doi:10.1016/j.jmb.2011.01.056 J. Mol. Biol. (2011) 409, 47–53

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

REVIEW

The Dinucleotide CG as a Genomic Signalling Module

Adrian Bird

The Wellcome Trust Centre for Cell Biology, University of Edinburgh, Michael Swann Building, The King's Buildings, Edinburgh EH9 3JR, UK

Received 13 December 2010; received in revised form 27 January 2011; accepted 28 January 2011
Available online 3 February 2011

Edited by M. Yaniv

Keywords:
CpG islands;
DNA methylation;
MeCP2;
Cfp1

The operon model proposed the existence of a category of proteins that control gene expression by interacting with specific DNA sequences. Since then, a large number of transcription factors recognizing a diversity of sequence motifs have been discovered. This article discusses an unusually short protein recognition sequence, 5′CG, which is read by multiple DNA binding proteins. CG exists in three distinct chemical states, two of which bind mutually exclusively to proteins that modulate chromatin structure. Non-methylated CG, which is highly concentrated at CpG island promoters, recruits enzymes that create the mark of promoter activity, trimethyl-lysine 4 of histone H3. Methylated CG, on the other hand, is a gene silencing mark and accordingly recruits enzymes that deacetylate histones. Thus, CG, despite its simplicity, has the properties of a genome-wide signalling module that adds a layer of positive or negative control over gene expression.

Introduction

Cells contain a complete set of genetic information encoded in their DNA, but not all genes are expressed at any one time. In bacteria, for example, certain genes become active only in response to an environmental (e.g., nutritional) signal. Mammalian cells have large numbers of silent genes, as many proteins are specific for a particular cell type and must not be expressed in other kinds of cells. β-Globin, growth hormone, and opsin, to name only a few, are each repressed in the vast majority of mammalian cell types. This gene expression programme is a key end point of the process of development, ensuring that irrelevant genes are shut down, whereas required genes are either active or poised for expression when called upon. Jacob and Monod were the first to make headway in understanding how differential gene expression is achieved, and the conclusions of their seminal paper in 1961[1] continue to reverberate through contemporary molecular biology. A major revelation was that there are proteins whose role is not to facilitate metabolic processes directly, but which recognise and interact in *trans* with specific DNA sequences to control gene expression. Even today, when chromatin structure and DNA/histone modifications are high on the research agenda, proteins that recognise specific DNA sequences are still considered the primary instigators of differential activity across the genome. To direct any process to a specific region of the genome, it is necessary to be able to distinguish the target DNA from a vast excess of non-target DNA. Sequence-specific DNA binding proteins and nucleic acid polymers can do this, but it is not clear that anything else can. Therefore, changing patterns of transcription, replication, and recombination with concomitant alterations in histone/DNA modification and chromatin conformation are almost certainly secondary to a targeted triggering event based on the recognition of DNA sequence.[2,3]

The genome of the bacterium *Escherichia coli* is estimated to encode ~250 DNA binding proteins. Not all the DNA sequences that they recognise are yet known, but the length of DNA that is read is often in the range of 20–30 bp, albeit with considerable

E-mail address: a.bird@ed.ac.uk.

Abbreviations used: CGIs, CpG islands; H3K4me3, trimethylation of lysine 4 of histone H3; MBD, methyl-CpG binding domain.

variability.[4] The lac repressor discovered by Jacob and Monod, for example, binds a 21-bp site.[5] As genomes get bigger across the phylogenetic spectrum, the discriminatory power needed to target specific genomic locations becomes more daunting. To cope with this, one might expect that DNA binding motifs would become longer. In fact, the opposite is the case: the average length of DNA recognised by mammalian transcription factors is usually 6–8 bp.[6] To circumvent this logistical problem, multiple factors bind to clustered sites and collaborate to regulate gene expression. The β-globin locus control region, for example, binds the erythroid DNA sequence-specific transcription factors GATA1, EKLF/KLF1, and NF-E2, among others.[7] The probability that an appropriate cluster of binding sites will occur by chance is low. Combinatorial binding of transcription factors thereby restores the missing specificity.

This article concerns a eukaryotic DNA binding motif that is atypically short by any standards—just 2 bp long. At first sight, this seems much too simple to be biologically interesting because a 2-bp sequence will occur very frequently by chance. For example, the sequence AG would be expected once every 17 bp, on average, in a genome with a base composition of 40% G+C. Here, I will discuss evidence that the dinucleotide sequence CG, despite its simplicity, has the properties of a genomic signalling module that participates in local and global regulation of gene expression through interaction with DNA binding proteins. The focus will be on animal genomes, but many of the same arguments apply to plants, where CG is also a genomic signal.[8]

CG exists in chemically distinct forms

CG is a self-complementary DNA sequence, but not uniquely so, as it shares this property with three other dinucleotides: AT, TA, and GC. CG differs from the others in that it exists in three chemically distinct forms: unmethylated, methylated, and hydroxymethylated.[9,10] CG methylation involves modification of the cytosine base at the 5 position of the pyrimidine ring. Viewed along the axis of the DNA double helix, it is evident that the methyl group lies in the major groove of B-form DNA and does not sterically interfere with the exquisite base pairing between G and C, which would compromise coding specificity. An important advantage of self-complementarity is that modifications at CG can be copied at DNA replication.[11,12] In the case of methylated CG, the maintenance DNA methyltransferase Dnmt1 only methylates the newly synthesised progeny strand if the parental strand bears a methyl group. Heritability means that DNA methylation patterns are stable through cell division.

The most important biological consequence of DNA methylation that has been observed in many systems is long-term silencing of transcription.[13] This can be accomplished in two ways: (1) repulsion of transcription factors that require non-methylated CG in their binding site[14]; (2) attraction of proteins that specifically bind m5CG (see below). Set against this useful property, there is a downside to the presence of 5-methylcytosine—its mutability. Both 5-methylcytosine and cytosine are prone to hydrolytic deamination, giving rise to thymine and uracil, respectively. Uracil, being an alien DNA base, is recognised by a dedicated DNA repair system and restored to cytosine.[15] Thymine, on the other hand, is an authentic DNA component, albeit incorrectly base-paired after m5C deamination, and this appears to interfere with its efficient repair. Glycosylases that can remove T from a T:G mismatch have been identified and, in the case of MBD4, shown to contribute to repair, but their efficiency appears to be inadequate to eliminate the problem altogether.[16] As a result, many C-to-T transitions arise at sites of CG methylation and cause mutations.

Given the existence of glycosylases such as MBD4[17] and TDG,[18,19] it is interesting to speculate about why they fail to rectify all 5mC deaminations. A possible reason is that enzymes removing T from DNA must not be overzealous or they will cause more mutations than they repair. After all, there are nearly 10^9 bona fide T residues in a haploid genome, only one of which is likely to be due to m5C deamination at any one time. The scope for potentially mutagenic error is therefore large. This may be why uracil glycosylases, which remove U from a U:G mismatch, are inert when faced with a T:G mismatch. Given that the double helix breathes continually, the presence of a mismatch may not be enough to distinguish which T residues should be excised. Interestingly, it has been suggested that replacement of U by T in the presumed ancestral RNA genome during the early evolution of life was driven by the need to distinguish the deamination product of C from a normal DNA base so that it could be recognised for repair.[20] By methylating C, genomes recreate the dilemma that was so effectively circumvented by the invention of T. Mutability is the legacy of this evolutionary step. Medically, for example, it is evident that about 30% of all point mutations causing human disease arise at CG sites, almost certainly as a result of DNA methylation.[21]

Hydroxymethylation also occurs at the 5 position of the cytosine ring.[9,10] This is not coincidental, as it is created by enzymes that specifically hydroxylate the methyl group of m5C to create this altered form.[10] Little is yet known about the functional significance of this modification, but two alternatives will be important to distinguish: (1) hmC is a chromatin signal in its own right with distinct biological consequences; (2) it is an intermediate in the demethylation of DNA. With respect to the latter possibility, it is clear that demethylation of m5C

without DNA replication occurs during animal development, but its mechanism is uncertain.[22] Breakage of carbon–carbon bonds is notoriously tricky chemically, but the presence of a potential formaldehyde leaving-group arguably would make this process more energetically favourable. Although this scenario is attractive, at the time of writing these two potential roles as a signal or a chemical intermediate have not been distinguished experimentally.

CG frequency varies nonrandomly across the genome

The frequency of CG moieties throughout the genome fluctuates widely, but follows a discernible pattern. The dominant influence on their distribution appears to be the mutability resulting from methylation of cytosine, as discussed above. Approximately 80% of all genomic CG sites are methylated in vertebrate animals. Consequently, there is a continuous and almost genome-wide mutational pressure tending to reduce the frequency of CGs. This does not of course mean that CGs will eventually disappear, as accelerated loss is accompanied by mutational events that create new CGs, albeit at a lower rate. The result is an equilibrium CG frequency throughout most of the genome that is less than what one would expect based on an average genomic base composition of 40% G+C. A CG would be anticipated every 25 bp of DNA on average, whereas the observed frequency is one CG per ~100 bp—about 4-fold less than expected. Calculations based on this observed equilibrium and on more direct measurements of CG mutability agree that m5C undergoes transition mutations about an order of magnitude more rapidly than non-methylated C and other bases.[23,24]

While most of the genome is CG-deficient because of its enhanced mutation rate, short regions of DNA called CpG islands (CGIs) are exceptional in retaining their expected CG density.[25] CGIs are ~1000 bp long on average, accounting for ~1% of the genome in total. Importantly, they encompass the transcription start sites of most mammalian genes. A lack of CG deficiency is not the only feature that distinguishes mammalian CGIs from the bulk of the genome. They are also significantly more G+C-rich than bulk genomic DNA (65% compared with 40% G+C in *Homo sapiens*). This altered base composition is not simply due to the increased density of CGs, but is an independent phenomenon. The combined absence of CG suppression and G+C richness mean that CGIs have a 10 times higher CG frequency than the surrounding bulk genome, or about 1 CG per 10 bp. CGIs are distinct in vertebrates predominantly because of their lack of DNA methylation and consequent absence of CpG deficiency, which sets them apart from bulk genomic DNA. The invertebrates *Drosophila melanogaster* and *Caenorhabditis elegans* and the fungus *Saccharomyces cerevisiae* have little or no DNA methylation, and as a result, CpG occurs at the expected frequency throughout the genome. CGIs are not detectable in these genomes because, in a sense, the whole genome is CGI-like.

What evolutionary tendencies have created CGIs? Lack of CG deficiency is very probably due to the absence of mutagenic cytosine methylation in germ cells and the totipotent stages of the embryo. Cells of these types, unlike somatic cells, have genomes that will live on for generations, and their methylation, or the lack of it, will influence CG frequency. In contrast, methylation of the somatic genome, while it may lead to loss of CGs by mutation in that organism, does not affect CG density across generations. The evolutionary explanation for G+C richness is less clear. Cumulative selection of single-nucleotide transitions from A/T to G/C within the 1000-bp CGI sounds unlikely, as each mutation would need to confer a selective advantage sufficient to provoke its spread throughout the species. An alternative scenario is that a passive mechanism leaves a footprint of biased DNA base composition at these regions. For example, CGIs are often origins of replication,[26,27] and it is conceivable that this role influences the local base composition at the site of replication initiation.[28] Recent results support colocalisation of CGIs and replication origins,[27] and there is evidence that nucleotide pool composition can influence origin usage[29] and might conceivably bias base composition in the long term. The trend towards G+C richness could still be under selection in this case, but the mechanism would not be based on an accumulation of individual stochastic point mutations. Thus far, this scenario remains speculative.

DNA binding proteins read the CG signal

The bipartite distribution and modification pattern of CG in the genome implies functional significance. For example, CG generally is 10-fold more abundant in CGIs than in genomic DNA as a whole; however, if only the non-methylated form of this sequence is considered, the difference becomes much more dramatic. CGs in the bulk genome are ~80% methylated, often in a stochastic pattern, and therefore the density of non-methylated CG in CGIs is ~50-fold higher than elsewhere. For this reason, CGIs were isolated as a discrete genomic fraction by cleavage with mCG-sensitive restriction endonucleases.[30] These shared properties suggest that CGIs have a common functional significance, despite the observation that their sequences are "unique" in the genome by DNA reassociation assays.[31] The close linkage between CGIs and sites

of transcription initiation hints at a causal interrelationship between these phenomena. Not only are CGIs found at the TSS of ~60% of human genes, but a similar number also occur at other locations within or between genes. These "orphan" CGIs often have TSS-like characteristics, suggesting that they too are promoters.[32–34] Promoter-like features include association with RNA polymerase II, the presence of transcripts detected by a variety of techniques, and colocalisation with the signature chromatin mark of CGIs: trimethylation of lysine 4 of histone H3 (H3K4me3). The latter is a diagnostic marker of active or transcriptionally competent promoters.[35]

Despite these clues, functional links between a high density of non-methylated CG and transcription have only recently begun to emerge. The existence of a zinc-finger protein motif that confers the ability to bind a single non-methylated CG was first detected a decade ago,[36,37] and the solution structure of the interaction has been solved.[38] DNA containing methylated CG or lacking CGs altogether fails to bind this "CXXC" motif. The list of proteins containing the CXXC domain is suggestive, as it includes Cfp1 and Mll1, both of which are causally linked with the formation of H3K4me3. Mll1 contains a functional SET (lysine methyltransferase) domain,[39] and Cfp1 is a component of the Set1/COMPASS complex.[40,41] Mapping of Cfp1 by chromatin immunoprecipitation places it at CGIs, and Cfp1 depletion reduces H3K4me3 at a set of CGI loci.[42] Importantly, creation of an artificial promoterless CG cluster generates a novel focus for Cfp1 binding together with a new H3K4me3 peak, demonstrating that dense CG clusters alone can direct this histone mark. Another member of the CXXC protein family is Kdm2a, a histone demethylase that removes the H3K36me3 mark. This histone modification is prevalent in transcribed regions, but depleted from TSS sites within CGIs. Again, Kdm2a maps to CGIs, and its presence is required to fully demethylate H3K36 at CGIs.[43] CGI chromatin is atypical in other respects too, as isolated CGI chromatin is hyperacetylated and deficient in histone H1.[44] Interestingly, CGIs are relatively nucleosome-deficient *in vivo* and reluctant to assemble nucleosomes *in vitro*.[45] The inference is that CGI promoters are relatively accessible to transcription factors without the need for ATP-dependent chromatin remodelling machines such as Swi/Snf to expose the DNA. Taken together, these findings indicate that CGIs predispose local chromatin to adopt a promoter-friendly chromatin structure. Indeed, this may be their functional role.

Methylated CGIs have been known for some time to cause silencing of the associated promoter (reviewed in Ref. [13]), and there is evidence that *m*ethyl-CpG *b*inding *d*omain (MBD) proteins are involved. The NMR and X-ray structures of the complex between the MBD and a single methylated CG pair provide a molecular explanation for the preference at atomic resolution.[46,47] Specificity for the cytosine methyl group is provided by multiple contacts, several of which are mediated by immobilized water molecules in the major groove of the

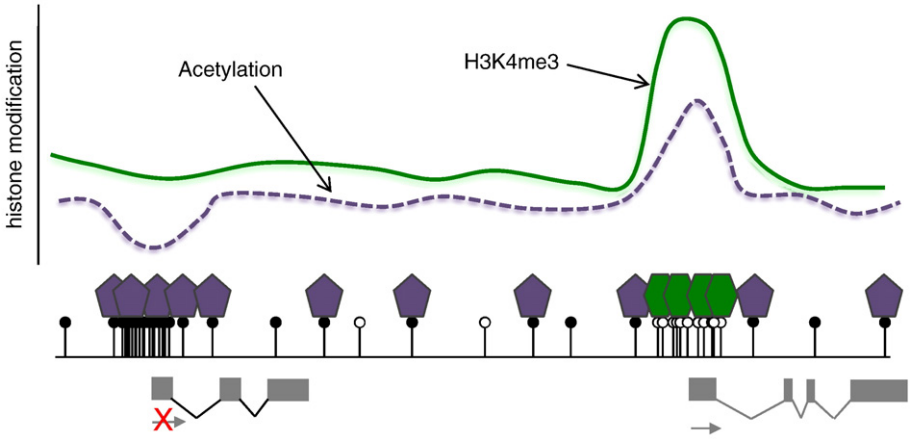

Fig. 1. Proteins read the CG signal to contribute to a local chromatin environment that is either "friendly" or "unfriendly" to promoter activity. The cartoon represents a stretch of genomic DNA with methylated (filled lollipops) and non-methylated CGs (open lollipops) that are clustered in two CGIs coincident with gene promoters. MeCP2 (purple pentagon) binds methylated CGs and recruits proteins that alter the chromatin modification state by deacetylation of histones (promoter-unfriendly). In neurons, there is enough MeCP2 to bind most available methyl-CGs, leading to genome-wide alterations in chromatin structure. Cfp1 (green hexagons) binds exclusively to non-methylated CGs and recruits the Set1/COMPASS complex, which methylates H3K4 to create a transcriptionally permissive chromatin configuration. Histones are present as nucleosomes throughout the domain (not shown). Their modification status is plotted to reflect the influence of these CG binding proteins.

double helix. Arginine side chains also precisely interact with G residues within the methyl-CpG dinucleotide, confirming the specificity of binding to 5-methylcytosine in this dinucleotide sequence context. The MBD can only bind to 5-methylcytosine in a symmetrically methylated CG context. Most MBD proteins are the antithesis of CXXC proteins. Not only do they bind specifically to methylated CG[46,47] but they also recruit marks of "inactive chromatin" structure.[48–50] Mbd2, for example, is part of the Mi2/NuRD histone deacetylase repressor complex.[51,52] Of particular current interest is MeCP2, which associates with histone deacetylase complexes and also has the properties of a transcriptional repressor.[53] MeCP2 is highly expressed in neurons, where it is only ~2-fold less abundant than the histone octomer.[54] Accordingly, MeCP2 binds genome-wide in neurons, tracking the density of mCG. Changes to neuronal chromatin in the absence of MeCP2 are informative. First, histone acetylation is more than 2.5-fold elevated, in keeping with the idea that MeCP2 normally depresses this mark of genome activity. Second, histone H1 abundance is doubled. The latter change may explain the long-standing observation that H1 in neurons is normally half as abundant as in other cell types.[55] It seems likely that MeCP2 competes with H1, reducing its abundance genome-wide, and there is evidence for this idea.[53,56] When MeCP2 is absent, however, histone H1 levels rise to the typical level of one molecule per nucleosome. These changes due to MeCP2 deficiency are not seen in cells where MeCP2 is much less abundant than in neurons, presumably because the changes are of lower magnitude or more localised and thus cannot be detected at a gross level.[54] Absence of MeCP2 also elevates "transcriptional noise."[54] Expression of transposable elements, for example, is relatively derepressed, leading to measurable increases in transposition.[57]

Concluding remarks

CG adds a dimension to genome regulation, as its properties are uniquely adapted for a role as a modulator of transcription and other genome-based activities. Non-methylated and methylated CG mutually exclusively attract protein families. These, in turn, recruit protein complexes bearing enzymatic activities that encourage the formation of transcriptionally permissive or nonpermissive chromatin structures, respectively (Fig. 1). The small size of the CG moiety allows it to be interspersed almost invisibly with other kinds of functional DNA. For example, coding and regulatory sequences can both accommodate multiple CGs without adverse effects. Non-methylated CGs are highly concentrated in ~25,000 promoter CGIs, whereas methylated CGIs are dispersed globally at 10-fold lower density. The distinctive pattern of CG sequences in the genome is of central importance in determining its biological significance. In spite of the coherence of this emerging picture, many issues remain to be resolved. We need to know all the molecular components relating to the reading and writing of CG signals and to gain a better grasp of the physiological readouts of the mCG and CG signals. There are grounds for optimism that these inquiries will have come to fruition before it is time to celebrate the operon's centenary.

Acknowledgements

I thank Matthew Lyst and Elisabeth Wachter for comments on the manuscript. Research in this laboratory is supported by the Wellcome Trust, the Medical Research Council (UK), and the Rett Syndrome Research Trust.

References

1. Jacob, F. & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**, 318–356.
2. Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, **98**, 1–4.
3. Cosma, M. P., Tanaka, T. & Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell*, **97**, 299–311.
4. Li, H., Rhodius, V., Gross, C. & Siggia, E. D. (2002). Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl Acad. Sci. USA*, **99**, 11772–11777.
5. Gilbert, W. & Maxam, A. (1973). The nucleotide sequence of the lac operator. *Proc. Natl Acad. Sci. USA*, **70**, 3581–3584.
6. Kadonaga, J. T. (2004). Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, **116**, 247–257.
7. Mahajan, M. C., Karmakar, S. & Weissman, S. M. (2007). Control of β-globin genes. *J. Cell. Biochem.* **102**, 801–810.
8. Henderson, I. R. & Jacobsen, S. E. (2007). Epigenetic inheritance in plants. *Nature*, **447**, 418–424.
9. Kriaucionis, S. & Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, **324**, 929–930.
10. Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y. *et al.* (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930–935.
11. Holliday, R. & Pugh, J. E. (1975). DNA modification mechanisms and gene activity during development. *Science*, **186**, 226–232.

12. Riggs, A. D. & Pfeifer, G. P. (1992). X-chromosome inactivation and cell memory. *Trends Genet.* **8**, *169–174.*
13. Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21.
14. Watt, F. & Molloy, P. L. (1988). Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus late promoter. *Genes Dev.* **2**, 1136–1143.
15. Barnes, D. E. & Lindahl, T. (2004). Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu. Rev. Genet.* **38**, 445–476.
16. Millar, C. B., Guy, J., Sansom, O. J., Selfridge, J., MacDougall, E., Hendrich, B. *et al.* (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science*, **297**, 403–405.
17. Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J. & Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature*, **401**, 301–304.
18. Hardeland, U., Bentele, M., Lettieri, T., Steinacher, R., Jiricny, J. & Schar, P. (2001). Thymine DNA glycosylase. *Prog. Nucleic Acid Res. Mol. Biol.* **68**, 235–253.
19. Neddermann, P. & Jiricny, J. (1993). The purification of a mismatch-specific thymine-DNA glycosylase from HeLa cells. *J. Biol. Chem.* **268**, 21218–21224.
20. Poole, A., Penny, D. & Sjoberg, B. M. (2001). Confounded cytosine! Tinkering and the evolution of DNA. *Nat. Rev. Mol. Cell Biol.* **2**, 147–151.
21. Cooper, D. & Krawczak, M. (1990). The mutational specturm of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum. Genet.* **85**, 55–74.
22. Reik, W., Dean, W. & Walter, J. (2001). Epigenetic reprogramming in mammalian development. *Science*, **293**, 1089–1093.
23. Sved, J. & Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl Acad. Sci. USA*, **87**, 4692–4696.
24. Nachman, M. W. & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, **156**, 297–304.
25. Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
26. Delgado, S., Gomez, M., Bird, A. & Antequera, F. (1998). Initiation of DNA replication at CpG islands in mammalian chromosomes. *EMBO J.* **17**, 2426–2435.
27. Sequeira-Mendes, J., Diaz-Uriarte, R., Apedaile, A., Huntley, D., Brockdorff, N. & Gomez, M. (2009). Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.* **5**, e1000446.
28. Antequera, F. & Bird, A. (1999). CpG islands as genomic footprints of promoter-replication origins. *Curr. Biol.* **9**, R661–R667.
29. Anglana, M., Apiou, F., Bensimon, A. & Debatisse, M. (2003). Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. *Cell*, **114**, 385–394.
30. Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of non-methylated, CpG-rich DNA. *Cell*, **40**, 91–99.
31. Cross, S. H., Charlton, J. A., Nan, X. & Bird, A. P. (1994). Purification of CpG islands using a methylated DNA binding column. *Nat. Genet.* **6**, 236–244.
32. Illingworth, R., Gruenwald-Schneider, U., Webb, S., Kerr, A., James, K. D., Turner, D. J., Smith, C., Harrison, D. J., Andrews, R. & Bird, A. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* **6**, e1001134.
33. Illingworth, R., Kerr, A., Desousa, D., Jorgensen, H., Ellis, P., Stalker, J. *et al.* (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* **6**, e22.
34. Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D. *et al.* (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, **466**, 253–257.
35. Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, **128**, 693–705.
36. Voo, K. S., Carlone, D. L., Jacobsen, B. M., Flodin, A. & Skalnik, D. G. (2000). Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Mol. Cell. Biol.* **20**, 2108–2121.
37. Lee, J. H., Voo, K. S. & Skalnik, D. G. (2001). Identification and characterization of the DNA binding domain of CpG-binding protein. *J. Biol. Chem.* **276**, 44669–44676.
38. Cierpicki, T., Risner, L. E., Grembecka, J., Lukasik, S. M., Popovic, R., Omonkowska, M. *et al.* (2010). Structure of the MLL CXXC domain–DNA complex and its functional role in MLL-AF9 leukemia. *Nat. Struct. Mol. Biol.* **17**, 62–68.
39. Cosgrove, M. S. & Patel, A. (2010). Mixed lineage leukemia: a structure–function perspective of the MLL1 protein. *FEBS J.* **277**, 1832–1842.
40. Lee J. S., Smith E. & Shilatifard A. The language of histone crosstalk. Cell 142, 682–685.
41. Lee, J. H. & Skalnik, D. G. (2005). CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J. Biol. Chem.* **280**, 41725–41731.
42. Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S. *et al.* (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, **464**, 1082–1086.
43. Blackledge, N. P., Zhou, J. C., Tolstorukov, M. Y., Farcas, A. M., Park, P. J. & Klose, R. J. (2010). CpG islands recruit a histone H3 lysine 36 demethylase. *Mol. Cell*, **38**, 179–190.
44. Tazi, J. & Bird, A. (1990). Alternative chromatin structure at CpG islands. *Cell*, **60**, 909–920.
45. Ramirez-Carrozzi, V. R., Braas, D., Bhatt, D. M., Cheng, C. S., Hong, C., Doty, K. R. *et al.* (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell*, **138**, 114–128.
46. Ohki, I., Shimotake, N., Fujita, N., Jee, J., Ikegami, T., Nakao, M. & Shirakawa, M. (2001). Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell*, **105**, 487–497.
47. Ho, K. L., McNae, I. W., Schmiedeberg, L., Klose, R. J., Bird, A. P. & Walkinshaw, M. D. (2008). MeCP2

binding to DNA depends upon hydration at methyl-CpG. *Mol. Cell*, **29**, 525–531.
48. Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L. & Bird, A. P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell*, **58**, 499–507.
49. Hendrich, B. & Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* **18**, 6538–6547.
50. Hendrich, B. & Tweedie, S. (2003). The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.* **19**, 269–277.
51. Zhang, Y., Ng, H. H., Erdjument-Bromage, H., Tempst, P., Bird, A. & Reinberg, D. (1999). Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev.* **13**, 1924–1935.
52. Feng, Q. & Zhang, Y. (2001). The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. *Genes Dev.* **15**, 827–832.
53. Nan, X., Campoy, J. & Bird, A. (1997). MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell*, **88**, 471–481.
54. Skene, P. J., Illingworth, R. S., Webb, S., Kerr, A. R., James, K. D., Turner, D. J. *et al.* (2010). Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol. Cell*, **37**, 457–468.
55. Thomas, J. O. & Thompson, R. J. (1977). Variation in chromatin structure in two cell types from the same tissue: a short DNA repeat length in cerebral cortex neurons. *Cell*, **10**, 633–640.
56. Ghosh, R. P., Horowitz-Scherer, R. A., Nikitina, T., Shlyakhtenko, L. S. & Woodcock, C. L. (2010). MeCP2 binds cooperatively to its substrate and competes with histone H1 for chromatin binding sites. *Mol. Cell. Biol.* **30**, 4656–4670.
57. Muotri, A. R., Marchetto, M. C., Coufal, N. G., Oefner, R., Yeo, G., Nakashima, K. & Gage, F. H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. *Nature*, **468**, 443–446.