EPIGENETICS

Transcription factors as readers and effectors of DNA methylation

Heng Zhu[1,2], Guohua Wang[3] and Jiang Qian[2,3]

Abstract | Recent technological advances have made it possible to decode DNA methylomes at single-base-pair resolution under various physiological conditions. Many aberrant or differentially methylated sites have been discovered, but the mechanisms by which changes in DNA methylation lead to observed phenotypes, such as cancer, remain elusive. The classical view of methylation-mediated protein–DNA interactions is that only proteins with a methyl-CpG binding domain (MBD) can interact with methylated DNA. However, evidence is emerging to suggest that transcription factors lacking a MBD can also interact with methylated DNA. The identification of these proteins and the elucidation of their characteristics and the biological consequences of methylation-dependent transcription factor–DNA interactions are important stepping stones towards a mechanistic understanding of methylation-mediated biological processes, which have crucial implications for human development and disease.

DNA methylation
A biological process in which a methyl group is covalently added to a cytosine.

[1]*Department of Pharmacology and Molecular Sciences, Johns Hopkins School of Medicine, Edward Miller Research Building, 733 North Broadway, Baltimore, Maryland 21205, USA.*
[2]*The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, Maryland 21287, USA.*
[3]*The Wilmer Eye Institute, Johns Hopkins School of Medicine, The Smith Building, 400 North Broadway, Baltimore, Maryland 21287, USA.*

Correspondence to H.Z. and J.Q.
hzhu4@jhmi.edu; jiang.qian@jhmi.edu

doi:10.1038/nrg.2016.83
Published online 1 Aug 2016

DNA methylation, one of the best-studied epigenetic marks in eukaryotes, is a biological process in which a methyl group is covalently added to a cytosine, yielding 5-methylcytosine (5mC)[1–3] (BOX 1). The methylation process is carried out by a set of enzymes called DNA methyltransferases (DNMTs)[4], which are encoded in many genomes, from bacteria to plants and mammals[5,6]. The evolutionary conservation of these enzymes suggests that DNA methylation provides a selective advantage to the organism. However, the percentage of methylated cytosine varies substantially across species. For example, vertebrates and plants often have a high percentage of methylated CpG dinucleotides outside CpG islands, whereas invertebrates typically exhibit intermediate levels or no methylation[7,8]. With the development of more sensitive methodological approaches, such as methylated DNA immunoprecipitation followed by bisulfite sequencing (MeDIP–BS-seq) — which sequences bisulfite-converted DNA species after enrichment for methylated DNA fragments using an anti-5mC antibody — some genomes previously considered not to have any DNA methylation (for example, that of *Drosophila melanogaster*) have now been found to be methylated at a limited number of cytosines[9–11]. In most animals, DNA is methylated predominantly at CpG dinucleotides, whereas in plants and fungi, a large fraction of DNA methylation also occurs at CHG or CHH (where H can be any nucleotide but G)[12–15]. That said, it was recently discovered that a small fraction of non-CpG methylation also occurs in animals (BOX 1).

DNA methylation has a critical role as a means to control gene expression; for example, during development to ensure X-chromosome inactivation or genomic imprinting[16,17] through various mechanisms. Furthermore, aberrant DNA methylation is a hallmark of many diseases, including various types of cancers[18]. Indeed, abnormal gains in methylation in normally unmethylated CpG islands have been linked to the inactivation of tumour suppressor genes[19–21]. Such abnormal promoter CpG island methylation is emerging as a potential biomarker for cancer detection, diagnosis and prognosis[22,23]. More recently, DNA methylation has also been implicated in non-cancerous diseases, such as schizophrenia[24] and autism spectrum disorders[25,26].

Thanks to rapid technological advances, especially the range of techniques based on deep sequencing, it is now possible to monitor the dynamics of the DNA methylome at single-nucleotide resolution[27–29]. These developments have provided new insights into how the epigenome is shaped and how it regulates different biological processes, such as cellular differentiation and cancer development. For instance, comparing methylation profiles under different physiological conditions revealed tissue-specific or disease-specific differentially methylated regions[30–32], suggesting that the role of DNA methylation in gene regulation is multifaceted and goes beyond simple repression of gene expression.

Despite the fast accumulating profiles of DNA methylomes in various biological processes and species, the interpretation of these data sets often falls short of

providing a mechanistic understanding of the dynamic changes in DNA methylation levels. It still remains a challenge to establish the causality between DNA methylation and physiological outcomes in the epigenetic field. In our view, the first step towards a mechanistic understanding of the DNA methylome is to determine the protein–DNA interactions associated with the dynamics of the DNA methylome. In other words, the identification of DNA methylation 'readers' and 'effectors', which translate methylation signals into biological actions, will be crucial to decipher the epigenetic 'code' of methylation-mediated biological processes.

DNA methyltransferases (DNMTs). Enzymes that catalyse the transfer of a methyl group to DNA.

CpG islands
A segment of DNA with a high frequency of CpG dinucleotides that often overlaps with promoters.

Genomic imprinting
A phenomenon by which some genes are expressed in an allele-specific manner; that is, alleles inherited either from the father or the mother are expressed.

Deep sequencing
A next-generation sequencing approach (for example, RNA sequencing or bisulfite sequencing) with high coverage.

Methylome
The collection of methylation status in an entire genome.

Epigenome
The collection of chemical modifications added to DNA or histones of a given genome, which do not alter the genetic codes but can be inherited and lead to changes in the function of the genome.

Differentially methylated regions
Regions of DNA with significant differences in methylation levels between two physiological conditions (for example, disease versus healthy) different developmental stages or different tissues.

Box 1 | Non-CpG methylation and methylcytosine derivatives

Methylated CpH

Cytosine methylation is usually considered to only occur at CpG sites. Recent advances in genome-wide single-nucleotide sequencing have led to a re-examination of this concept. Interestingly, non-CpG methylation (that is, CpH; where H can be any nucleotide but G) was observed in mammalian stem cells and neuronal cells[27,28,124]. A recent deep-sequencing survey on 18 human tissues revealed an unexpected presence of methylation at non-CpG sites in almost all tissues tested[125]. Several lines of evidence suggest that non-CpG methylation might be functional. First, the flanking sequences of the methylated CpH (mCpH) showed similar motifs to 5′-TNCA(C/G)[125] (where N can be any nucleotide). Second, the position of DNA methylation is highly conserved across different cell types[27]. Third, gene expression level is negatively correlated with the methylation level in the gene body[125]. To understand the biological functions of the modification, identification of the proteins that interact with these modifications would be a crucial step.

One such protein is methyl-CpG-binding protein 2 (MeCP2), which is known to interact with mCpG sites and negatively regulates gene expression. Superimposing MeCP2 chromatin immunoprecipitation followed by sequencing (ChIP–seq) and mCpH profiles in neurons showed an enrichment of mCpH around the binding peak of MeCP2. MeCP2 ChIP followed by bisulfite sequencing (ChIP– BS-seq) confirmed its ability to bind mCpH sites *in vivo*[124]. *In vitro* electrophoretic mobility shift assays (EMSAs) demonstrated a direct interaction between MeCP2 and mCpH. The relative affinity of MeCP2 with mCpA is similar to that with mCpG[126,127]. However, the affinity of MeCP2 for mCpT and mCpC is markedly lower than for mCpA and mCpG[126,127].

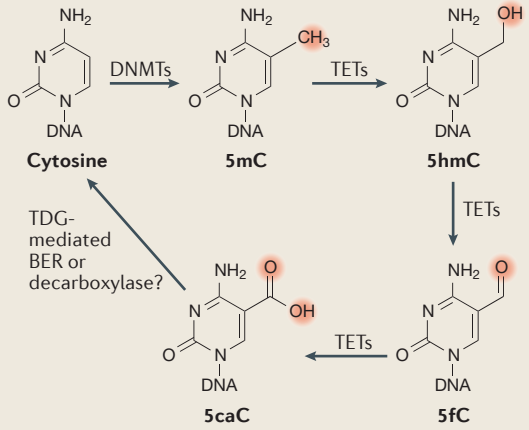
Methylcytosine derivatives

It is well known that DNA methyltransferases (DNMTs) are the enzymes responsible for cytosine methylation, although it long remained elusive which enzymes could reverse DNA methylation in metazoans. In 2009, it was discovered that DNA demethylation might be a multistep process that involves TET (ten-eleven translocation) methylcytosine dioxygenase enzymes that convert 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC)[128,129] (see the figure). These enzymes can further oxidize 5hmC to 5-formylcytosine (5fC) and to 5-carboxylcytosine (5caC)[130,131]. Thymine-DNA glycosylase (TDG)-mediated base excision repair (BER) of 5fC and 5caC can regenerate unmethylated cytosines[132,133].

One important question is whether these oxidized derivatives of 5mC are simply the intermediate products of the demethylation process, or whether they have a functional role themselves. Genome-wide sequencing approaches have generated 5hmC, 5fC and 5caC profiles and have revealed the distribution of these modifications across the genome[130,131,134–137]. The modification levels for these three derivatives are substantially lower than the mCpG levels. For example, the level of 5hmC (that is, 5hmCG/CG) varies from 1% to 30% depending on the cell type[27,134,135,138], whereas the levels of 5fC and 5caC range from 8% to 10% (REFS 139,140). In comparison, the methylation level for mCpG typically ranges from 80% to 90% (REF. 28).

These modifications are not randomly distributed in the genome, but show a preference for certain genomic regions. For example, 5hmC is enriched at distal regulatory elements, such as enhancers and DNase I hypersensitivity sites[134], 5fC is enriched in poised enhancers[137] and a large fraction of 5fC sites are located in intragenic regions with a particular enrichment in exons[141]. By contrast, 5caC was found to be preferentially enriched at major satellite repeats[136]. Interestingly, different modifications showed distinct patterns surrounding protein–DNA binding sites[135].

To understand the function of these modifications, researchers have started to identify proteins that interact with these modifications using various techniques, including mass spectrometry-based approaches[51,142]. For example, MeCP2 was recently found to bind to 5hmC[127,143–145], and the binding affinity seems to be context-dependent. The binding of MeCP2 to 5hmCG, 5hmCC and 5hmCT is substantially weaker than their corresponding methylated probes. However, the conversion of mCpA to 5hmCA does not alter the high affinity binding to MeCP2 (REFS 127,143–145). Interestingly, the binding of these readers is often modification-specific and cell type-specific[51,142]. For example, THAP domain-containing protein 11 (THA11), a transcriptional repressor that plays a central part in embryogenesis, was identified as a brain-specific 5hmC reader[51]. In addition, a number of forkhead box proteins (FOXK1, FOXK2, FOXP1, FOXP4 and FOXI3) were found to interact with 5fC[142]. The dynamic nature of such interactions suggests specific and complex biological roles for these modifications. We expect that more proteins remain to be discovered because these studies only used one or two DNA probes, and because the binding of many proteins could depend on the sequence context surrounding the modifications.



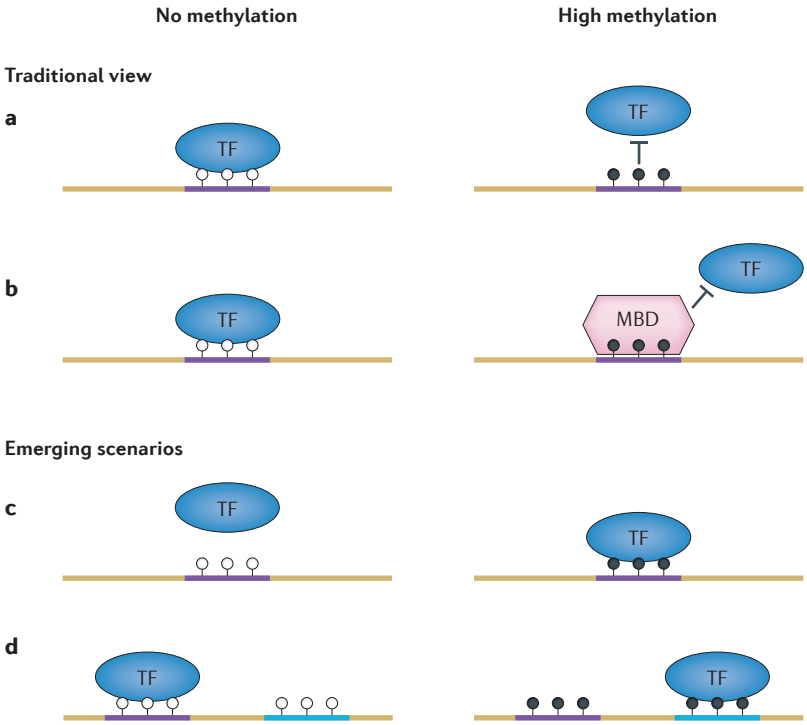


Figure 1 | **Interaction modes between proteins and DNA. a**,**b** | The traditional view of the protein–DNA interaction patterns. Transcription factors (TFs) usually bind to non-methylated DNA motifs (open circles, left panels) in open chromatin regions (part **a**). However, such interactions can be directly disrupted by methylation on the CpG sites in the motifs (filled circles, right panels). Alternatively, methyl-CpG binding-domain (MBD) proteins can be recruited to the methylated DNA motifs and compete off TFs through their higher affinity to the mCpG site in a sequence-independent fashion (part **b**). **c**,**d** | Newly emerging scenarios for protein–DNA interactions. DNA methylation could create a new binding site for TFs (part **c**). TFs may be able to recognize different sequences with or without DNA methylation (part **d**).

In this Analysis article, we review the discovery of a new class of methylated-DNA-binding proteins, namely transcription factors (TFs), and the approaches used to discover these interactions. We focus on the interaction partners with methylated CpG sites in mammals, with a brief discussion of other methylation derivatives (BOX 1). We then summarize the specific properties of methylation-dependent interactions between TFs and DNA, and discuss the causal relationship between TF–DNA interactions and DNA methylation, before concluding with an overview of potential biological consequences of methylation-dependent protein–DNA interactions.

Readers of methylated DNA

The classical view of methylation-mediated protein–DNA interactions is that only proteins with a methyl-CpG (mCpG)-binding domain (MBD) can recognize and bind to methylated CpG dinucleotides[3,6,33–35] (FIG. 1). The MBD protein family has five known members in mammals, including MeCP2 (methyl-CpG-binding protein 2), MBD1, MBD2, MBD3 and MBD4. Except for MBD3, which does not bind to methylated DNA, all MBD proteins bind to methylated DNA in a non-sequence-specific manner[36,37]. Comparison of MBD proteins from different species showed conservation and divergence in terms of the number of MBD genes and the composition of the MBD domains[6,38]. Interestingly, the extent of genomic methylation generally correlates with the number of MBD proteins in a species[6]. Dysfunction of MBD proteins is associated with human diseases. For example, mutations in the gene encoding MeCP2 cause the neurodevelopmental disorder Rett syndrome[39,40].

Over the past 15 years, evidence has emerged that suggests that some TFs lacking MBDs are able to interact with methylated DNA[23–27] (FIG. 1). Unlike MBD proteins, a handful of mammalian TFs were found to possess sequence-dependent mCpG-binding activity in a few studies. For example, the transcriptional regulator Kaiso, which contains POZ (pox virus and zinc-finger) and zinc-finger domains, was found to bind to a specific methylated sequence with its C2H2 zinc-finger domains[41]. In other studies, the basic leucine zipper (bZIP) CCAAT/enhancer-binding protein-α (CEBPα)[42], the zinc-finger protein ZFP57 and its cofactor KRAB-associated protein 1 (KAP1; also known as TIF1β)[43,44] were shown to interact with specific methylated sequences. In addition to mammalian proteins, a bZIP herpesvirus protein, Zta, was found to bind to methylated regulatory elements and control the epigenetic landscape during the latency-to-lytic phase transition in infected mammalian cells[45]. Two amino acids, one cysteine and one serine, were found to interact with 5mC[45]. In rice, the nuclear protein MVBP (methylated VBE-binding protein) was shown to bind to a rice tungro bacilliform virus promoter region only when the promoter was methylated[46].

As the discovery of these mCpG-binding proteins was often serendipitous, whether TFs represent a new class of DNA methylation readers and, potentially, effectors, and whether sequence-specific mCpG-dependent binding activity is a widespread phenomenon or merely an exception, remained questionable. In addition, recent large-scale analyses of gene expression profiles and DNA methylomes showed that a substantial portion of DNA methylation sites is positively correlated with gene expression[31]. This finding may result from the high levels of DNA methylation in the gene body of highly expressed genes; however, it also raises the possibility that some TFs bind to methylated regulatory elements and activate gene expression. It should be noted that DNA methylation of a promoter or an enhancer has been shown to be correlated with increased transcription of a target gene[47–49], although most of the evidence showing a positive correlation between methylation and expression seems to result from methylation downstream of the transcription start site. Intrigued by these observations, several research groups have conducted unbiased, high-throughput screens to search for such a correlation in higher eukaryotes (TABLE 1).

High-throughput reader discovery

Tandem mass spectrometry. One systematic approach for the discovery of mCpG-binding proteins is based on tandem mass spectrometry (MS/MS)[50,51]. A recent study

Table 1 | **Comparison of high-throughput approaches for mCpG reader discovery**

Method	Bait	Prey	Advantage	Disadvantage	Refs
MS/MS	Generic DNA sequences	Nuclear extracts	A comprehensive survey against nuclear proteins; tissue-specific interactions can be detected	Uses generic DNA probes with no sequence specificity; limited to high abundant proteins	51,142
Protein microarray	Sequence-specific DNA motifs	~1,500 human TF proteins	A comprehensive survey against the entire TF repertoire and is not limited to protein abundance in cells	Limited to a few hundred DNA probes	53
DNA microarray	Individual proteins	All possible combinations of 8–10-nucleotide-long DNA sequences	Accurate mapping of binding consensus	Candidate approach, prior knowledge required	56
ChIP–BS-seq	Antibodies against TF of interest	TF–DNA complexes in cultured cells	Genome-wide survey for *in vivo* binding events	Limited by antibody quality and availability	57,58

ChIP–BS-seq, chromatin immunoprecipitation followed by bisulfite sequencing; MS/MS, tandem mass spectrometry; TF, transcription factor.

used a generic DNA sequence harbouring an mCpG site to pull down interacting proteins from nuclear extracts of cultured cells[51]. Proteins bound to methylated DNA sequences were then identified by MS/MS. Based on this approach, 19 proteins were identified that interact preferentially with the methylated DNA probe rather than the non-methylated counterpart in mouse embryonic stem (ES) cell nuclear extracts. Besides the known MBD proteins (such as MeCP2, MBD1 and MBD4), many TFs (such as MHC class II regulatory factor RFX1, zinc-finger homeobox 3 (ZFHX3), lysine-specific histone demethylase 1A (LSD1), zinc-finger and BTB domain-containing protein 44 (ZBT44) and thymocyte nuclear protein 1 (THYN1; also known as THY28), and the Krüppel-like factors (for example, KLF2, KLF4 and KLF5)), were identified as new mCpG-binding proteins (TABLE 2). The authors applied the same approach to neuronal progenitor cells and found that a large and distinct set of proteins showed preferential binding to mCpG sites, suggesting that the interaction with mCpG is dynamic and thus varies under different physiological conditions. A similar approach was also used to identify nucleosome-interacting proteins that are affected by DNA methylation[50]. Although proteins from cell extracts are in a more native state in the MS/MS-based approach, the DNA probes used in this approach are typically generic and, therefore, sequence specificity of observed interactions remains elusive.

Functional protein microarray. Functional protein microarrays have been used as a powerful tool to profile protein–DNA interactions in the past[52]. A comprehensive examination of sequence-specific mCpG-binding activities was conducted by sequentially probing a human protein microarray containing 1,321 TFs and 210 cofactors with 154 DNA motifs that each carried at least one mCpG site[53]. To identify human TFs that preferentially bind to methylated DNA motifs, each methylated motif was mixed with its unlabelled and unmethylated counterpart in tenfold excess in the binding assays. This competition assay ensures that the identified interactions are indeed methylation-dependent, rather than due to CpG-flanking sequences. Of the 154 methylated motifs examined, 150 showed strong binding signals to at least one protein on the microarray. In total, 41 TFs and 6 cofactors were found to bind to at least one methylated sequence. Most of these factors were found to bind to only a few methylated sequences, suggesting that the interactions are not only methylation-dependent but also sequence-specific. Interestingly, the factors that showed binding activity to methylated sequences were widespread among various TF subfamilies, such as zf-C2H2, homeobox, bHLH (basic helix–loop–helix), forkhead, bZIP and HMG (high-mobility group) box. Many of these factors are known to be involved in tissue development or have been associated with cancer. A subsequent validation assay showed that some of these TFs indeed bind to methylated DNA *in vivo* and regulate gene expression[53].

DNA microarray. DNA (or protein-binding) microarray technology has been used to determine the binding specificity of TFs[54,55]. A double-stranded DNA microarray, typically comprising 40,000 unique DNA sequences that cover all possible combinations of 8–10-nucleotide-long DNA sequences that could constitute a binding motif, is incubated with a purified TF so that its binding preference can be accurately determined. In a recent study, the bacterial DNMT SssI was used to methylate the CpG sites of the sequences on the array[56], followed by individual probing with eight purified proteins containing bZIP domains. By comparing the binding profiles of each protein obtained on the methylated and unmethylated microarrays, proteins that preferentially bind to specific sequences were determined. Among the eight bZIP proteins, CEBPα and CEBPβ were found to specifically bind to a methylated sequence[56]. This approach enables a large amount of DNA sequences to be surveyed for protein–DNA interactions; accurate sequence specificity can, therefore, be determined for a given protein. However, prior knowledge of a candidate TF

is required because it can be cumbersome to survey an entire TF family. Therefore, this approach is ideally used for fine-mapping sequence specificity of a previously identified mCpG-binding protein.

ChIP–BS-seq. To determine methylation-dependent protein–DNA interactions *in vivo*, chromatin immunoprecipitation followed by bisulfite sequencing (ChIP–BS-seq) is an ideal approach. ChIP is first performed to obtain the DNA sequences that are bound by a protein of interest, and then the methylation level is sequentially determined using BS-seq. This approach was developed recently to determine the crosstalk between histone modifications and DNA methylation[57–59]. However, it requires prior knowledge of mCpG-binding proteins and the availability of antibodies that are directed against the proteins of interest. For example, after KLF4 was determined to bind to mCpG sites, ChIP–bisulfite conversion followed by PCR was used to validate the methylated DNA–protein interactions *in vivo*[53]. It is important to note that ChIP–BS-seq is the only approach that does not use naked DNA fragments to identify the TFs that bind to methylated DNA. Therefore, TFs identified using the other methods may not necessarily recognize mCpGs *in vivo*, and further studies are needed to dissect the functionality of the interactions.

Methylated DNA–TF interactions *in vivo*

Several studies have demonstrated that methylated DNA–TF interactions can occur in a cellular context (*in vivo*). For example, ZFP57 and its cofactor KAP1 were shown to bind selectively to nine DNA-methylated alleles of imprinting control regions (ICRs) in ES cells[43]. In another study, ZFP57 binding sites were mapped in hybrid ES cells, and ZFP57 was found to interact with the methylated parental-origin allele[60]. Similarly, Kaiso was shown to bind to the methylated promoter of the *MTA2* gene in HeLa cells[61]. Another study, which used a quantitative ChIP–PCR assay, demonstrated that Kaiso binds to the methylated promoters of *CDKN2A*, *MGMT* and *HIC1* in both HCT116 and Colo320 human colon cancer cell lines[62]. The finding that Kaiso binds to the methylated promoter of *CDKN2A* was recently reproduced in an independent study[63]. By contrast, a different study discovered that Kaiso was not associated with highly methylated promoters in GM12878 lymphoblastoid cells or in K562 human myeloid leukaemia cell lines[64]. Of note, this observation does not necessarily rule out the possibility that Kaiso binds to methylated DNA motifs in other cell types; rather, it suggests that methylation-dependent TF–DNA interactions may be cell type-specific. That is, some TFs might bind to methylated DNA motifs in certain cell types but not in others, presumably owing to variations in accessibility to methylated motifs in different cell types and/or dynamics of the DNA methylomes during differentiation and development.

Although these studies may suggest that some TFs can bind to methylated DNA *in vivo*, one important question remains: how prevalent are methylated-DNA–TF interactions in a given genome? For example, the studies showing that Kaiso binds to methylated DNA *in vivo*[58–60] were focused on a few genes or genomic regions rather than genome-wide surveys. To determine to what extent these TFs interact with methylated loci in cells, we globally evaluated the accessibility of highly methylated regions in the H1 human ES cell line. Integration of the DNA accessibility data obtained by mapping DNase I hypersensitivity sites (DHSs)[65] and the DNA methylome data obtained from the same cell type[28] revealed that numerous open chromatin regions (that is, accessible regions) indeed contain highly methylated CpG sites. Overall, 258,188 DHS peaks were determined in the H1 human ES cell line by The ENCODE Consortium. By superimposing the DHS peaks with the DNA methylome of the H1 cells determined by whole-genome BS-seq[28], we calculated the average methylation level (m) of CpG sites within a DHS peak, defined as:

$$m = \frac{\Sigma_{i=1}^{N} m_i}{N} \quad (1)$$

where N is the number of CpG sites within a peak, and m_i is the methylation level for CpG site i. Overall, 77,124 (29.9%) of the 258,188 DHSs detected in H1 cells had an average methylation level greater than 80% at CpG sites (FIG. 2a), suggesting that many methylated CpG sites are accessible to TFs.

We then examined whether the TFs listed in TABLE 2 could interact with methylated DNA *in vivo*. We obtained TF ChIP–seq data sets in H1 ES cells from The ENCODE Consortium, and uniform peaks were called using the Irreproducible Discovery Rate (IDR) method[66]. The ChIP–seq peaks were superimposed with the methylome data set and the average methylation levels within each ChIP–seq peak were calculated using the method described above. For each TF, we obtained the distribution of the methylation level for each ChIP–seq peak. Although the availability of ChIP–seq data sets was limited, six TFs (namely CEBPβ, E2F6, BACH1, RFX5, KLF4 (REF. 28) and retinoic acid receptor RXRα) had ChIP–seq data in H1 cells (TABLE 2). The DNA methylation levels within the ChIP–seq peaks showed a bimodal distribution for all TFs except RXRα, indicating that a substantial fraction of their binding sites are located in highly methylated regions (FIG. 2a). For example, of the 15,557 ChIP–seq peaks identified for CEBPβ, 6,675 (42.9%) had a methylation level greater than 80%. As a comparison, we selected two TFs (nuclear respiratory factor 1 (NRF1) and transcription initiation factor TFIID subunit 1 (TAF1)), which are known not to interact with methylated DNA (based on our current knowledge), as negative controls: neither NRF1 (FIG. 2a) nor TAF1 (Supplementary information S1 (figure)) showed a bimodal distribution, demonstrating that these TFs only bind to regions with low levels of methylation.

We further examined whether the methylated CpG sites located exactly at the TF binding sites (~10–20 bp) within the ChIP–seq peaks (200–500 bp), using CEBPβ as an example (FIG. 2b). We first used the MEME (multiple EM for motif elicitation) algorithm to predict significantly enriched sequence motifs using the sequences of

Table 2 | **Representative mCpG-binding transcription factors**

Protein name*	DNA-binding domain‡	Canonical motif	Methylated motif	*In vivo* evidence?	Pioneer TF?	Refs
Kaiso	Zinc-finger	TCCTGCNA	TCTmCGmCGAGA	Yes	Yes	41,50,51
ZFP57	Zinc-finger	Unknown	TGCmCGC	Yes	No	43,60
GATAD2A	Zinc-finger	Unknown	Unknown	No	No	50,51
GATAD2B	Zinc-finger	Unknown	Unknown	No	No	50,51
RFX5	Other	HYRDBVMCH	Unknown	Yes	No	50,53
KLF4	Zinc-finger	GCCMCRCCC	CCmCGCC	Yes	Yes	51,53
HOXA5	Homeobox	CCYCATTAKTGN	Non-specific	No	No	51,53
CEBPα	Zinc-finger	TTKCNYMA	mCGTCA	Yes	No	42,56
CEBPβ	Zinc-finger	TTKCNYMA	TTGmCGYMA	Yes	No	56
ZBTB40	Zinc-finger	Unknown	Unknown	No	No	50
ZBTB9	Zinc-finger	Unknown	Unknown	No	No	50
ZHX1	Zinc-finger homeobox	Unknown	Unknown	No	No	50
ZHX2	Zinc-finger homeobox	Unknown	Unknown	No	No	50
ZHX3	Zinc-finger homeobox	Unknown	Unknown	No	No	50
HOMEZ	Homeobox	YTCGYYY	Unknown	No	No	50
FOXA1	Forkhead	TRTTTGYTYWN	Unknown	No	Yes	50
GZF1	Zinc-finger	TGCGCKTMTATA	Unknown	No	No	51
KLF10	Zinc-finger	Unknown	Unknown	No	No	51
KLF2	Zinc-finger	Unknown	Unknown	No	No	51
KLF3	Zinc-finger	CAGGGTGTG	Unknown	No	No	51
KLF5	Zinc-finger	YYMCDCCC	Unknown	No	No	51
RREB1	Zinc-finger	CCCCAAACMMCCCC	Unknown	No	No	51
SALL2	Zinc-finger	Unknown	Unknown	No	No	51
ZBTB4	Zinc-finger	CNNTCACTGGNA	Unknown	No	No	51
ZBTB44	Zinc-finger	Unknown	Unknown	No	No	51
ZCHC8	Zinc-finger	Unknown	Unknown	No	No	51
ZFP597	Zinc-finger	Unknown	Unknown	No	No	51
ZN513	Zinc-finger	NNAACATCTGGA	Unknown	No	No	51
ZNF710	Zinc-finger	Unknown	Unknown	No	No	51
ZFHX2	Zinc-finger homeobox	Unknown	Unknown	No	No	51
ZFHX3	Zinc-finger homeobox	ATTAAYTRCAC	Unknown	No	No	51
ZFHX4	Zinc-finger homeobox	Unknown	Unknown	No	No	51
DLX1	Homeobox	NTGNNNTAATTANY	Unknown	No	No	51
DLX5	Homeobox	NNRGYAATTRNYK	Unknown	No	No	51
DLX6	Homeobox	YAATTA	Unknown	No	No	51
HOXB8	Homeobox	NNNNGYAATTAATANW	Unknown	No	No	51
HOXB9	Homeobox	NRRGCMATAAAA	Unknown	No	No	51
MEIS1	Homeobox	NNNTGACAG	Unknown	No	No	51
PBX1	Homeobox	ATCAATCAW	Unknown	No	No	51
PBX3	Homeobox	Unknown	Unknown	No	No	51
BACH1	bZIP	NNSATGAGTCATGNT	Unknown	Yes	No	51
TFCP2	CP2	DWCYRGH	Unknown	No	No	51
UBIP1	CP2	SCAGYB	Unknown	No	No	51
FOXK1	Forkhead	AATGTAAACAAA	Unknown	No	Yes	51
FOXK2	Forkhead	Unknown	Unknown	No	Yes	51

Table 2 (cont.) | **Representative mCpG-binding transcription factors**

Protein name*	DNA-binding domain‡	Canonical motif	Methylated motif	*In vivo* evidence?	Pioneer TF?	Refs
RFX4	Other	CCNTAGCAACS	Unknown	No	No	51
RFXAP	Other	Unknown	Unknown	No	No	51
DIDO1	Zinc-finger	Unknown	GCAGmCGAGC	No	No	53
FEZF2	Zinc-finger	Unknown	SYmCGCC	No	No	53
GATA3	Zinc-finger	NNGATARNG	Non-specific	No	Yes	53
GATA4	Zinc-finger	AGATADMAGGGA	AAAmCGCTTCC	No	Yes	53
PF21A	Zinc-finger	Unknown	Unknown	No	No	53
PPARγ	Zinc-finger	NNWGRGGTCAAAGGTCA	Unknown	No	No	53
RN138	Zinc-finger	Unknown	Unknown	No	No	53
RXRA	Zinc-finger	NNNNNTGACCCC	TCmCGVN	No	No	53
SCAPE	Zinc-finger	Unknown	Non-specific	No	No	53
ZCHC7	Zinc-finger	Unknown	BKmCGDS	No	No	53
ZKSC5	Zinc-finger	Unknown	Unknown	No	No	53
ZMYM3	Zinc-finger	TTTGAAA	GAmCGTC	No	No	53
ZN114	Zinc-finger	Unknown	Non-specific	No	No	53
ZNF22	Zinc-finger	HYDCCYMCD	Unknown	No	No	53
ZNF28	Zinc-finger	Unknown	TTTAmCGTGCAG	No	No	53
ZN416	Zinc-finger	Unknown	Unknown	No	No	53
ZN461	Zinc-finger	Unknown	VHmCGHM	No	No	53
ZN695	Zinc-finger	Unknown	DNmCGCY	No	No	53
CERS4	Homeobox	Unknown	Unknown	No	No	53
CRX	Homeobox	YNNNTAATCYSMN	CCCmCGTAA	No	No	53
HOXA9	Homeobox	NCGGYCATWAAAWTANW	Unknown	No	No	53
TGIF1	Homeobox	AGCTGTCANNA	RVmCGMM	No	No	53
ATF6β	bZIP	Unknown	Unknown	No	No	53
E2F3	E2F TDP	GGCGGGN	Non-specific	No	No	53
E2F6	E2F TDP	CNTTTCNT	Unknown	Yes	No	53
FOXC1	Forkhead	GTAAATAAACA	HVmCGBS	No	Yes	53
ARNT2	HLH	Unknown	AAAmCGCTTCC	No	No	53
NPAS2	HLH	VCAMRTR	AAACmCGGCTC	No	No	53
ARI3B	Other	HWTAWW	AAAmCGCTTCC	No	No	53
PMS1	Other	Unknown	ATGAmCGTCAC	No	No	53
RBPJ	Other	CGTGGGAA	AAACmCGAGAAC	No	No	53
SMAD4	Other	GKSRKKCAGMCANCY	NCmCGGG	No	No	53
SUB1	Other	Unknown	Unknown	No	No	53
AP2α	Other	GCCNNNRGS	GTCAmCGCCC	No	No	53

AP2α, activating enhancer-binding protein 2α; ARI3B, AT-rich interactive domain-containing protein 3B; ARNT2, aryl hydrocarbon receptor nuclear translocator 2; ATF6β, activating transcription factor 6β; B, any nucleotide except A; bZIP, basic leucine zipper; CEBPα, CCAAT/enhancer-binding protein-α; CERS4, ceramide synthase 4; CP2, CCAAT box binding protein 2; CRX, cone-rod homeobox; D, any nucleotide except C; DIDO1, death-inducer obliterator 1; FEZF2, Fez family zinc-finger protein 2; FOXA1, forkhead box A1; GATA3, GATA-binding factor 3; GATAD2A, GATA zinc-finger domain-containing protein 2A; GZF1, GDNF-inducible zinc-finger protein 1; H, any nucleotide except G; HLH, helix–loop–helix; HOMEZ, homeobox and leucine zipper-containing protein; HOXA5, homeobox protein A5; K, G or T; KLF4, Krüppel-like factor 4; M, A or C; m, methyl; N, any nucleotide; NA, not available; NPAS2, neuronal PAS domain-containing protein 2; PBX1, pre-B cell leukaemia transcription factor 1; PF21A, PHD finger protein 21A; PPARγ, peroxisome proliferator-activated receptor-γ; R, A or G; RFXAP, regulatory factor X-associated protein; RN138, RING finger protein 138; RREB, Ras-responsive element-binding protein 1; RXRα, retinoic acid receptor RXRα; S, G or C; SALL2, sal-like protein 2; SCAPE, S phase cyclin A-associated protein in the endoplasmic reticulum; SMAD4, mothers against decapentaplegic homologue 4; TDP, transcription factor E2F dimerization partner; TF, transcription factor; TFCP2, α-globin transcription factor CP2; UBIP1, upstream-binding protein 1; V, any nucleotide except T; W, A or T; Y, C or T; ZBTB40, zinc-finger and BTB domain-containing protein 44; ZCHC, zinc-finger CCHC domain-containing protein 8; ZFHX2, zinc-finger homeobox protein 2; ZFP57, zinc-finger protein 57; ZHX1, zinc-fingers and homeoboxes protein 1; ZKSC5, zinc-finger protein with KRAB and SCAN domains 5; ZMYM3, zinc-finger MYM-type protein 3. *TFs are sorted by assay type performed (see reference indicated); the TFs identified by multiple studies are ranked on top. ‡The DNA-binding domains that are found only in a small number of TFs are denoted as 'Other'.

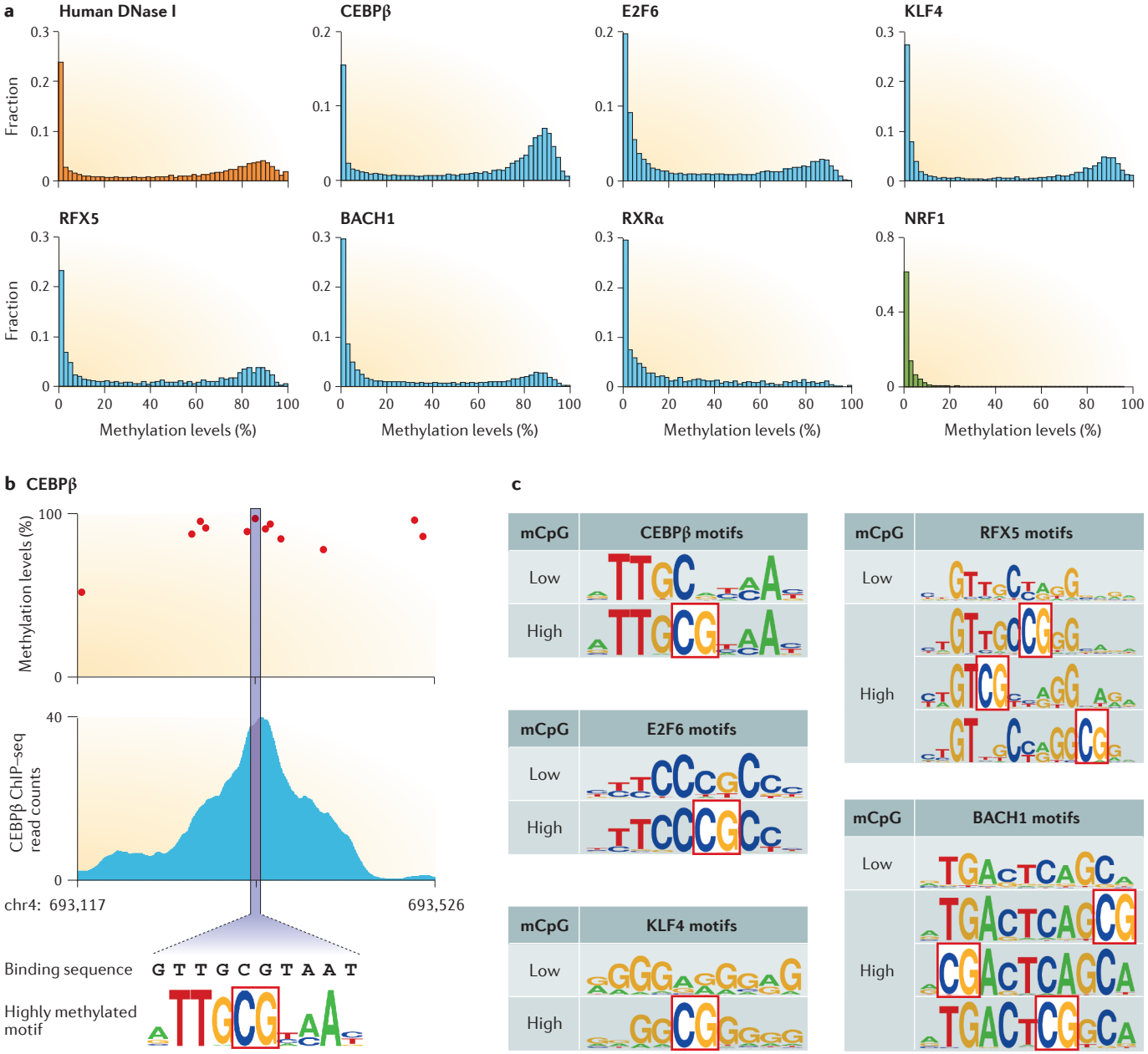


Figure 2 | **Methylated-DNA–TF interactions *in vivo*. a** | Integration of DNA methylome data[28] with transcription factor (TF) chromatin immunoprecipitation followed by sequencing (ChIP–seq) data[28] (see also Further Information) or with DNA accessibility data[65] obtained from mapping DNase I hypersensitivity sites (DHSs) from the same cell lines suggests that some TFs bind to methylated DNA *in vivo*. All of the data, including DHS, ChIP–seq and DNA methylome, were obtained from the H1 embryonic stem cell line. The *x* axis shows the average methylation level (percentage) of the CpG sites within a DHS or TF ChIP–seq peak, and the *y* axis shows the fraction of peaks with a certain average methylation level. The methylation levels within each DHS peak show bimodal distribution, indicating that a large portion of highly methylated CpGs (mCpGs) are accessible to TFs. The methylation levels within the ChIP–seq peaks for five TFs (namely CCAAT/enhancer-binding protein-β (CEBPβ), E2F6, Krüppel-like factor 4 (KLF4), RFX5 and BACH1) also show a bimodal distribution, suggesting that they can interact with methylated DNA *in vivo*. Retinoic acid receptor RXRα shows no binding activity in highly methylated regions, similar to the negative control nuclear respiratory factor 1 (NRF1). **b** | We examined whether methylation occurred in the binding sites within the ChIP–seq binding peaks. DNA sequences within each ChIP–seq peak were extracted and grouped based on their average methylation level. The low methylation group contains peaks with a methylation level of <60% and the high methylation group contains peaks with a methylation level of >80%. MEME (multiple EM for motif elicitation) analysis was performed on each of the top 500 peaks with the highest ChIP–seq intensities for the two groups and the most significant motifs were identified. The motif was then used to scan the DNA sequence within each peak and the DNA segment with highest match score to the motif was recorded. The methylation level was examined for the CpG sites within the identified DNA segment. As illustrated with CEBPβ, a matched segment (for example, 5′-GTTGCGTAAT) containing a highly methylated CpG site in the middle was identified within a ChIP–seq peak. **c** | Motifs were identified separately for binding peaks with low-level (<60%) methylation and high-level (>80%) methylation. To obtain the most reliable methylated motifs, the DNA sequences that match to the MEME motif were further grouped based on the position of the mCpG, and the sequences with a methylation level of >80% were assembled to generate the methylated motif for each subgroup. The mCpG sites for each subgroup are outlined with red boxes.

the ChIP–seq peaks that have a low methylation level[67]. The most significantly enriched motif did not contain a CpG site. Interestingly, when the same analysis was applied to those peaks that have high methylation levels, a significantly enriched motif containing a CpG site at position 4 was discovered (FIG. 2b). We next examined the methylation level of the CpG sites within the motif in each ChIP–seq peak (FIG. 2b). Among the 6,675 peaks with a high methylation level, 3,894 carried a highly methylated (>80%) CpG site within the enriched motifs. A motif could be reconstructed with these 3,894 binding peaks, which represented the methylated motif for CEBPβ (FIG. 2c). The same analysis was performed for the other four TFs. In summary, 25.0% (3,894 out of 15,557), 7.7% (1,103 out of 14,396), 5.2% (88 out of 1,695), 3.0% (115 out of 3,793) and 1.6% (186 out of 11,457) of binding sites were highly methylated for CEBPβ, E2F6, RFX5, KLF4 and BACH1, respectively (FIG. 2c). Note that this is a conservative estimate because we used a stringent definition of highly methylated sites (that is, >80%).

Taken together, the above analysis suggests that many TFs shown to bind methylated DNA *in vitro* are also able to interact with methylated DNA *in vivo*, although further *in vivo* genome-wide characterization of TF binding patterns and high-resolution DNA methylation analyses are needed to strengthen the evidence base. The list of TFs that interact with methylated DNA (TABLE 2) provides a foundation for further functional characterization of methylated DNA–TF interactions in various biological processes.

Features of methylated-DNA–protein interaction

Protein domains that interact with methylated DNA. Identification of the protein domains that recognize mCpG sites is important to characterize mCpG-dependent protein–DNA interactions. Such knowledge will enable the mutation of critical residues within these domains that abolish the mCpG-dependent binding activity of these proteins, while maintaining their ability to bind non-methylated DNA. Therefore, mutated proteins can be useful tools to dissect the physiological roles of mCpG-dependent protein–DNA interactions.

Besides the well-known MBDs, other protein domains seem to interact with mCpG sites. For example, the recent crystal structure of mouse ZFP57 in complex with a methylated DNA sequence demonstrated that its two zinc-fingers interact with methylated DNA, and that an arginine (Arg178), which is involved in hydrophobic interactions, plays a crucial part in mCpG binding[44]. A separate study suggested that an arginine and glutamate pair in KLF4 recognizes the mCpG site[68]. A structural comparison of MeCP2 and KLF4 indeed showed a common structural feature involving one arginine and one asparagine[53]. A global survey of methylated-DNA-binding proteins suggests that many other protein domains might also be able to interact with mCpG sites, including homeobox, HLH and E2F domains[53].

There are currently no general rules of evolutionary conservation for the domains that interact with methylated sites, owing to a lack of data from multiple species. However, in a comparison of mCpG-binding proteins between humans and mice[51,53], a few proteins such as KLF4 and homeobox A5 (HOXA5) were shown to bind mCpG sites in both species, which is indicative of the functional importance of methylation-dependent protein–DNA interactions.

Sequence specificity. Notably, many proteins can bind both non-methylated and methylated sequences in a different sequence context. For example, CEBPα is known to bind a particular sequence element, 5′-TGACGTCA[42]. However, when the CpG is methylated, CEBPα can effectively recognize half of the motif: 5′-mCGTCA[42]. Similarly, although KLF4 recognizes a non-methylated canonical motif of 5′-TTTACGCC, it has been demonstrated that KLF4 specifically recognizes a 5′-TCCmCGCCC motif only when the CpG is methylated[53]. If the methylation status of these two sequences is exchanged, KLF4 loses the ability to bind to either sequence[53]. Indeed, for many newly discovered methylated DNA-binding proteins, the methylated motifs differ from the non-methylated motifs[53]. Therefore, it is reasonable to speculate that 5mC might represent the fifth nucleotide that further fine-tunes the specificity of protein–DNA interactions; that is, 5mC acts as an additional regulatory layer to remove, create and/or change TF binding sites (FIG. 1).

Several recent studies have started to provide the structural basis for the altered sequence specificity due to DNA methylation. Both *in vitro* DNase I digestion assays and structural studies indicate that methylation has a profound impact on DNA structure and shape (for an in-depth review see REF. 69). Adding a methyl group to the cytosine could affect the local DNA shape, as evidenced by the altered DNase I digestion rate and patterns[70,71]. Similarly, based on a few reported crystal structures of double-stranded DNA fragments with 5mC bases[68,72,73], the presence of a bulky methyl group in the major groove leads to a subtle widening of the major groove and a subtle narrowing of the minor groove. Consequently, 5mC can affect the access of a given TF to the affected motifs in both major and minor grooves in genomic DNA and thus change the sequence specificity of protein–DNA interactions.

Binding affinity. One important question is whether methylated-DNA–protein interactions have a similar binding affinity to the interactions between the same protein and a non-methylated DNA, or whether they are just labile interactions. Using an *in vitro* pulldown-coupled MS/MS approach, the relative affinity of protein–mCpG interactions can be estimated[51]. For example, for a particular sequence (5′-GGGCGTG), which was determined on the basis of the KLF4 ChIP–seq data sets[74], KLF4 showed higher affinity when the cytosine in the motif was methylated compared with the unmethylated sequence[51]. The protein microarray approach[53], which uses the concept of relative affinity to identify proteins that preferentially bind to methylated DNA, revealed proteins with strong fluorescent signals, which are expected to bind tighter to the

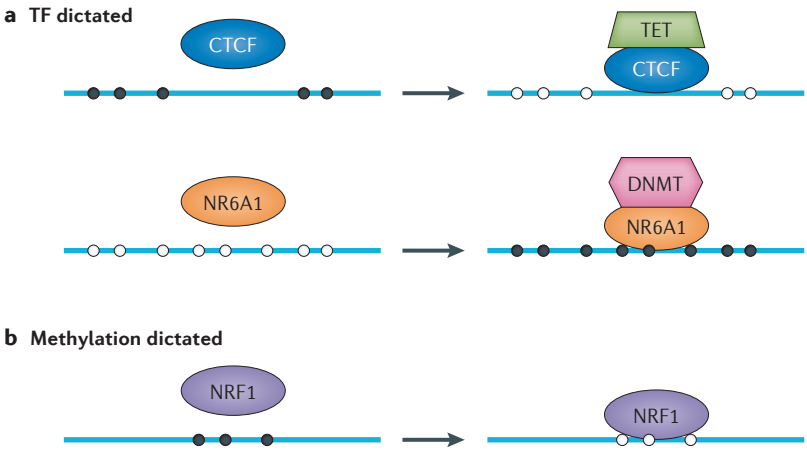


Figure 3 | **Two action models between TF and methylated DNA interactions.** **a** | The binding of transcription factors (TFs) dictates the methylation status surrounding the binding sites. The filled circles represent methylated DNA and the open circles represent unmethylated DNA. The binding of the transcriptional repressor CTCF reduces the local methylation level, presumably by recruitment of the TET (ten-eleven translocation) enzymes, which can demethylate surrounding CpG sites. The interaction between CTCF and TET remains to be experimentally validated. Conversely, the binding of NR6A1 (nuclear receptor subfamily 6 group A member 1) induces DNA methylation by interacting with DNA methyltransferase (DNMT) proteins. **b** | The DNA methylation status dictates TF binding activity. NRF1 (nuclear respiratory factor 1) only binds to DNA when its consensus sequence is non-methylated.

methylated motif than to the unmethylated counterpart. The absolute binding affinity (that is, K_d values) can be measured by applying the oblique incidence reflectivity difference (OIRD), which is a real-time, label-free method to measure the kinetics of a binding event[53,75]. Three proteins (ZMYM3, AP2α and KLF4) were selected to determine the K_{on} and K_{off} values with their corresponding motifs in methylated forms. The deduced K_d values of ZMYM3, AP2α and KLF4 were determined as 460 nM, 399 nM and 479 nM, respectively. Importantly, no obvious affinity could be detected when these tested motifs were unmethylated. As a comparison, the K_d values of the short isoform of MBD2, MBD2b, for the same motifs ranged from 97 nM to 197 nM, suggesting that MBD-lacking TFs bind to methylated DNA motifs nearly as strongly as MBD2b.

Cis-regulatory elements. To better understand the physiological role of mCpG-binding proteins, it is important to determine which methylated regions in the genome can be specifically recognized by these proteins. Although MBD family proteins tend to bind to regions with a high methylation density (that is, high methylation level and high CpG density)[76], it is interesting to examine whether the same is true for sequence-specific mCpG-binding proteins.

As protein–DNA interactions are dynamic, differentially methylated regions might be possible candidates for methylation-dependent interactions. Analysis of the methylomes obtained from 17 adult mouse tissues at single base-pair resolution showed that approximately 6.7% of the mouse genome is differentially methylated, mostly at distal *cis*-regulatory regions[77]. Another study discovered that regions with a low level of methylation, ranging from 10% to 50%, often occur at distal regulatory regions[78]; that is, regions that are enriched for enhancer marks, including high levels of histone H3 lysine 4 monomethylation (H3K4me1) as well as binding sites for p300 histone acetyltransferase and other regulatory factors. Similarly, extensive DNA methylation was found to coexist with active H3K27 acetylation (H3K27ac) marks in a large number of enhancers[79]. More importantly, the reduction of DNA methylation led to a decrease in H3K27ac marks, suggesting an active role of DNA methylation in regulating enhancer activity. Based on the analysis of KLF4 binding in ES cells, we also found that KLF4 binds to methylated enhancer regions[53], which may suggest that sequence-specific mCpG-binding proteins interact preferentially with distal enhancer regions.

Cause or consequence?

Although many proteins have been found to recognize methylated DNA, the causality between DNA methylation and TF binding is far from clear. On the one hand, DNA methylation could dictate the interaction between proteins and DNA, but on the other hand, the binding of certain proteins may affect the methylation of DNA. Recent studies suggest that both scenarios can occur in different contexts (FIG. 3).

Protein binding affects the DNA methylation status. The binding of methyltransferases or methylcytosine dioxygenases (for example, DNMTs and TETs (ten-eleven translocation proteins)) affects the status of DNA methylation, but recent studies suggest that many non-enzymatic proteins, such as TFs, could regulate the establishment and maintenance of the local DNA methylation levels in a sequence-specific fashion. One such regulator is the transcriptional repressor CTCF, which is known to have an essential role in imprinting control; that is, to achieve allele-specific gene regulation[80]. CTCF binds to the unmethylated ICRs in maternal alleles, which prevents distal enhancers from activating downstream genes[81]. By contrast, when the paternal ICR is methylated, CTCF cannot bind to the ICR, thus allowing the activation of downstream genes by distal enhancers. One study suggests that CTCF itself contributes to the maintenance of the non-methylated status of maternal ICRs, as maternally transmitted mutant ICRs in neonatal mice that harbour point mutations in CTCF binding sites acquire a heterogeneous degree of methylation[82]. Although the traditional view of imprinting control is that differential methylation leads to differential binding of CTCF, and thus yields allele-specific gene regulation, this study suggests that CTCF binding itself is necessary to maintain differential methylation of ICRs.

Moreover, a recent report confirmed that the binding of some proteins (for example, CTCF and RE1-silencing transcription factor (REST)) can affect local methylation patterns[78]. The authors first created a reporter construct with a CTCF-binding motif that was inserted into a genomic locus in mouse ES cells. Insertion of the

K_d
The dissociation constant K_d is defined by the K_{off}/K_{on} ratio, which has the unit of concentration.

Oblique incidence reflectivity difference
(OIRD). A form of polarization-modulated imaging ellipsometer for label-free, high-throughput detection of binding events on protein microarrays.

K_{on} and K_{off}
In a simple binding event, K_{on} and K_{off} refer to the on-rate and off-rate constants, which have units of 1/(concentration time) and 1/time, respectively.

TETs
(Ten-eleven translocation proteins). The TET family of methylcytosine dioxygenases is made of TET1, TET2, TET3 and TET4, which catalyse the conversion of the modified DNA base 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC).

binding site induced CTCF binding and resulted in a reduced methylation level in local genomic regions. A single-nucleotide mutation in the CTCF binding motif had no effect on the DNA methylation level. To test the effect in an endogenous setting, the authors generated a *Rest*$^{-/-}$ mouse ES cell line and, as expected, observed that the REST binding regions were highly methylated. Most importantly, they found that the methylation levels at these sites were much reduced after reintroduction of wild-type *Rest* into the cells. Altogether, these results support a model whereby the binding of certain proteins can directly affect DNA methylation levels (FIG. 3).

Another study examined the methylation levels of hundreds of sequences that were individually inserted at the same genomic site in mouse ES cells[83]. Using this approach, the contribution of various sequence motifs to methylation levels could be quantified. They found that CpG density showed a negative correlation with methylation level, which is consistent with the previously established view that CpG islands are generally unmethylated. Interestingly, when the sequences of binding motifs were altered, overall methylation levels decreased[83]. This work suggests that protein binding has a general role in reducing DNA methylation levels, perhaps by preventing DNMT enzymes from gaining access to these sites, which is consistent with previous findings[84].

As TFs have no enzymatic activity to methylate or demethylate a CpG dinucleotide, a possible model would be that these proteins provide sequence-specific guidance and recruit methyltransferases or methylcytosine dioxygenases to these specific sites (FIG. 3). A recent study showed that the nuclear receptor PPARγ (peroxisome proliferator-activated receptor-γ) recruits TET1, resulting in a reduced methylation level around its binding sites through the interaction with TET1 (REF. 85). The reverse can also happen. DNMTs have been found to form protein complexes with various TFs or chromatin modification enzymes. For instance, Sato *et al.*[86] demonstrated that DNMT3A and DNMT3B interact with an orphan nuclear receptor, NR6A1 (nuclear receptor subfamily 6 group A member 1), and that this interaction induced the methylation of the *OCT4* (also known as *OCT3* and *POU5F1*) promoter carrying the NR6A1 binding site.

DNA methylation dictates protein–DNA interactions. It is well known that DNA methylation can affect the binding of some TFs[87]. The manipulation of the methylation status of DNA sequences has been shown (mostly in *in vitro* studies) to result in the differential binding of TFs, including E2F, AP2α, MYC and MYN[88–95]. Specifically, hypermethylation is often associated with a depletion of TF binding. Recently, a few studies examined the effect of DNA methylation on TF binding *in vivo*[96,97].

Using a gene-editing approach, Domcke *et al.*[97] generated a genetic deletion of three methyltransferases (*Dnmt3a*, *Dnmt3b* and *Dnmt1*) in a mouse ES cell line. A large number of novel binding sites for the TF NRF1 were created as a result of the triple knockout (TKO). These binding sites often correlated with novel DHSs in the TKO cells, which exhibited predominantly low methylation levels due to their generation in cells lacking DNMTs. Interestingly, novel NRF1 binding sites were hypermethylated in the wild-type cell line[97], suggesting that the removal of DNA methylation in TKO cells generated new binding sites for NRF1. These new binding sites had poor sequence conservation, indicating that these sites are non-functional in the wild-type background. In an earlier study, the same group showed that CTCF was able to reduce the DNA methylation level near its binding sites[78]. In this work, the authors tested whether CTCF binding could affect NRF1 binding by reducing the methylation level of NRF1 binding sites. Reporter constructs harbouring an NRF1 binding motif and a CTCF motif were introduced into the ES cell line. Deletion of CTCF motifs within the construct led to hypermethylation and thus decreased NRF1 binding, suggesting that NRF1 binding *in vivo* depends on both methylation levels and co-occurring TFs, such as CTCF[97].

The examples above represent the two major mechanisms by which protein–DNA interactions and DNA methylation influence each other (FIG. 3). Of note, these two mechanisms are not mutually exclusive. In some cases, the two mechanisms have been found to coexist for the same TFs. For example, although CTCF is known to change the local methylation status[78], it has been shown that the binding of CTCF is also methylation sensitive[96]. Finally, it is worth noting that the crosstalk between TF–DNA interaction and DNA methylation is not restricted to the TFs whose binding motifs contain CpGs. Changes in DNA methylation are often associated with chromatin status, resulting in increased or decreased DNA accessibility[96,97]. Differences in chromatin states will either create or eliminate TF binding sites and thus lead to differential TF binding. Although only approximately 25% of known TF binding motifs contain at least one CpG site[98], through such indirect crosstalk mechanisms, the binding of TFs without CpGs in their binding motifs could also be influenced by DNA methylation.

Biological consequences

Activation or repression. Methylation in promoters is often considered the hallmark for gene repression[99]. However, large-scale analyses of gene expression profiles and DNA methylomes have revealed that a substantial proportion of DNA methylation sites is positively correlated with gene expression[31]. This analysis was performed on methylation sites located within 300 bp upstream from transcriptional start sites, which raises the possibility that methylation in promoters could also be positively correlated with increased transcription of a target gene[31] (FIG. 4). Of course, whether these methylation sites fall exactly within the regulatory elements and whether they are recognized by TFs remains to be tested. Single-gene studies have also demonstrated that DNA methylation can activate gene expression[47–49]. For example, the sequence-specific DNA-binding protein RFX activates a methylated promoter[49]. Interestingly, this protein was previously shown to bind to methylated DNA[47,51]. Moreover, it was found that methylation at the 3′ end of the CpG island confers tissue-specific transcriptional activation during human ES cell differentiation[100].

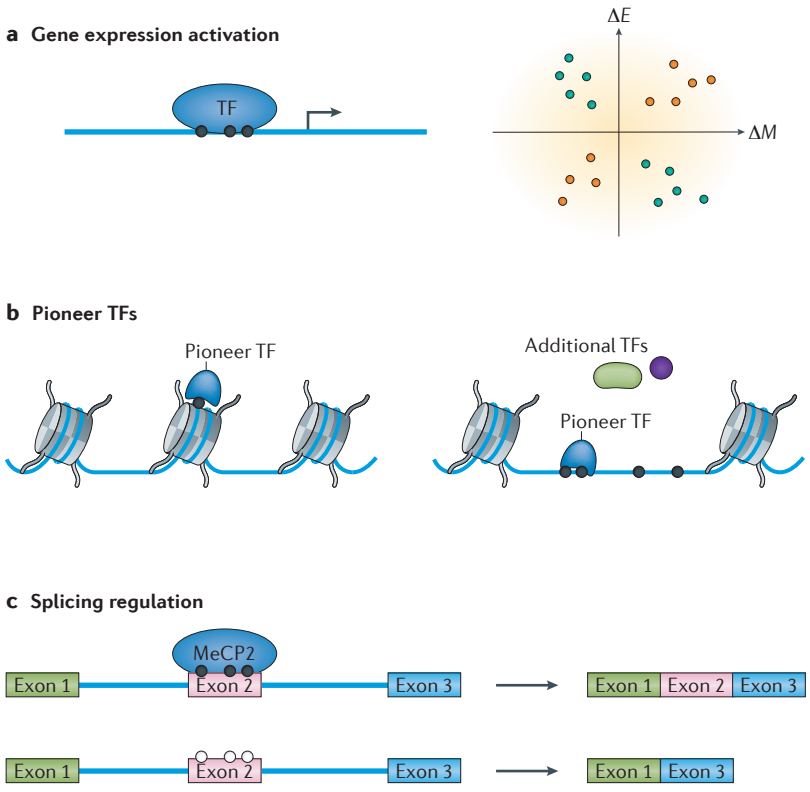


Figure 4 | **Possible biological consequences of methylated DNA–TF interactions.** **a** | Some transcription factors (TFs) can bind to methylated DNA and activate gene expression. Genome-wide profiling of gene expression and DNA methylome data have revealed that many methylation sites positively correlate with gene expression. The orange and blue dots are the CpG sites with methylation levels that are positively or negatively correlated with gene expression, respectively. Δ*E* is the difference of gene expression, and Δ*M* is the difference in methylation level between two conditions. **b** | Pioneer TFs bind to DNA sequences wrapped around the nucleosomes. Consequently, other cofactors (such as chromatin remodelling enzymes and other TFs) are recruited to open up the chromatin regions. As condensed chromatin is often associated with DNA methylation, TFs that can bind to methylated DNA might be good candidates as pioneer TFs. **c** | Factors bound to exons or introns can affect the splicing activity. For example, MeCP2 (methyl-CpG-binding protein 2) binds to a methylated (filled circles) exon (Exon 2), which results in the inclusion of the exon. When the exon is not methylated (open circles), the MeCP2 does not bind, which leads to exclusion of the exon.

A recent comparative study of mouse retina and brain explicitly explored the possible role of methylation sites whose methylation levels were positively correlated with gene expression[101]. Among the differentially methylated regions located within 4 kb upstream of transcriptional start sites, approximately 47% showed a positive correlation with the expression of their putative target genes. These methylation regions are overrepresented in DHSs and are evolutionarily conserved, suggesting that these sites are likely to be functional[101]. More importantly, a distinct set of sequence motifs was discovered in these regions, suggesting that some TFs bind preferentially to these regions[101].

Pioneer TFs. The human genome is not made of linear, naked DNA strands. Instead, it is mainly organized into two forms. One is heterochromatin (or condensed chromatin), in which DNA sequences and histones are highly condensed, and genes in these regions are inactive. The other form is euchromatin (or open chromatin), in which DNA sequences are largely accessible to TFs, and genes in these regions can be activated[102,103]. Chromatin organization is dynamic, and the different types of chromatin can change from one form to another during development or differentiation[104] (FIG. 4).

Pioneer TFs are a unique subset of TFs that drive these chromatin changes. A typical characteristic of pioneer TFs is their ability to bind directly to heterochromatic DNA and recruit other factors to change the status to euchromatin to initiate transcription[105,106]. As DNA in heterochromatin is wrapped tightly around the nucleosomes and is often methylated, it is inaccessible to most TFs; pioneer TFs must possess special features to enable protein–DNA interactions. For example, a handful pioneer TFs (such as OCT4, SOX2 and KLF4) were shown to bind only partial motifs displayed on the nucleosome surface[107].

It could be speculated that the ability to bind mCpG sites might prove a useful property for pioneer TFs. If a pioneer TF can interact with an mCpG site, such an interaction would provide an anchor point for the pioneer TF to open up the closed chromatin. Indeed, we observed a large overlap between known pioneer TFs and proteins that bind to methylated DNA (for example, forkhead box protein A (FOXA) and GATA families, which are the best-studied pioneer factors)[106,108–110]. Interestingly, several of their members showed the ability to bind to methylated DNA, including HOXA5, HOXA9, GATA3 and GATA4 (REF. 53). The mCpG-binding protein KLF4 was also shown to be a pioneer factor[105,111]. Although there is no simple assay to identify pioneer TFs, evidence that TFs are able to bind methylated DNA would provide a short list of candidate pioneer TFs for future tests. Notably, as methylation-binding proteins participate in multiple biological processes, including gene regulation and splicing regulation, not all methylation-binding proteins are likely to be pioneer factors. Yet, binding to an mCpG site is just one approach for a pioneer TF to access condensed chromatin. Other pioneer TFs might have alternative approaches such as binding to partial motifs.

Splicing regulation. Historically, RNA splicing was considered to be regulated only at the post-transcriptional level. On the basis of this idea, DNA methylation was not expected to have any substantial role in splicing regulation. However, it is now well-established that splicing occurs co-transcriptionally, which means that DNA modification could influence RNA splicing. In one study, the authors observed that the binding of CTCF in an exon region created a roadblock for RNA polymerase II elongation and thus promoted the inclusion of the exon[112]. Importantly, the binding of CTCF to the exon or intron was dependent on DNA methylation, suggesting that the methylation status surrounding the spliced exons could affect the inclusion level of these exons. Similarly, the mCpG-binding protein MeCP2 was found to play a part in regulating exon splicing[113]. In this case, a high methylation level led to MeCP2 binding to alternatively spliced exons, which resulted in exon inclusion (FIG. 4). The same trend was observed in a study of the brain

Topologically associated domains
3D spatial organization units of mammalian genomes, within which most enhancer–promoter interactions occur.

methylome of honeybees[114]. A comparison of methylation levels between queen and worker bees revealed that intron-containing histone genes were highly methylated, whereas intronless histone genes were not methylated, suggesting that mCpG-binding proteins might play a part in splicing regulation. This observation is consistent with a global correlation analysis of DNA methylation and differential splicing events between the brain and the retina[115]. Although CTCF motifs were significantly enriched in differentially methylated regions associated with alternative splicing, other motifs were also enriched, suggesting that other TFs might also participate in splicing regulation. Interestingly, the methylation levels in some of the regions were positively associated with the inclusion level of the spliced exons, indicating that other mCpG-binding proteins are involved in regulating the splicing process.

Human diseases. Many studies have shown that aberrant DNA methylation is associated with various human diseases, including some types of cancer[19–21]. For example, profiling the DNA methylation status in the promoters of 272 glioblastoma tumours showed that a distinct subset of samples displayed hypermethylation at a large number of loci, a phenotype termed 'CpG island methylator phenotype' (REF. 116). However, the mechanism by which the altered epigenetic state causes disease remains elusive. A recent study analysed the effect of methylation-dependent protein–DNA interactions on gliomas[117]. The *IDH* genes (*IDH1* and *IDH2*) encode isocitrate dehydrogenases, and mutations in these genes are among the most frequent found in diffuse gliomas[118,119]. Mutant IDH protein is a competitive inhibitor of hydroxylases, including the TET family of 5mC hydroxylases[120–122]. As a result, the *IDH* mutation leads to a remodelling of DNA methylation profiles. Specifically, owing to the interference with TET family proteins, the mutation causes the CpG island methylator phenotype[116,123]. *IDH* mutant gliomas have been shown to exhibit hypermethylation at CTCF binding sites, which leads to a reduction in CTCF binding; loss of CTCF in topologically associated domains removed the domain boundary and caused aberrant gene activation[117].

Conclusions

Similar to genome-wide association studies (GWAS), the profiling of epigenomes (including DNA methylomes) has been extensively carried out under various physiological conditions and in many different biological systems. Transitioning to a post-epigenome era, it is time to elucidate the functional consequences of the observed changes in DNA methylation status and link these changes to phenotypes. Although the role of MBD proteins, as non-sequence-specific methylation readers, has been fairly well-studied, the biological functions of an emerging class of sequence-specific methylation readers and/or effectors remain elusive.

To fully understand the biological processes that are mediated by DNA methylation, many challenges and unanswered questions regarding the methylation readers and/or effectors remain to be tackled in future research. First, we need a more comprehensive catalogue of methylation readers and effectors. Although a few studies have provided more than 100 proteins that can interact with methylated DNA in humans and mice, more readers remain to be discovered in these and other species. An evolutionary conservation analysis of these proteins will provide critical insights into their functional importance. In addition, the identification of the readers for 5mC derivatives (BOX 1) will greatly facilitate the elucidation of their roles in epigenetics. Second, these newly identified methylation readers require more and detailed characterization. For example, it is imperative to understand whether these TFs actually interact with genomic DNA *in vivo*. As we demonstrated above, superimposing ChIP–seq and DNA methylome data sets can be an effective approach to validate mCpG-dependent DNA–TF interactions *in vivo*. Although more technically challenging, ChIP-coupled genome-wide BS-seq is a more direct approach to map the *in vivo* protein–mCpG interactions. Another possible approach is to observe genome-wide changes in TF binding sites by perturbing DNA methylation; for example, by knocking out DNMTs or by pharmacologically removing DNA methylation. Finally, the physiological relevance of protein–mCpG interactions will need to be established. Given a lack of adequate assays or approaches, this could well be a daunting task. A methylation reader usually interacts with both methylated sites and unmethylated sites. Therefore, simply knocking down a methylation reader will not help reveal its role. Identification of the key residues that interact with mCpG sites and the effects of mutations of these residues will provide the next step to dissect the functional role of methylation readers.

Taken together, the notion that TFs may act as DNA methylation readers is an emerging concept supported by predominantly *in vitro* but also by emerging *in vivo* evidence. Of note, this new concept does not refute the conventional view that most TFs do not interact with methylated DNA. Instead, these two scenarios may well coexist in cells. Here, we have focused on this exciting and novel concept with a full awareness that it may apply only to a subset of TFs and to a subset of their binding sites.

1. Bestor, T. H. DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Phil. Trans. R. Soc. Lond. B* **326**, 179–187 (1990).
2. Bird, A. P. & Wolffe, A. P. Methylation-induced repression — belts, braces, and chromatin. *Cell* **99**, 451–454 (1999).
3. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33**, 245–254 (2003).
4. Goll, M. G. & Bestor, T. H. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.* **74**, 481–514 (2005).
5. Bestor, T. H. The DNA methyltransferases of mammals. *Hum. Mol. Genet.* **9**, 2395–2402 (2000).
6. Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.* **19**, 269–277 (2003).
7. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
8. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
9. Krauss, V. & Reuter, G. DNA methylation in *Drosophila* — a critical evaluation. *Prog. Mol. Biol. Transl. Sci.* **101**, 177–191 (2011).
10. Lyko, F., Ramsahoye, B. H. & Jaenisch, R. DNA methylation in *Drosophila melanogaster*. *Nature* **408**, 538–540 (2000).
11. Takayama, S. *et al.* Genome methylation in *D. melanogaster* is found at specific short motifs and is independent of DNMT2 activity. *Genome Res.* **24**, 821–830 (2014).
12. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl Acad. Sci. USA* **107**, 8689–8694 (2010).

13. Selker, E. U. Epigenetic phenomena in filamentous fungi: useful paradigms or repeat-induced confusion? *Trends Genet.* **13**, 296–301 (1997).
14. Jeon, J. *et al.* Genome-wide profiling of DNA methylation provides insights into epigenetic regulation of fungal development in a plant pathogenic fungus, *Magnaporthe oryzae*. *Sci. Rep.* **5**, 8567 (2015).
15. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
16. Bird, A. The essentials of DNA methylation. *Cell* **70**, 5–8 (1992).
17. Jones, P. A. & Takai, D. The role of DNA methylation in mammalian epigenetics. *Science* **293**, 1068–1070 (2001).
18. Robertson, K. D. DNA methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).
19. Jones, P. A. & Baylin, S. B. The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* **3**, 415–428 (2002).
20. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
21. Jones, P. A. & Laird, P. W. Cancer epigenetics comes of age. *Nat. Genet.* **21**, 163–167 (1999).
22. Laird, P. W. The power and the promise of DNA methylation markers. *Nat. Rev. Cancer* **3**, 253–266 (2003).
23. Li, M. *et al.* Sensitive digital quantification of DNA methylation in clinical samples. *Nat. Biotechnol.* **27**, 858–863 (2009).
24. Gavin, D. P. & Sharma, R. P. Histone modifications, DNA methylation, and schizophrenia. *Neurosci. Biobehav. Rev.* **34**, 882–888 (2010).
25. Jiang, Y. H. *et al.* A mixed epigenetic/genetic model for oligogenic inheritance of autism with a limited role for *UBE3A*. *Am. J. Med. Genet. A* **131A**, 1–10 (2004).
26. Nagarajan, R. P., Hogart, A. R., Gwye, Y., Martin, M. R. & LaSalle, J. M. Reduced MeCP2 expression is frequent in autism frontal cortex and correlates with aberrant *MECP2* promoter methylation. *Epigenetics* **1**, e1–e11 (2006).
27. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
28. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
29. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
30. Doi, A. *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* **41**, 1350–1353 (2009).
31. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
32. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12**, 529–541 (2011).
33. Wade, P. A. Methyl CpG binding proteins: coupling chromatin architecture to gene regulation. *Oncogene* **20**, 3166–3173 (2001).
34. Hendrich, B. & Bird, A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* **18**, 6538–6547 (1998).
35. Zhang, X. Y. *et al.* Binding sites in mammalian genes and viral gene regulatory regions recognized by methylated DNA-binding protein. *Nucleic Acids Res.* **18**, 6253–6260 (1990).
36. Saito, M. & Ishikawa, F. The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. *J. Biol. Chem.* **277**, 35434–35439 (2002).
37. Zhang, Y. *et al.* Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* **13**, 1924–1935 (1999).
38. Springer, N. M. & Kaeppler, S. M. Evolutionary divergence of monocot and dicot methyl-CpG-binding domain proteins. *Plant Physiol.* **138**, 92–104 (2005).
39. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked *MECP2*, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23**, 185–188 (1999).
40. Robertson, K. D. & Wolffe, A. P. DNA methylation in health and disease. *Nat. Rev. Genet.* **1**, 11–19 (2000).
41. Prokhortchouk, A. *et al.* The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev.* **15**, 1613–1618 (2001).
42. Rishi, V. *et al.* CpG methylation of half-CRE sequences creates C/EBPα binding sites that activate some tissue-specific genes. *Proc. Natl Acad. Sci. USA* **107**, 20311–20316 (2010).
43. Quenneville, S. *et al.* In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Mol. Cell* **44**, 361–372 (2011).
This paper demonstrates that ZFP57 and its cofactor KAP1 affect chromatin by interacting with methylated ICRs in embryonic stem cells.
44. Liu, Y., Toh, H., Sasaki, H., Zhang, X. & Cheng, X. An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes Dev.* **26**, 2374–2379 (2012).
45. Karlsson, Q. H., Schelcher, C., Verrall, E., Petosa, C. & Sinclair, A. J. Methylated DNA recognition during the reversal of epigenetic silencing is regulated by cysteine and serine residues in the Epstein-Barr virus lytic switch protein. *PLoS Pathog.* **4**, e1000005 (2008).
46. He, X., Futterer, J. & Hohn, T. Sequence-specific and methylation-dependent and -independent binding of rice nuclear proteins to a rice tungro bacilliform virus vascular bundle expression element. *J. Biol. Chem.* **276**, 2644–2651 (2001).
47. Bahar Halpern, K., Vana, T. & Walker, M. D. Paradoxical role of DNA methylation in activation of FoxA2 gene expression during endoderm development. *J. Biol. Chem.* **289**, 23882–23892 (2014).
48. Hantusch, B., Kalt, R., Krieger, S., Puri, C. & Kerjaschki, D. Sp1/Sp3 and DNA-methylation contribute to basal transcriptional activation of human podoplanin in MG63 versus Saos-2 osteoblastic cells. *BMC Mol. Biol.* **8**, 20 (2007).
49. Niesen, M. I. *et al.* Activation of a methylated promoter mediated by a sequence-specific DNA-binding protein, RFX. *J. Biol. Chem.* **280**, 38914–38922 (2005).
50. Bartke, T. *et al.* Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143**, 470–484 (2010).
51. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy) methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–1159 (2013).
This paper describes the identification of proteins that interact with mCpG sites, 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) in ES cells and neuronal progenitor cells using a MS/MS-based approach.
52. Hu, S. *et al.* Profiling the human protein–DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell* **139**, 610–622 (2009).
53. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription factors. *eLife* **2**, e00726 (2013).
This study identifies the transcription factors that preferentially bind to methylated DNA using a protein microarray-based approach and verified that endogenous KLF4 binds to methylated DNA in human ES cells.
54. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
55. Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276 (2008).
56. Mann, I. K. *et al.* CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active *in vivo*. *Genome Res.* **23**, 988–997 (2013).
This paper describes the use of DNA microarrays to identify proteins that interact with methylated DNA.
57. Brinkman, A. B. *et al.* Sequential ChIP–bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* **22**, 1128–1138 (2012).
58. Statham, A. L. *et al.* Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP–seq) directly informs methylation status of histone-modified DNA. *Genome Res.* **22**, 1120–1127 (2012).
59. Gao, F. *et al.* Direct ChIP–bisulfite sequencing reveals a role of H3K27me3 mediating aberrant hypermethylation of promoter CpG islands in cancer cells. *Genomics* **103**, 204–210 (2014).
60. Strogantsev, R. *et al.* Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome Biol.* **16**, 112 (2015).
61. Yoon, H. G., Chan, D. W., Reynolds, A. B., Qin, J. & Wong, J. N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. *Mol. Cell* **12**, 723–734 (2003).
62. Lopes, E. C. *et al.* Kaiso contributes to DNA methylation-dependent silencing of tumor suppressor genes in colon cancer cell lines. *Cancer Res.* **68**, 7258–7263 (2008).
63. Qin, S. *et al.* Kaiso mainly locates in the nucleus *in vivo* and binds to methylated, but not hydroxymethylated DNA. *Chin. J. Cancer Res.* **27**, 148–155 (2015).
64. Blattler, A. *et al.* ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes. *Epigenetics Chromatin* **6**, 13 (2013).
65. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
66. Li, J. J., Jiang, C. R., Brown, J. B., Huang, H. & Bickel, P. J. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *Proc. Natl Acad. Sci. USA* **108**, 19867–19872 (2011).
67. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
68. Liu, Y. *et al.* Structural basis for Klf4 recognition of methylated DNA. *Nucleic Acids Res.* **42**, 4859–4867 (2014).
This study determined the crystal structure of the KLF4-methylated DNA complex and provided the structural basis for mCpG–TF interactions.
69. Dantas Machado, A. C. *et al.* Evolving insights on how cytosine methylation affects protein–DNA binding. *Brief. Funct. Genom.* **14**, 61–73 (2015).
70. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **11**, 73–78 (2014).
71. Lazarovici, A. *et al.* Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl Acad. Sci. USA* **110**, 6376–6381 (2013).
72. Buck-Koehntop, B. A. *et al.* Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso. *Proc. Natl Acad. Sci. USA* **109**, 15229–15234 (2012).
73. Tippin, D. B. & Sundaralingam, M. Nine polymorphic crystal structures of d(CCGGGCCCGG), d(CCGGGCCm^{5}CGG), d(Cm^{5}CGGGCCm^{5}CGG) and d(CCGGGCC(Br)^{5}CGG) in three different conformations: effects of spermine binding and methylation on the bending and condensation of *A*-DNA. *J. Mol. Biol.* **267**, 1171–1185 (1997).
74. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
75. Liu, S. *et al.* Characterization of monoclonal antibody's binding kinetics using oblique-incidence reflectivity difference approach. *MAbs.* **7**, 110–119 (2015).
76. Baubec, T., Ivanek, R., Lienert, F. & Schubeler, D. Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* **153**, 480–492 (2013).
77. Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.* **45**, 1198–1206 (2013).
78. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
This paper demonstrates that some proteins, such as CTCF and REST, can reduce DNA methylation levels at the genomic regions near their binding regions.
79. Charlet, J. *et al.* Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. *Mol. Cell* **62**, 422–431 (2016).
80. Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.* **17**, 520–527 (2001).
81. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405**, 482–485 (2000).
82. Schoenherr, C. J., Levorse, J. M. & Tilghman, S. M. CTCF maintains differential methylation at the *Igf2/H19* locus. *Nat. Genet.* **33**, 66–69 (2003).
83. Krebs, A. R., Dessus-Babus, S., Burger, L. & Schubeler, D. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *eLife* **3**, e04094 (2014).

84. Han, L., Lin, I. G. & Hsieh, C. L. Protein binding protects sites on stable episomes and in the chromosome from *de novo* methylation. *Mol. Cell. Biol.* **21**, 3416–3424 (2001).
85. Fujiki, K. *et al.* PPARγ-induced PARylation promotes local DNA demethylation by production of 5-hydroxymethylcytosine. *Nat. Commun.* **4**, 2262 (2013).
86. Sato, N., Kondo, M. & Arai, K. The orphan nuclear receptor GCNF recruits DNA methyltransferase for *Oct-3/4* silencing. *Biochem. Biophys. Res. Commun.* **344**, 845–851 (2006).
87. 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).
88. Bednarik, D. P. *et al.* 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 (1991).
89. Comb, M. & Goodman, H. M. CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. *Nucleic Acids Res.* **18**, 3975–3982 (1990).
90. Ehrlich, K. C., Cary, J. W. & Ehrlich, M. A broad bean cDNA clone encoding a DNA-binding protein resembling mammalian CREB in its sequence specificity and DNA methylation sensitivity. *Gene* **117**, 169–178 (1992).
91. Falzon, M. & Kuff, E. L. Binding of the transcription factor EBP-80 mediates the methylation response of an intracisternal A-particle long terminal repeat promoter. *Mol. Cell. Biol.* **11**, 117–125 (1991).
92. Iguchi-Ariga, S. M. & Schaffner, W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. *Genes Dev.* **3**, 612–619 (1989).
93. Inamdar, N. M., Ehrlich, K. C. & Ehrlich, M. CpG methylation inhibits binding of several sequence-specific DNA-binding proteins from pea, wheat, soybean and cauliflower. *Plant Mol. Biol.* **17**, 111–123 (1991).
94. Kovesdi, I., Reichel, R. & Nevins, J. R. Role of an adenovirus *E2* promoter binding factor in E1A-mediated coordinate gene control. *Proc. Natl Acad. Sci. USA* **84**, 2180–2184 (1987).
95. Prendergast, G. C., Lawe, D. & Ziff, E. B. Association of Myn, the murine homolog of max, with c-Myc stimulates methylation-sensitive DNA binding and ras cotransformation. *Cell* **65**, 395–407 (1991).
96. Maurano, M. T. *et al.* Role of DNA methylation in modulating transcription factor occupancy. *Cell Rep.* **12**, 1184–1195 (2015).
97. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).
This paper shows that the removal of DNA methylation would create novel binding sites for NRF1 and thus affect the NRF1–DNA interactions *in vivo*, whereas other studies showed that DNA methylation could affect TF–DNA interactions *in vitro*.
98. Blattler, A. & Farnham, P. J. Cross-talk between site-specific transcription factors and DNA methylation states. *J. Biol. Chem.* **288**, 34287–34294 (2013).
99. Baylin, S. B. DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.* **2**, S4–S11 (2005).
100. Yu, D. H. *et al.* Developmentally programmed 3′ CpG island methylation confers tissue- and cell-type-specific transcriptional activation. *Mol. Cell. Biol.* **33**, 1845–1858 (2013).
101. Wan, J. *et al.* Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. *BMC Genomics* **16**, 49 (2015).
102. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868–871 (1974).
103. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
104. Ho, L. & Crabtree, G. R. Chromatin remodelling during development. *Nature* **463**, 474–484 (2010).
105. Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes Dev.* **28**, 2679–2692 (2014).
106. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25**, 2227–2241 (2011).
107. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015).
108. Bossard, P. & Zaret, K. S. GATA transcription factors as potentiators of gut endoderm differentiation. *Development* **125**, 4909–4917 (1998).
109. Laverriere, A. C. *et al.* GATA-4/5/6, a subfamily of three transcription factors transcribed in developing heart and gut. *J. Biol. Chem.* **269**, 23177–23184 (1994).
110. Liu, J. K., DiPersio, C. M. & Zaret, K. S. Extracellular signals that regulate liver transcription factors during hepatic differentiation in vitro. *Mol. Cell. Biol.* **11**, 773–784 (1991).
111. Buganim, Y., Faddah, D. A. & Jaenisch, R. Mechanisms and models of somatic cell reprogramming. *Nat. Rev. Genet.* **14**, 427–439 (2013).
112. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011).
113. Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res.* **23**, 1256–1269 (2013).
This study demonstrates that MeCP2 affects splicing events through its interaction with methylated DNA *in vivo*.
114. Lyko, F. *et al.* The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol.* **8**, e1000506 (2010).
115. Wan, J. *et al.* Integrative analysis of tissue-specific methylation and alternative splicing identifies conserved transcription factor binding motifs. *Nucleic Acids Res.* **41**, 8503–8514 (2013).
116. Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510–522 (2010).
117. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* (2015).
118. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
119. Yan, H. *et al. IDH1* and *IDH2* mutations in gliomas. *N. Engl. J. Med.* **360**, 765–773 (2009).
120. Cairns, R. A. & Mak, T. W. Oncogenic isocitrate dehydrogenase mutations: mechanisms, models, and clinical opportunities. *Cancer Discov.* **3**, 730–741 (2013).
121. Dang, L. *et al.* Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739–744 (2009).
122. Xu, W. *et al.* Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. *Cancer Cell* **19**, 17–30 (2011).
123. Turcan, S. *et al.* IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479–483 (2012).
124. Guo, J. U. *et al.* Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* **17**, 215–222 (2014).
This paper describes genome-wide methylation profiling in adult mammalian brain and the discovery of MeCP2 as a reader of non-CpG methylation.
125. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212–216 (2015).
126. Gabel, H. W. *et al.* Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature* **522**, 89–93 (2015).
127. Kinde, B., Gabel, H. W., Gilbert, C. S., Griffith, E. C. & Greenberg, M. E. Reading the unique DNA methylation landscape of the brain: non-CpG methylation, hydroxymethylation, and MeCP2. *Proc. Natl Acad. Sci. USA* **112**, 6800–6806 (2015).
128. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
129. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
130. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
131. Pfaffeneder, T. *et al.* The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed Engl.* **50**, 7008–7012 (2011).
132. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
133. Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286**, 35334–35338 (2011).
134. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).
135. Sun, Z. *et al.* A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. *Mol. Cell* **57**, 750–761 (2015).
136. Shen, L. *et al.* Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**, 692–706 (2013).
137. Song, C. X. *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* **153**, 678–691 (2013).
138. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
139. Neri, F. *et al.* Single-base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter DNA methylation dynamics. *Cell Rep.* (2015).
140. Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32**, 1231–1240 (2014).
141. Xia, B. *et al.* Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **12**, 1047–1050 (2015).
142. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14**, R119 (2013).
This paper identifies proteins that interact with 5hmC and 5fC using promoter sequences as bait in an MS/MS-based screens. Numerous 5fC interaction partners were discovered, including transcriptional regulators, DNA repair factors and chromatin regulators.
143. Khrapunov, S. *et al.* Unusual characteristics of the DNA binding domain of epigenetic regulatory protein MeCP2 determine its binding specificity. *Biochemistry* **53**, 3379–3391 (2014).
144. Mellen, M., Ayata, P., Dewell, S., Kriaucionis, S. & Heintz, N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151**, 1417–1430 (2012).
145. Valinluck, V. *et al.* Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Res.* **32**, 4100–4108 (2004).

Acknowledgements
The authors thank J. Wan, Y. Zhao and other laboratory members from the Zhu and Qian groups for their discussions. The authors are supported in part by the NIH (EY024580, EY023188 to J.Q. and GM111514 to H.Z.).

Competing interests statement
The authors declare no competing interests.

DATABASES
ENCODE: http://encodeproject.org
ENCSR000EBQ | ENCSR000EBV | ENCSR000BSI | ENCSR000ECC | ENCSR000ECF | ENCSR000BJW | ENCSR000BHQ

FURTHER INFORMATION
Irreproducible Discovery Rate: https://www.encodeproject.org/software/idr

SUPPLEMENTARY INFORMATION
See online article: S1 (figure)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF