatin accessibility between WT and TKO neurons, it is crucial that gene expression changes are limited and highly reproducible. Drastic gene deregulation would introduce many secondary effects and thus make it inherently difficult to identify those accessible sites that derive directly from methylation-sensitive binding events. We therefore compared gene expression in WT and TKO neurons in more detail. In line with the repressive effects of DNA methylation, more genes are upregulated than downregulated

in TKO neurons compared to WT (Fig. 3-12b). Gene ontology term analysis revealed strong enrichment of gamete-specific genes within the upregulated set (Fig. 3-13a,b). The CpG-rich promoters of these genes are unique in that they are normally highly methylated in somatic cells (Weber et al., 2007) and activation of these germline-specific genes has been previously reported in TKO ES cells (Karimi et al., 2011). Indeed we observe activation of the same gene class already in our ES and NP TKO cell lines (Fig. 3-13a,b).

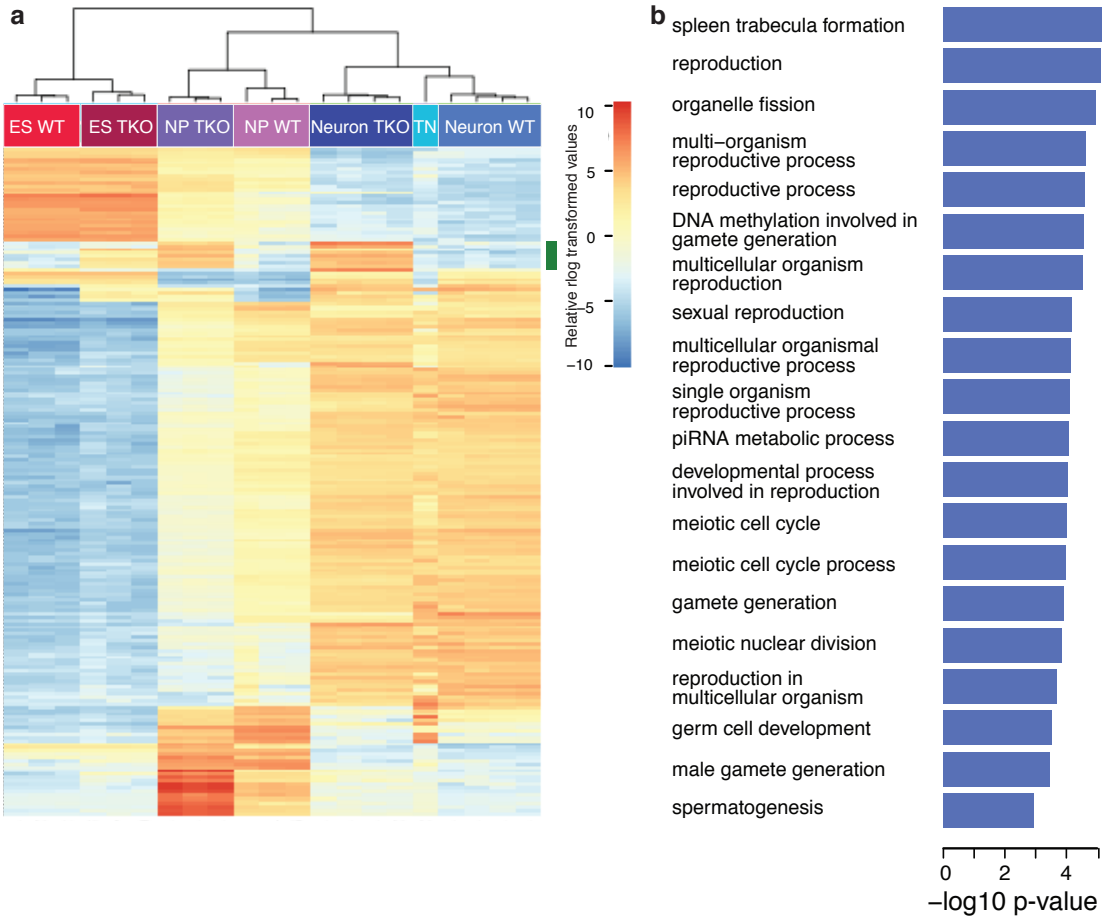


Figure 3-13. Germline genes deregulated in TKO neurons are already upregulated before differentiation.
a) Hierarchical clustering of 200 most differentially expressed genes across neuronal differentiation stages of WT and TKO cells. WT and TKO ES were differentiated either with retinoic acid treatment into neuronal progenitors (NP) and terminal neurons (TN, only WT) according to Bibel et al., 2007 or directly into neurons by inducible expression of NGN2. Regularised-logarithm (rlog) transformation was applied to normalised read counts for all samples. The gene cluster that shows most differential expression between the WT and TKO neurons generated by inducible expression of NGN2 (green bar) shows upregulation also in NP and ES TKO compared to matching WT stages. **b)** Gene ontology (GO) categories overrepresented in the gene cluster that is upregulated in TKO cells across various neuronal differentiation stages (annotated by green bar in (a)). The twenty GO categories with the lowest p-value are shown (hypergeometric test), revealing many germline-specific genes.

Strikingly, we did not detect an enrichment of genes involved in apoptosis or cell stress among the upregulated group that could be linked to a global genic response and the ensuing cell death. Overall, gene expression changes between TKO and WT neurons are remarkably limited, with only 1.7 times as many genes upregulated compared to the ES stage. WT and TKO neurons lie in close proximity in a principal component analysis of genic RNA-seq (Fig. 3-14). In fact, WT inducible-NGN2 neurons resemble the TKO neurons more closely than WT neurons generated by the other differentiation protocol (Fig. 3-14), although both methods yield the same subtype of glutamatergic neurons. Taken together, WT and TKO neurons are very similar in terms of morphology and gene expression, implying that these cells can indeed serve as a model system to identify methylation-sensitive TF binding in differentiated cells.

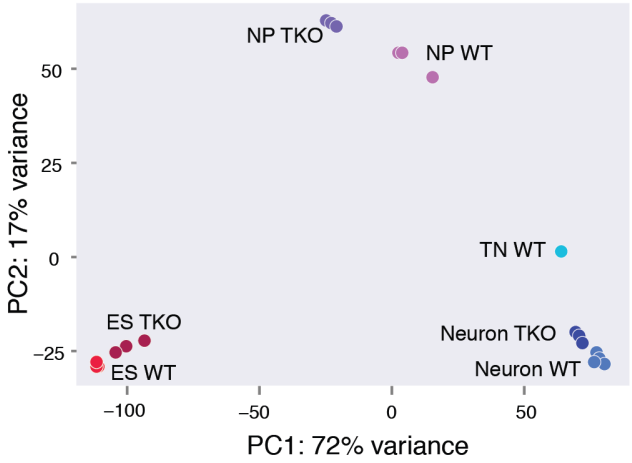


Figure 3-14. TKO neurons resemble WT neurons in gene expression.
Principal component analysis of total gene expression across the differentiation stages described in Figure 3-13a (colour code), showing three or four biological replicates (normalised exonic read counts).

3.3.3.3 A subset of sites are only accessible in TKO neurons

DNase-seq enables measurement of differential TF occupancy at high resolution (Domcke et al., 2015; Neph et al., 2012). However, it requires millions of cells per condition, which can be challenging to obtain in differentiation protocols. ATAC-seq has the advantage that it can be

performed on a few thousand cells and the workflow requires substantially less time (Buenrostro et al., 2013), but its sensitivity has not been compared to DNase-seq in detail. Therefore, we first determined whether these methods can be used interchangeably to identify differential TF binding at high sensitivity and specificity.

Accessibility measured by ATAC-seq in WT ES cells is indeed in good agreement with DNase-seq (Pearson correlation = 0.83), with high replicate reproducibility (Pearson correlation = 0.96) (Fig. 3-15a,b).

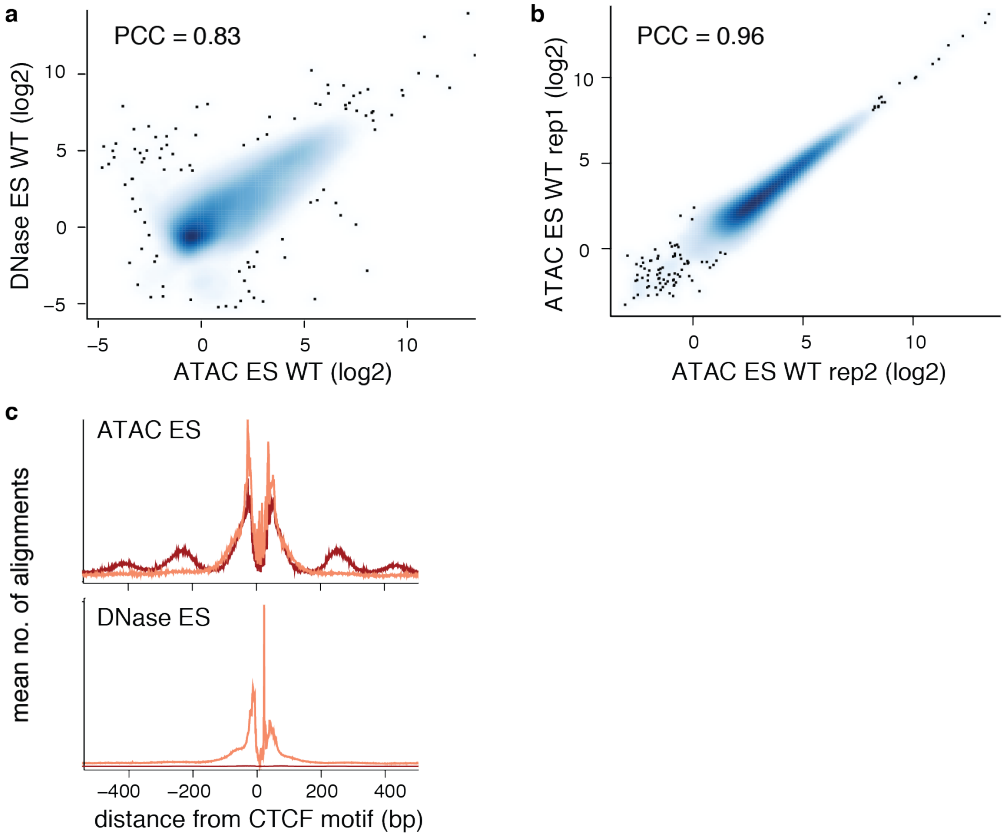


Figure 3-15. Comparison of DNase-seq and ATAC-seq in WT ES cells.
a) Comparison of ATAC-seq and DNase-seq signal in WT ES (mean of two replicates normalised to library and region size) at all accessible regions identified with at least one of the two methods (n = 267,538). PCC = Pearson correlation coefficient. **b)** ATAC-seq signal (normalised to library and region size) for two biological replicates of WT ES cells at all highly accessible regions identified with ATAC-seq (n = 65,564). **c)** Metaplot of ATAC-seq and DNase-seq signal over bound CTCF motifs (called based on CTCF ChIP-seq in WT ES, see Chapter 3.1.3; n = 30,422) for long (> 100 bp, dark) and short DNA fragments (<= 100 bp, light). In DNase-seq longer fragments are experimentally removed prior to sequencing whereas they are still present in ATAC-seq and inform on nucleosome positioning.

While ATAC-seq and DNase-seq libraries were sequenced to a similar depth (~ 200 mio. reads) it should be noted that more reads fall into the genomic background for ATAC-seq compared to the DNase-seq protocol, since the latter only sequences the fraction of sequences that correspond to TF binding sites in terms of size. Nonetheless, both methods reveal clear footprints at TF binding sites (Fig. 3-15c). Sequence bias of the enzymes is likely responsible for the different footprint shapes.

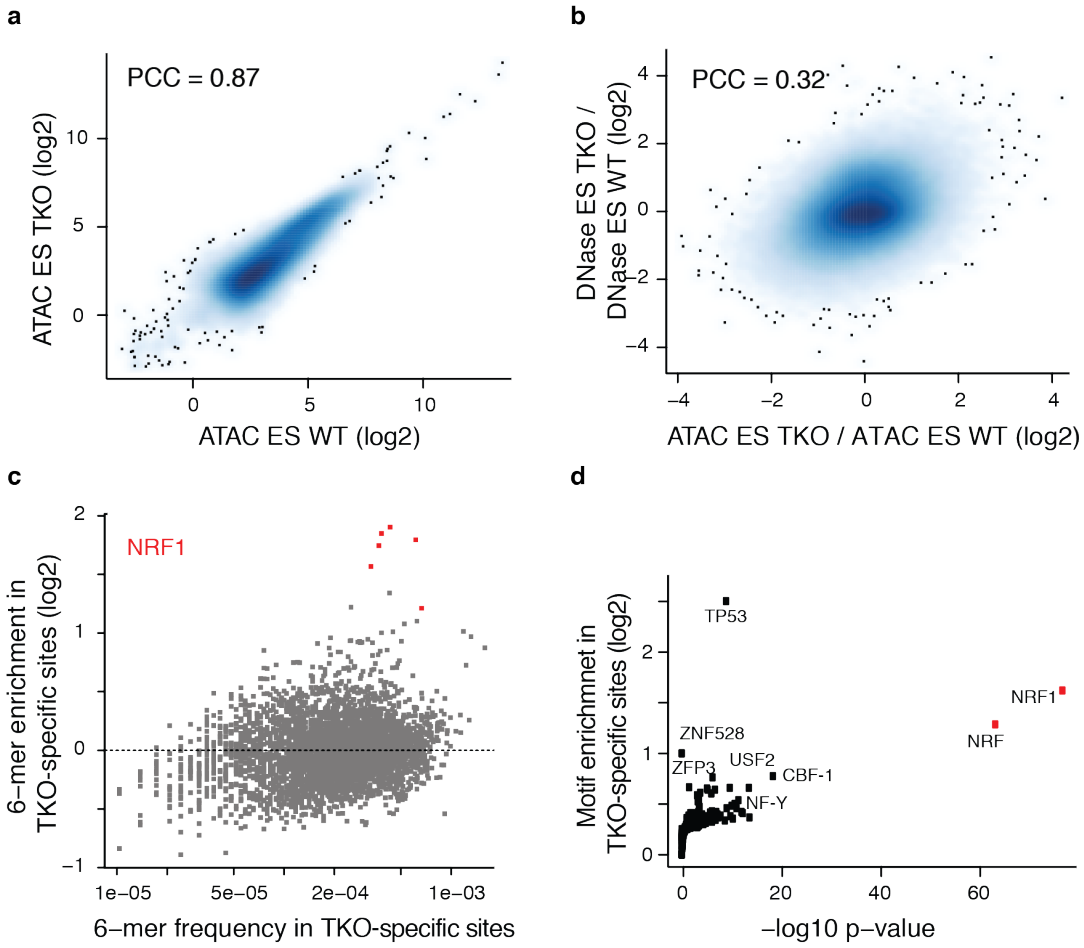


Figure 3-16. Comparison of ATAC-seq in TKO and WT ES cells as indicator of differential TF binding.
a) Comparison of ATAC-seq signal (normalised to library and region size) for TKO and WT at all highly accessible regions identified with ATAC-seq in ES cells (n = 65,564). PCC = Pearson correlation coefficient. **b)** Comparison of changes in ATAC-seq and DNase-seq between TKO and WT at all accessible regions identified in at least one method and cell line (n = 267,538). **c)** Occurrence of all possible hexamers in the 500 ATAC-seq sites that show most accessibility gain in TKO ES over shared sites (abs(log2 fold-change TKO/WT ES) < 0.3) (see Methods for details). Hexamers representing the NRF1 motif are coloured in red. **d)** Known TF motifs enriched in TKO-specific ATAC-seq sites in ES cells compared to shared sites. Motifs representing the NRF1 motif are coloured in red. P-values are from a binomial test. See Methods for details.

As with DNase-seq, changes between WT and TKO ES are limited and occur to a similar extent in both cell lines (Fig. 3-16a). This differential accessibility correlates for the two methods (Fig. 3-16b). Searching for hexamer sequences or TF motifs enriched in TKO-specific ATAC-seq sites compared to shared sites yielded NRF1 as the top candidate, as we already observed with DNase-seq in Chapter 3.2.2 (Fig. 3-16c,d). Taken together, these results imply that differential TF binding can be detected with either method and the identification of factors is not majorly distorted by the sequence or other preferences of the respective enzymes.

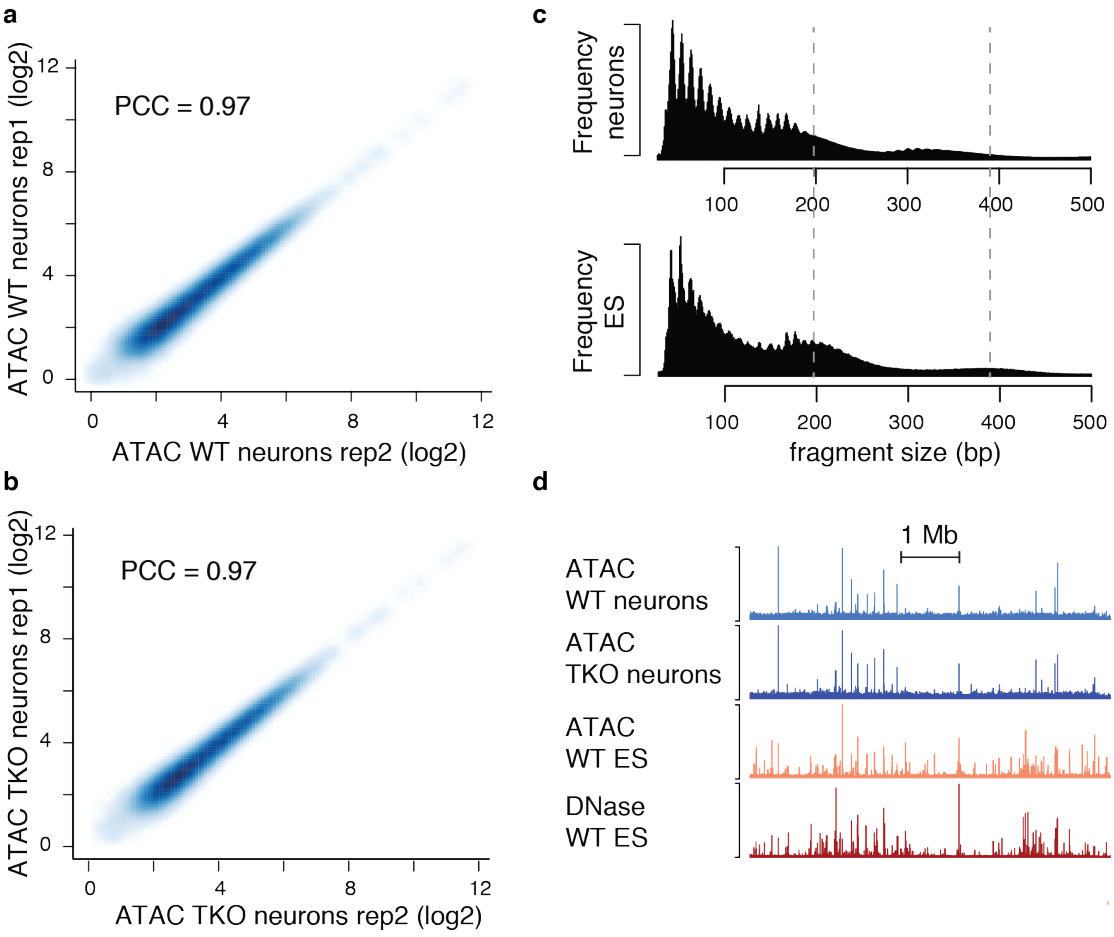


Figure 3-17. Profiling chromatin accessibility by ATAC-seq in neurons.
a, b) ATAC-seq signal for two biological replicates of WT (a) or TKO (b) neurons at all highly accessible regions in neurons (n = 26,972). PCC = Pearson correlation coefficient. **c)** Fragment size distribution of all sequenced ATAC-seq libraries in neurons (top) or ES cells (bottom) showing differential nucleosomal spacing. Dotted grey lines mark summit positions of the distribution in ES cells. Ten base pair periodicity is likely due to background cuts of the transposase in DNA wrapped around nucleosomes. **d)** Chromatin accessibility measured by ATAC-seq or DNase-seq in WT and TKO neurons and ES cells at a representative genomic region (chr19: 27,517,485-33,736,294).

Accordingly, we profiled chromatin accessibility in WT and TKO neurons by ATAC-seq and called accessible sites (Fig. 3-17a,b). Interestingly, both WT and TKO neurons showed a shorter nucleosomal spacing, as has been previously observed in neurons (Pearson et al., 1984), and overall fewer highly accessible sites than ES cells (26972 vs. 70609; Fig. 3-17c,d). These differences could in part be due to the fact that neurons are postmitotic, whereas ES cells represent a heterogeneous mixture of cells in different stages of the cell cycle.

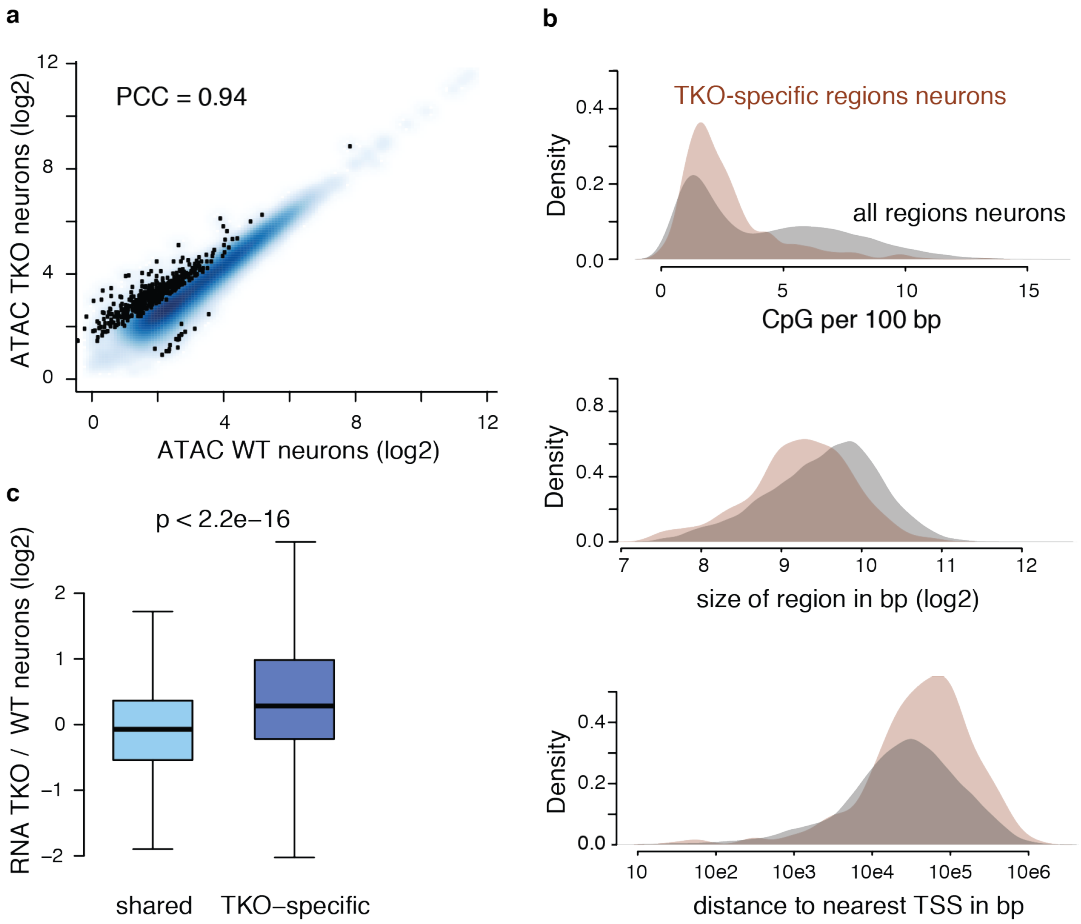


Figure 3-18. Characterisation of TKO-specific ATAC-seq sites in neurons.
a) ATAC-seq signal in WT and TKO neurons at all 26,972 sites reproducibly accessible in at least one of the two cell lines. The mean of two replicates was normalised to peak size and across samples with DESeq2. Sites that are differentially accessible (at least twofold change, adjusted p-value < 0.05) are marked in black (436 TKO-specific, 18 WT-specific). PCC = Pearson correlation coefficient. **b)** Comparison of CpG density (top), region size (middle) and distance to nearest TSS (bottom) for TKO-specific (red, n = 436) and all (grey, n = 26,972) ATAC-seq sites identified in neurons. **c)** Expression change between TKO and WT neurons for genes whose TSS is closest to shared or TKO-specific ATAC-seq peaks. P-value results from a two-sided Mann-Whitney test.

Similar to the observation in ES cells, changes between TKO and WT neurons were limited but highly reproducible (Fig. 3-18a). As observed in ES, TKO-specific sites tend to have lower CpG density, be further away from TSS and less broad than shared sites (Fig. 3-18b). This is in line with most changes occurring at distal regulatory regions or non-functional sites which are methylated and un-occupied in WT conditions. Interestingly however, changes are more unidirectional than in ES cells, with the vast majority of accessible sites being gained in TKO neurons. This is in line with a repressive effect of DNA methylation and in good agreement with the tendency towards upregulation in differential gene expression (Fig. 3-12b). The genes closest to differentially accessible neuronal sites indeed show a significant increase in expression in TKO neurons compared to genes next to constitutively open regions (Fig. 3-18c).

3.3.3.4 HNF6 is a candidate methylation-sensitive TF

Next we sought to identify the TFs responsible for the observed gain in accessibility in TKO neurons. To this end, we asked which TF motifs are enriched in the TKO-specific ATAC-seq peaks compared to the shared sites. The top three most significantly enriched known TF motifs identified by homer2 are NRF1, NFY and HNF6 ($p\text{-value} < 0.01$), with several further motifs showing slighter yet still significant enrichment (Fig. 3-19a). Unbiased enrichment analysis of all possible hexamers revealed that the top enriched hexamers can all be assigned to these same three motifs (Fig. 3-19b). Importantly, the factors reported to bind to these motifs are expressed at similar levels in WT and TKO neurons (Fig. 3-19c). One of the top candidates, NRF1, accounted for the majority of TKO-specific sites in our study in ES cells and was found to indeed bind in a methylation-sensitive manner there, as validated e.g. by ChIP-seq (see Chapter 3.2.2). In contrast to the situation in ES cells, differential NRF1 binding is able to explain only a small percentage of TKO-specific sites in neurons. The absolute enrichments are low for all identified motifs, implying that many TFs contribute slightly to the differences in chromatin accessibility. That said, several of the significantly enriched

motifs appearing in the neurons besides NRF1 also belong to the most enriched motifs in ES cells, such as ARNT/ HIF1A, ETS1 or CRE-like motifs (Fig. 3-19a and Domcke et al., 2015).

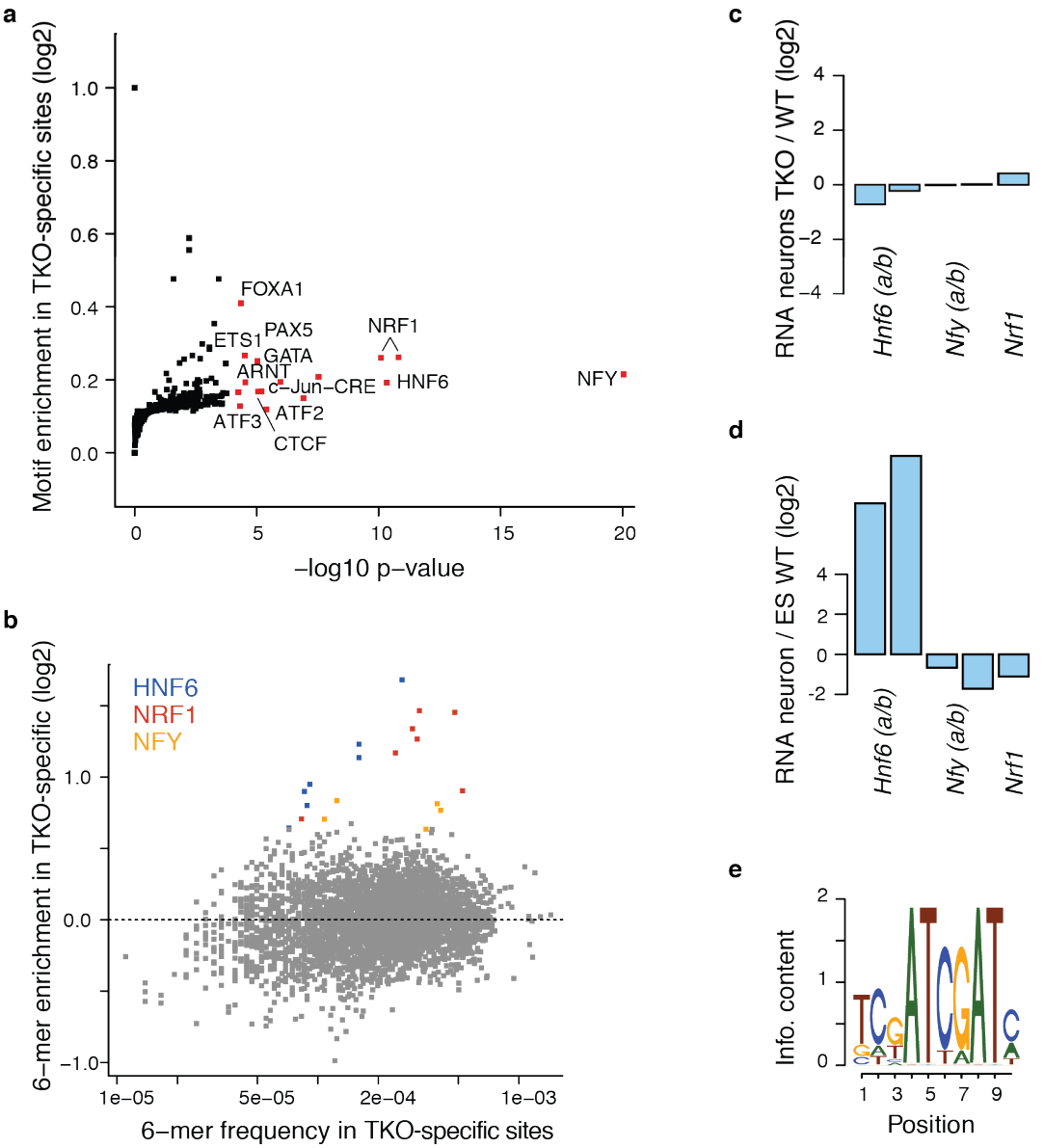


Figure 3-19. Candidate methylation-sensitive TFs in neurons.
a) Known TF motifs enriched in TKO-specific ATAC-seq sites in neurons compared to shared sites. Motifs that are significantly enriched with a p-value of less than 0.01 are marked in red and labelled. P-values are from a binomial test. See Methods for details. **b)** Occurrence of all possible hexamers in TKO-specific neuronal ATAC-seq sites compared to shared neuronal ATAC-seq sites (see Methods for details). The top enriched hexamers could be manually assigned to three TF motifs (blue, red, yellow). **c)** Change in gene expression levels (RPKM) for different isoforms of candidate methylation-sensitive TFs identified in (a) and (b) between TKO and WT neurons. **d)** Change in gene expression levels (RPKM) for different isoforms of candidate methylation-sensitive TFs identified in (a) and (b) during neuronal differentiation. **e)** Position-weight matrix for the HNF6 motif variant enriched in TKO-specific ATAC-seq sites in neurons.

HNF6 (also known as ONECUT1), one of the top candidates in neurons, is an especially interesting case. This factor is not expressed in ES cells – and accordingly was not among the methylation-sensitive factors identified there – but is strongly upregulated in the neuronal lineage (Fig. 3-19d). It is expressed upon differentiation in several tissues and is considered a key regulator for gene expression in liver, pancreas and the nervous system (Audouard et al., 2013). Interestingly, the motif enriched in TKO-specific neuron ATAC-seq sites is only the second most commonly bound motif for this factor in ChIP-seq experiments: This particular motif variant contains a prominent CpG (Fig. 3-19e), in contrast to the canonical motif found at most HNF6 binding sites (Ballester et al., 2014; Wang et al., 2014). Likely with HNF6 we have thus identified another methylation-sensitive TF, although this needs to be further validated by ChIP-seq in WT and TKO neurons.

3.3.3.5 Specific retrotransposons are strongly activated in TKO neurons

The chromatin accessibility differences between WT and TKO neurons suggest that some TF binding events are restricted by DNA methylation. Nevertheless overall changes in chromatin accessibility are rather limited, which is in line with the modest gene expression changes, but at odds with the ensuing cell death. Of note, while gene expression is remarkably similar, unique mapping efficiencies for RNA-seq samples in TKO neurons were drastically lower than in WT (0.56 vs. 0.85), since many reads aligned to multiple regions in the genome. Given this observation and the known importance of DNA methylation for silencing repeats (Walsh et al., 1998), we analysed non-genic transcripts from repetitive regions of the genome. To this end, we re-sequenced the RNA-seq samples using 100 bp paired-end reads, which increased unique mapping efficiency, especially for the TKO sample (0.68 vs. 0.88). We observed a striking deregulation of repetitive elements in TKO neurons (Fig. 3-20a), especially of the intracisternal A particles (IAPs) (Fig. 3-20b,c). An estimated 15% of all RNA transcripts within TKO neurons stems from this repeat type. This is in line with a landmark study reporting strong IAP derepression in DNMT1-deficient embryos (Walsh et al., 1998). A

recent paper argued that the strong activation in *Dnmt1* -/- embryos is the result of UHRF1 binding hemimethylated DNA and thus impeding SETDB1 in setting of H3K9me3 (Sharif et al., 2016). Of note, this is unlikely to be the case here since TKO neurons are generated from stable TKO ES cells, which do not contain any hemimethylated DNA.

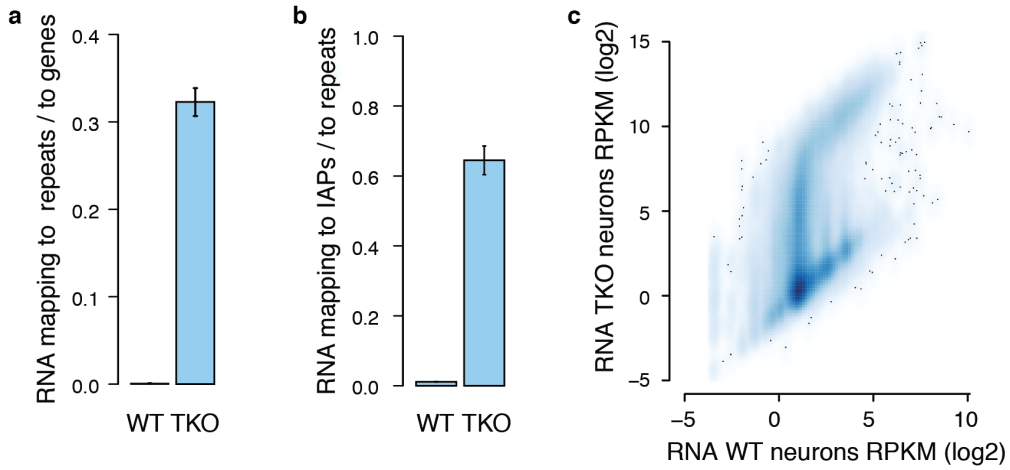


Figure 3-20. Strong activation of IAP elements in TKO neurons.
a, b) Ratio of 100 bp paired-end RNA-seq reads mapping to repeats over genes (a) or IAP elements over all repetitive elements as annotated by RepeatMasker (b). Reads mapping at multiple locations were randomly assigned to one position. Error bars represent standard deviation of three biological replicates. **c)** Comparison of 100 bp paired-end RNA-seq reads uniquely mapping to IAP elements in WT and TKO neurons. The mean read count of three biological replicates was normalised to element and library size.

IAPs belong to the evolutionarily youngest family of ERV-K endogenous retroviruses and are the most active repeat type in rodents (Maksakova et al., 2006). Transcription starts within the long terminal repeats (LTRs), which flank internal viral genes or occur as solo LTRs. In addition to being transcribed themselves, these elements have also been reported to be able to drive strong overexpression of downstream host genes by forming chimeric transcripts (Karimi et al., 2011), as we also observe here (Fig. 3-21a,b).

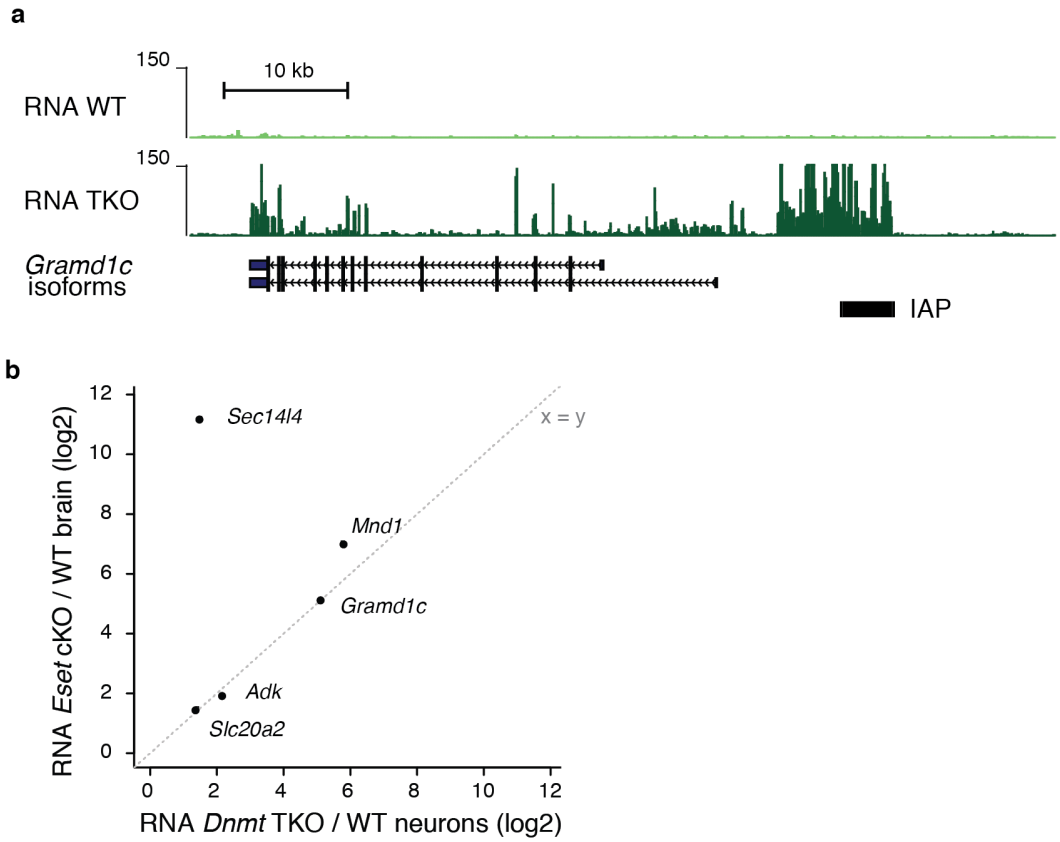


Figure 3-21. Gene expression changes next to activated IAPLTRs in TKO neurons.
a) Example of a gene that is activated in TKO neurons and lies immediately downstream of a transcribed IAP element. Only uniquely mapping 100 bp paired-end RNA-seq reads are shown. **b)** Comparison of expression changes in TKO neurons and ESET KO brain samples versus matching WT at genes that were previously found to form chimeric transcripts with IAPs (Tan et al., 2012). Expression in neurons was measured with RNA-seq and in brain tissue by microarray (data from Tan et al., 2012).

3.3.3.6 Comparison of repeat activation with other chromatin mutants

Silencing of ERV-K elements has been attributed to both DNA methylation and H3K9me3 pathways (Rowe and Trono, 2011). We wondered how the derepression we observe quantitatively compares to previously published datasets in mutants for these pathways. To this end, we compared total RNA-seq reads from repetitive elements across different samples, allowing for multiple mappers and collapsing reads to family level (see Methods for details) (Fig. 3-22).

Figure 3-22. Comparison of repeat activation across mutants in the DNA methylation or H3K9me3 pathway.
Hierarchical clustering of expression change in all repeat families for different chromatin mutants compared to the matching WT. *Dnmt* TKO cells at various differentiation stages (underlaid with light grey boxes) were compared to available published data sets in other cell lines or tissues deficient in the DNA methylation (purple) or H3K9me3 (green) pathways (see Methods for details and sources of published data sets). The ERV-K family (arrow) contains the IAP elements. MEF = mouse embryonic fibroblast, NPC = neural precursor cell.

We do not observe strong upregulation either in methylation-deficient stem cells or mutants of the H3K9me3 pathway in differentiated cells. In contrast, methylation mutants in differentiated cells as well as *Setdb1 cDnmt1* DKO ES cells show pronounced derepression mainly of the ERV-K family, reminiscent of what we observe in the TKO neurons. The extent of derepression is however clearly strongest in TKO neurons. This is likely due to the fact that these are the only differentiated cells that have completely lost DNA methylation, whereas the other cell lines and tissues all still contain functional DNMTs. In TKO neurons, transcripts can be detected especially at the best-conserved IAPLTR1 and 1a types within the ERV-K family (Fig. 3-23a). Importantly, the same elements are also upregulated in other

methylation mutants in differentiated cells, including a recently published conditional *Uhrf1* KO in postnatal mouse cortex (Ramesh et al., 2016) (Fig. 3-23b). This implies that the repeat activation we observe is not a clonal or tissue culture artefact but also occurs *in vivo* upon methylation loss. Therefore, our system provides an interesting cellular model to study the effects of DNA methylation loss on repeat activation and an opportunity to understand the mechanisms underlying specific activation of a subset of repetitive elements.

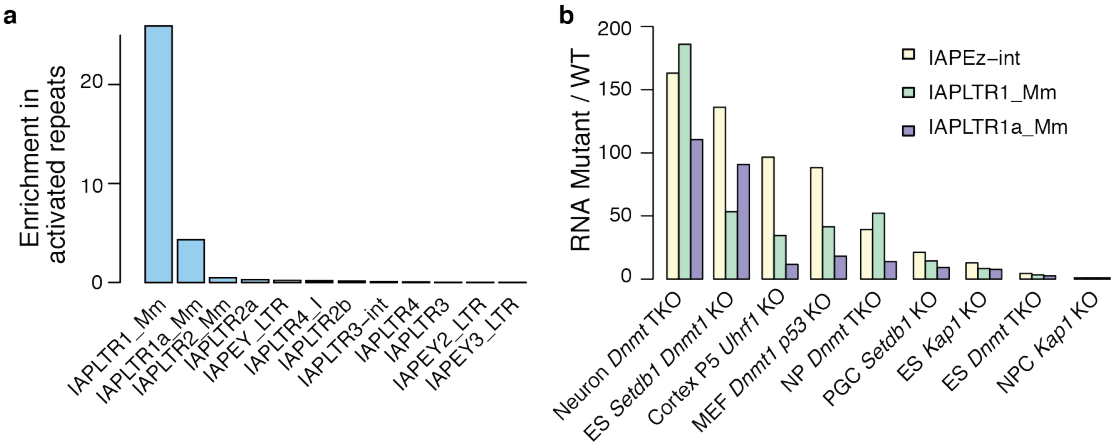


Figure 3-23. Activation of different IAP subtypes across mutants in the DNA methylation or H3K9me3 pathway.
a) Enrichment of different IAPLTR subtypes in activated (log2 fold-change > 1 in TKO/WT neurons) over silent elements (abs(log2 fold-change in TKO/WT neurons) < 0.3). **b)** Expression change of the IAPLTR subtypes most enriched in (a) (IAPLTR1 and IAPLTR1) as well as their internal viral genes (IAPEz-int) compared to the matching control for the different chromatin mutants analysed in Figure 3-22.

The observed derepression in different chromatin mutants is in line with the current thinking of how repeat repression is accomplished in ES and differentiated cells. In stem cells, silencing is thought to be mediated mainly by the H3K9me3-machinery through KRAB zinc-finger protein-mediated recruitment of KAP1 and SETDB1 (Karimi et al., 2011; Rowe et al., 2010). In differentiated cells on the other hand, DNA methylation rather than H3K9me3 is deemed necessary for silencing (Leung and Lorincz, 2012). In line with H3K9me3 losing importance as a silencing mechanism, the proteins involved in setting this mark are downregulated upon neuronal differentiation

(Fig. 3-24a). Accordingly, H3K9me3 is lost at repetitive elements including IAPs, as is observed when comparing H3K9me3 enrichment during ES to NP differentiation (Fig. 3-24b) (Bulut-Karslioglu et al., 2014). Of note, this loss occurs to a similar extent at elements that are activated and silent upon methylation removal (Fig. 3-24b,c). Differential H3K9me3 levels are therefore unlikely to be the explanation for dissimilar activation levels of repeat elements, although this still needs to be experimentally tested in our model system.

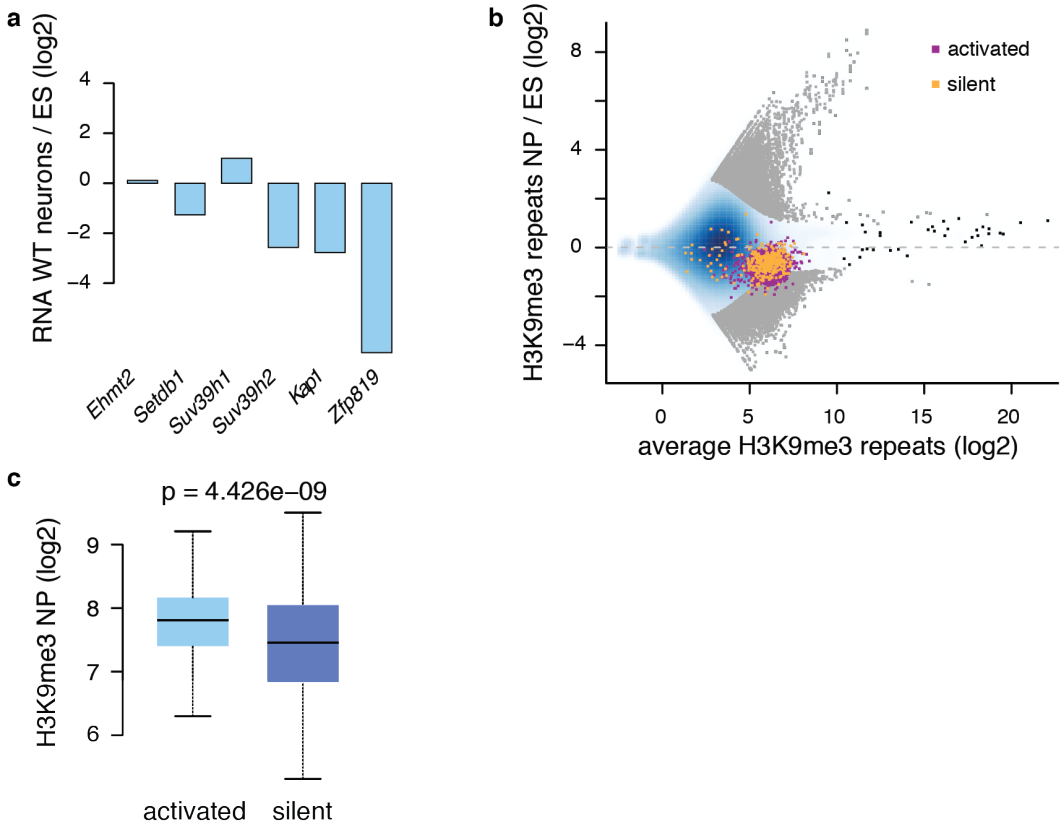


Figure 3-24. The H3K9me3 mark is reduced at IAPLTRs during differentiation.
a) Gene expression changes during neuronal differentiation for enzymes involved in setting methylation of histone H3 at lysine 9. EHMT2 mono- and dimethylates H3K9, SETDB1 and SUV39H1/2 trimethylate, KAP1 (aka TRIM28) recruits SETDB1 and ZFP281 is an exemplary KRAB zinc-finger protein that interacts with KAP1 and is involved in repressing IAP elements in stem cells (Tan et al., 2013). The mean of three biological RNA-seq replicates is shown. **b)** MA plot showing changes in H3K9me3 ChIP-seq signal between neuronal progenitor cells (NP) and ES cells at 1 kb windows centred on all repetitive elements (data from Bulut-Karslioglu et al., 2014). Regions changing significantly (adjusted p-value < 0.05) are marked in grey, IAPLTR1/1a elements activated in TKO neurons in purple and IAPLTR1/1a elements that remain silent in TKO neurons in orange. Samples were normalised with DESeq2, see Methods for details. **c)** H3K9me3 ChIP-seq signal in NPs at IAPLTR1/1a elements activated in TKO neurons and IAPLTR1/1a elements with the same CpG and GC content that remain silent in TKO neurons (data from Bulut-Karslioglu et al., 2014; normalised with DESeq2 and to region size). P-value from a Mann-Whitney test.

3.3.3.7 The CRE motif is highly predictive of transposon activation

While the necessity of DNA methylation rather than H3K9me3 for repeat repression in differentiated cells has been proposed previously (Leung and Lorincz, 2012), it remains unresolved how DNA methylation actually silences repeats. In principle, this could occur in an indirect way by recruitment of histone deacetylases through methyl-CpG binding domain proteins (MBDs), leading to chromatin compaction (Nan et al., 1996; 1998). Another appealing hypothesis is that methylation of motifs within the LTRs directly inhibits binding of TFs that would otherwise drive repeat expression. Importantly the two models are not mutually exclusive. We hypothesised that comparing LTR sequences of activated and silent repeats of the same subtype might identify features linked to methylation-dependent repression.

First, we asked whether TF motifs are enriched in the IAPLTRs that are activated in TKO neurons compared to those that remain silent (Fig. 3-25).

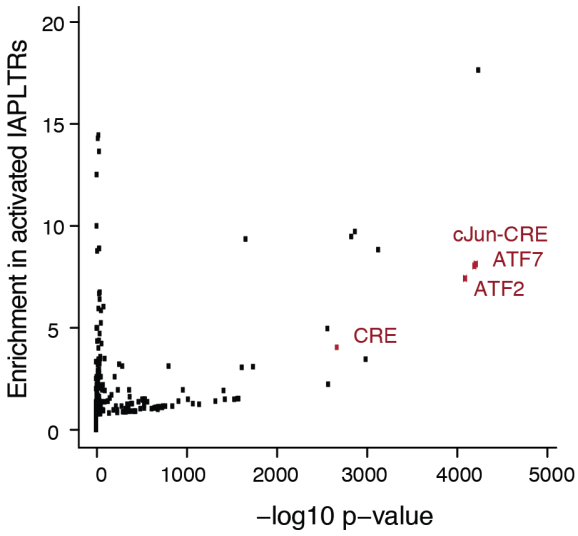


Figure 3-25. The CRE motif is strongly enriched in IAPLTRs that are activated in TKO neurons.
Uniquely mapping 100 bp paired-end RNA-seq reads from WT and TKO neurons were counted in all IAPLTRs in the genome and normalised for library size. Enrichments of known TF motifs in activated (log2 fold-change TKO/WT > 1, n = 3,948) over silent (abs(log2 fold-change TKO/WT) <= 0.3, n = 5,025) LTRs with matching CpG content were calculated using homer2. Motifs that closely resemble the cAMP-responsive element (CRE) are marked in red and labelled. P-values are from a binomial test. See Methods for details. Counting RNA in windows downstream of the LTR rather than the LTR itself gave the same results (data not shown).

Strikingly, several TF motifs are strongly and highly significantly enriched in the activated IAPLTRs. Among the top six enriched motifs, five belong to the same motif family, the cAMP-responsive element, or CRE. It has indeed been previously reported that *in vitro* binding of unknown TFs at this element is impeded by DNA methylation (Iguchi-Ariga and Schaffner, 1989). When we compared the location and strength of the CRE motif in activated and silent members of the same repeat subtype, we observed that the CRE motif is less conserved in silent repeats (Fig. 3-26a,b).

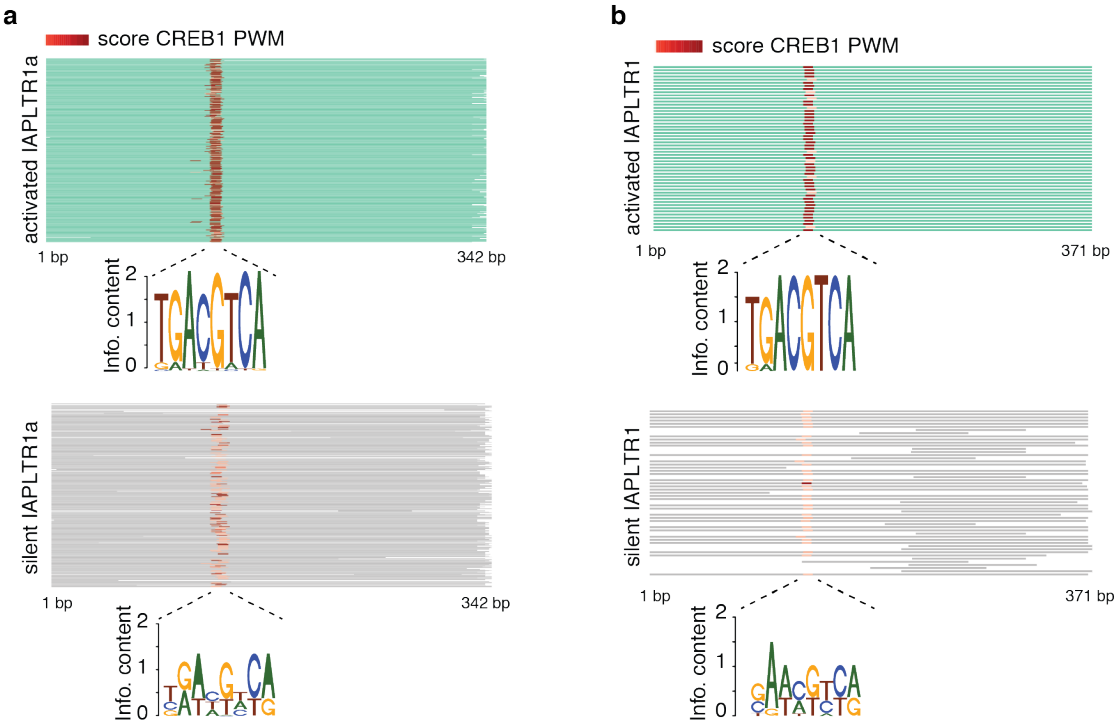


Figure 3-26. The CRE motif is more conserved in members of the IAPLTR1a/1 subtype that are activated in TKO neurons.
a, b) CRE motif location and score in activated (green) and silent (grey) elements of the IAPLTR1a (a) and IAPLTR1 (b) subtype. All silent IAPLTR1a/1 elements and a random sample of the same number of activated IAPLTR1a/1 elements were aligned to the repeat reference sequence using the coordinates given by RepeatMasker. Both groups were scanned for the best match to the JASPAR CREB1 PWM; the position and the absolute score are annotated in shades of red. The consensus PWM was built over these regions for all elements of a group.

This raises the question whether silent repeats are generally less conserved, thus losing their transcriptional competence, or whether local loss of conservation at the CRE motif is associated with silencing. To address this

issue, we compared the identity to the repeat reference sequence across both activated and silent repeats. Interestingly, the sequence is locally less conserved over the CRE motif in silent members (Fig. 3-27). This is true for both IAPLTR1 and 1a types, whereas many other local dips in conservation are not shared between the two subtypes (e.g. around the TSS). In contrast to most TF motifs detected in these LTR sequences, the CRE motif also falls within the region found to be required for transcriptional activity in reporter assays (Christy and Huang, 1988).

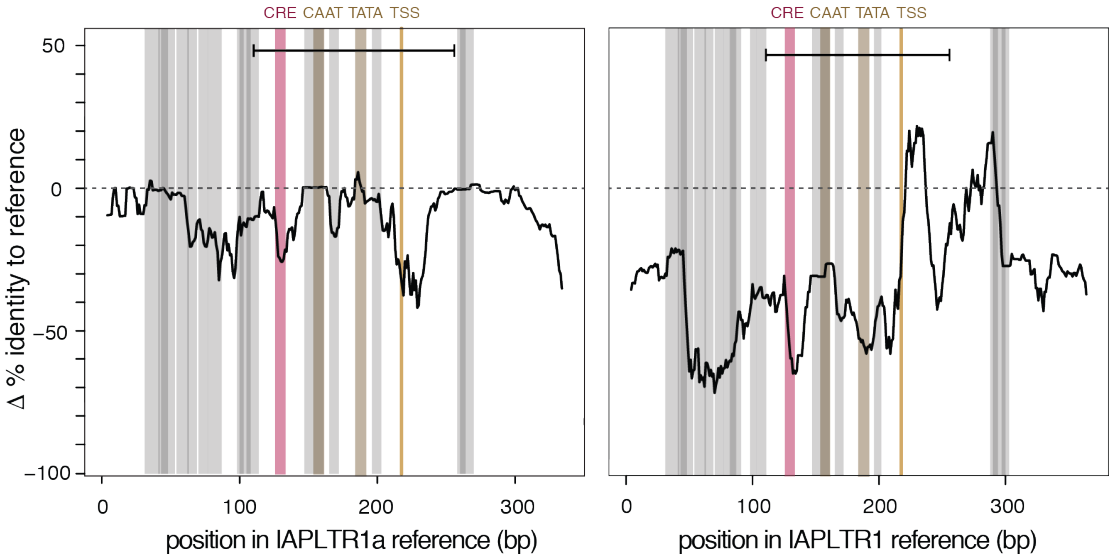


Figure 3-27. Local differences in sequence conservation between active and silent IAPLTR1a/1 elements.
The percent identity of activated IAPLTR1a (left)/ IAPLTR1 (right) elements to the repeat reference sequence at each position was calculated from a multiple sequence alignment and subtracted from the value for silent elements. Accordingly, values below zero mean this position is less conserved in the silent group. The positions of TF motifs from the JASPAR database that have at least a relative score of 95% in the reference sequence are annotated in grey. The region of the IAPLTR found to be required for driving expression in a transient reporter assay (Christy and Huang, 1988) is marked on top as a black line. The location of core promoter elements within this region (CAAT-box, TATA-box and TSS) as well as of the CRE motif are indicated as coloured boxes.

So far the analysis divided the IAPLTRs into two binary categories of activated and silent repeats depending on an arbitrary cut-off (at least twofold upregulation in TKO neurons). We asked if we could also quantitatively predict the level of expression in TKO neurons purely from the sequence of the LTR. Therefore we constructed a linear model for all IAPLTR1 and 1a elements in

the mouse genome, using the scores for all 134 TF motifs from JASPAR that have at least a weak match in the reference sequences as input. In addition we included the CpG content, the GC content and the number of insertions, deletions and mismatches to the reference sequence for each LTR. After performing elastic net regression with fivefold cross-validation for selection of the tuning parameter, 20 non-zero coefficients remained: 19 TF motifs and the number of mismatches to the reference sequence (Fig. 3-28a).

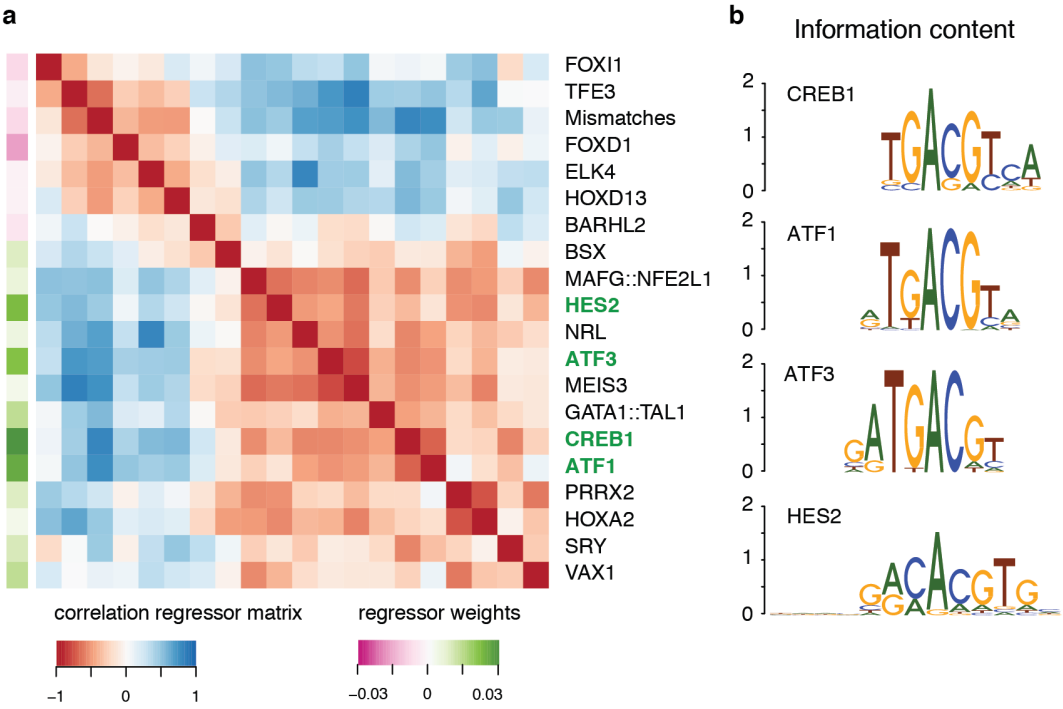


Figure 3-28. The CRE motif score is highly predictive of IAPLTR1/1a expression in TKO neurons.
a) Hierarchical clustering of the correlated regressor matrix for the 20 non-zero coefficients output by the linear model. The model predicts IAPLTR1/1a expression changes between TKO and WT neurons (R^2 = 0.53). See Methods for details. Regressor weights are annotated on the left. The names of the coefficients with the highest absolute weights are marked in green. **b)** JASPAR PWMs for the TF motifs with the highest absolute weights from (a).

This model performs very well, with predicted and actual expression levels of the IAPLTRs correlating at R = 0.73 (R^2 = 0.53). Strikingly, the five motifs with the highest positive weight in the model are all slight variations of the CRE motif (Fig. 3-28b). Other positively weighted TF motifs include SRY, PRRX2 and BSX, which resemble the CAAT box core promoter element. As expected, the number of mismatches receives a negative weight in the model,

so repeats with more mismatches are less likely to be transcriptionally active. Importantly however, the absolute value is lower than for the CRE-like motifs, implying that the score of the CRE motif is highly predictive of the degree of expression in TKO neurons, more so than the overall conservation of the element.

The CRE element has been reported to be bound by homo- or heterodimers of CREB, CRE-BP (also known as ATF2), ATF factors and JUN/FOS (AP-1) (Hai and Hartman, 2001). While CREB and ATF2/4/6 are already highly expressed in stem cells, ATF5, JUN and FOS are upregulated during differentiation both in WT and TKO neurons (Fig. 3-29a,b).

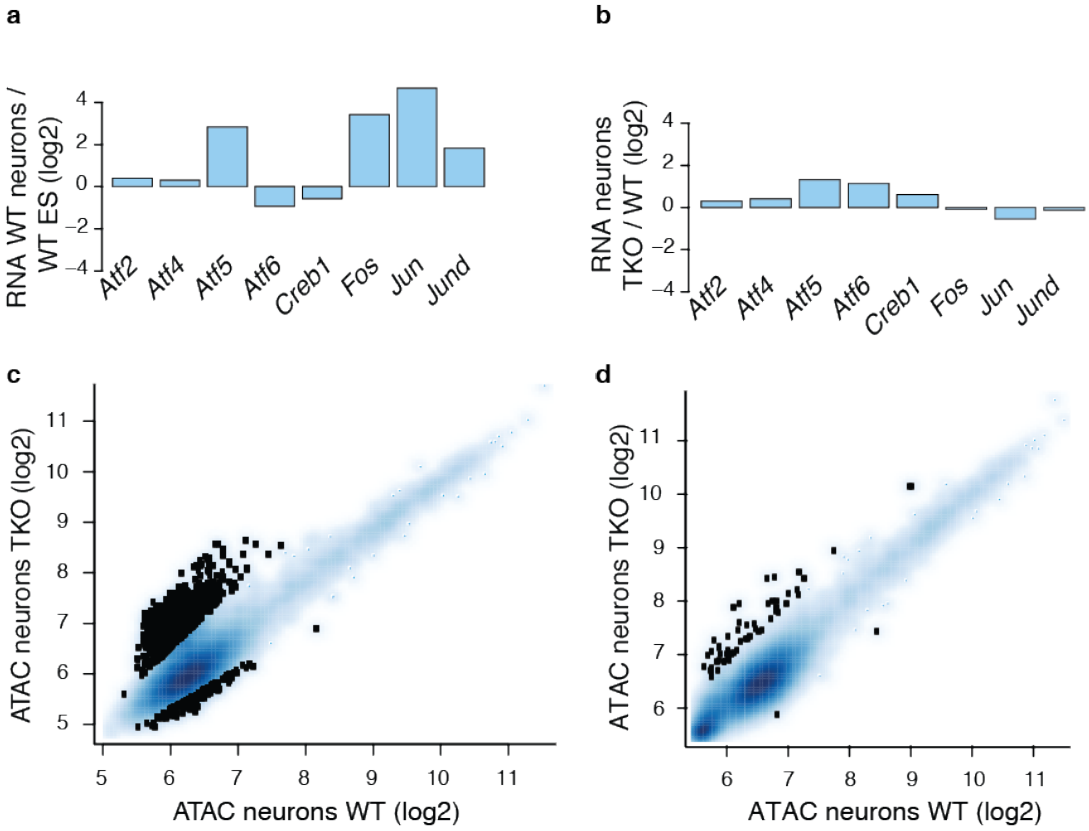


Figure 3-29. Expression of candidate binding factors and accessibility of the CRE motif.
a, b) Gene expression changes during neuronal differentiation (a) and for TKO compared to WT neurons (b) for all TFs that have been reported to bind the CRE motif and are substantially expressed in neurons, as measured by RNA-seq. ATF1 and CREM are further factors that can bind the CRE motif but are not expressed in neurons. **c, d)** ATAC-seq signal at all exact CRE motifs (TGACGTCA) in the genome. ATAC-seq reads were counted in a 400 bp window around the CRE motif, using random assignment of multiple mappers (c) or only uniquely mapping reads (d). Normalisation was performed with DESeq2; significantly differential sites (log2 fold-change TKO/WT > 1, adjusted p-value < 0.05) are marked in black.

As a first step towards analysing differential TF binding at the CRE motif in WT and TKO neurons and to further verify our findings, we analysed chromatin accessibility at all CRE motifs in the genome, either counting randomly assigned multiple mappers or only uniquely mapping reads (Fig. 3-29c,d). We observed an increase in accessibility around the CRE motif in TKO neurons, especially at repetitive regions (Fig. 3-29c), implying that there is indeed differential TF binding at these sites. A random sequence motif with similar occurrence in IAPLTRs did not show such differences in chromatin accessibility (data not shown). Importantly, the CRE motif is also the sixth most enriched motif in all TKO-specific ATAC-seq sites (Fig. 3-19a), implying that methylation sensitivity is not limited to repetitive regions. Differential accessibility is not as apparent at non-repetitive regions however (Fig. 3-29d), likely because the majority of CRE motifs that are not in IAPLTRs fall into unmethylated CGI and promoter regions and thus do not contribute to a strong TKO-specific signal (Smith et al., 2007).

Taken together, different lines of evidence point towards TF binding at the CRE motif being crucial for IAP activation, yet highly sensitive to DNA methylation.

3.3.4 Discussion

For this part of the thesis, we generated a differentiated cell state completely devoid of DNA methylation, in the form of TKO neurons. To our knowledge, such a system has not been previously described and thus it provides an interesting opportunity to investigate the role of DNA methylation in TF binding and repeat silencing in differentiated cells. Similar to our work in ES cells (Chapter 3.2.2), we observe overall limited changes in gene expression and chromatin accessibility in TKO neurons. However, the changes are more unidirectional than in ES cells, in line with a repressive effect of DNA methylation. Whereas one motif, NRF1, explained the majority of TKO-specific accessible sites in ES cells, in neurons many motifs seem to contribute slightly to the differential regions. Cell death in methylation-deficient differentiated cells is likely due to the striking upregulation of certain repeat families, especially IAP elements. Within the IAP family, we find the CRE motif to be a strong predictor for activation.

These differences in repeat activation across repeat families and differentiation stages could be explained by the following hypothesis (Fig. 3-30): In pluripotent cells, endogenous retroviruses are recognised by KRAB zinc-finger proteins or the piRNA pathway and are both H3K9 as well as DNA methylated. Even when undergoing a period of low methylation, as occurs naturally after fertilisation and in primordial germ cells (PGCs), or upon complete loss in TKO ES cells, the repeats remain mostly silent. Nonetheless, IAP elements are in fact those regions in the genome that are most resistant to methylation loss during preimplantation development and in PGCs (Hajkova et al., 2002; Lane et al., 2003; Popp et al., 2010). Since IAPLTRs are activated upon simultaneous loss of both DNA methylation and H3K9me3 in ES cells (Sharif et al., 2016), it is likely that the TFs responsible for driving the expression are already expressed at this stage, but prevented from binding through H3K9me3. During differentiation, H3K9me3 is depleted at repeats, but these are still silenced since DNA methylation prevents binding of TFs, e.g. at the CRE motif. If DNA methylation is removed at these elements, due to genetic manipulation or disease, TFs can bind and induce transcription.

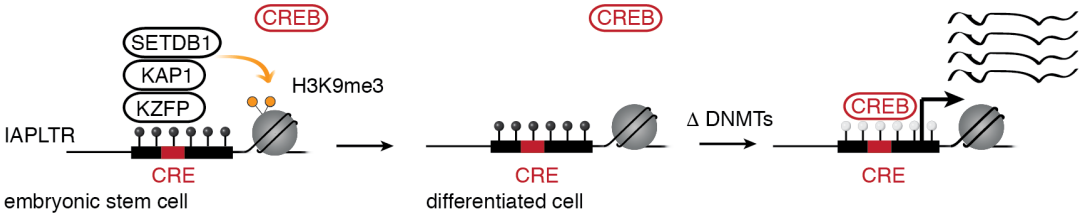


Figure 3-30. Proposed model for regulation of IAP expression in stem and differentiated cells.
In stem cells, IAPLTRs are recognised by KRAB-zinc finger proteins (KZFP), which recruit KAP1 and SETDB1, leading to deposition of the repressive H3K9me3 mark (left). Thus, elements are silenced even upon loss of DNA methylation and in presence of activating TFs associated with binding the CRE motif (e.g. CREB). During differentiation, the H3K9me3 mark is depleted but elements remain silent, since DNA methylation prevents binding of TFs at the CRE motif (middle). Upon loss of methylation in differentiated cells, TFs can bind the CRE motif and trigger high levels of IAP expression (right).

Although DNA methylation has long been associated with repeat repression (Walsh et al., 1998), it remarkably remains unclear to date how this is actually brought about. Since DNA methylation likely evolved as a means to repress repetitive elements, gaining mechanistic insight into this process would also educate us on how and whether methylation-mediated silencing might have been co-opted at other regulatory regions. Repression by cytosine methylation has been suggested to act directly, by interfering with sequence-specific DNA-TF interactions, or indirectly e.g. through MBD binding at CpG-dense regions and chromatin compaction (Klose and Bird, 2006). Of note, a combination of the two models could also be at work, since they are not mutually exclusive. Differential activity of IAPs in TKO neurons is strongly linked to the presence of a CpG-containing CRE motif, which argues for direct blocking of binding. However, if the TEs rely mainly on this motif for driving their expression, this observation would also be in line with indirect repression. While the CRE motif does also come up as a methylation-sensitive candidate in CpG-poor regions outside of repeats, it should be noted that IAPLTRs are relatively CpG-rich sequences and thus could feasibly recruit MBDs. The MBD MeCP2 has in fact been shown to bind the Moloney murine leukaemia-based provirus (Lorincz et al., 2001). Genome-wide mapping of MBDs in mouse ES cells recently revealed low enrichment at

repetitive elements in general, but increased enrichment at CpG-rich repeats including IAPLTRs (Baubec et al., 2013). Within the IAPLTR group however, CpG density is not predictive of the extent of activation upon DNA methylation loss. Also, there is currently no conclusive evidence that MBD binding at repeats is actually responsible for their repression. The issue of direct or indirect repression will be discussed in more detail in the final chapter of this thesis in the context of our other findings.

One might speculate that it is even advantageous for a transposable element (TE) to be regulated by a methylation-sensitive TF. It is only beneficial for a TE to replicate in the germline. Strong activation in somatic cells is deleterious for the host without enhancing virus survival, rather increasing the selective pressure on virus removal out of the population. Indeed ERV families are active to some extent in both mouse and human germlines, which undergo periods of low methylation. Being regulated by a methylation-sensitive TF would thus ensure just this kind of expression pattern: active in the germline but silent in somatic cells. It would also enable use of a ubiquitously expressed strong activator such as CREB without having deleterious effects in somatic cells and might be one reason why IAP elements are the most successful TEs in mouse genomes (Maksakova et al., 2006). In fact, these elements are so active that inbred mouse strains only share roughly 40% of their IAP insertions (Ray et al., 2011). A comparison of IAPLTR sequences across strains which are being sequenced as part of the Mouse Genomes Project might provide insight into the importance of the conserved CRE motif for successful spreading of an ERV. Apart from strong and broad activation of IAPs, individual elements from other repeat families are also upregulated in TKO neurons. In contrast to rodent-specific IAPs, these elements are likely found in more species, enabling cross-species comparisons. This could reveal whether they, too, contain characteristic methylation-sensitive TF motifs in contrast to the majority of other family members and whether this feature is indeed associated with evolutionary success.

It will be interesting to experimentally test the importance of the CRE motif for repeat activation and how TF binding there is impacted by DNA methylation and H3K9me3 marks during differentiation. To this end, we are currently performing reporter assays integrating the wildtype IAPLTR1a reference sequence and one where the CRE motif has been deleted into the same ectopic genomic site in ES cells. Measuring relative activity of these constructs in WT and TKO ES cells as well as neurons should shed light on the importance of the CRE motif for repeat activation. Since these elements are likely to be silenced in ES cells, their relative activity can be tested after reducing H3K9me3 levels by knockdown of enzymes setting this mark (Maksakova et al., 2011). Should these experiments confirm the importance of the CRE motif for repeat activation, it would be exciting to identify the actual bZIP protein binding there by ChIP-seq of likely candidates in the different cell lines. One of the most promising candidates is CREB1, which is expressed in both ES cells and neurons and is known to bind highly conserved palindromic CRE motifs as a homodimer (Benbrook and Jones, 1994). Knowledge of the exact factor(s) responsible for IAP activation would enable us to knockdown this TF and assess whether this can rescue repeat activation and ultimately perhaps even survival of TKO neurons.

The mechanisms we derive in TKO neurons likely also apply in other cell types and species. Although neurons have been described to have distinctive DNA methylation profiles compared to non-neuronal cells (Iwamoto et al., 2011), we importantly see upregulation of the same elements not only in mouse cortex but also in the other methylation-deficient differentiated cell line we analysed, namely fibroblasts. Thus we anticipate that our findings can be transferred to other cell types. While IAP elements do not exist in humans, the larger family of ERV-K elements has a human counterpart, the HERVK LTR retrotransposons. This is indeed the only ERV family member that has continued to replicate in the human population (Marchi et al., 2014). It is lowly but detectably expressed during normal human embryogenesis as well as in many cancers, some autoimmune/ inflammatory diseases and HIV-infected cells (Grow et al., 2015; Wildschutte et al., 2016). Interestingly, several human

LTR retrotransposons contain CRE motifs, and CREB or ATF/AP-1 factors have been implicated in driving expression of human ERVs, Human T cell leukaemia virus type 1 and HIV (Caselli et al., 2012; Grant et al., 2006; Toufaily et al., 2015). CRE methylation has also been associated with promoter silencing of the Epstein-Barr virus genome (Tierney et al., 2000). Insights into how retroviruses are repressed and what leads to their activation could therefore be highly valuable in a broader context, not only for understanding the evolutionary origins of methylation-mediated silencing but also for human disease.

4. General discussion

The goal of this thesis was to investigate the influence of DNA methylation on TF binding in the cellular context. Our findings reveal that most TF binding events are not restricted by DNA methylation in either stem or differentiated neuronal cells. However, DNA methylation is likely capable of preventing binding for a subset of factors at CpG-containing motifs, such as CTCF, NRF1, HNF6 or CREB/ATF. We validate this in detail for NRF1, which binds twice as many sites in the absence of DNA methylation and relies on other TFs to keep its motif in an unmethylated state. While loss of DNA methylation has an overall modest impact on chromatin accessibility and gene transcription in both cell types, it initiates a vast and potentially lethal derepression of endogenous retroviruses in neurons. This appears to be linked to methylation-sensitive binding of TFs to the CRE motif within viral LTRs.

4.1 Extent of binding site restriction by DNA methylation

The vast majority of accessible sites bound by TFs do not change upon removal of DNA methylation in either ES cells or neurons. In both cases, only 2-3% of all detected sites are specific to either the cell line with or without methylation. Whereas NRF1 binding accounts for the majority of differential sites in ES cells, other top TF motifs that are enriched in ES TKO-specific sites have indeed been linked to methylation-sensitive binding in historic *in vitro* gel shift experiments. These include USF, E2F, MYC and CREB, with only NF-κB from these reports not showing any enrichment in our study (Bednarik et al., 1991; Campanero et al., 2000; Iguchi-Ariga and Schaffner, 1989; Prendergast and Ziff, 1991; Watt and Molloy, 1988). In neurons, none of the enriched motifs explains a large portion of differential sites: In contrast to ES cells, many factors seem to contribute to a small subset of differential sites. Still there is agreement with factors identified in ES cells, namely for

NRF1, NFY and CREB, with the other most strongly enriched motifs belonging to neuron-specific HNF6 as well as CTCF.

Although the small change in chromatin accessibility observed in both cell types is in line with the overall limited changes in gene expression, it stands in contrast to the substantial number of TFs suggested to be prevented from binding by DNA methylation based on *in vitro* studies (O'Malley et al., 2016; Spruijt et al., 2013). Importantly, while we observe good agreement with *in vitro* studies for the top identified TFs, there are several reasons that could explain the smaller number of methylation-sensitive factors called with our approach. First, the general lack of binding site restriction by DNA methylation observed here does not preclude methylation-sensitive binding behaviour of TFs in other contexts that cannot be tested in this setup, e.g. at CpG-rich sequences occurring in unmethylated CGIs. In order to be enriched in our approach, a motif needs to occur many times in a methylated state in the genome. This could explain why more factors are reported to be methylation-sensitive at certain motifs *in vitro*. In fact, apart from the TF motifs that are strongly enriched in TKO-specific accessible regions and that were studied here in more detail, many additional motifs are slightly enriched. Thus the matching factors could be restricted by DNA methylation in some contexts, as is the case for CTCF. This TF motif is not enriched in TKO-specific accessible sites in ES cells, yet CTCF known to occasionally bind in a methylation-dependent manner, e.g. at imprinted regions. Second, differential sites are most likely to be detected for TFs that have characteristics of pioneer factors, in that they can bind on their own to previously closed chromatin, thus creating an accessible site *de novo*. Such behaviour has indeed been assigned to most of the top factors identified here, namely NRF1, CTCF and CREB (Sherwood et al., 2014). Factors that are unable to create new accessible sites upon removal of DNA methylation, rather leading to broadening or deepening of existing sites or requiring the presence of certain co-factors, are much harder to detect in this context. Third, some TF binding events might not form highly accessible sites as measured by DNase-/ ATAC-seq due to sequence or other bias of the enzymes used in these methods

(Madrigal, 2015). Both independent chromatin accessibility measures yielded similar results in ES cells, so enzyme sequence bias is likely not a major distorting factor. However, the ability of some TFs to form particularly strong DNase-/ ATAC-seq sites and defined footprints has been associated with longer residency time on the DNA (Sung et al., 2014). NRF1 is the poster child for such a factor generating strong sites (Neph et al., 2012), which likely contributed to the high enrichment of this motif in TKO-specific sites.

Of note, recent *in vitro* studies have also reported several TFs that preferentially bind methylated sites (Hu et al., 2013; Spruijt et al., 2013). We did not observe a strong enrichment of any TF motifs in WT-specific accessible sites. Indeed WT-specific sites are nearly non-existent in neurons and not heavily methylated in ES. However, we cannot exclude that some TFs preferentially bind methylated sites and induce their demethylation.

In spite of these limitations, the results obtained in this thesis argue that DNA methylation is unlikely to have a key role in determining most TF binding events in both pluripotent and differentiated cells.

4.2 Comparison of identified methylation-sensitive transcription factors

Notable exceptions to this rule are the factors discussed in more detail here, namely NRF1, HNF6, CRE-binding factors such as CREB and to some extent CTCF. This raises the question whether these methylation-sensitive TFs have something in common that sets them apart from other members of the TF family. In the following I will compare these proteins in terms of function, structure and motifs.

4.2.1 Comparison of expression and target genes

NRF1 (Schaefer et al., 2000), CREB (Mayr and Montminy, 2001) and CTCF (Nakahashi et al., 2013) are all ubiquitously expressed proteins, whereas HNF6 is only found in a small subset of tissues including liver, pancreas and

brain (Audouard et al., 2013). In line with these expression patterns, HNF6 regulates developmental genes during differentiation in these tissues (Audouard et al., 2013), whereas NRF1 binds the promoters of several house-keeping genes, including those of the respiratory chain (Evans and Scarpulla, 1990). CREB regulates a vast array of biological processes, with some target genes involved in specialised processes such as neurotransmission and others more generally in metabolism and signal transduction (Lonze and Ginty, 2002). CTCF has multiple roles in transcriptional regulation, including a key function in the three-dimensional organisation of the genome (Holwerda and de Laat, 2013). Accordingly, CTCF, NRF1 and CREB deletion are all lethal at the peri-implantation, embryonic or perinatal stage (Bleckmann et al., 2002; Huo and Scarpulla, 2001; Moore et al., 2012), whereas HNF6-null mice are viable (Jacquemin et al., 2000). Aberrant binding of these factors in the absence of methylation differs in terms of scale and effect. TKO-specific sites with NRF1 and HNF6 motifs occur largely distal to gene transcription start sites and although the differences in TF occupancy upon methylation removal are substantial at least for NRF1, most of these binding sites are likely non-functional. Differences in CTCF binding are so minimal that they are only confidently detected in an isogenic setting and using several independent methods. Nonetheless, given the crucial role of CTCF in chromatin looping, altered binding due to methylation changes e.g. at the *H19/Igf2* ICR is associated with growth disorders and other diseases in humans (Herold et al., 2012). In turn, methylation-sensitive binding at CRE motifs in repeat regions has potentially devastating effects by activating these TEs. Interestingly, both CREB and NRF1 are known as especially strong transcriptional activators (Ernst et al., 2016; Mayall et al., 1997).

4.2.2 Comparison of DNA-binding domains

The largest TF families in vertebrates are C2H2 zinc-finger proteins, followed by the homeodomain and the basic superfamily, which includes the leucine zipper (bZIP) and helix-loop-helix (bHLH) families. Together they account for more than 80% of TFs (Weirauch and Hughes, 2011). Each of these classes

is suggested to have arisen from a single common ancestral protein followed by multiple rounds of duplication and divergence (Weirauch and Hughes, 2011). The TFs found to be methylation-sensitive here represent all three major classes: CTCF is a zinc-finger protein that binds divergent long motifs by combinatorial clustering of its eleven zinc fingers (Nakahashi et al., 2013), HNF6 is a homeodomain protein that binds DNA through both its homeodomain and a cut domain (Iyaguchi et al., 2007), whereas CREB/ATF are classical bZIP proteins (Schumacher et al., 2000) and NRF1 has a novel DNA-binding domain that is not shared by any other known TF (Schaefer et al., 2000; Virbasius et al., 1993). NRF1 and CREB/ATF factors dimerise to bind palindromic motifs, while CTCF and HNF6 bind on their own to non-palindromic sequences (Benbrook and Jones, 1994; Gugneja and Scarpulla, 1997). Crystal structures of the protein in complex with (unmethylated) DNA are available for CREB1 (Schumacher et al., 2000) and HNF6 (Iyaguchi et al., 2007), although the latter template does not contain a CpG. Both proteins interact mainly with the major groove of the DNA, where the methyl-group of the cytosine is positioned (Dantas Machado et al., 2015). Few crystal structures of DNA stretches with methylated CpGs have been reported, but they imply that the methyl group leads to widening of the major groove and thus might impact specific protein interactions there (Tippin and Sundaralingam, 1997). Interestingly, HNF6 homologs share distinct residues involved in the interaction with the DNA major groove that set them apart from all other known homeodomains (Lannoy et al., 1998). For CREB1 the interaction between a certain arginine residue and the CpG was found to be crucial for high-affinity binding (Schumacher et al., 2000). Indeed CREB1 is unable to bind a motif variant where the central cytosine is replaced with a thymidine (TGA**T**GTCA), in contrast to other closely related bZIP proteins capable of binding this CRE variant (Benbrook and Jones, 1994). Of note, methyl-cytosine closely resembles thymidine. For NRF1 and CTCF no crystal structures are available, although other C2H2 zinc-finger protein-DNA interactions have been characterised. Some of these are indeed capable of differentiating between methylated and unmethylated DNA, as seen for Kaiso.

This methyl-CpG binding protein has been crystallised in complex with methylated DNA (Buck-Koehntop et al., 2012).

Since the TFs identified here all have very different structures, methylation sensitivity likely did not arise from a common DNA-binding domain or common evolutionary ancestor. This raises the question how methylation sensitivity of these TFs evolved. The DNA-binding domain of NRF1, while unique within the TF family, is highly conserved between *Drosophila* and human and recognises the exact same motif in both cases, although the remaining part of the amino acid chain is quite divergent (Fazio et al., 2001). For HNF6 the DNA-binding cut and homeodomains are conserved more than 80% between homologs in species with and without DNA methylation (Poustka et al., 2004). *Drosophila* CREB has 88% identity to mammalian CREB in the bZIP domain (Yin et al., 1995), while the activator domain is less conserved, and indeed CRE-binding factors go as far back as yeast (Nehlin et al., 1992). The *Drosophila* homolog of CTCF also contains eleven zinc fingers highly similar to the vertebrate version (Moon et al., 2005). This high conservation of DNA-binding domains observed for all identified methylation-sensitive factors in organisms without DNA methylation raises the question whether their invertebrate homologs are methylation-sensitive or not (Fig. 4-1a). Answering this question might provide valuable insight on whether methylation sensitivity at regulatory regions has only evolved in vertebrates as part of the co-option of methylation for other means than TE silencing. For instance, it is feasible to replace the mouse DNA-binding domain of NRF1 with the *Drosophila* homolog and measure binding in WT and TKO murine ES cells. If the *Drosophila* homolog is indeed *in*sensitive to DNA methylation, it would be interesting to identify and determine methylation sensitivity of the common vertebrate and invertebrate ancestor TF, to learn whether methylation sensitivity was lost in invertebrates or gained in vertebrates.

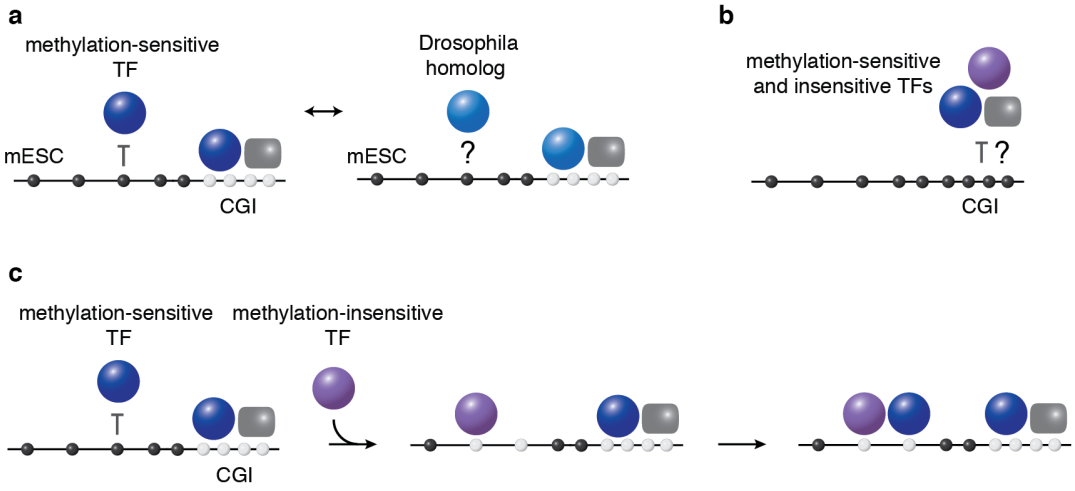


Figure 4-1. Differences in methylation sensitivity across species, regions and factors. **a)** Evolution of methylation sensitivity. Expressing homologs with highly conserved DNA-binding domains from species without DNA methylation in WT and TKO mouse ES cells (mESC) to determine their methylation sensitivity might educate us on how this feature evolved. **b)** Methylated CpG island (CGI) promoters are generally stably repressed. It is unclear how and whether TF binding site restriction differs here from CpG-poor regions. In contrast to CpG-poor regions, methylation-insensitive TFs do not seem to induce dynamic local hypomethylation at methylated CGIs. **c)** DNA methylation mediates TF hierarchies. Methylation-sensitive TFs can only bind unmethylated sequences, which are mainly found in CpG islands (CGI) (left). However, methylation-insensitive TFs can bind methylated CpG-poor regions and induce their local demethylation (middle). This enables downstream binding of sensitive TFs to distal CpG-poor regulatory regions that are active in a given cell type (right).

4.2.3 Comparison of methylation-sensitive motifs

The nature and conservation is not only of interest for the DNA-binding domains of methylation-sensitive TFs but also for the sequence motifs they interact with. With the exception of the NFY motif, a core promoter element known as CCAAT box, all top identified methylation-sensitive motifs in this study contain at least one prominent CpG.

CTCF binds degenerate motifs that do not necessarily contain CpGs in vertebrates but mostly do in *Drosophila* (Ni et al., 2012; Stadler et al., 2011). In fact, the consensus CTCF motif in *Drosophila* is more defined and closely resembles one of the CpG-containing methylation-sensitive motif variants identified here. In contrast to *Drosophila*, where CTCF motifs lie mostly in promoter regions (Ni et al., 2012), the majority of canonical CTCF motifs in vertebrates reside in intergenic regions, in line with their CpG-poor nature (Kim et al., 2007). Interestingly, as for HNF6, the canonical vertebrate CTCF motif closely resembles a deaminated version of the CpG-containing

methylation-sensitive variant. CpGs might have been lost in most CTCF motifs in vertebrates to circumvent binding obstruction by DNA methylation, as a means to achieve the high agreement of robust CTCF binding profiles observed across cell types (Chen et al., 2012). However, this remains pure speculation without a closer look at the evolutionary constraints underlying changes in CTCF motif preference.

In contrast to CTCF, CRE and NRF1 motifs belong to the most commonly found TF motifs in mammalian promoters (Weirauch and Hughes, 2011; Zhang et al., 2005). Both have a strong positional bias, in that their distribution peaks shortly upstream of the transcriptional start site in different vertebrate species from fish to human (FitzGerald et al., 2006; Smith et al., 2007). Although CREB and NRF1 bind the same highly specific motifs in both vertebrates and invertebrates, a preferential enrichment or positional bias is not observed for either motifs in promoters of species without blanket methylation, such as *Drosophila melanogaster*, *Caenorhabditis elegans* or *Ciona intestinalis* (Smith et al., 2007). This is not the case for all CpG-containing motifs, since some, such as the E-box, can have positional bias and enrichment within both mammalian and *Drosophila* promoters (FitzGerald et al., 2006). For NRF1 and CRE motifs however, deamination of methylated CpGs likely contributed to depletion of existing motifs outside of unmethylated CGI promoters and an increasingly unequal genomic distribution in vertebrates. Indeed the human genome only contains 25% of the expected number of CREs, with 62% of them occurring in CGIs (Smith et al., 2007). Interestingly, around 10% of the roughly 28,000 fully conserved CRE motifs in mice reside in IAPLTRs, which in contrast to CGI promoters are methylated at almost all times. These distributions imply that there is no strong selection to maintain methylation-regulated motifs at regulatory regions outside of CGIs, although it would be necessary to compare the exact rate and location of CpG loss in these motifs with other CpG-containing sequences.

Deamination of CpGs is indeed rather frequent in TF binding sites genome-wide, as was observed when comparing human, chimp and rhesus genomes (Zemojtel et al., 2011). In fact based on position frequency matrices

an estimated 85% of human TFs recognize a motif containing TpG, whereas 25% of TF motifs in the JASPAR database contain a CpG (Blattler and Farnham, 2013; Zemojtel et al., 2011). Amongst this latter group, the CpG is only crucial for recognition in barely a dozen cases (Blattler and Farnham, 2013). Of note, most of these motifs are either strongly (NRF1, CREB) or slightly (USF, HIF1A/ ARNT, ETS and E2F factors) enriched in our approach. Indeed both NRF1 and CREB depend on the presence of a central CpG for high affinity binding (Benbrook and Jones, 1994) and are thus likely methylation-sensitive at all strong motifs. Importantly, these observations imply that methylation could restrict many more binding events in genomes that contain methylation but are not yet depleted of CpGs, such as in the African clawed frog *Xenopus laevis*.

Taken together, the identified methylation-sensitive TFs do not share a mutual function or structure. While their DNA-binding domains are remarkably conserved across species without DNA methylation, their CpG-containing motifs are depleted over time. In-depth evolutionary comparisons of factors and motifs involved in methylation sensitivity might indicate when and how this feature evolved independently in different TF families and which selective pressures are acting on it.

4.3 Direct or indirect blocking of binding by DNA methylation

A key question in the field is whether binding site restriction by DNA methylation is a direct process, where methylation in the motif interferes with the DNA-TF interaction (Dantas Machado et al., 2015), or an indirect process, by recruitment of MBDs and chromatin compaction independent of specific TF motifs (Nan et al., 1996; 1998). As detailed above, the methylation-sensitive TFs identified here contain prominent CpGs in their motif. This is particularly striking for CTCF and HNF6, for which only CpG-containing motif variants are enriched in TKO-specific accessible sites, in contrast to the canonical motifs.

As far as can be conjectured from the existing crystal structures, it is feasible that the methyl-group directly impacts the interaction between the identified TFs and the major groove of the DNA. We find that NRF1 occupancy is reduced at a sequence that has locally higher methylation levels in the CpGs of the motif but is otherwise identical in terms of chromosomal location and overall methylation levels. These observations all point towards direct blocking of binding by DNA methylation within the motif. However, it cannot be excluded that indirect mechanisms are also involved and the two are not mutually exclusive. MBDs are supposed to preferentially bind regions with dense and methylated CpGs (Baubec et al., 2013). Indeed increased CpG density of the flanking regions characterises TKO-specific CTCF sites. It is unclear to which extent this is linked to a higher likelihood of encountering a CpG-containing CTCF motif variant in regions of higher CpG density. Genomic editing approaches as suggested in Chapter 3.1.4 could shed light on the relative importance of CpGs within the motif and the flanking regions for methylation-sensitive CTCF binding. Importantly, we do not observe an influence of CpG density on methylation-sensitive binding of NRF1. Indeed, the vast majority of TKO-specific accessible sites in both ES cells and neurons fall into CpG-poor regions. This is likely due to the fact that most CpG-dense regions are already unmethylated in the WT cells. Nonetheless, it represents one of the first descriptions of methylation having an inhibitory effect on transcription in CpG-poor regions of the genome.

Thus, while direct repression is a likely scenario at least in CpG-poor regions, this question still warrants further study. Of note, it also remains unclear if different mechanisms mediate repression at methylated CGIs (Fig. 4-1b). For example, simultaneous deletion of all MBDs should reveal if these proteins are indeed involved in the repression of methylated CGIs and repeats. In addition, crystal structures of methylation-sensitive TFs in complex with methylated and unmethylated DNA would provide further insight into the mechanism of repression. Again, comparison with DNA-binding domain structures of closely related proteins from species without genome

methylation could be very valuable to understand which evolutionary constraints act at key amino acids involved in the DNA interaction.

4.4 Transcription factor hierarchies mediated by DNA methylation

In principle, DNA methylation-mediated TE silencing mechanisms could have been co-opted by vertebrates to introduce another layer in gene regulation. Here we present evidence for a TF hierarchy, where methylation-insensitive TFs can bind CpG-poor methylated regulatory regions and induce their demethylation, thus enabling binding of methylation-sensitive TFs like NRF1 (Fig. 4-1c). In this scenario, methylation mediates cooperativity between TFs but in a remarkably indirect manner. This cooperativity does not depend on direct interaction between TFs or even on sequence context in terms of specific co-occurring motifs, but rather only on the ability of some TFs to remove methylation around their binding sites. Indeed NRF1 sites bound in WT cells are strongly enriched for co-binding of all tested TFs compared to TKO-specific sites. This makes NRF1 binding unlikely to depend on a certain factor for demethylating its motif rather than just on the presence of an active, unmethylated region. In this manner, especially TFs that are expressed ubiquitously across tissues, such as NRF1 or CREB, could be guided to regulatory regions that are active in a given cell type, independently of their expression level. This might be especially important for TFs that act as strong transcriptional activators and in principle can bind closed chromatin on their own. DNA methylation could thus provide an additional layer of regulation for certain pioneer factors, which are otherwise thought to bind independently of chromatin conformation.

However, both NRF1 and CREB lose high affinity binding when their central CpGs are deaminated (Benbrook and Jones, 1994). Comparison with ancestral genomes reveals ongoing depletion of these and other CpG-containing motifs (Smith et al., 2007; Zemojtel et al., 2011). Those methylated NRF1 motifs that still contain central CpGs and are bound in the TKO cells lie

in lowly conserved regions compared to WT. Together this suggests that methylation-protected sites are being eroded over evolutionary times, further raising the question which of these regions are actually functional in other cell types. Analysing the conservation of methylation-restricted TF binding sites in more detail might inform on whether these regions are likely to have important roles in gene regulation or whether methylation prevents binding mostly at non-functional sites. Of note, this could still have an effect on gene regulation, by avoiding dilution of free TF levels through binding at thousands of irrelevant sites.

4.5 DNA methylation and cell survival

It is currently unclear if the essential nature of DNA methylation in differentiated cells is driven by aberrant gene expression (Jackson-Grusby et al., 2001) or repeat activation in its absence (Walsh et al., 1998; Yoder et al., 1997). Both might be linked to mitotic catastrophe, which has been suggested to be at the centre of rapid cell death in the absence of DNA methylation (Chen et al., 2007). Of note, TKO neurons are in fact postmitotic for several days before dying, making such a direct link to a cell cycle checkpoint unlikely. Since gene expression is remarkably similar between WT and TKO cells, our results imply that repeat activation is at the core of the matter. Repeat activation is also the key feature that distinguishes TKO neurons, which are unable to survive for many days, from TKO ES cells, which do not show any obvious phenotype. Activation of TEs can potentially induce cell death in several ways, e.g. by sheer transcriptional load or insertion of active ERVs into genes or promoter regions, thus producing mutants or high levels of chimeric transcripts (Bestor, 2003). Interestingly, although we were unable to generate TKO neurons with a retinoic acid-based differentiation protocol, we in fact observed deregulation of the same repeat families and cell death on the same time-scale (~ 10 days) as for the rapid NGN2-induced neuronal differentiation protocol. Of note, cell death for deletions of *Dnmt1* in human ES cells, which represent a slightly more differentiated stage than mouse ES cells

(Nichols and Smith, 2009), cancer cells and even mouse embryos occurs on a similar temporal scale (Chen et al., 2007; Li et al., 1992; Liao et al., 2015). This implies that cell death in the absence of DNA methylation is not dependent on reaching a specific differentiated cell state but might be due to activation of endogenous retroviruses and accumulation of either their transcripts or genomic insertions over time. It will be interesting to test this hypothesis in other cell types and to align these observations with temporal changes in H3K9me3 enrichment at repetitive regions during differentiation.

4.6 Transferability of the approach to studying other chromatin features

At least for the methylation-sensitive TFs identified here, the ability to predict binding will markedly improve if DNA methylation profiles in a given cell type are taken into account. Twice as many NRF1 motifs representing an exact match to the consensus sequence are bound in the absence of DNA methylation. This is a striking increase, although it will be less pronounced for factors that are not affected by methylation at all of their strong binding sites, such as CTCF. That said, roughly one third of these high-confidence NRF1 motifs remain unbound in TKO cells, so that other factors must contribute to binding site restriction even for this TF.

This raises the question if the approach applied here can be transferred to other chromatin components potentially involved in binding site restriction, such as nucleosomes and histone modifications (Fig. 4-2a). The crucial point here will be to find conditions that allow altering these chromatin features without leading to rapid cell death and massive transcriptional changes. Certain histone modifications were found to be dispensable in ES cells and their removal leads to limited changes in gene expression, as observed for H3K27me3 (Riising et al., 2014). This mark is likely less important for silencing of TEs than DNA methylation (with the exception of murine leukaemia virus elements) and thus the observed cell death upon differentiation could indeed be due to differential TF binding at crucial regions

(Leeb et al., 2010). Applying a similar approach to study binding site restriction by Polycomb-group proteins is thus possible. While ES cells have also been at least partially depleted of H3K9me3, full removal has been reported to be lethal (Dodge et al., 2004; Matsui et al., 2010; Peters et al., 2003; Walter et al., 2016). Still it is conceivable to generate complete knockouts of all H3K9 methylating enzymes with CRISPR that at least survive for a certain amount of time. If gene expression changes are not too drastic, this could still be informative, as seen here for methylation mutants of differentiated cells similarly reported to be lethal.

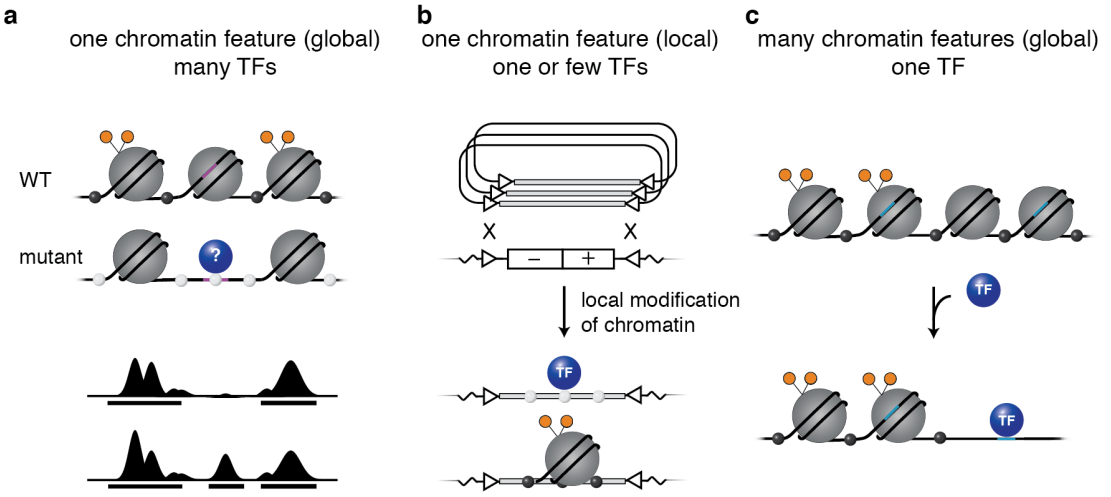


Figure 4-2. Complementary strategies for investigating the role of chromatin in binding site restriction.
a) Global exploratory approach. Comparing chromatin accessibility measured by DNase-seq/ ATAC-seq as an indicator of TF binding in WT cells and cells mutant for a chromatin feature (e.g. DNA methylation/ histone modification) can reveal regions that are only bound in the absence of this mark. This approach relies on the ability of the cell to survive global depletion of the chromatin feature and subsequent identification of sensitive TFs by sequence analysis of differentially accessible regions. **b)** Context-specific approach. Chromatin features can also be interrogated locally at one specific site in the genome. Libraries of sequence variants can be inserted into the same ectopic site by RMCE and TF binding compared between modified and unmodified states. These chromatin states can be achieved by inserting differentially modified sequences (as for DNA methylation) or by locally recruiting modifying enzymes post-insertion (as for histone marks). This approach controls for chromosomal context and changes due to global loss of a certain chromatin feature. **c)** Factor-specific approach. Expression of TFs in non-native contexts and measurement of resulting TF binding and changes in chromatin profiles can educate on which chromatin states are instructive for binding and which are adopted downstream. For example, a TF might be only able to bind sites not possessing a certain repressive histone mark. This approach can be applied in either WT or mutant cells and has the potential elucidate the relative importance of different chromatin features and their combinations in binding site restriction.

Depleting cells of nucleosomes is likely more complicated. While a reduction of nucleosome levels has been observed in some settings, such as in HMGB1 mutant MEFs (Celona et al., 2011), complete removal is not feasible and the reproducibility of nucleosome positions for a partial removal is questionable. Investigating binding site restriction by nucleosomes thus requires an alternative strategy. Rather than depleting repressive features genome-wide (Fig. 4-2a), one could focus on modifying chromatin at one controlled locus and assaying many sequences/ TF motifs there (Fig. 4-2b). Such an approach was recently taken in yeast to study the effect of nucleosomes on binding of TFs with a massively parallel reporter assay (MPRA) of nucleosome-favouring or -disfavouring sequences, using NOMe-seq as a readout to simultaneously call nucleosome positions and TF binding (Levo et al., 2017). It is feasible to study the impact of not only nucleosomes but also repressive histone modifications in vertebrates in a similar manner. This could be achieved by either integrating a library of TF motifs into WT and mutant cells, choosing heterochromatic and euchromatic regions as integration site or actively recruiting the histone mark in question to the sequence library by tethering modifying enzymes in the vicinity. While conceptually appealing, we learnt the limitations of such an approach when studying the influence of DNA methylation on CTCF binding. It was exceedingly difficult to maintain differential methylation states for the same sequence within cells. This caveat might also apply to other chromatin modifications, making it challenging to investigate the impact of a certain mark at the exact same sequence and location. That said, we were nevertheless able to use this strategy to compare NRF1 binding at identical sequences in WT ES cells that only differed in core motif methylation levels. The sensitivity to methylation of CpGs in the core motif observed here strongly supported our findings in TKO ES cells, especially since it was independent of global changes in methylation levels.

In a reverse approach, instead of locally tampering with chromatin modifications, it is also possible to express individual TFs in non-native contexts and study how existing chromatin modifications restrict their binding

(Fig. 4-2c). For example, overexpression of NRF1 in WT cells revealed binding to many weak motifs in unmethylated regions of the genome but not at strong methylated sites. A cursory analysis was unable to identify those chromatin features that prevent binding at the roughly one third of high-confidence NRF1 motifs that remain unbound even in the absence of DNA methylation. However, expressing NRF1 in different cell types with variable well-annotated chromatin landscapes and employing more sophisticated analytical strategies such as machine learning might uncover which features besides DNA methylation restrict binding of this factor.

A major drawback to both local sequence variations as in MPRAs or ectopic expression of individual TFs is that they only educate us on a single genomic context or factor. We anticipate that the relationship between TF binding and chromatin modifications is on the contrary highly factor and context dependent, since we observe a remarkable variety in structure and sequence-dependency for DNA methylation-sensitive TFs in this study. As Slattery and colleagues put it, 'the only common thread in the world of TF–DNA interactions and transcriptional regulation is that no single model is sufficient to explain all the mechanisms used to achieve regulatory specificity' (Slattery et al., 2014). Thus, combining genome-wide exploratory approaches as employed here with more in-depth and high-resolution factor- and context-specific studies can complement each other and will be necessary to reveal the complex rules determining TF binding in the context of chromatin. The ability to modify regulatory sequences and create gene knockouts with CRISPR has opened up unprecedented avenues to explore this fascinating interaction. In the future, this should enable us to markedly improve our predictions of TF binding in a given cell type and represent a substantial advancement in the quest for understanding and predicting gene regulation in eukaryotic development and disease.

5. Materials and methods

Cell culture

Mouse ES cells were cultivated without feeders on 0.2% gelatine-coated dishes in DMEM, supplemented with 15% fetal calf serum, 1× non-essential amino acids, 2 mM L-glutamine, LIF and 0.001% β-mercaptoethanol (37°C, 7% CO_2). For insertion of *H19/Igf2* fragments, TC-1 cells with a RMCE site in the beta-globin locus (Lienert et al., 2011) were used. Mouse HA36 ES cells (mixed 129-C57Bl/6 strain) with a stable integration of the *Neurogenin2* gene under control of pTRE-tight were a kind gift from the Jeff Chao lab (FMI). The three DNA methyltransferases *Dnmt1*, *Dnmt3a* and *Dnmt3b* were deleted in these cells by CRISPR-Cas9 gene editing as previously described (Domcke et al., 2015) to generate a TKO line without DNA methylation. *Dnmt* genes were sequenced to confirm successful targeting of all six alleles and residual methylation levels were measured by Zymo Research (www.zymoresearch.com), using high-pressure liquid chromatography coupled to mass spectrometry.

Neuronal differentiation

ES cells HA36 were differentiated into embryoid bodies and neuronal progenitors using LIF withdrawal and retinoic acid treatment as previously described (Bibel et al., 2007). For lines containing the pTRE-*Ngn2* construct, differentiation was carried out by inducing expression of NGN2 with 1 *μ*g/mL doxycycline as previously described (Thoma et al., 2012). Neurons were harvested 8 or 9 days after induction.

Recombinase-mediated cassette exchange

Fragments of the *H19/Igf2* ICR to be inserted into the ectopic genomic site in TC-1 cells were PCR amplified from genomic DNA and cloned into a plasmid flanked by two inverted L1 Lox sites. Fragments contained one to four CTCF

sites (coordinates in Table 5-1). For the RMCE reaction we used both unmethylated plasmids and plasmids that were *in vitro* methylated with M.SssI (NEB) (Schubeler et al., 2000). Complete *in vitro* methylation of the plasmids was confirmed by digestion with HpaII/MspI (NEB) and gel electrophoresis. RMCE was performed in TC-1 ES cells as previously described (Jermann et al., 2014; Lienert et al., 2011). Clones were picked 12 days after the nucleofection reaction and tested for successful insertion events by PCR. For removal of inserts, RMCE was performed with a plasmid containing the hytk (hygromycin-phosphotransferase thymidine kinase fusion gene) expression cassette under the control of the Pgk promoter; clones that had exchanged the insert were selected with hygromycin.

Table 5-1. Genomic coordinates of ICR fragments inserted into the ectopic site.

Name	Type	Start	End	Width (bp)
H19 ICR	ICR	chr7: 149,765,874	chr7: 149,768,737	2,863
CTCF motif 1	CTCF motif	chr7: 149,766,211	chr7: 149,766,230	19
CTCF motif 2	CTCF motif	chr7: 149,766,666	chr7: 149,766,685	19
CTCF motif 3	CTCF motif	chr7: 149,767,692	chr7: 149,767,711	19
CTCF motif 4	CTCF motif	chr7: 149,767,936	chr7: 149,767,955	19
CTCF1234	ICR fragment insert	chr7: 149,766,016	chr7: 149,768,088	2,072
CTCF12	ICR fragment insert	chr7: 149,766,129	chr7: 149,766,765	636
CTCF34	ICR fragment insert	chr7: 149,767,567	chr7: 149,768,089	522
CTCF1	ICR fragment insert	chr7: 149,766,020	chr7: 149,766,387	367
CTCF2	ICR fragment insert	chr7: 149,766,465	chr7: 149,766,830	365
CTCF3	ICR fragment insert	chr7: 149,767,567	chr7: 149,767,842	275
CTCF4	ICR fragment insert	chr7: 149,767,818	chr7: 149,768,081	263

Bisulfite sequencing

Bisulfite conversion was performed on 2 µg of the RNaseA-treated genomic DNA (EpiTect Bisulfite kit, Qiagen). Converted DNA was amplified with the designed specific primers using following cycling conditions: 20 touch-down cycles from 55 to 50°C with 30 s at 95°C, 30 s at 55/ 50°C and 30 s at 72°C, followed by 36 cycles of 30 s at 95°C, 30 s at 50°C and 30 s at 72°C and a final 5 min extension step at 72°C (AmpliTaq Gold, Thermofisher). PCR products were gel-purified and cloned into OneShot *E. coli* using blue/white

colony screening for selection of recombinants (Topo TA, Thermofisher). For 15 to 20 positive bacterial colonies per PCR reaction, DNA was isolated and amplified using rolling circle amplification (Templiphi, GE Healthcare) and the inserts were sequenced with Sanger sequencing. Reads were analysed with BISMA (Rohde et al., 2010). Identical bisulfite sequencing reads were discarded in the analysis, since they are likely to be PCR duplicates.

Methylation-sensitive qPCR

For bisulfite-independent measurement of DNA methylation states, genomic DNA was isolated using the DNeasy Blood & Tissue kit (Qiagen) and 2 μg were homogenised by passing through a 27 1/2-gauge needle at a concentration of 12.5 ng/μL. The homogenised sample was split into 40 μL reactions. Four of these aliquots were digested for 5 h at 37°C with 25 units of either HpaII/HhaI (cut at unmethylated site) or MspI/McrBC (cut at methylated site). A fifth aliquot without enzyme was used as mock control and otherwise treated in the same manner. Digestion reactions were performed in triplicates. After the incubation step, each sample was diluted eightfold with water and standard qPCR was performed with 2.5 μL template DNA. Validated primer sequences for the endogenous *H19/Igf2* ICR were used (Oakes et al., 2009), which amplify a region that is not inserted in the ectopic site for the tested clones. The mean Ct values of the triplicate digested samples were subtracted from the mean Ct value of the mock-digested sample to produce a deltaCt value for each digest. From this the percentage of methylation of a given CpG site within the amplicons was calculated (Oakes et al., 2009).

Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) was carried out essentially as previously described (Jermann et al., 2014), using a polyclonal antibody against CTCF (SantCruz, sc-15914) and H3K9me3 (Abcam, ab8898). CTCF ChIP-seq libraries were prepared according to standard Illumina library preparation protocols using 12 cycles of PCR (NEB Q5 Hot Start HiFi PCR)

and sequenced on an Illumina HiSeq 2500 machine (50 bp read length, single end). ChIP-qPCRs were performed according to standard protocols using one primer within the insert and one outside of the L1 sites for insert-specific detection and a primer pair that extends into a non-inserted region of the *H19/Igf2* ICR for endogenous-specific detection.

RNA-seq

For neuronal progenitors, cellular aggregates were pelleted eight days after starting LIF withdrawal and after four days of retinoic acid treatment. Neurons generated by NGN2 induction were collected by removing most of the medium, scraping the cells off the plate with a cell scraper and centrifuging briefly to remove cell debris before proceeding directly with RNA isolation. RNA was isolated from the cell pellets with the RNeasy mini kit (Qiagen) using on-column DNA digestion. For RNA-seq, two micrograms of total RNA from three to four independent cultures were depleted from ribosomal RNA using the Ribo-Zero rRNA removal kit (Epicentre). Strand-specific total RNA-seq libraries were prepared from rRNA depleted samples using the ScriptSeq v2 protocol (Epicentre). Libraries were sequenced on an Illumina 2500 HiSeq with 50 bp single-end reads. Three neuron replicates per cell line were resequenced with 100 bp paired-end reads to allow better mapping at repetitive regions.

ATAC-seq

ATAC-seq was performed essentially as previously described (Buenrostro et al., 2015). For neurons, nuclei were isolated directly from scraped cell pellets (see above) and different cell numbers (10-100,000) were tested on a low-coverage MiSeq run. Since profiles for these conditions were very similar, 50,000 cells were used for further experiments, as for ES cells. Libraries were PCR amplified for 10 cycles using the NEBNext Q5 Hot Start HiFi PCR Master Mix and sequenced on an Illumina 2500 HiSeq with 50 bp paired-end reads.

ChIP-seq and DNase-seq data analysis

For identification of differential CTCF sites, ChIP-seq and DNase-seq samples were aligned single-end to the mm9 mouse genome using bowtie (-m 1 --best --strata) (Langmead et al., 2009). ChIP-seq peaks were called with peakzilla with default parameters (Bardet et al., 2013). For comparison of fold-changes between ChIP-seq and DNase-seq data, the ChIP-seq reads extended to 200 bp (average estimated fragment length) and the first bp (5'-end) of the DNase-seq reads (DNase I cut site) were used to calculate raw read counts in 300 bp windows around 391,862 low confidence CTCF motifs, excluding CpG islands, and normalised to library size. Motif enrichments within TKO-specific sites (log2 TKO/WT > 1 for both ChIP-seq and DNase-seq signal) were calculated using homer2 (Heinz et al., 2010).

For analysis of H3K9me3 enrichment at repetitive regions in ES cells and NPs, 36 bp paired-end ChIP-seq data was retrieved from GEO and aligned with bowtie allowing for multiple mappers and randomly assigning them to one location (-m 300 --best --strata) (Langmead et al., 2009). Reads were counted in 1 kb windows centred on repeat instances annotated by RepeatMasker and samples were normalised using DESeq2 (Love et al., 2014).

RNA-seq analysis of genes and repeats

For gene expression analysis, single-end 50 bp RNA-seq reads were aligned to the mm9 genome using QuasR (splicedAlignment=T, bowtie parameters -m 1 --best --strata) (Gaidatzis et al., 2015). Reads were counted in all exons and normalised across samples with DESeq2 (Love et al., 2014). Genes were considered differentially expressed if they changed more than two-fold with an adjusted p-value of p < 1e-5. Overrepresentation of gene ontology categories in selected gene sets was analysed using the GOstats R package (Falcon and Gentleman, 2007).

To measure expression at repetitive elements, 100 bp paired-end RNA-seq data were aligned using STAR (--outFilterMultimapNmax 300 --outMultimapperOrder Random) (Dobin et al., 2013). Repeat locations in

the mm9 genome were downloaded from the RepeatMasker track in the UCSC table browser (http://genome.ucsc.edu/) (Karolchik et al., 2004). For quantification of total RNA transcribed at repeats, multiple mappers were taken into account, but randomly assigned to only one region (i.e. quality score < 255). For comparison of published datasets, all of our own and published total RNA-seq samples were trimmed to 50 bp read length and aligned single-end using bowtie allowing up to 300 matches for reads, but with random assignment of multiple mappers to only one position in the genome (-m 300 --best --strata) (Langmead et al., 2009). Since individual repeat occurrences can then no longer be distinguished, repeats were collapsed to the family or name level for further analysis. Counts were normalised to library size and to the number of bases per repeat family that are mappable with these alignment parameters. Enrichment of transcript counts in repeat families was calculated for each chromatin mutant compared to the matching WT using the mean of all available replicates. Only uniquely mapping 100 bp paired-end reads (i.e. quality score = 255) were considered for comparing activated and silent IAPLTRs in neurons. Uniquely mapping RNA-seq reads were counted in all 13,810 IAPLTR occurrences in the genome and normalised to LTR width and library size. LTRs with a log2 fold-change (neurons TKO/WT) of more than 1 were considered activated, those with a log2 fold-change of less than 0.3 were considered silent. Reference sequences of IAPLTRs were downloaded from repbase (Bao et al., 2015). For calculation of conservation of IAPLTRs, we performed a multiple sequence alignment with kalign (-gapopen 1200.0 -gapextension 25.0 -tgpe 100.0 -bonus 283.0) (Lassmann and Sonnhammer, 2005) and calculated the percent identity to the reference sequence at each position. The JASPAR2016 and TFBSTools R packages were used to analyse the position and scores of TF motifs in the repeat reference sequences (Mathelier et al., 2016; Tan and Lenhard, 2016).

ATAC-seq data analysis

50 bp paired-end reads were trimmed for Illumina adapters using skewer and aligned to the mm9 genome using bowtie with the options -m 1 --best --strata (Jiang et al., 2014; Langmead et al., 2009). Accessible regions were called with MACS2 in paired-end mode (Zhang et al., 2008), using two replicates of WT and TKO each. A q-value cutoff of 10^{-9} or 10^{-8} was used to identify significant peaks in ES and neuron ATAC-seq, respectively. DESeq2 was used to normalise samples and identify significantly differential peaks between conditions (at least twofold change, adjusted p-value < 0.05) (Love et al., 2014). ATAC-seq signal and DNase-seq signal were compared across all regions identified either as ATAC-seq peaks called by MACS2 or as DNase hypersensitive sites called with a sliding window approach as described previously (Domcke et al., 2015). For analysis of chromatin accessibility at CRE motifs (TGACGTCA), samples were realigned with bowtie allowing up to 300 matches for reads, but with random assignment of multiple mappers to only one position in the genome (-m 300 --best --strata), to retrieve signal in repetitive regions.

Motif and hexamer enrichment

De novo and known TF motifs in TKO-specific ATAC-seq sites were identified by homer2 using shared sites (abs(log2 fold-change TKO/WT) < 0.3) that contain the same CpG content as the TKO-specific sites as background (Heinz et al., 2010). For IAPLTRs, activated IAPLTRs were used as foreground and all silent IAPLTRs (abs(log2 fold-change TKO/WT) < 0.3) with matched CpG content were used as background.

For identification of enriched hexamers in TKO-specific ATAC-seq sites, the background was defined as a random sample of four times as many accessible regions that are shared between WT and TKO (abs(log2 fold-change TKO/WT) < 0.3). These regions were selected to have identical CpG and GC content as the foreground sequences.

Linear model

TF motifs present in the IAPLTR1 and 1a reference sequences were identified with the TFBStools and JASPAR2016 R packages (score > 90%, absScore > 6) (Mathelier et al., 2016; Tan and Lenhard, 2016). For these 134 TF motifs the sum of the score of all motif occurrences with a score of at least 75% was calculated for each of the 3,456 LTR instances. We also tested using the maximum score or the number of occurrences of each motif, but these models performed slightly worse than with the sum of the score. The number of mismatches, deletions and insertions relative to the reference sequence were downloaded from the UCSC table browser (Karolchik et al., 2004), the CpG and GC content were calculated and these values were added to a 139 x 3,456 regressor matrix. Elastic net regression was performed using the glmnet R package (Friedman et al., 2010), with an alpha of 0.5. Five-fold cross-validation was used to choose the tuning parameter lambda.

Published data sets

Whole genome bisulfite sequencing data was downloaded from GEO for WT ES cells (GSM748786) (Stadler et al., 2011). The following total RNA-seq datasets were obtained from GEO: terminal neurons (GSM687306) (Tippmann et al., 2012), ES *cSetdb1 cDnmt1* KO (GSM2059172/3) and matching WT (GSM2059170/1) (Sharif et al., 2016), ES *Kap1* KO (GSM1032183) and matching WT (GSM1032182) (Rowe et al., 2013), NPC *Kap1* KO (GSM1119765/6/7) and matching WT (GSM1119762/3/4) (Fasching et al., 2015), PGC E13.5 *Setdb1* KO (GSM1477419/20) and matching WT (GSM1477414) (Liu et al., 2014), MEF *Tp53* Dnmt1 KO (GSM1089794) and MEF *Tp53* KO (GSM1089793) (Reddington et al., 2013), P5 mouse cortex *cUhrf1* KO (GSM2241736/9) and matching heterozygote (GSM2241735/7/8) (Ramesh et al., 2016). The following ChIP-seq data sets were obtained from GEO: H3K9me3 in ES (GSM1375155) and NP (GSM1375164) (Bulut-Karslioglu et al., 2014).

6. References

Adams, C.C., and Workman, J.L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Mol. Cell. Biol. *15*, 1405–1421.

Aravin, A.A., Hannon, G.J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science *318*, 761–764.

Auclair, G., and Weber, M. (2012). Mechanisms of DNA methylation and demethylation in mammals. Biochimie *94*, 2202–2211.

Audouard, E., Schakman, O., Ginion, A., Bertrand, L., Gailly, P., and Clotman, F. (2013). The Onecut transcription factor HNF-6 contributes to proper reorganization of Purkinje cells during postnatal cerebellum development. Mol. Cell. Neurosci. *56*, 159–168.

Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A.J., Funnell, A.P.W., Goncalves, A., et al. (2014). Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. eLife *3*, e02626.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11.

Bardet, A.F., Steinmann, J., Bafna, S., Knoblich, J.A., Zeitlinger, J., and Stark, A. (2013). Identification of transcription factor binding sites from ChIP-seq data at high resolution. Bioinformatics *29*, 2705–2713.

Baubec, T., and Schübeler, D. (2014). Genomic patterns and context specific interpretation of DNA methylation. Curr. Opin. Genet. Dev. *25*, 85–92.

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

Bednarik, D.P., Duckett, C., Kim, S.U., Perez, V.L., Griffis, K., Guenthner, P.C., and Folks, T.M. (1991). DNA CpG methylation inhibits binding of NF-kappa B proteins to the HIV-1 long terminal repeat cognate DNA motifs. New Biol. *3*, 969–976.

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

Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell *98*, 387–396.

Benbrook, D.M., and Jones, N.C. (1994). Different binding specificities and transactivation of variant CRE's by CREB complexes. Nucleic Acids Res. *22*, 1463–1469.

Bestor, T.H. (1990). DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. Philos. Trans. R. Soc. Lond., B, Biol. Sci. *326*, 179–187.

Bestor, T.H. (2003). Cytosine methylation mediates sexual conflict. Trends Genet. *19*, 185–190.

Bestor, T.H., Edwards, J.R., and Boulard, M. (2015). Notes on the role of dynamic DNA methylation in mammalian development. Proc. Natl. Acad. Sci. USA *112*, 6796–6799.

Bibel, M., Richter, J., Lacroix, E., and Barde, Y.-A. (2007). Generation of a defined and uniform population of CNS progenitors and neurons from mouse embryonic stem cells. Nat. Protoc. *2*, 1034–1043.

Biggin, M.D. (2011). Animal transcription networks as highly connected, quantitative continua. Dev. Cell *21*, 611–626.

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

Bird, A.P. (1995). Gene number, noise reduction and biological complexity. Trends Genet. *11*, 94–100.

Blattler, A., and Farnham, P.J. (2013). Cross-talk between site-specific transcription factors and DNA methylation states. J. Biol. Chem. *288*, 34287–34294.

Bleckmann, S.C., Blendy, J.A., Rudolph, D., Monaghan, A.P., Schmid, W., and Schütz, G. (2002). Activating transcription factor 1 and CREB are important for cell survival during early mouse development. Mol. Cell. Biol. *22*, 1919–1925.

Boller, S., Ramamoorthy, S., Akbas, D., Nechanitzky, R., Burger, L., Murr, R., Schübeler, D., and Grosschedl, R. (2016). Pioneering Activity of the C-Terminal Domain of EBF1 Shapes the Chromatin Landscape for B Cell Programming. Immunity *44*, 527–541.

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.

Boulard, M., Edwards, J.R., and Bestor, T.H. (2015). FBXL10 protects Polycomb-bound genes from hypermethylation. Nat. Genet. *47*, 497-85.

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.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. (1994). Sp1 elements protect a CpG island from de novo methylation. Nature *371*, 435–438.

Brûlet, P., Kaghad, M., Xu, Y.S., Croissant, O., and JACOB, F. (1983). Early differential tissue expression of transposon-like repetitive DNA sequences of the mouse. Proc. Natl. Acad. Sci. USA *80*, 5641–5645.

Bucceri, A., Kapitza, K., and Thoma, F. (2006). Rapid accessibility of nucleosomal DNA in yeast on a second time scale. EMBO J. *25*, 3123–3132.

Buck-Koehntop, B.A., Stanfield, R.L., Ekiert, D.C., Martinez-Yamout, M.A., Dyson, H.J., Wilson, I.A., and Wright, P.E. (2012). Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso. Proc. Natl. Acad. Sci. USA *109*, 15229–15234.

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.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. *109*, 21.29.1–.29.9.

Bulut-Karslioglu, A., La Rosa-Velázquez, De, I.A., Ramirez, F., Barenboim, M., Onishi-Seebacher, M., Arand, J., Galán, C., Winter, G.E., Engist, B., Gerle, B., et al. (2014). Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. Mol. Cell *55*, 277–290.

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

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. USA *97*, 6481–6486.

Caselli, E., Benedetti, S., Gentili, V., Grigolato, J., and Di Luca, D. (2012). Short communication: activating transcription factor 4 (ATF4) promotes HIV type 1 activation. AIDS Res. Hum. Retroviruses *28*, 907–912.

Cedar, H. (1988). DNA methylation and gene activity. Cell *53*, 3–4.

Cedar, H., Stein, R., Gruenbaum, Y., Naveh-Many, T., Sciaky-Gallili, N., and Razin, A. (1983). Effect of DNA methylation on gene expression. Cold Spring Harb. Symp. Quant. Biol. *47 Pt 2*, 605–609.

Celona, B., Weiner, A., Di Felice, F., Mancuso, F.M., Cesarini, E., Rossi, R.L., Gregory, L., Baban, D., Rossetti, G., Grianti, P., et al. (2011). Substantial histone reduction modulates genomewide nucleosomal occupancy and global transcriptional output. PLoS Biol. *9*, e1001086.

Chen, H., Tian, Y., Shu, W., Bo, X., and Wang, S. (2012). Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. PLoS ONE *7*, e41374.

Chen, T., Hevi, S., Gay, F., Tsujimoto, N., He, T., Zhang, B., Ueda, Y., and Li, E. (2007). Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells. Nat. Genet. *39*, 391–396.

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.

Christy, R.J., and Huang, R.C. (1988). Functional analysis of the long terminal repeats of intracisternal A-particle genes: sequences within the U3 region determine both the efficiency and direction of promoter activity. Mol. Cell. Biol. *8*, 1093–1102.

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

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.

Cooper, D.N., and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. Hum. Genet. *78*, 151–155.

Dantas Machado, A.C., Zhou, T., Rao, S., Goel, P., Rastogi, C., Lazarovici, A., Bussemaker, H.J., and Rohs, R. (2015). Evolving insights on how cytosine methylation affects protein–DNA binding. Brief. Funct. Genomics *14*, 61–73.

Dellino, G.I., Schwartz, Y.B., Farkas, G., McCabe, D., Elgin, S.C.R., and Pirrotta, V. (2004). Polycomb silencing blocks transcription initiation. Mol. Cell *13*, 887–893.

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

Dodge, J.E., Kang, Y.-K., Beppu, H., Lei, H., and Li, E. (2004). Histone H3-K9 methyltransferase ESET is essential for early development. Mol. Cell. Biol. *24*, 2478–2486.

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 *528*, 575–579.

Dong, K.B., Maksakova, I.A., Mohn, F., Leung, D., Appanah, R., Lee, S., Yang, H.W., Lam, L.L., Mager, D.L., Schübeler, D., et al. (2008). DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity. EMBO J. *27*, 2691–2701.

Dupressoir, A., and Heidmann, T. (1996). Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice. Mol. Cell. Biol. *16*, 4495–4503.

Ecco, G., Cassano, M., Kauzlaric, A., Duc, J., Coluccio, A., Offner, S., Imbeault, M., Rowe, H.M., Turelli, P., and Trono, D. (2016). Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. Dev. Cell *36*, 611–623.

Elgin, S.C. (1981). DNAase I-hypersensitive sites of chromatin. Cell *27*, 413–415.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Ernst, J., and Kellis, M. (2013). Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. Genome Res. *23*, 1142–1154.

Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S., and Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat. Biotech. *34*, 1180–1190.

Evans, M.J., and Scarpulla, R.C. (1990). NRF-1: a trans-activator of nuclear-encoded respiratory genes in animal cells. Genes Dev. *4*, 1023–1034.

Falcon, S., and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. Bioinformatics *23*, 257–258.

Fasching, L., Kapopoulou, A., Sachdeva, R., Petri, R., Jönsson, M.E., Männe, C., Turelli, P., Jern, P., Cammas, F., Trono, D., et al. (2015). TRIM28 represses transcription of endogenous retroviruses in neural progenitor cells. Cell Rep. *10*, 20–28.

Fazio, I.K., Bolger, T.A., and Gill, G. (2001). Conserved regions of the Drosophila erect wing protein contribute both positively and negatively to transcriptional activity. J. Biol. Chem. *276*, 18710–18716.

Feldmann, A., Ivanek, 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.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. Proc. Natl. Acad. Sci. USA *107*, 8689–8694.

Feng, Y.Q., Seibler, J., Alami, R., Eisen, A., Westerman, K.A., Leboulch, P., Fiering, S., and Bouhassira, E.E. (1999). Site-specific chromosomal integration in mammalian cells: highly efficient CRE recombinase-mediated cassette exchange. J. Mol. Biol. *292*, 779–785.

Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell *143*, 212–224.

Filippova, G.N. (2008). Genetics and epigenetics of the multifunctional protein CTCF. Curr. Top. Dev. Biol. *80*, 337–360.

FitzGerald, P.C., Sturgill, D., Shyakhtenko, A., Oliver, B., and Vinson, C. (2006). Comparative genomics of Drosophila and human core promoters. Genome Biol. *7*, R53.

Flavahan, W.A., Drier, Y., Liau, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suvà, M.L., and Bernstein, B.E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature *529*, 110–114.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

Frost, J., Monk, D., Moschidou, D., Guillot, P.V., Stanier, P., Minger, S.L., Fisk, N.M., Moore, H.D., and Moore, G.E. (2011). The effects of culture on genomic imprinting profiles in human embryonic and fetal mesenchymal stem cells. Epigenetics *6*, 52–62.

Gaidatzis, D., Lerch, A., Hahne, F., and Stadler, M.B. (2015). QuasR: quantification and annotation of short reads in R. Bioinformatics *31*, 1130–1132.

Gal-Yam, E.N., Egger, G., Iniguez, L., Holster, H., Einarsson, S., Zhang, X., Lin, J.C., Liang, G., Jones, P.A., and Tanay, A. (2008). Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proc. Natl. Acad. Sci. USA *105*, 12979–12984.

Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. Genes Dev. *30*, 881–891.

Girton, J.R., and Johansen, K.M. (2008). Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila. Adv. Genet. *61*, 1–43.

Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. Annu. Rev. Biochem. *74*, 481–514.

Göke, J., Lu, X., Chan, Y.-S., Ng, H.-H., Ly, L.-H., Sachs, F., and Szczerbinska, I. (2015). Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. Cell Stem Cell *16*, 135–141.

Grant, C., Jain, P., Nonnemacher, M., Flaig, K.E., Irish, B., Ahuja, J., Alexaki, A., Alefantis, T., and Wigdahl, B. (2006). AP-1-directed human T cell leukemia virus type 1 viral gene expression during monocytic differentiation. J. Leukoc. Biol. *80*, 640–650.

Grow, E.J., Flynn, R.A., Chavez, S.L., Bayless, N.L., Wossidlo, M., Wesche, D.J., Martin, L., Ware, C.B., Blish, C.A., Chang, H.Y., et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature *522*, 221–225.

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.

Hai, T., and Hartman, M.G. (2001). The molecular biology and nomenclature of the activating transcription factor/cAMP responsive element binding family of transcription factors: activating transcription factor proteins and homeostasis. Gene *273*, 1–11.

Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J., and Surani, M.A. (2002). Epigenetic reprogramming in mouse primordial germ cells. Mech. Dev. *117*, 15–23.

Han, L., Lin, I.G., and Hsieh, C.L. (2001). Protein binding protects sites on stable episomes and in the chromosome from de novo methylation. Mol. Cell. Biol. *21*, 3416–3424.

Han, M., and Grunstein, M. (1988). Nucleosome loss activates yeast downstream promoters in vivo. Cell *55*, 1137–1145.

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.

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.

Heitz, E. (1928). Das Heterochromatin der Moose. Jahrb. Wiss. Bot. *69*, 762–818.

Hendrich, B., Guy, J., Ramsahoye, B., Wilson, V.A., and Bird, A. (2001). Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. Genes Dev. *15*, 710–723.

Hermann, A., Gowher, H., and Jeltsch, A. (2004). Biochemistry and biology of mammalian DNA methyltransferases. Cell. Mol. Life Sci. *61*, 2571–2587.

Herold, M., Bartkuhn, M., and Renkawitz, R. (2012). CTCF: insights into insulator function during development. Development *139*, 1045–1057.

Hickey, D.A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. Genetics *101*, 519–531.

Ho, J.W.K., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.-A., Minoda, A., Tolstorukov, M.Y., Appert, A., et al. (2014). Comparative analysis of metazoan chromatin organization. Nature *512*, 449–452.

Holwerda, S.J.B., and de Laat, W. (2013). CTCF: the protein, the binding partners, the binding sites and their chromatin loops. Philos. Trans. R. Soc. Lond., B, Biol. Sci. *368*, 20120369–20120369.

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.

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.

Huo, L., and Scarpulla, R.C. (2001). Mitochondrial DNA Instability and Peri-Implantation Lethality Associated with Targeted Disruption of Nuclear Respiratory Factor 1 in Mice. Mol. Cell. Biol. *21*, 644–654.

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.

Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature *16*, 669.

Ito, Y., Nativio, R., and Murrell, A. (2013). Induced DNA demethylation can reshape chromatin topology at the IGF2-H19 locus. Nucleic Acids Res. *41*, 5290–5302.

Iwafuchi-Doi, M., and Zaret, K.S. (2014). Pioneer transcription factors in cell reprogramming. Genes Dev. *28*, 2679–2692.

Iwamoto, K., Bundo, M., Ueda, J., Oldham, M.C., Ukai, W., Hashimoto, E., Saito, T., Geschwind, D.H., and Kato, T. (2011). Neurons show distinctive DNA methylation profile and higher interindividual variations compared with non-neurons. Genome Res. *21*, 688–696.

Iyaguchi, D., Yao, M., Watanabe, N., Nishihira, J., and Tanaka, I. (2007). DNA recognition mechanism of the ONECUT homeodomain of transcription factor HNF-6. Structure *15*, 75–83.

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-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.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.

Jacquemin, P., Durviaux, S.M., Jensen, J., Godfraind, C., Gradwohl, G., Guillemot, F., Madsen, O.D., Carmeliet, P., Dewerchin, M., Collen, D., et al. (2000). Transcription factor hepatocyte nuclear factor 6 regulates pancreatic endocrine cell differentiation and controls expression of the proendocrine gene ngn3. Mol. Cell. Biol. *20*, 4445–4454.

Jaenisch, R., and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat. Genet. *33 Suppl*, 245–254.

Jähner, D., Stuhlmann, H., Stewart, C.L., Harbers, K., Löhler, J., Simon, I., and Jaenisch, R. (1982). De novo methylation and expression of retroviral genomes during mouse embryogenesis. Nature *298*, 623–628.

Jermann, P., Hoerner, L., Burger, L., and Schübeler, D. (2014). Short sequences can efficiently recruit histone H3 lysine 27 trimethylation in the absence of enhancer activity and DNA methylation. Proc. Natl. Acad. Sci. USA *111*, E3415–E3421.

Jessen, W.J., Dhasarathy, A., Hoose, S.A., Carvin, C.D., Risinger, A.L., and Kladde, M.P. (2004). Mapping chromatin structure in vivo using DNA methyltransferases. Methods *33*, 68–80.

Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics *15*, 182.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.-H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. *43*, 264–268.

Johnson, P.F., and McKnight, S.L. (1989). Eukaryotic transcriptional regulatory proteins. Annu. Rev. Biochem. *58*, 799–839.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genetics *13*, 484–492.

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 chimeric transcripts in mESCs. Cell Stem Cell *8*, 676–687.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, D493–D496.

Kelly, T.K., Liu, Y., Lay, F.D., Liang, G., Berman, B.P., and Jones, P.A. (2012). Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. Genome Res. *22*, 2497–2506.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Klose, R.J., and Bird, A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends Biochem. Sci. *31*, 89–97.

Knezetic, J.A., and Luse, D.S. (1986). The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. Cell *45*, 95–104.

Korf, I., Fan, Y., and Strome, S. (1998). The Polycomb group in Caenorhabditis elegans and maternal control of germline development. Development *125*, 2469–2478.

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

Krebs, A.R., Dessus-Babus, S., Burger, L., and Schübeler, D. (2014). High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. eLife *3*, e04094.

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.

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., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lannoy, V.J., Bürglin, T.R., Rousseau, G.G., and Lemaigre, F.P. (1998). Isoforms of hepatocyte nuclear factor-6 differ in DNA-binding properties, contain a bifunctional homeodomain, and define the new ONECUT class of homeodomain proteins. J. Biol. Chem. *273*, 13552–13562.

Lassmann, T., and Sonnhammer, E.L.L. (2005). Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics *6*, 298.

Le, T., Kim, K.-P., Fan, G., and Faull, K.F. (2011). A sensitive mass spectrometry method for simultaneous quantification of DNA methylation and hydroxymethylation levels in biological samples. Anal. Biochem. *412*, 203–209.

Lechner, M., Marz, M., Ihling, C., Sinz, A., Stadler, P.F., and Krauss, V. (2013). The correlation of genome size and DNA methylation rate in metazoans. Theory Biosci. *132*, 47–60.

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.

Leung, D.C., and Lorincz, M.C. (2012). Silencing of endogenous retroviruses: when and why do histone marks predominate? Trends Biochem. Sci. *37*, 127–133.

Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nature Rev. Genetics *15*, 453–468.

Levo, M., Avnit-Sagi, T., Lotan-Pompan, M., Kalma, Y., Weinberger, A., Yakhini, Z., and Segal, E. (2017). Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. Mol. Cell *65*, 604–617.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707–719.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. Nature *366*, 362–365.

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, G., Levitus, M., Bustamante, C., and Widom, J. (2005). Rapid spontaneous accessibility of nucleosomal DNA. Nat. Struct. Mol. Biol. *12*, 46–53.

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. *47*, 469-78.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. *43*, 1091–1097.

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.

Liu, S., Brind'Amour, J., Karimi, M.M., Shirane, K., Bogutz, A., Lefebvre, L., Sasaki, H., Shinkai, Y., and Lorincz, M.C. (2014). Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. Genes Dev. *28*, 2041–2055.

Liu, X.S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R.A., and Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. Cell *167*, 233–247.

Liu, X., Lee, C.-K., Granek, J.A., Clarke, N.D., and Lieb, J.D. (2006). Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. Genome Res. *16*, 1517–1528.

Lonze, B.E., and Ginty, D.D. (2002). Function and regulation of CREB family transcription factors in the nervous system. Neuron *35*, 605–623.

Lorch, Y., Maier-Davis, B., and Kornberg, R.D. (2010). Mechanism of chromatin remodeling. Proc. Natl. Acad. Sci. USA *107*, 3458–3462.

Lorincz, M.C., Schubeler, D., and Groudine, M. (2001). Methylation-mediated proviral silencing is associated with MeCP2 recruitment and localized histone H3 deacetylation. Mol. Cell. Biol. *21*, 7913–7922.

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.

Loyola, A., LeRoy, G., Wang, Y.H., and Reinberg, D. (2001). Reconstitution of recombinant chromatin establishes a requirement for histone-tail modifications during chromatin assembly and transcription. Genes Dev. *15*, 2837–2851.

Lupien, M., Eeckhoute, J., Meyer, C.A., Wang, Q., Zhang, Y., Li, W., Carroll, J.S., Liu, X.S., and Brown, M. (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell *132*, 958–970.

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

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.

Madrigal, P. (2015). On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions. Front. Bioeng. Biotechnol. *3*, 144.

Maksakova, I.A., Goyal, P., Bullwinkel, J., Brown, J.P., Bilenky, M., Mager, D.L., Singh, P.B., and Lorincz, M.C. (2011). H3K9me3-binding proteins are dispensable for SETDB1/H3K9me3-dependent retroviral silencing. Epigenetics Chromatin *4*, 12.

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*, e2.

Maniatis, T., Goodbourn, S., and Fischer, J.A. (1987). Regulation of inducible and tissue-specific gene expression. Science *236*, 1237–1245.

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 CEBPBlATF4 heterodimer that is active in vivo. Genome Res. *23*, 988–997.

Marchi, E., Kanapin, A., Magiorkinis, G., and Belshaw, R. (2014). Unfixed endogenous retroviral insertions in the human population. J. Virol. *88*, 9529–9537.

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.-Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *44*, D110–D115.

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.

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. *12*, 1184–1195.

Mayall, T.P., Sheridan, P.L., Montminy, M.R., and Jones, K.A. (1997). Distinct roles for P-CREB and LEF-1 in TCR alpha enhancer assembly and activation on chromatin templates in vitro. Genes Dev. *11*, 887–899.

Mayr, B., and Montminy, M. (2001). Transcriptional regulation by the phosphorylation-dependent factor CREB. Nat. Rev. Mol. Cell Biol. *2*, 599–609.

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.

Miller, J.A., and Widom, J. (2003). Collaborative competition mechanism for gene activation in vivo. Mol. Cell. Biol. *23*, 1623–1632.

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.

Molaro, A., and Malik, H.S. (2016). Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. Curr. Opin. Genet. Dev. *37*, 51–58.

Moon, H., Filippova, G., Loukinov, D., Pugacheva, E., Chen, Q., Smith, S.T., Munhall, A., Grewe, B., Bartkuhn, M., Arnold, R., et al. (2005). CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. EMBO Rep. *6*, 165–170.

Moore, J.M., Rabaia, N.A., Smith, L.E., Fagerlie, S., Gurley, K., Loukinov, D., Disteche, C.M., Collins, S.J., Kemp, C.J., Lobanenkov, V.V., et al. (2012). Loss of maternal CTCF is associated with peri-implantation lethality of Ctcf null embryos. PLoS ONE *7*, e34915.

Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M.P., and Chambon, P. (1981). The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. Nucleic Acids Res. *9*, 6047–6068.

Mukhopadhyay, R., Yu, W., Whitehead, J., Xu, J., Lezcano, M., Pack, S., Kanduri, C., Kanduri, M., Ginjala, V., Vostrov, A., et al. (2004). The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. Genome Res. *14*, 1594–1602.

Nakahashi, H., Kwon, K.-R.K., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A., et al. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. Cell Rep. *3*, 1678–1689.

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.

Nan, X., Tate, P., Li, E., and Bird, A. (1996). DNA methylation specifies chromosomal localization of MeCP2. Mol. Cell. Biol. *16*, 414–421.

Naveh-Many, T., and Cedar, H. (1981). Active gene sequences are undermethylated. Proc. Natl. Acad. Sci. USA *78*, 4246–4250.

Nehlin, J.O., Carlberg, M., and Ronne, H. (1992). Yeast SKO1 gene encodes a bZIP protein that binds to the CRE motif and acts as a repressor of transcription. Nucleic Acids Res. *20*, 5271–5278.

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.

Ni, X., Zhang, Y.E., Nègre, N., Chen, S., Long, M., and White, K.P. (2012). Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. PLoS Biol. *10*, e1001420.

Nichols, J., and Smith, A. (2009). Naive and primed pluripotent states. Cell Stem Cell *4*, 487–492.

O'Malley, R.C., Huang, S.-S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell *165*, 1280–1292.

Oakes, C.C., La Salle, S., Trasler, J.M., and Robaire, B. (2009). Restriction digestion and real-time PCR (qAMP). Methods Mol. Biol. *507*, 271–280.

Ohlsson, R., Renkawitz, R., and Lobanenkov, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends Genet. *17*, 520–527.

Padeken, J., Zeller, P., and Gasser, S.M. (2015). Repeat DNA in genome organization and stability. Curr. Opin. Genet. Dev. *31*, 12–19.

Panning, B., and Jaenisch, R. (1996). DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes Dev. *10*, 1991–2002.

Pearson, E.C., Bates, D.L., Prospero, T.D., and Thomas, J.O. (1984). Neuronal nuclei and glial nuclei from mammalian cerebral cortex. Nucleosome repeat lengths, DNA contents and H1 contents. Eur. J. Biochem. *144*, 353–360.

Peters, A.H.F.M., Kubicek, S., Mechtler, K., O'Sullivan, R.J., Derijck, A.A.H.A., Perez-Burgos, L., Kohlmaier, A., Opravil, S., Tachibana, M., Shinkai, Y., et al. (2003). Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. Mol. Cell *12*, 1577–1589.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194–1211.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. *21*, 447–455.

Polach, K.J., and Widom, J. (1995). Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. J. Mol. Biol. *254*, 130–149.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101–1105.

Poustka, A.J., Kühn, A., Radosavljevic, V., Wellenreuther, R., Lehrach, H., and Panopoulou, G. (2004). On the origin of the chordate central nervous system: expression of onecut in the sea urchin embryo. Evol. Dev. *6*, 227–236.

Prendergast, G.C., and Ziff, E.B. (1991). Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. Science *251*, 186–189.

Ramesh, V., Bayam, E., Cernilogar, F.M., Bonapace, I.M., Schulze, M., Riemenschneider, M.J., Schotta, G., and Götz, M. (2016). Loss of Uhrf1 in neural stem cells leads to activation of retroviral elements and delayed neurodegeneration. Genes Dev. *30*, 2199–2212.

Ray, A., Rahbari, R., and Badge, R.M. (2011). IAP display: a simple method to identify mouse strain specific IAP insertions. Mol. Biotechnol. *47*, 243–252.

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.

Renda, M., Baglivo, I., Burgess-Beusse, B., Esposito, S., Fattorusso, R., Felsenfeld, G., and Pedone, P.V. (2007). Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. J. Biol. Chem. *282*, 33336–33345.

Richmond, T.J., and Davey, C.A. (2003). The structure of DNA in the nucleosome core. Nature *423*, 145–150.

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

Rohde, C., Zhang, Y., Reinhardt, R., and Jeltsch, A. (2010). BISMA--fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. BMC Bioinformatics *11*, 230.

Rowe, H.M., and Trono, D. (2011). Dynamic control of endogenous retroviruses during development. Virology *411*, 273–287.

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.

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.

Schaefer, L., Engman, H., and Miller, J.B. (2000). Coding sequence, chromosomal localization, and expression pattern of Nrf1: the mouse homolog of Drosophila erect wing. Mamm. Genome *11*, 104–110.

Schmitges, F.W., Radovani, E., Najafabadi, H.S., Barazandeh, M., Campitelli, L.F., Yin, Y., Jolma, A., Zhong, G., Guo, H., Kanagalingam, T., et al. (2016). Multiparameter functional diversity of human C2H2 zinc finger proteins. Genome Res. *26*, 1742–1752.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., et al. (2013). Patterns of population epigenomic diversity. Nature *495*, 193–198.

Schoenherr, C.J., Levorse, J.M., and Tilghman, S.M. (2003). CTCF maintains differential methylation at the Igf2/H19 locus. Nat. Genet. *33*, 66–69.

Schubeler, 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.

Schumacher, M.A., Goodman, R.H., and Brennan, R.G. (2000). The structure of a CREB bZIP.somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding. J. Biol. Chem. *275*, 35242–35247.

Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321–326.

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.

Shaknovich, R., Cerchietti, L., Tsikitas, L., Kormaksson, M., De, S., Figueroa, M.E., Ballon, G., Yang, S.N., Weinhold, N., Reimers, M., et al. (2011). DNA methyltransferase 1 and DNA methylation patterning contribute to germinal center B-cell differentiation. Blood *118*, 3559–3569.

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. Stem Cell 1–15.

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. Biotech. *32*, 171–178.

Shilatifard, A. (2006). Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. Annu. Rev. Biochem. *75*, 243–269.

Singh, P., Wu, X., Lee, D.-H., Li, A.X., Rauch, T.A., Pfeifer, G.P., Mann, J.R., and Szabó, P.E. (2011). Chromosome-wide analysis of parental allele-specific chromatin and DNA methylation. Mol. Cell. Biol. *31*, 1757–1770.

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., Tomizawa, S.-I., Krueger, F., Ruf, N., Carli, N., Segonds-Pichon, A., Sato, S., Hata, K., Andrews, S.R., and Kelsey, G. (2011). Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. Nat. Genet. *43*, 811–814.

Smith, B., Fang, H., Pan, Y., Walker, P.R., Famili, A.F., and Sikorska, M. (2007). Evolution of motif variants and positional bias of the cyclic-AMP response element. BMC Evol. Biol. *7 Suppl 1*, S15.

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.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genetics *13*, 613–626.

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., Ivanek, 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.

Stein, R., Razin, A., and Cedar, H. (1982). In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. Proc. Natl. Acad. Sci. USA *79*, 3418–3422.

Sung, M.-H., Guertin, M.J., Baek, S., and Hager, G.L. (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol. Cell *56*, 275–285.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genetics *9*, 465–476.

Svaren, J., and Hörz, W. (1997). Transcription factors vs nucleosomes: regulation of the PHO5 promoter in yeast. Trends Biochem. Sci. *22*, 93–97.

Szabó, P., Tang, S.H., Rentsendorj, A., Pfeifer, G.P., and Mann, J.R. (2000). Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function. Curr. Biol. *10*, 607–610.

Szabó, P.E., Tang, S.-H.E., Silva, F.J., Tsark, W.M.K., and Mann, J.R. (2004). Role of CTCF binding sites in the Igf2/H19 imprinting control region. Mol. Cell. Biol. *24*, 4791–4800.

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.

Tan, G., and Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics *32*, 1555–1556.

Tan, S.-L., Nishi, M., Ohtsuka, T., Matsui, T., Takemoto, K., Kamio-Miura, A., Aburatani, H., Shinkai, Y., and Kageyama, R. (2012). Essential roles of the histone methyltransferase ESET in the epigenetic control of neural progenitor cells during development. Development *139*, 3806–3816.

Tan, X., Xu, X., Elkenani, M., Smorag, L., Zechner, U., Nolte, J., Engel, W., and Pantakani, D.V.K. (2013). Zfp819, a novel KRAB-zinc finger protein, interacts with KAP1 and functions in genomic integrity maintenance of mouse embryonic stem cells. Stem Cell Res. *11*, 1045–1059.

Tanay, A., O'Donnell, A.H., Damelin, M., and Bestor, T.H. (2007). Hyperconserved CpG domains underlie Polycomb-binding sites. Proc. Natl. Acad. Sci. USA *104*, 5521–5526.

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.

Thoma, E.C., Wischmeyer, E., Offen, N., Maurus, K., Sirén, A.-L., Schartl, M., and Wagner, T.U. (2012). Ectopic expression of neurogenin 2 alone is sufficient to induce differentiation of embryonic stem cells into mature neurons. PLoS ONE *7*, e38651.

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.

Tierney, R.J., Kirby, H.E., Nagra, J.K., Desmond, J., Bell, A.I., and Rickinson, A.B. (2000). Methylation of transcription factor binding sites in the Epstein-Barr virus latent cycle promoter Wp coincides with promoter down-regulation during virus-induced B-cell transformation. J. Virol. *74*, 10468–10479.

Tippin, D.B., and Sundaralingam, M. (1997). Nine polymorphic crystal structures of d(CCGGGCCCGG), d(CCGGGCCm5CGG), d(Cm5CGGGCCm5CGG) and d(CCGGGCC(Br)5CGG) in three different conformations: effects of spermine binding and methylation on the bending and condensation of A-DNA. J. Mol. Biol. *267*, 1171–1185.

Tippmann, S.C., Ivanek, R., Gaidatzis, D., Schöler, A., Hoerner, L., van Nimwegen, E., Stadler, P.F., Stadler, M.B., and Schübeler, D. (2012). Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. Mol. Syst. Biol. *8*.

Todeschini, A.-L., Georges, A., and Veitia, R.A. (2014). Transcription factors: specific DNA binding and specific gene regulation. Trends Genet. *30*, 211–219.

Toufaily, C., Lokossou, A.G., Vargas, A., Rassart, É., and Barbeau, B. (2015). A CRE/AP-1-like motif is essential for induced syncytin-2 expression and fusion in human trophoblast-like model. PLoS ONE *10*, e0121468.

Tschiersch, B., Hofmann, A., Krauss, V., Dorn, R., Korge, G., and Reuter, G. (1994). The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes. EMBO J. *13*, 3822–3831.

Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S.-I., 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.

Tudor, M., Akbarian, S., Chen, R.Z., and Jaenisch, R. (2002). Transcriptional profiling of a mouse model for Rett syndrome reveals subtle transcriptional changes in the brain. Proc. Natl. Acad. Sci. USA *99*, 15536–15541.

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.

Voss, T.C., and Hager, G.L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat. Rev. Genet. *15*, 69–81.

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*, R87.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012a). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. *22*, 1680–1688.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012b). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. *22*, 1798–1812.

Wang, L., Chen, J., Wang, C., Uusküla-Reimand, L., Chen, K., Medina-Rivera, A., Young, E.J., Zimmermann, M.T., Yan, H., Sun, Z., et al. (2014). MACE: model based analysis of ChIP-exo. Nucleic Acids Res. *42*, e156–e156.

Wang, Y., Xiao, M., Chen, X., Chen, L., Xu, Y., Lv, L., Wang, P., Yang, H., Ma, S., Lin, H., et al. (2015). WT1 recruits TET2 to regulate its target gene expression and suppress leukemia cell proliferation. Mol. Cell *57*, 662–673.

Watt, F., and Molloy, P.L. (1988). Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Genes Dev. *2*, 1136–1143.

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.

Weih, F., Nitsch, D., Reik, A., Schütz, G., and Becker, P.B. (1991). Analysis of CpG methylation and genomic footprinting at the tyrosine aminotransferase gene: DNA methylation alone is not sufficient to prevent protein binding in vivo. EMBO J. *10*, 2559–2567.

Weintraub, H., and Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. Science *193*, 848–856.

Weirauch, M.T., and Hughes, T.R. (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Subcell. Biochem. *52*, 25–73.

Wildschutte, J.H., Williams, Z.H., Montesion, M., Subramanian, R.P., Kidd, J.M., and Coffin, J.M. (2016). Discovery of unfixed endogenous retrovirus insertions in diverse human populations. Proc Natl Acad Sci USA *113*, E2326–E2334.

Workman, J.L., and Kingston, R.E. (1992). Nucleosome core displacement in vitro via a metastable transcription factor-nucleosome complex. Science *258*, 1780–1784.

Wu, C., Wong, Y.C., and Elgin, S.C. (1979). The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. Cell *16*, 807–814.

Wunderlich, Z., and Mirny, L.A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. *25*, 434–440.

Yin, J.C., Wallach, J.S., Wilder, E.L., Klingensmith, J., Dang, D., Perrimon, N., Zhou, H., Tully, T., and Quinn, W.G. (1995). A Drosophila CREB/CREM homolog encodes multiple isoforms, including a cyclic AMP-dependent protein kinase-responsive transcriptional activator and antagonist. Mol. Cell. Biol. *15*, 5123–5130.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. *13*, 335–340.

Yuan, G.-C., Liu, Y.-J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Rando, O.J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. Science *309*, 626–630.

Zemach, A., and Zilberman, D. (2010). Evolution of eukaryotic DNA methylation and the pursuit of safer sex. Curr. Biol. *20*, R780–R785.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916–919.

Zemojtel, T., Kielbasa, S.M., Arndt, P.F., Behrens, S., Bourque, G., and Vingron, M. (2011). CpG deamination creates transcription factor-binding sites with high efficiency. Genome Biol. Evol. *3*, 1304–1311.

Zhang, B., Zhou, Y., Lin, N., Lowdon, R.F., Hong, C., Nagarajan, R.P., Cheng, J.B., Li, D., Stevens, M., Lee, H.J., et al. (2013a). Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome Res. *23*, 1522–1540.

Zhang, T., Cooper, S., and Brockdorff, N. (2015). The interplay of histone modifications - writers that read. EMBO Rep. *16*, 1467–1481.

Zhang, X., Odom, D.T., Koo, S.-H., Conkright, M.D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E., Jacobsen, E., et al. (2005). Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. Proc. Natl. Acad. Sci. USA *102*, 4459–4464.

Zhang, Y., Pak, C., Han, Y., Ahlenius, H., Zhang, Z., Chanda, S., Marro, S., Patzke, C., Acuna, C., Covy, J., et al. (2013b). Rapid single-step induction of functional neurons from human pluripotent stem cells. Neuron *78*, 785–798.

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.