Review TRENDS in Genetics Vol.19 No.5 May 2003

The methyl-CpG binding domain and the evolving role of DNA methylation in animals

Brian Hendrich and Susan Tweedie

Institute for Stem Cell Research, The University of Edinburgh, Roger Land Building, The King's Buildings, Edinburgh EH9 3JQ, UK

DNA methylation occurs in bacteria, fungi, plants and animals, however its role varies widely among different organisms. Even within animal genomes, methylation patterns vary substantially from undetectable in nematodes, to global methylation in vertebrate genomes. The number and variety of proteins containing methyl-CpG binding domains (MBDs) that are encoded in animal genomes also varies, with a general correlation between the extent of genomic methylation and the number of MBD proteins. We describe here the evolution of the MBD proteins and argue that the vertebrate MBD complement evolved to exploit the benefits and protect against the dangers of a globally methylated genome.

DNA methylation might have evolved to protect bacterial genomes from invasion by foreign DNA. Two sets of enzymes exist in numerous different bacterial strains, both of which recognize a specific sequence present in the bacterial genome. Upon recognition of this sequence, one of these enzymes, a DNA methyltransferase (DNMT), will covalently modify one position in this sequence by methylating it. The other enzyme, a restriction endonuclease, will cleave the DNA at this sequence, but only if the sequence is unmethylated. Thus, bacteria have devised a way to distinguish their own DNA from that of an invader; in the bacterial genome the sequence will contain the methylated base, whereas in foreign DNA the same sequence will be unmethylated and therefore digested by the restriction endonuclease [1].

The DNMT protein motif is evolutionarily ancient, occurring in all known DNA methyltransferases from bacteria to plants and humans [2]. This remarkable degree of conservation indicates that the ability to methylate one's own DNA provides a major selective advantage to the organism in question. The reason for this advantage is likely to be found in the diversity of functions attributed to DNA methylation [3]. Even within the animals, where DNA methylation is predominantly associated with transcriptional repression, the presence or absence of DNA methylation and of the DNMTs varies, as does the apparent use of DNA methylation within animal genomes [4] (Box 1).

A major change in the distribution of DNA methylation occurred at the invertebrate–vertebrate boundary [5,6]. In invertebrate genomes, DNA methylation is patchy and does not necessarily correlate with the location of genes and/or selfish DNA elements [5,7] (Box 1). By contrast, the genomes of all vertebrates studied thus far are globally methylated [5]. This expansion in methylation appears to have been coupled with an increase in the numbers of both DNMTs and of methyl-CpG binding proteins [8].

It has been suggested that the expansion of the extent of genome methylation that occurred at the onset of vertebrate evolution was accompanied by an increased dependence upon methylated DNA-mediated silencing as a means of transcriptional control [6,9]. We propose that to increase the fidelity of DNA methylation-mediated silencing, and to protect against extensive mutation, there was also a coordinate increase in the number and diversity of methyl-CpG binding proteins encoded in the proto-vertebrate genome. By providing both enhanced transcriptional control and protection against mutation, the methyl-CpG binding proteins could have facilitated the expansion of the methylated DNA compartment within the evolving vertebrate genome. We will review the function of methyl-CpG binding proteins in mammals, and point out subtle differences between mammalian and non-mammalian vertebrates that might indicate ongoing evolution of the use of DNA methylation in the vertebrate lineage.

MBD2 and MBD3: direct descendants of the original MBD

MBD2 and MBD3 are the only vertebrate methyl-CpG binding proteins for which homologues can be identified in invertebrate genomes or expressed sequence tag (EST) collections, and thus we predict that MBD2/3 represents the original methyl-CpG binding protein (Figs 1, 2; Table 1). The putative ancestral MBD2/3 protein is encoded by a single gene in invertebrate genomes, in contrast to the distinct *Mbd2* and *Mbd3* genes present in vertebrates [10]. This trend even holds for the invertebrate *Ciona intestinalis* which, despite being a chordate, contains one *Mbd2/3* gene characteristic of the other invertebrates rather than distinct *Mbd2* and *Mbd3* genes as in the vertebrates (Figs 1, 2). The mammalian *Mbd2* and *Mbd3* genes have an identical genomic structure, differing only in the sizes of their introns, and they encode proteins that are 70% identical. These observations are consistent with a gene duplication event producing distinct *Mbd2* and *Mbd3* genes at a time approximately coincident with the formation of the vertebrate lineage (Figs 2, 3).

Corresponding author: Brian Hendrich (Brian.Hendrich@ed.ac.uk).

http://tigs.trends.com 0168-9525/03/$ - see front matter doi:10.1016/S0168-9525(03)00080-5

Box 1. Genomic methylation patterns

Genomic methylation patterns differ substantially between species and the variation occurs at several levels. Detailed analysis of more species is required to give a clear picture of how these various aspects of methylation have changed during evolution but some general trends are already emerging.

Methyltransferase recognition sequences

Methylation in most animals occurs at cytosines within the sequence CpG with additional low levels of non-CpG methylation reported in some species. Plants are additionally methylated extensively at CpNpG sequences. However there are exceptions to these general rules: CpT was recently identified as the preferred recognition sequence for *Drosophila* methylation [71] and non-CpG methylation is the norm in methylated fungi [72].

Methylation levels

The percentage of methylated cytosines varies substantially between species from no detectable methylation (e.g. the nematode *Caenorhabditis elegans*, the flat worm *Schistosoma mansoni* and the yeast species, *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*) to very high levels in typical vertebrates (60–90% of all CpGs methylated) and most plants. It is assumed that methylation has been lost in some lineages, but the details of which and when remain sketchy because of the paucity of data. All invertebrates tested have either no methylation or some intermediate level of methylation. Direct comparison between species is complicated by the fact that there are several different ways to estimate methylation levels. These methods differ in sensitivity, which partly accounts for the conflicting reports about whether there is methylation in *Drosophila* as methylation in this species is at the limits of detection. To complicate matters further, some methods test only a subset of cytosines and the levels of methyl cytosine are alternatively reported as the percentage of all bases, percentage of all cytosines, or fraction of the subset of sites tested. Regev *et al.* [73] attempt to rationalize the published figures and provide a summary of methylation levels in invertebrates.

Distribution of methylated sites

Vertebrate genomes are globally methylated; methylated cytosines are found over the entire genome apart for short (average ~1-kb) stretches. This unmethylated DNA, the CpG island fraction, accounts for around 1% of the genome and frequently coincides with promoter regions. It is interesting to note that the lamprey, although having a lower percentage of methylated CpG than other vertebrates (20%), shares the same global methylation distribution and has a similar density of methylation [5]. This is significant because methylation density is a factor in methylation dependent silencing.

In all the substantially methylated invertebrates tested the distribution of methylated bases is quite different from that of vertebrates. Rather than global distribution with increased distances between methylated sites, the genome is separated into alternating compartments of methylated and unmethylated DNA. Although the vast majority of invertebrates have not been analysed, it is striking that this distribution pattern is seen in representatives from a diverse cross-section of invertebrate groups (Cnidaria, Orthoptera, Nemertea, Priapulida, Mollusca, Annelida, Bryozoa, Echinodermata) and also includes the invertebrate chordates amphioxus and *Ciona intestinalis*. A switch from compartmentalized to globally distributed methylation is proposed to occur at the invertebrate–vertebrate boundary [5]. The distribution of very low-level methylation in *Drosophila* is not yet known.

What sequences get methylated?

Consistent with widespread distribution, methylation in vertebrates is found in all classes of sequences: exons, introns and regulatory elements, as well as repetitive DNA, transposons and other foreign DNA. A more clear-cut distribution of sequences was originally proposed for the fractionally methylated genomes with genes confined to the unmethylated compartments and the methylated compartment composed of silenced invading elements such as retroviruses and transposons. In some species this seems to be the case; methylation in fungi is confined to repeated sequences and associated with a genome defense function [72]. However, the correlation breaks down in completely in *Ciona intestinalis*, where transposons and retroviruses are frequently unmethylated and genes are found in methylated compartments [7]. There is still much debate about whether genes or invading elements are the primary functional targets and to what extent this might have altered during the course of evolution.

Methylation changes during development

Methylation patterns can also change during the course of development. For instance in mammals there is loss of methylation in early development and then the pattern is re-established [74] whereas methylation in *Drosophila* is only present during early development [65]. Changes also occur locally with loss of methylation at some sites in some tissues; such local demethylation often correlates with expression.

Mammalian MBD2 is a methyl-CpG binding protein, but mammalian MBD3 is not [11]. By contrast, frogs contain two forms of MBD3, one binds specifically to methylated DNA and one does not [12], as well as an MBD2 orthologue. The difference in DNA binding activity between the amphibian and mammalian MBD3 is due to the sequence of their methyl-CpG binding domains (MBDs): mammalian MBD3 differs from other methyl-CpG binding

Table 1. Accession numbers of representative MBD genes sequences in invertebrate (top) and vertebrate (bottom) genomes

Organism	MeCP2	MBD1	MBD2	MBD3	MBD2/3	MBD4
Caenorhabditis elegans	ND[a]	ND	ND	ND	ND	ND
Drosophila melanogaster	ND	ND	ND	ND	AF171098	ND
Anopheles gambiae	ND	ND	ND	ND	BX027478	ND
Ciona intestinalis	ND	ND	ND	ND	BW220395	BW201270
Branchiostoma floridae	ND	ND	ND	ND	AY238339	ND
Takifugu rubripes	SINFRUG00000121582[b]	SINFRUG00000154896[e]	SINFRUG00000147916	CA591604/Y238341[c]	N/A[d]	SINFRUG00000125105
Danio rerio	BE201619	AW422262	AY238336	AY238337/AY238338[c]	N/A	AL921290
Xenopus laevis	AF106951	BC043835	AB061672	AB084168	N/A	AW641890
Gallus gallus	Y14166	ND	BM489705	AJ447899	N/A	AAF68981
Mus musculus	NM_010788	NM_013594	NM_010773	NM_013595	N/A	NM_010774
Homo sapiens	NM_004992	NM_015846	NM_003927	NM_003926	N/A	NM_003925

[a]ND, not detected.
[b]Gene annotation from Ensembl Fugu Genome Database: http://www.ensembl.org/Fugu_rubripes/.
[c]MBD3a/MBD3b.
[d]N/A, not applicable as these genomes have distinct MBD2 and MBD3 genes.
[e]This is incorrectly annotated as MBD2 in the Ensembl Fugu Genome Database.

Fig. 1. MBD2/3: the original methyl-CpG binding protein. Multiple sequence alignment of MBD2/3 proteins from *Drosophila melanogaster* (Dm), mosquito (*Anopheles gambiae*; Ag), cricket (*Achetea domesticus*; Ad), bee (*Apis mellifera*; Am), human (*Homo sapiens*; Hs), mouse (*Mus musculus*; Mm), frog, (*Xenopus laevis*; Xl); zebrafish (*Danio rario*; Dr), pufferfish (*Takifugu rubripes*; Tr), amphioxus (*Branchiostoma floridae*; Bf); sea squirt (*Ciona intestinalis*; Ci) and tapeworm (*Schistosoma mansoni;* Sm). The methyl-CpG binding domain is marked by red lines. The position of insertions within the *Drosophila* sequence are indicated by an 'X' with the sequence shown above. Sequences that are frequently removed due to alternative splicing are shown in red text. The extended N-terminal sequences found only in murine (137aa) and human (134aa) MBD2 are not shown. The AmMbd2/3 is deduced from partial EST sequences and is incomplete at the C-terminus as indicated by the dots. Residues identical in at least 40% of sequences are highlighted in dark blue, similar residues in light blue. Residues acids identical in all sequences are highlighted in purple.

family members at two crucial amino acid residues (Fig. 1, HsMbd3 and MmMbd3, ruler positions 43 and 47). One of these changes is a seemingly mild change of tyrosine to phenylalanine, but the MBD solution structure reveals that the hydroxyl group of that tyrosine makes a crucial hydrogen bond with one of the methyl groups in the 5mCpG base pair [13]. Thus, the substitution destroys the ability to recognize the methyl group. Notably, Tyr → Phe (Fig. 1, ruler position 47) and Lys → His (ruler position 43) are only found in the mammalian MBD3 proteins (Fig. 1). Although it's clear that this change eliminates methyl-CpG binding activity, it is unclear whether it also provides some other, mammalian-specific activity for MBD3.

The genomes of two other non-mammalian vertebrates, pufferfish (*Takifugu rubripes*) and zebrafish (*Danio rerio*), both contain two *Mbd3* genes in addition to an *Mbd2* gene. (These fish proteins are highly similar to mammalian MBD3 at the amino acid level, and are unlikely to be orthologues of the 'MBD3-like' proteins of humans and mice, which show a very low level of similarity to MBD3 [14].) In both fish genomes, one gene encodes an MBD3 with an intact MBD that, based upon sequence conservation, we predict to bind methylated DNA, and the other encodes a version of MBD3 that completely lacks anything resembling an MBD (Fig. 1). Thus, non-mammalian MBD3 probably retains some methyl-binding function, but this function has been lost in mammals (Fig. 3). Yet if mammalian MBD3 has completely lost the ability to bind methylated DNA, then why does the mammalian *Mbd3* gene encode two isoforms of the MBD3 protein, one with an intact MBD and the other with a deleted MBD (Fig. 1) [11]? Furthermore, why should *Mbd3* and *Mbd2* show a genetic interaction in mice [15]? Perhaps the division of labour between MBD2 and MBD3 is not as strict as the existing evidence indicates (see below).

MeCP1 and NuRD

MBD3 is a component of the nucleosome remodelling and histone deacetylation (NuRD) co-repressor complex in humans, frogs and flies [12,16,17]. NuRD is an abundant co-repressor complex that can be recruited to DNA by several different repressor proteins [18]. NuRD component orthologues are found in numerous animals and plants, and deletion or depletion of component gene products results in abnormal (or absence of) embryonic development [15,18]. Exactly what role MBD3 has within the NuRD complex is not clear.

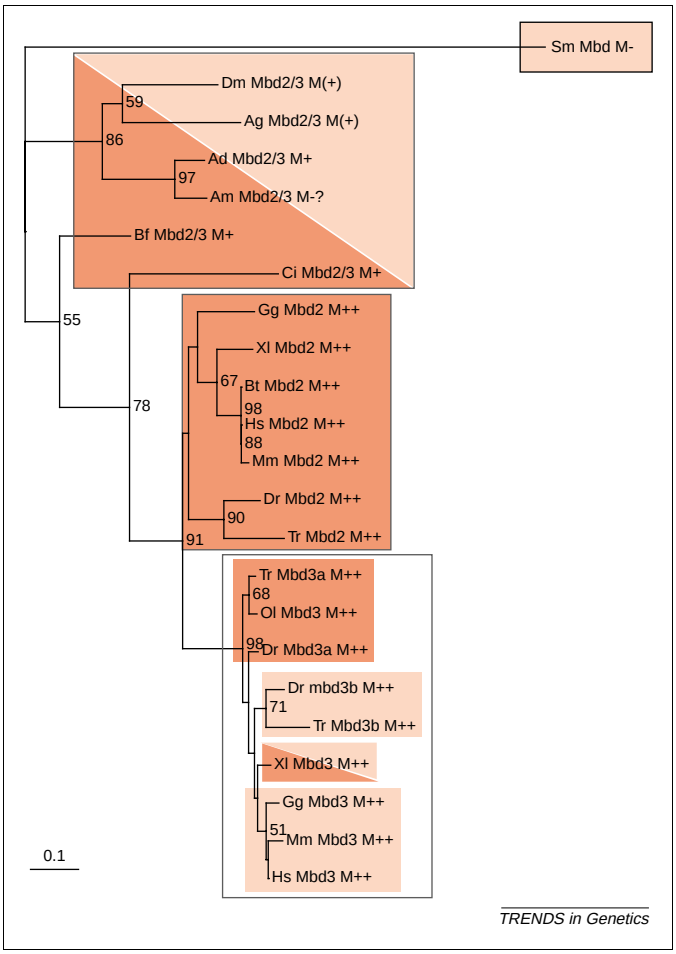

Fig. 2. Phylogenetic analysis of the MBD2/3 proteins suggests that a duplication to produce MBD2 and MBD3 in the vertebrate lineage occurred after *Ciona* branched from the common ancestor (branch marked with '78'). This neighbour-joining tree was generated from an alignment of the above MBD2/3 sequences together with additional sequences from cow (*Bos taurus*; Bt), chicken (*Gallus gallus*; Gg) and medaka (*Oryzias latipes*; Ol) using clustal X (with correcting for multiple substitutions). The tree was constructed with 1000 bootstrap replicates and values of >50% are shown. The scale bar indicates the number of amino acid substitutions per site. The MBD2/3, MBD2 and MBD3 proteins are grouped into boxes, but it is unclear which class the Sm Mbd protein belongs to. The colour scheme reflects the known or predicted functions of the proteins (as in Fig. 3): one function depends on the ability to bind to methylated DNA (dark pink) and one does not (pale pink). The ancestral *Mbd2/3* and *Xenopus Mbd3* genes produce both types of proteins through alternate splicing. The genomic methylation status of each species is indicated: M − unmethylated; M(+) very low level methylation; M + patchy methylation; M++ global methylation.

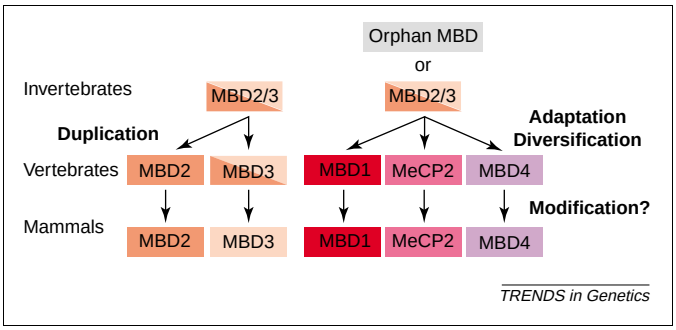

Fig. 3. A model for the evolution of the MBD protein family. The ancestor of MBD2 and MBD3 is proposed to have two functions: one that depends on the ability to bind to methylated DNA (dark pink) and one, produced by alternate splicing, that does not (pale pink). Duplication and divergence has separated these functions in mammals but a methyl-binding activity is retained in MBD3 in other vertebrates. The other MBDs appear to be unique to vertebrates but it is not yet clear whether they arose from MBD2/3 or from another related sequence such as the orphan *Ciona* MBD. The different colours represent diversification of function and in the case of MBD4 a change in binding preference.

MBD2 is a component of the methyl-CpG binding protein 1 (MeCP1) complex [19,20]. MeCP1 [21] was the first methyl-CpG binding activity to be described, and represses transcription in a methylation density-dependent fashion [8]. Biochemical purification revealed that MeCP1 in HeLa cells consists of the NuRD complex and MBD2 [19]. Thus, MeCP1 is very similar to NuRD in *Xenopus* oocytes, where the complex contains both the DNA binding and non-binding forms of MBD3 [12]. It is not yet known whether *Xenopus* MBD2 can also associate with the NuRD complex. Although mammalian MBD2 is a near-stoichiometric component of mammalian MeCP1, most of the NuRD detectable in mammalian cells is not in the MeCP1 complex [16]. Thus, MeCP1 contains only a small fraction of total NuRD activity. It is notable that in gel filtration experiments using HeLa nuclei, a small proportion of MBD2 is detectable in a complex of 100 kD; this complex is distinct from the 1-MD MeCP1 complex [19]. Whether this fraction of MBD2 protein has some MeCP1-independent function remains to be determined.

Functionally redundant, or functions unknown?

MBD2-null mice, despite lacking the MeCP1 complex (as originally defined), show no misexpression of imprinted or X-linked genes [15]. At the time of writing, only one example exists of MBD2 (or any methyl-CpG binding protein) directly influencing the expression of an endogenous gene *in vivo*. Recent work from the Reiner and Bird laboratories demonstrated that the *Il4* gene is inappropriately silenced in MBD2-null T-progenitor cells undergoing TH2 differentiation [22]. This case provides a slightly different view of the repression mediated by methyl-CpG binding proteins from those shown by numerous *in vitro* or *ex vivo* experiments. First, the effect on *Il4* expression is not all-or-none: absence of MBD2 results in leaky repression, not a lack of repression. Second, the site of MBD2 action is not a densely methylated CpG-island promoter, rather it is an enhancer located in an intron of the *Il4* gene that contains relatively few methyl-CpG sites [22]. Although several different researchers have demonstrated binding of MBD2 or other methyl-CpG binding proteins to endogenous methylated CpG islands *in vivo* by chromatin immunoprecipitation [23–27], in no case has a methyl-CpG binding protein been shown to be functionally important for silencing at any of these loci [15,28].

Why doesn't MBD2 deficiency result in a global failure of silencing at methylated loci, given both its impressive credentials as a methylation-dependent transcriptional repressor and the fact that numerous groups have detected its presence at a variety of methylated loci? One possibility is that in the best-known cases of DNA methylation-mediated repression (e.g. imprinted genes, genes on the inactive X chromosome), chromatin changes induced by DNA methyltransferases themselves [29] might be sufficient to keep methylated genes silent, and that MBD proteins are used for fine-tuning as in the case of the *Il4* gene [22]. Another, non-exclusive explanation is that

there might be considerable functional redundancy among the methyl-CpG binding proteins. Although tissues of *Mbd2* $^{-/-}$ mice lack MeCP1, they still contain one to three other methyl-CpG binding activities that could compensate for the lack of MeCP1 [15]. If this is indeed the case, then why are there so many different activities that can apparently do exactly the same thing? Perhaps it is because of the importance of DNA methylation in silencing. Mice unable to maintain their DNA methylation levels die in early postgastrulation development [30,31]. By contrast, deletion of methyl-CpG binding proteins has thus far produced relatively mild phenotypes [15,28,32,33]. By encoding several different methyl-CpG binding proteins, our genomes have evolved several overlapping mechanisms to ensure that the methylation signal is interpreted correctly. Thus, when we experimentally remove one of these layers through gene targeting, several other layers remain in place and the overall effect is minimized. If we remove several of these layers at once, however, then we might reveal a failure of methylation-mediated repression. If not, then it could be that the biochemical data has led us astray, and that the methyl-CpG binding proteins have some other, as yet unidentified, function.

Vertebrate MBDs: new kids on the evolutionary block

The expansion of the methylated compartment of the genome that accompanied the invertebrate-to-vertebrate transition could have provided vertebrates with a way of fine-tuning their genetic programmes [6,9] (Box 1). By globally methylating its genome, this proto-vertebrate genome could reduce unscheduled transcription, decreasing transcriptional 'noise'. This newly acquired ability allowed for (or possibly paralleled) an increase in gene number, thereby increasing biological complexity. However, a greater dependence upon methylation-mediated silencing would provide evolutionary pressure for a concordant increase in the number and diversity of proteins capable of repressing transcription from methylated DNA. Correspondingly, vertebrates have, in addition to MBD2 and MBD3, two other methyl-CpG binding transcriptional repressor proteins containing an MBD: MBD1 and MeCP2 (Figs 3, 4; Table 1). Additionally, vertebrate genomes have evolved the MBD4 protein, which appears to be dedicated specifically to reducing the mutational risk posed by covering the genome with 5-methylcytosine.

MeCP2

The first methyl-CpG binding protein to be cloned was the second methyl-CpG binding activity to be discovered and was thus named MeCP2 [34]. It is also the protein in which the MBD was defined [35] and for which the molecular link between DNA methylation and histone modification was elucidated [36–38]. MeCP2 is a highly abundant chromosomal protein that colocalizes with methylated DNA in nuclei. MeCP2 expression in mice can be detected in most tissues, although expression is particularly high in postmitotic neurons [39]. Its high affinity for methylated DNA and general abundance provide MeCP2 with the right qualifications for a protein dedicated to maintaining transcriptional silence in genomic DNA [40]. As was the case for *Mbd2*, targeting of the *MeCP2* gene in mice revealed that MeCP2 is required neither for embryonic development nor for silencing of imprinted or X-inactivated genes [28,32]. Comparison of global gene expression patterns has revealed no large-scale differences between wild-type and *MeCP2*-null murine brains [41] or human fibroblasts [42]. After

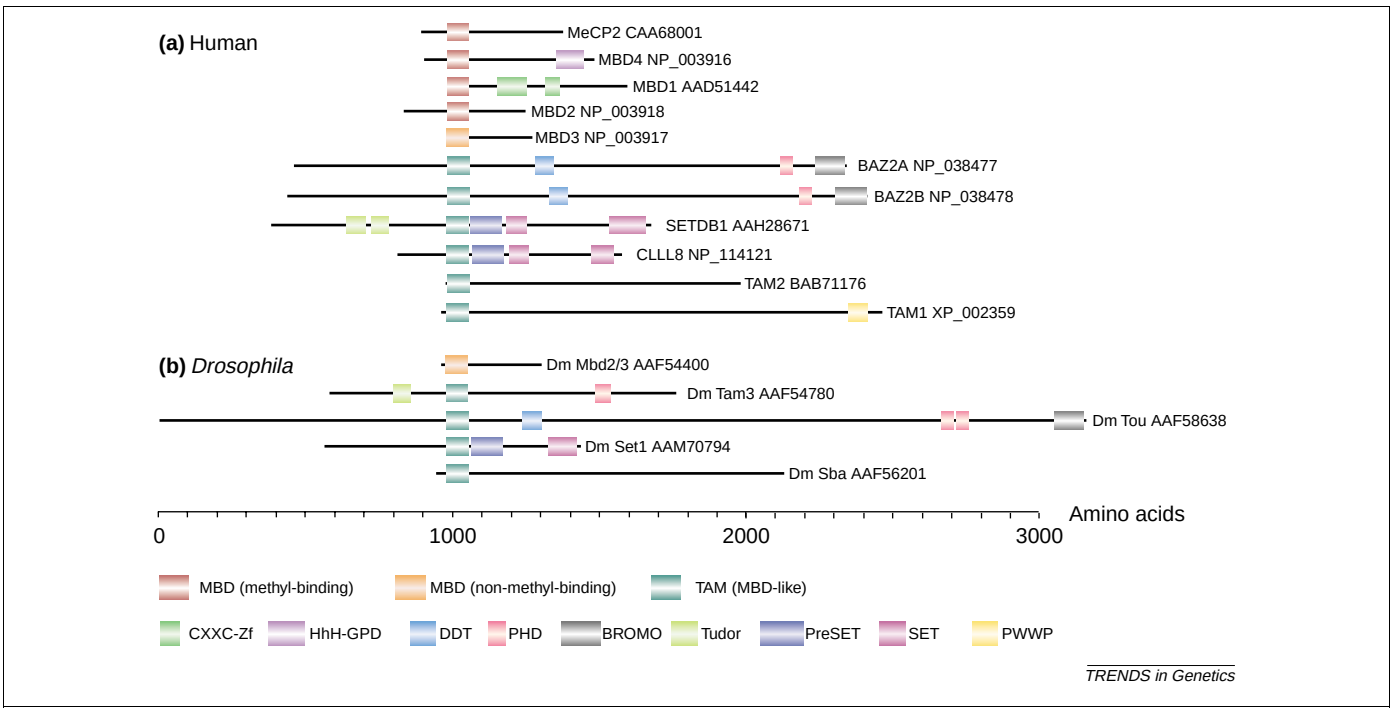

Fig. 4. A box diagram showing the conserved sequence motifs as predicted by DART (http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd = rps) from the human (a) and *Drosophila* (b) MBD/TAM containing proteins. Domains known to bind to methylated DNA are shown in red, and those with unknown function are orange. MBDs known to lack methyl-CpG binding activity (those of MBD3 and DmMBD2/3) are shown in brown. Other conserved domains are indicated.

comparing brain expression patterns using extensive microarray experiments, Tudor *et al.* identified a small number of genes with expression levels that were altered by only 10–20% in the absence of MeCP2 [41]. The results of this comprehensive study are consistent with a 'noise reduction' function for MeCP2, assuming that most increased noise would not be identified using microarrays.

Although we do not yet know the molecular function of MeCP2, we do know that MeCP2 is required in neurons for normal brain function. Mutations in the *MECP2* gene cause Rett syndrome (RTT) in humans [43], and result in a RTT-like phenotype in mice [28,32]. RTT is a progressive neurological disorder that affects girls almost exclusively [44]. Girls suffering from RTT are heterozygous for inactivating mutations in the X-linked *MECP2* gene. Random X chromosome inactivation in these girls means that, on average, half of their cells have inactivated the X chromosome carrying the mutant *MECP2* allele and are should thus be normal. The remaining 50% of cells will have inactivated the X chromosome bearing the wild-type copy of *MECP2* and will thus be capable of expressing no functional MeCP2 protein. Thus, MeCP2-null cells can contribute to the formation and initial function of a mammalian brain, but at some point this mosaic brain starts to be functionally deficient, resulting in the degeneration characteristic of RTT. It is important to note that onset of RTT (or RTT-like symptoms in mice) is accompanied by neither neuronal death nor profound changes in brain morphology, indicating that the physiological defect is likely to be very subtle. Could noise-reduction be this vertebrate-specific function for MeCP2 that, when absent from half of the brain's cells, results in such a devastating disease? If so, then techniques more sensitive than microarray hybridization will have to be used to identify this proposed increase in transcriptional noise.

MBD1

MBD1 is unique among the methyl-CpG binding proteins in that it is capable of repressing transcription from both methylated and unmethylated promoters in cell transfection experiments [45–48]. MBD1 is also unique among the methyl-CpG binding repressors in that it contains within it two or three copies of a cysteine-rich motif (CXXC) that is also found in the DNA methyltransferase protein DNMT1, the histone methyltransferase MLL, and numerous other uncharacterized ESTs (e.g. PFAM 02008; InterPro IPR002857) [45] (Fig. 4). The ability of MBD1 to associate with unmethylated promoters in reporter assays depends upon one CXXC motif [48]. Similar to MeCP2, MBD1 contains a powerful transcriptional repression domain [47,48], and a proportion of its repression activity appears to rely upon the recruitment of histone deacetylases (HDACs), although this activity is less dependent upon HDAC1 and HDAC2 than is that of MeCP1 or MeCP2 [47]. Also similar to MeCP2, MBD1 is an abundant, chromosomal protein [47], that has been detected at the methylated allele of an imprinted gene [27]. Hopefully, gene-targeting experiments will provide insight into the physiological function of MBD1, and it's functional relationships with other methyl-CpG binding proteins.

MBD4

MBD4 is the only member of the MBD family of proteins that does not appear to be involved in transcriptional repression. Rather, MBD4 minimizes the mutability of methyl-CpG in the genome [33,49]. 5-methylcytosine is inherently mutagenic as it spontaneously deaminates to form thymine, resulting in a mismatched T–G basepair. This is in contrast to deamination at unmethylated cytosine, which produces uracil. Uracil is not a base normally found in DNA, so any U–G mismatches are quickly recognized and repaired [50]. The presence of a T–G mismatch presents a problem for the DNA repair machinery: is it supposed to be a T–A basepair, or a C–G basepair? Although T–G mismatches are preferentially repaired to a C–G basepair [51], this remains an imperfect process. The functional consequences of this is that C → T or, on the other strand, G → A transition mutations account for more than 20% of all base substitutions in human genetic disease [52].

At the sequence level, MBD4, with both an MBD and a glycosylase domain [11] (Fig. 4) appeared to be a good candidate for a protein that had co-evolved with the spread of DNA methylation throughout vertebrate genomes to counteract the mutability of 5mC. The initial biochemical characterization of the MBD4 protein lent experimental support to this idea, in that it was demonstrated that the glycosylase domain did indeed remove T or U when present in a mismatched base pair with G without cleaving the DNA strand [53]. Further, it was demonstrated *in vitro* that although its MBD can bind symmetrically methylated CpG sites, it has a higher affinity for 5mCpG/TpG mismatches, i.e. the products of deamination at methyl-CpG in the genome [53].

To test functionally the hypothesis that MBD4 reduces mutability of 5-methylcytosine *in vivo*, mutation frequencies were measured in *Mbd4*-mutant mice. As predicted by the biochemistry, a 3.3-fold increase in C → T transitions at CpG sites was found at a transgenic test locus in *Mbd4*-mutant mice [33,49]. Furthermore, increased mutation frequency at the *adenomatous polyposis coli* (*Apc*) gene [54] in $Mbd4^{-/-}$, $Apc^{Min/+}$ mice resulted in significantly reduced survival and an increased tumour burden compared to their $Mbd4^{+/-}$, $Apc^{Min/+}$ littermates. Thus, these studies provide evidence to support the notion that MBD4 acts to reduce mutation at methylated CpG sites *in vivo*. Consistent with this is the observation that 26–43% of human colorectal tumours showing microsatellite instability also contain mutations in *MBD4* [55,56].

MBD4 consists of two well-conserved, functional domains separated by a poorly conserved spacer region [53] (Fig. 5a). Although the intact MBD4 protein has thus far only been identified in vertebrates, the genome of *Ciona intestinalis* is capable of encoding a protein highly homologous to the glycosylase domain of MBD4, but which apparently does not contain an MBD (Table 1). Notably, this 'MBD-less' MBD4 resembles the only MBD4-like protein identified thus far in chicken [57] or zebrafish (Table 1). No MBD-like sequence can be identified in any of the *Ciona*, chicken or zebrafish MBD4-like ESTs identified thus far, and the recent publication of the draft genomic sequence of *Ciona* reveals no evidence for a hidden MBD for this gene [58]. It is tempting to speculate that vertebrate MBD4 arose as a fusion between some primitive MBD-containing protein and

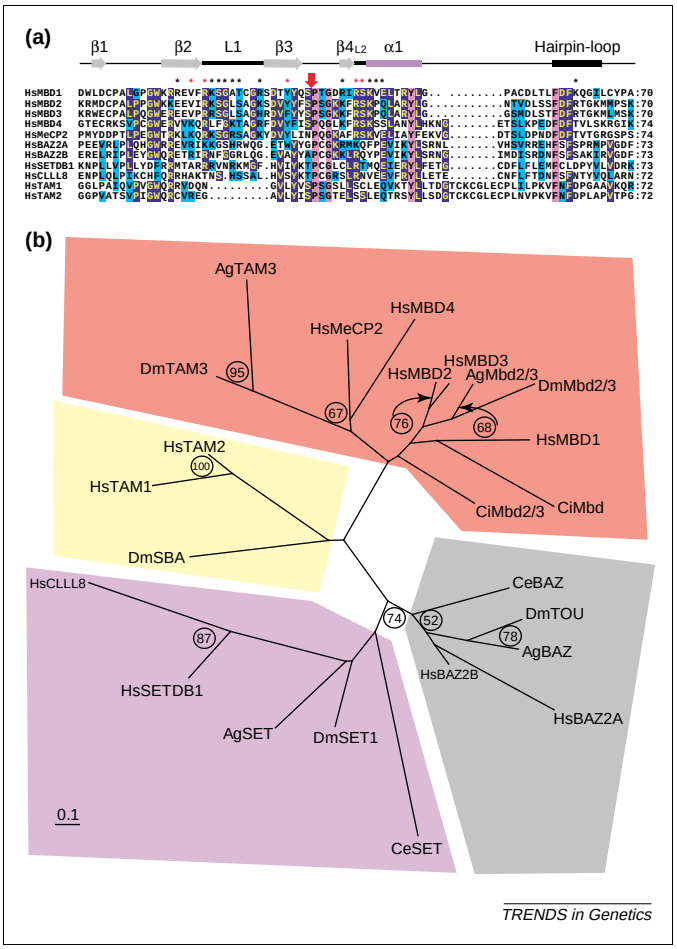

Fig. 5. Evolutionary relationship between MBD and TAM domains. (a) Multiple sequence alignment of the human MBD/TAM domains. The structural elements identified in MBD1 are indicated above the sequence [13]. Asterisks indicate residues in MBD1 that interact with DNA; red for residues interacting with DNA bases and black for residues interacting with the DNA backbone. The position corresponding to the intron conserved in MBDs 1–4, MeCP2 and TAM1–2 is indicated by an arrow. (b) Phylogenetic analysis of the MBD/TAM domains does not provide strong support for these sequences sharing a common ancestor. This neighbour-joining tree was generated from an alignment of the MBD/TAM sequences [as shown in (a)] encoded by the draft genome sequences of human (Hs), *C. intestinalis* (Ci), *D. melanogaster* (Dm), *A. gambiae* (Ag), and *C. elegans* (Ce) using clustal X (with correcting for multiple substitutions). The tree was constructed with 1000 bootstrap replicates and values of >50% are shown. The scale bar indicates the number of amino acid substitutions per site. The functional methyl-CpG binding proteins cluster with the novel MBD identified in *Ciona* and the novel insect sequences that contain a tudor domain. The different branches of the tree are shown within different coloured boxes: The MBD-containing proteins are boxed in orange, the BAZ proteins in grey, SET domain-containing proteins in purple, and SBA/TAM1/TAM2 in yellow. Only MBD1, MBD2, MBD4 and MeCP2 have been shown to be true methyl-CpG binding proteins.

this 'headless' glycosylase. Perhaps this glycosylase helped reduce mutation in the *Ciona* genome, but the addition of an MBD to target it to mutated methyl-CpG basepairs in some early vertebrate genome made it even more efficient at reducing mutations, and it was thus fixed in the vertebrate lineage by natural selection. It will be interesting to discover what form of this protein exists in other chordate genomes on both sides of the vertebrate–invertebrate boundary.

When is an MBD not an MBD?

Since the description of these classical MBD family members, MBD-like sequences have been identified in several other proteins present in several different species [59–62] (Figs 4, 5). A consensus for the MBD is defined in the conserved domain databases (e.g. Pfam01429, SMART00391.5 and LOAD_tam.6). However, not every sequence with similarity to MBD is necessarily a methyl-CpG binding domain, as illustrated by mammalian MBD3. Rather, it seems very likely that a general DNA binding fold has evolved specific binding preferences that in some cases are methyl-CpG specific. Because sequences that were used to build these domain consensuses include a mixture of proteins that do and do not bind to methyl-CpG, homology to these domains is not a reliable predictor of function.

How many additional MBD-like sequences are there? In fact a total of six new 'MBD' containing sequences can be identified in humans and mice (Fig. 3). These new proteins group into pairs of related sequences both by phylogenetic analysis using the MBD domain and to some extent by virtue of other shared domains (Fig. 4). A single homologue exists for each pair from several invertebrate genomes (Fig. 5b). It should be noted that the degree of similarity to the original MBDs is modest and could reflect convergent evolution. However, one compelling argument for a common ancestor between the MBDs and the TAM1/2 pair is the presence of the conserved intron position also found in the MBDs (Fig. 5a). None of the proteins has been extensively characterized functionally (with the exception of TIP5, see below), but they are predicted to be chromatin-associated based on their conserved domains. The key question is whether these proteins are actually involved in interpreting DNA methylation, but in most cases this has not yet been tested. However, comparison with the other methyl-binding proteins and a knowledge of the key residues involved in methyl cytosine recognition from structural work [13,63,64] leads us predict that they will not bind specifically to methyl-CpG, but could have other DNA binding activity (Fig. 5a). Data from the functional analysis of TIP5, the mouse homologue of BAZ2A, supports this notion: a fusion protein containing the TIP5 MBD-like domain bound weakly to DNA but showed no preference for a methylated probe [59]. The lack of evidence for specific methylated DNA binding activity for any of these additional proteins emphasizes that a degree of caution is required in ascribing function on the basis of homology alone. From this point of view, we choose to use the Library of Ancient Domains (LOAD) name for the domain TAM (for 'TIP5, ARBP, MBD') that reflects the similarity without implying homology or methyl-binding activity. Numerous MBD-like sequences can also be identified in the genomes of various plant species, although it is currently unclear which of them, if any, are bona fide methyl-CpG binding proteins.

The complete genome sequence of *Ciona intestinalis* encodes two MBDs: the MBD2/3 related sequence shown in Fig. 1 and a completely novel protein (encoded by Ci0100150311; http://genome.jgi-psf.org/ciona4/ciona4.home.html). This protein clusters with the functional MBDs (CiMbd; Fig. 5b) but it remains to be determined whether it actually binds methylated DNA.

MBDs and methylation in insects

Drosophila melanogaster has an additional gene capable of encoding an MBD-like sequence, we have called this TAM3 (Figs 4, 5b). Homologues of this protein are not detected in vertebrates, but there is a closely related sequence in

mosquito, suggesting an insect-specific function. DNA methylation in insects has been the subject of some debate. It now seems clear that there are small amounts of DNA methylation in *Drosophila*, but it appears only transiently during a short window in early development and even then is at the limits of detection [65] (Box 1). DNA extracted from one mosquito (*Aedes albopictus*) cell line was also reported to possess very low levels of cytosine methylation (0.03%) [66]. It remains to be determined whether this insect also displays developmentally regulated changes in DNA methylation levels. A single DNMT can be identified in the genome of a different mosquito, *Anopheles gambiae* (EAA08679) as well as that of *D. melanogaster* [17,67]. These insect DNMTs are closely related to mammalian DNMT2. This is the Cinderella of the DNMT family; being largely ignored after no activity could be demonstrated and a mouse knockout lacked a phenotype [68]. However, the enzyme is clearly very well conserved throughout evolution – evidence that it must be doing something.

Is low level insect methylation 'read' by MBDs, or does it have a direct effect? There is some controversy about whether the *Drosophila* MBD2/3 protein has the ability to bind to methylated DNA. The general consensus is that it does not [17,69], and this fits with the fact that the MBD region of the protein is severely disrupted (Fig. 1). However, a short form of the protein generated by alternative splicing was reported to show methyl-specific binding activity [70] despite the fact that the splice removes almost half of the MBD. *Anopheles gambiae* also has a homologue of MBD2/3 that does not appear to be as obviously inactive as does the *Drosophila* one (based upon its sequence; Fig. 1); it is not yet known whether it has methyl-binding activity. One insect known to have both a methyl-CpG binding MBD2/3 protein and higher levels of DNA methylation is cricket (*Acheta domesticus*) [17], although there has been no analysis of the sequence context of this methylation and nothing is known about cricket DNMTs. The identification of the active DNA methyltransferase and its targets in these insects, and of the role(s) of insect MBD-containing proteins will certainly provide some vital insights into the evolving role of DNA methylation in animals.

TAM chickens and MBD eggs?

Where did the MBD come from? Both the MBD and TAM domain can be found in numerous animal genomes (e.g. Fig. 5b), with one exception being *Caenorhabditis elegans*, which does not appear to have an MBD. So which came first? Was the nonspecific DNA binding ability of the TAM domain converted into an MBD? Or was the methyl-CpG binding activity of the MBD recruited to other purposes by the TAM domain proteins? Or do they have separate origins? What is clear is that the MBD/TAM domain is adaptable: it is probably a DNA binding domain in the TAM proteins, although whether it has any sequence specificity remains to be determined. The MBDs from MBD1, MBD2 and MeCP2 are capable of binding to a single, symmetrically methylated CpG *in vitro*. The MBD in MBD4, although most similar to that in MeCP2, has yet a different DNA binding specificity in that it prefers mCpG when paired with TpG. Furthermore, the MBD of mammalian MBD3 has nonspecific DNA binding activity in *in vitro* assays, but whether this domain ever actually sees the light of DNA is unclear.

What is also clear is evidence pointing to a changing role for DNA methylation and methyl-CpG binding proteins in animal evolution (Fig. 3). However it might have arisen, the ancestral MBD appears to have been exploited to provide greater transcriptional control as well as protection from mutation in the expanding, methylated proto-vertebrate genome, possibly facilitating the expansion of methylation through the genome. Thus, the study of these proteins in invertebrates, non-mammalian vertebrates, and mammals will probably provide distinct, but related snapshots of the changing role of DNA methylation in animal evolution.

Acknowledgements

We thank Adrian Bird and Miho Suzuki for critical reading of the manuscript. Cloning and initial characterization of amphioxus and Fugu cDNAs was done in Adrian Bird's laboratory, and we acknowledge the help of Callum Brandon in that work. Research in the authors' laboratories is funded by the Wellcome Trust (B.H.) and the BBSRC (S.T.).

References

1 Arber, W. and Linn, S. (1969) DNA modification and restriction. *Annu. Rev. Biochem.* 38, 467–500
2 Kumar, S. *et al.* (1994) The DNA (cytosine-5) methyltransferases. *Nucleic Acids Res.* 22, 1–10
3 Colot, V. and Rossignol, J.L. (1999) Eukaryotic DNA methylation as an evolutionary device. *BioEssays* 21, 402–411
4 Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.* 16, 6–21
5 Tweedie, S. *et al.* (1997) Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol. Cell. Biol.* 17, 1469–1475
6 Bird, A.P. (1995) Gene number, noise reduction and biological complexity. *Trends Genet.* 11, 94–100
7 Simmen, M.W. *et al.* (1999) Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* 283, 1164–1167
8 Bird, A.P. and Wolffe, A.P. (1999) Methylation-induced repression–belts, braces, and chromatin. *Cell* 99, 451–454
9 Bestor, T.H. (1990) DNA methylation: evolution of a bacterial immune function into a regulator of gene expression and genome structure in higher eukaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 326, 179–187
10 Hendrich, B. *et al.* (1999) Genomic structure and chromosomal mapping of the murine and human *Mbd1*, *Mbd2*, *Mbd3*, and *Mbd4* genes. *Mamm. Genome* 10, 906–912
11 Hendrich, B. and Bird, A. (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* 18, 6538–6547
12 Wade, P.A. *et al.* (1999) Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat. Genet.* 23, 62–66
13 Ohki, I. *et al.* (2001) Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell* 105, 487–497
14 Jiang, C.L. *et al.* (2002) MBD3L1 and MBD3L2, two new proteins homologous to the methyl-CpG-binding proteins MBD2 and MBD3: characterization of MBD3L1 as a testis-specific transcriptional repressor. *Genomics* 80, 621–629
15 Hendrich, B. *et al.* (2001) Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. *Genes Dev.* 15, 710–723
16 Zhang, Y. *et al.* (1999) Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* 13, 1924–1935
17 Tweedie, S. *et al.* (1999) Vestiges of a DNA methylation system in *Drosophila melanogaster*? *Nat. Genet.* 23, 389–390
18 Ahringer, J. (2000) NuRD and SIN3 histone deacetylase complexes in development. *Trends Genet.* 16, 351–356
19 Feng, Q. and Zhang, Y. (2001) The MeCP1 complex represses

transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. *Genes Dev.* 15, 827–832
20 Ng, H.H. *et al.* (1999) MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat. Genet.* 23, 58–61
21 Meehan, R.R. *et al.* (1989) Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* 58, 499–507
22 Hutchins, A.S. *et al.* (2002) Gene silencing quantitatively controls the function of a developmental trans-activator. *Mol. Cell* 10, 81–91
23 El-Osta, A. *et al.* (2002) Precipitous release of methyl-CpG binding protein 2 and histone deacetylase 1 from the methylated human multidrug resistance gene (MDR1) on activation. *Mol. Cell. Biol.* 22, 1844–1857
24 Nguyen, C.T. *et al.* (2001) Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: correlation of accessibility, methylation, MeCP2 binding and acetylation. *Nucleic Acids Res.* 29, 4598–4606
25 Magdinier, F. and Wolffe, A.P. (2001) Selective association of the methyl-CpG binding protein MBD2 with the silent p14/p16 locus in human neoplasia. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4990–4995
26 Drewell, R.A. *et al.* (2002) Methylation-dependent silencing at the H19 imprinting control region by MeCP2. *Nucleic Acids Res.* 30, 1139–1144
27 Fournier, C. *et al.* (2002) Allele-specific histone lysine methylation marks regulatory regions at imprinted mouse genes. *EMBO J.* 21, 6560–6570
28 Guy, J. *et al.* (2001) A mouse *Mecp2*-null mutation causes neurological symptoms that mimic Rett syndrome. *Nat. Genet.* 27, 322–326
29 Burgers, W.A. *et al.* (2002) DNA methyltransferases get connected to chromatin. *Trends Genet.* 18, 275–277
30 Li, E. *et al.* (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 69, 915–926
31 Okano, M. *et al.* (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. *Cell* 99, 247–257
32 Chen, R.Z. *et al.* (2001) Deficiency of methyl-CpG binding protein-2 in CNS neurons results in a Rett-like phenotype in mice. *Nat. Genet.* 27, 327–331
33 Millar, C.B. *et al.* (2002) Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science* 297, 403–405
34 Lewis, J.D. *et al.* (1992) Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* 69, 905–914
35 Nan, X. *et al.* (1993) Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Res.* 21, 4886–4892
36 Nan, X. *et al.* (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* 393, 386–389
37 Jones, P.L. *et al.* (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat. Genet.* 19, 187–191
38 Fuks, F. *et al.* (2003) The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J. Biol. Chem.* 278, 4035–4040
39 Shahbazian, M.D. *et al.* (2002) Insight into Rett syndrome: MeCP2 levels display tissue- and cell-specific differences and correlate with neuronal maturation. *Hum. Mol. Genet.* 11, 115–124
40 Nan, X. *et al.* (1997) MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* 88, 471–481
41 Tudor, M. (2002) Transcriptional profiling of a mouse model for Rett syndrome reveals subtle transcriptional changes in the brain. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15536–15541
42 Traynor, J. *et al.* (2002) Gene expression patterns vary in clonal cell cultures from Rett syndrome females with eight different MECP2 mutations. *BMC Med. Genet.* 3, 12
43 Amir, R.E. *et al.* (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* 23, 185–188
44 Kerr, A.M. *et al.* (2001) Guidelines for reporting clinical features in cases with MECP2 mutations. *Brain Dev.* 23, 208–211
45 Cross, S. *et al.* (1997) A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. *Nat. Genet.* 16, 256–259
46 Fujita, N. *et al.* (1999) Methylation-mediated transcriptional silencing in euchromatin by methyl-CpG binding protein MBD1 isoforms. *Mol. Cell. Biol.* 19, 6415–6426
47 Ng, H.H. *et al.* (2000) Active repression of methylated genes by the chromosomal protein MBD1. *Mol. Cell. Biol.* 20, 1394–1406
48 Fujita, N. *et al.* (2000) Mechanism of transcriptional regulation by methyl-CpG binding protein MBD1. *Mol. Cell. Biol.* 20, 5107–5118
49 Wong, E. *et al.* (2002) Mbd4 inactivation increases C → T transition mutations and promotes gastrointestinal tumor formation. *Proc. Natl. Acad. Sci. U. S. A.* 99, 14937–14942
50 Nilsen, H. *et al.* (2001) Excision of deaminated cytosine from the vertebrate genome: role of the SMUG1 uracil-DNA glycosylase. *EMBO J.* 20, 4278–4286
51 Brown, T.C. and Jiricny, J. (1987) A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell* 50, 945–950
52 Krawczak, M. *et al.* (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.* 63, 474–488
53 Hendrich, B. *et al.* (1999) The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* 401, 301–304
54 Su, L.K. *et al.* (1992) Multiple intestinal neoplasia caused by a mutation in the murine homolog of the APC gene. *Science* 256, 668–670
55 Bader, S. *et al.* (1999) Somatic frameshift mutations in the MBD4 gene of sporadic colon cancers with mismatch repair deficiency. *Oncogene* 18, 8044–8047
56 Riccio, A. *et al.* (1999) The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability. *Nat. Genet.* 23, 266–268
57 Zhu, B. *et al.* (2000) 5-Methylcytosine DNA glycosylase activity is also present in the human MBD4 (G/T mismatch glycosylase) and in a related avian sequence. *Nucleic Acids Res.* 28, 4157–4165
58 Dehal, P. *et al.* (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 298, 2157–2167
59 Strohner, R. *et al.* (2001) NoRC–a novel member of mammalian ISWI-containing chromatin remodeling machines. *EMBO J.* 20, 4892–4900
60 Schultz, D.C. *et al.* (2002) SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* 16, 919–932
61 Mabuchi, H. *et al.* (2001) Cloning and characterization of *CLLD6*, *CLLD7*, and *CLLD8*, novel candidate genes for leukemogenesis at chromosome 13q14, a region commonly deleted in B-cell chronic lymphocytic leukemia. *Cancer Res.* 61, 2870–2877
62 Roloff, T.C. *et al.* (2003) Comparative study of methyl-CpG-binding domain proteins. *BMC Genomics* 4, 1
63 Wakefield, R.I. *et al.* (1999) The solution structure of the domain from MeCP2 that binds to methylated DNA. *J. Mol. Biol.* 291, 1055–1065
64 Ohki, I. *et al.* (1999) Solution structure of the methyl-CpG-binding domain of the methylation-dependent transcriptional repressor MBD1. *EMBO J.* 18, 6653–6661
65 Lyko, F. (2001) DNA methylation learns to fly. *Trends Genet.* 17, 169–172
66 Adams, R.L. *et al.* (1979) Methylation of mosquito DNA. *Biochim. Biophys. Acta* 563, 72–81
67 Lyko, F. *et al.* (2000) The putative *Drosophila* methyltransferase gene dDnmt2 is contained in a transposon-like element and is expressed specifically in ovaries. *Mech. Dev.* 95, 215–217
68 Okano, M. *et al.* (1998) Dnmt2 is not required for *de novo* and maintenance methylation of viral DNA in embryonic stem cells. *Nucleic Acids Res.* 26, 2536–2540
69 Ballestar, E. *et al.* (2001) A *Drosophila* MBD family member is a transcriptional co-repressor associated with specific genes. *Eur. J. Biochem.* 268, 5397–5406
70 Roder, K. *et al.* (2000) Transcriptional repression by *Drosophila* methyl-CpG-binding proteins. *Mol. Cell. Biol.* 20, 7401–7409
71 Lyko, F. *et al.* (2000) DNA methylation in *Drosophila melanogaster*. *Nature* 408, 538–540
72 Selker, E.U. (1997) Epigenetic phenomena in filamentous fungi: useful paradigms or repeat-induced confusion? *Trends Genet.* 13, 296–301
73 Jablanka, E. and Regev, A. (1995) Gene number, methylation and biological complexity. *Trends Genet.* 11, 383–384
74 Reik, W. *et al.* (2001) Epigenetic reprogramming in mammalian development. *Science* 293, 1089–1093