CpG islands and the regulation of transcription

Citation for published version:

Digital Object Identifier (DOI):
10.1101/gad.2037511

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Genes & Development

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
CpG islands and the regulation of transcription

Aimée M. Deaton and Adrian Bird

The Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh EH9 3JR, United Kingdom

Vertebrate CpG islands (CGIs) are short interspersed DNA sequences that deviate significantly from the average genomic pattern by being GC-rich, CpG-rich, and predominantly nonmethylated. Most, perhaps all, CGIs are sites of transcription initiation, including thousands that are remote from currently annotated promoters. Shared DNA sequence features adapt CGIs for promoter function by destabilizing nucleosomes and attracting proteins that create a transcriptionally permissive chromatin state. Silencing of CGI promoters is achieved through dense CpG methylation or polycomb recruitment, again using their distinctive DNA sequence composition. CGIs are therefore generically equipped to influence local chromatin structure and simplify regulation of gene activity.

Vertebrate genomes are methylated predominantly at the dinucleotide CpG, and consequently are CpG-deficient owing to the mutagenic properties of methylcytosine (Coulondre et al. 1978; Bird 1980). The globally methylated, CpG-poor genomic landscape is punctuated, however, by CpG islands (CGIs), which are, on average, 1000 base pairs (bp) long and show an elevated G+C base composition, little CpG depletion, and frequent absence of DNA methylation. These shared properties have allowed CGIs to be isolated as a relatively homogeneous fraction of the genome, despite the heterogeneity of their individual nucleotide sequences (Bird et al. 1985; Cross et al. 1994; Illingworth et al. 2010). Approximately 70% of annotated gene promoters are associated with a CGI, making this the most common promoter type in the vertebrate genome (Saxonov et al. 2006). Included are virtually all housekeeping genes, as well as a proportion of tissue-specific genes and developmental regulator genes (Larsen et al. 1992; Zhu et al. 2008). Recent work has uncovered a large class of CGIs that are remote from annotated transcription start sites (TSSs), but nevertheless show evidence for promoter function (Illingworth et al. 2010; Maunakea et al. 2010). These findings emphasize the strong correlation between CGIs and transcription initiation.

In spite of their link with transcription, the functional significance of CGIs is only just beginning to emerge. CGI promoters turn out to have distinctive patterns of transcription initiation and chromatin configuration. Their regulation involves proteins (some of which specifically bind nonmethylated CpG) that influence the modification status of CGI chromatin. In addition, the CpG moieties themselves are sometimes subject to cytosine methylation, which correlates with stable shutdown of the associated promoter. Here we examine the properties shared by vertebrate CGIs and how transcription is regulated at these sites. Recent related reviews include Illingworth and Bird (2009), Mohn and Schubeler (2009), and Blackledge and Klose (2011).

Evolutionary conservation of CGIs

CGIs are distinct in vertebrates due to their lack of DNA methylation and absence of CpG deficiency, which sets them apart from bulk genomic DNA. Organisms such as the invertebrates *Drosophila melanogaster* and *Caenorhabditis elegans* and the fungus *Saccharomyces cerevisiae* have little or no DNA methylation and, as a result, CGI 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. On the other hand, many plant genomes are very highly methylated, and early research detected a nonmethylated CGI-like genomic fraction (Antequera and Bird 1988). Lack of DNA methylation in plants is evidently linked to genes, as isolation of nonmethylated DNA by “methylation filtration” greatly enriches for transcribed sequences (Palmer et al. 2003), despite the presence of gene body methylation in these organisms (Zhang et al. 2006; Zilberman et al. 2007). Extensive DNA methylome analysis has documented nonmethylated regions at both extremities of plant transcription units [Feng et al. 2010; Zemach et al. 2010]. Whether these can be considered equivalent to or distinct from vertebrate CGIs is not presently known.

Unlike vertebrates, most invertebrate animals exhibit mosaically methylated genomes comprising alternating methylated and nonmethylated domains [Bird et al. 1979; Tweedie et al. 1997]. The persistence over evolutionary time of discrete genomic regions with different DNA methylation states has partitioned DNA sequences into
two fractions: [1] methylated and, consequently, CpG-deficient; and [2] nonmethylated, with the expected frequency of CpG. The origin of vertebrates appears to have coincided with a transition from mosaic to global DNA methylation, accompanied by concomitant CpG depletion throughout most of the genome (Tweedie et al. 1997). Well-studied among the mosaic genomes is the invertebrate Ciona intestinalis (sea squirt), which is evolutionarily close to the invertebrate–vertebrate boundary. C. intestinalis genes within methylated domains are sometimes associated with short nonmethylated CGI-like regions that colocalize with TSSs (Suzuki et al. 2007). CGIs may therefore predate the evolution of vertebrates.

Until recently, it was not clear that CGIs were conserved in either number or genomic location between different vertebrates. Initially, far fewer CGIs were bioinformatically predicted in the mouse genome than in the human genome (Waterston et al. 2002), and this apparent lack of conservation called into question their regulatory importance. CGI prediction algorithms by necessity employ thresholds for detection, alteration of which dramatically changes the number predicted (Illingworth et al. 2008; Illingworth and Bird 2009). The algorithms are also unable to take into account the methylation status of CGIs. A biochemical approach has shed fresh light on the issue, using affinity purification with the CXXC protein domain to isolate clusters of unmethylated CpGs from genomic DNA (Illingworth et al. 2008). High-throughput DNA sequencing of this fraction identified a comprehensive CGI complement from both humans and mice (Illingworth et al. 2010) and revealed similar numbers of CGIs per haploid genome: 25,495 and 23,021, respectively. The reason for the initial discrepancy is that mouse CGIs show, on average, a slightly lower CpG content compared with human CGIs. Biochemical purification of clusters of nonmethylated CpG overcomes this weakness. Also conserved between humans and mice was the proportion of CGIs associating with annotated TSSs (~50%). Moreover, the remaining half of CGIs were distributed equally between locations within gene bodies (intragenic) or between genes (intergenic) in both species (Fig. 1). The genomic position of many of these additional CGIs appears to have been maintained since the divergence of humans and mice ~75 million years ago, implying functional importance.

CGIs are sites of transcriptional initiation

About half of all CGIs self-evidently contain TSSs, as they coincide with promoters of annotated genes. The other half are either within or between characterized transcription units and have been termed “orphan” CGIs to reflect uncertainty over their significance (Illingworth et al. 2010). Do orphan CGIs weaken the correlation between CGIs and transcriptional activity, or do they also mark hitherto unsuspected promoters? The available evidence shows that many orphan CGIs are also sites of transcriptional initiation. Specific examples of intragenic CGI promoters have been known for many years. For example, CGIs at the 3’ end of the Pomp gene and exon 2 of the MHC class II I-Aβ gene both initiate transcripts of unknown function whose coding potential is minimal (Gardiner-Garden and Frommer 1994; Macleod et al. 1998). More is known functionally about the role played by the Air transcript in imprinting of the Igf2r gene. Air is a noncoding RNA (ncRNA) that initiates at a CGI within intron 2 of Igf2r and is essential for silencing of the paternal allele (Sleutels et al. 2002). Similarly, analysis of a CGI in intron 10 of the imprinted Kcnq1 gene identified it as the origin of a noncoding transcript [Kcnq1ot1] that is required for imprinting of several genes within this domain (Mancini-DiNardo et al. 2003, 2006). In these and other cases, the presence of a CGI in an unexpected location stimulated a successful search for an associated transcript.

Genome-wide analyses have confirmed that many orphan CGIs represent novel promoters (Illingworth et al. 2010),
Criteria for transcript annotation include colocalization of a CGI with bound RNA polymerase II (RNAPII) as detected by chromatin immunoprecipitation (ChIP) or, more tellingly, RNA sequence data showing that transcripts originate within orphan CGIs (Fig. 2; Illingworth et al. 2010, Maunakea et al. 2010). Examples of the latter data include cap analysis of gene expression (CAGE), which uses the 5’ cap to isolate and sequence full-length transcripts (Shiraki et al. 2003), and global run-on sequencing (GRO-seq) (Core et al. 2008), which detects in vitro elongation of engaged RNAPII. Altogether, evidence for transcriptional initiation has been found at ~40% of orphan CGIs (~5000) (Illingworth et al. 2010). Less directly, CGIs are frequently marked by trimethylation of histone H3 (H3K4me3), which is a signature of active promoters (see Fig. 2 and below). A subset of intergenic H3K4me3 peaks (~1600), many of which are likely to correspond to orphan CGIs, were found to represent TSSs for long ncRNAs (Guttman et al. 2009).

Many orphan CGI promoters are active in a tissue-specific manner (Illingworth et al. 2010), suggesting that they are tightly regulated. Because only a few tissues have been investigated so far, it is likely that most, if not all, orphan CGIs will be associated with a novel transcript. What is the functional significance of this transcription? Some orphan CGIs probably represent alternative promoters of nearby annotated genes [Maunakea et al. 2010]. Others may initiate ncRNAs that regulate gene expression. For example, the study of imprinted genes has already uncovered key examples of the regulatory relevance of CGI-derived ncRNAs (see above). In addition to Air and Kcnq1ot1, the Xist and Tsix ncRNAs function in X-chromosome inactivation (Herzing et al. 1997; Lee et al. 1999), and the ncRNA HOTAIR has been reported to regulate Hox gene expression (Rinn et al. 2007). Given the hitherto unsuspected abundance of CGI promoters in the genome, it seems likely that involvement of ncRNAs in gene regulation may be widespread. Recent studies have shown that enhancers are often associated with transcription of ncRNAs, although these have not been linked so far to CGIs [Kim et al. 2010]. In a different vein, CGI promoters within MHC class II genes coincide precisely with the hypervariable exon 2, whose polymorphism is caused by gene conversion between members of the class II gene family. It has been suggested that the “open” CGI chromatin structure in germ cells enhances gene conversion in this exon, thereby benefiting the immune system [Macleod et al. 1998]. Potential roles as boundaries or insulators of transcriptional units remain as yet unexplored. In summary, the biological roles played by CGI transcripts may be diverse, but the stage is now set for rigorous testing of hypotheses concerning their function.

It is clear from the above discussion that CGIs act as promoters in mammalian genomes. This leaves two possibilities regarding their significance: [1] CGIs are evolutionary footprints of molecular events that occur at many eukaryotic promoters, but are only visible in organisms that have extensive genomic DNA methylation; or [2] CGIs are important regulatory structures that have evolved under selection in genomes where DNA methylation plays a regulatory role. The weight of evidence now appears to support the second of these possibilities. In what follows, we summarize this emerging data and discuss ways in which CGIs are adapted for promoter function.

Characteristics of CGI promoters

**DNA sequence motifs**

CGIs colocalize with the majority of promoters in both the human and mouse genomes. Early studies suggested that CGI promoters may often lack TATA boxes and display heterogeneous TSSs [Reynolds et al. 1984]. CAGE analysis has shown, on a genome-wide scale, the broad correlation between this distributive pattern of transcription initiation, typically over a region of 50–100 bp, and the presence of CGIs (Carninci et al. 2006). These observations are compatible with the idea that CGI promoters adopt a transcriptionally permissive state within which initiation can occur at a number of locations. In general,
TATA boxes, along with other core promoter elements (such as the BRE, DPE, and DCE), tend to be associated with focused transcriptional initiation, whereas CGIs tend to lack these elements and display dispersed initiation patterns [for review, see Juven-Gershon et al. 2008]. There are, however, exceptions to this generalization. The human genes for \(\alpha\)-globin, MyoD1, and erythropoietin, for example, have CGI promoters, yet possess TATA boxes.

The idea that many CGI promoters are transcriptionally permissive is supported by genome-wide ChIP and transcriptome analysis. RNAPII is bound at the CGI promoters of many inactive genes in embryonic stem (ES) cells (Guenther et al. 2007) and at silent lipopolysaccharide (LPS)-inducible genes in unstimulated macrophages (Hargreaves et al. 2009). Also, global nuclear run-on analysis examining the products of transcriptionally engaged RNAPII molecules detected bidirectional transcription of short nonproductive RNAs as well as full-length transcripts at many CGI promoters (Core et al. 2008; Seila et al. 2008). Much emphasis had been placed previously on the recruitment of RNAPII as the rate-limiting step in transcription, but these results suggest that regulation of many CGI promoters takes place downstream from polymerase binding. One way of achieving this is by regulating transcriptional elongation and mRNA processing, as in the case of CGI-associated inducible genes in macrophages (Hargreaves et al. 2009). In the latter study, RNAPII recruitment was dependent on the transcription factor Sp1, but the initiating form of RNAPII (phosphorylated at Ser 5 of the C-terminal domain [CTD]) was promoter-bound even when the genes were inactive and low levels of full-length transcripts were produced (Hargreaves et al. 2009). A switch from nonproductive to productive transcription was triggered by inducible transcription factor-dependent recruitment of P-TEFb and subsequent Ser 2 phosphorylation of the RNAPII CTD. This resulted in the production of mature, processed transcripts. Similarly, gene regulation at the level of transcriptional elongation via RNAPII pausing (controlled by the pause factors DSIF and NELF) and release of RNAPII mediated by P-TEFb (for review, see Peterlin and Price 2006) is reported to be widespread in ES cells (Rahl et al. 2010). CGI promoters, it seems, attract RNAPII and, unless actively restrained, will engage in transcription of some kind.

**Transcription factor binding at CGIs**

CGIs share little long-range sequence conservation, apart from an elevated CpG density and G+C content, and often lack core promoter elements such as the TATA box (Carninci et al. 2006; Juven-Gershon et al. 2008). What features adapt them to promoter function? A simplistic possibility is that GC richness increases the probability that ubiquitous transcription factors will bind. In general, mammalian transcription factor-binding sites are more GC-rich than the bulk genome (see Fig. 3) and many contain CpG in their recognition sequence. These include the general transcription factor Sp1, which has been shown to recruit TATA-binding protein (TBP) to promoters lacking a TATA box (Butler and Kadonaga 2002).

Transcription factor binding at CGIs

CGIs share little long-range sequence conservation, apart from an elevated CpG density and G+C content, and often lack core promoter elements such as the TATA box (Carninci et al. 2006; Juven-Gershon et al. 2008). What features adapt them to promoter function? A simplistic possibility is that GC richness increases the probability that ubiquitous transcription factors will bind. In general, mammalian transcription factor-binding sites are more GC-rich than the bulk genome (see Fig. 3) and many contain CpG in their recognition sequence. These include the general transcription factor Sp1, which has been shown to recruit TATA-binding protein (TBP) to promoters lacking a TATA box (Butler and Kadonaga 2002).

**Figure 3.** Transcription factor-binding sites are, on average, GC-rich. The G+C content of binding sites for 46 mouse transcription factors was calculated using fasta sequences of actual binding sites obtained from the JASPAR database (Bryne et al. 2008). G+C frequency is given on the X-axis, while the Y-axis shows the number of each of these 46 binding sites with a given G+C frequency.

Transient reporter gene assays examining the activity of 4575 human promoters found that ubiquitously active CGI promoters tended to be enriched for Sp1, Nrf-1, E2F, and ETS transcription factor-binding motifs, each of which contains a CpG (Landolin et al. 2010). Consistent with this, potential ETS and E2F family binding sites, as characterized in vitro, are overrepresented in mouse CGIs (Jaeger et al. 2010). Similarly, CpG-containing ETS, NRF-1, BoxA, SP1, CRE, and E-box motifs are enriched in the CGI promoters of housekeeping genes (Rozenberg et al. 2008), and Sp1 binding on human chromosomes 21 and 22 is focused predominantly at CGIs (Cawley et al. 2004). In the case of the \(\alpha\)-globin gene, DNA footprinting studies detected little difference in transcription factor-binding patterns whether the gene was active or inactive (Cuadrado et al. 2001), suggesting that, even in the case of a highly tissue-specific gene, transcription factor binding can be constitutive.

**The chromatin signature of CGIs**

**Unstable nucleosomes**

There is evidence that nonmethylated CGIs are organized in a characteristic chromatin structure that predisposes them toward promoter activity. A study of chromatin at LPS-inducible genes in macrophages found that CGIs are relatively nucleosome-deficient (Ramirez-Carrozzi et al. 2009). Inducible “primary response genes” fall into two classes: those that require SWI/SNF chromatin remodeling complexes for their activation and those that do not. It was noticed that these groups corresponded with non-CGI and CGI promoters, respectively, suggesting that DNA in CGI chromatin is intrinsically accessible without the need for ATP-dependent nucleosome displacement.
The ability of CpG density alone to directly influence the unmethylated CpG residues, Cfp1, and H3K4me3 comes (Thomson et al. 2010). Further support for a mechanistic link between the abundance of CpG in CGIs and H3K4me3, mediated by a CXXC domain protein that binds specifically to nonmethylated CpG. Accordingly, Kdm2a is bound in vivo at ~90% of CGIs in mouse ES cells and mediates demethylation of H3K36me2 at these regions (Blackledge et al. 2010). Why H3K36me2 should be depleted at CGIs is uncertain, but this modification has been reported to inhibit transcriptional initiation through histone deacetylase [HDAC] recruitment in yeast [Strahl et al. 2002, Youdell et al. 2008; Li et al. 2009]. H3K36me2 depletion may therefore contribute to a transcriptionally permissive state at CGIs (Fig. 4A). Both Cfp1 and Kdm2a have the characteristics of CGI-specific proteins that use CpG density to influence chromatin modification. It is likely that more factors of this kind are yet to be identified, for example, by a comprehensive characterization of the CGI proteome.

Characteristic histone modifications

Early biochemical studies of isolated CGI chromatin showed high levels of histone H3 and H4 acetylation, which are characteristic of transcriptionally active chromatin. Histone H1, on the other hand, which is regarded as antagonistic to transcription, was depleted in this fraction [Tazi and Bird 1990]. Genome-wide studies have confirmed this association at high resolution [Birney et al. 2007; Wang et al. 2008] and have revealed that H3K4me3 is a signature histone mark of CGI promoters, often persisting even when the associated gene is inactive [Guenther et al. 2007; Mikkelsen et al. 2007]. Recent work has established a biochemical connection between the abundance of CpG in CGIs and H3K4me3, mediated by a CXXC domain protein that binds specifically to nonmethylated CpG [Voo et al. 2000]. Cfp1 [CXXC finger protein 1; also known as CGBP] is an integral component of the Setd1 H3K4 methyltransferase complex [Lee and Skalnik 2005] and localizes to the vast majority of CGIs in the mouse genome, suggesting dependence of this histone modification on the DNA sequence [Thomson et al. 2010]. In keeping with this model, depletion of Cfp1 reduces H3K4me3 at many CGIs. Importantly, insertion of an artificial CGI-like DNA sequence into the genome results in recruitment of Cfp1 and creates a novel peak of H3K4me3 in the absence of RNAPII [Thomson et al. 2010]. Further support for a mechanistic link between unmethylated CpG residues, Cfp1, and H3K4me3 comes from the finding that CpG density in CGIs correlates positively with H3K4me3 levels [Illingworth et al. 2010]. The ability of CpG density alone to directly influence the chromatin modification state [Thomson et al. 2010] is likely to be a key function of CGIs.

The presence of H3K4me3 appears to facilitate transcription in a number of ways. The H3K4me3 tail has been shown to interact with the NuRf chromatin remodeling complex [Li et al. 2006; Wysocka et al. 2006; Ruthenburg et al. 2007] as well as ING4-containing histone acetyltransferase complexes [Saksouk et al. 2009]. Also, H3K4me3 interacts with the transcriptional machinery directly, as the core transcription factor TFIIID has an affinity for the H3K4me3 mark [Vermeulen et al. 2007; van Ingen et al. 2008]. Core transcriptional machinery can recruit H3K4 methyltransferases to chromatin [for review, see Ruthenburg et al. 2007], so it is likely that transcription also contributes to H3K4me3 at CGIs. The relative contributions of CpG-mediated and transcription-mediated H3K4me3 to CGI chromatin modification have yet to be determined.

Another distinctive feature of CGI chromatin is depletion of histone H3K36 dimethylation [H3K36me2] compared with non-CGI promoters and gene bodies [Blackledge et al. 2010]. The H3K36me2 histone demethylase Kdm2a [Tsukada et al. 2006] is a CXXC domain protein that, like Cfp1, binds specifically to nonmethylated CpG. Accordingly, Kdm2a is bound in vivo at ~90% of CGIs in mouse ES cells and mediates demethylation of H3K36me2 at these regions [Blackledge et al. 2010]. Why H3K36me2 should be depleted at CGIs is uncertain, but this modification has been reported to inhibit transcriptional initiation through histone deacetylase [HDAC] recruitment in yeast [Strahl et al. 2002, Youdell et al. 2008; Li et al. 2009]. H3K36me2 depletion may therefore contribute to a transcriptionally permissive state at CGIs (Fig. 4A). Both Cfp1 and Kdm2a have the characteristics of CGI-specific proteins that use CpG density to influence chromatin modification. It is likely that more factors of this kind are yet to be identified, for example, by a comprehensive characterization of the CGI proteome.

CGI promoter silencing by DNA methylation

CGIs are typically in a nonmethylated state in an otherwise heavily methylated genome, even when the corresponding gene is transcriptionally inactive. There are, however, well-known examples of CGIs that become methylated during normal development, leading to stable silencing of the associated promoter [Stein et al. 1982; Mohn et al. 2008; Payer and Lee 2008]. Silencing is thought to be either due to direct inhibition of transcription factor binding by DNA methylation or mediated by methyl-binding domain [MBD] proteins that recruit chromatin-modifying activities to methylated DNA (see Fig. 4B; for reviews, see Klose and Bird 2006; Bogdanovic and Veenstra 2009). It seems that CGI methylation is not the initiating event in gene silencing, but acts to lock in the silent state. For example, during X-chromosome inactivation in female eutherian mammals, X-linked CGIs do not become methylated until after gene silencing and the acquisition of several silencing chromatin modifications, such as H3K27me3 [for reviews, see Payer and Lee 2008; Okamoto and Heard 2009]. CGI methylation is, however,
essential for maintenance of leak-proof X-chromosome inactivation, as inhibition of DNA methylation leads to gene reactivation in a fraction of cells (Sado et al. 2000; Csankovszki et al. 2001). As discussed above, CGI methylation also has well-characterized roles in genomic imprinting where parent-of-origin monoallelic expression is controlled by CGI methylation marks (for review, see Edwards and Ferguson-Smith 2007). In several instances, the CGIs concerned act as promoters for ncRNAs whose expression is silenced by DNA methylation. Expression of these ncRNAs—including Air and Kcnq1ot1—is responsible for the silencing of neighboring genes (Sleutels et al. 2002; Mancini-DiNardo et al. 2003, 2006).

Until recently, methylation of CGIs during imprinting and X-chromosome inactivation were suspected to be special cases, most CGIs remaining nonmethylated regardless of gene expression. However, genome-wide studies focusing on CGIs at annotated TSSs have uncovered numerous instances of CGI methylation in normal somatic cells. CGIs in the germline are almost invariably nonmethylated, but a small proportion acquire methylation in somatic tissues (Schilling and Rehli 2007; Shen et al. 2007; Weber et al. 2007). Similarly, a small number of CGI promoters acquire methylation during differentiation of ES cells into neurons, with most of the changes taking place during the early stages of differentiation (Mohn et al. 2008). The majority of CGIs that gain methylation during differentiation are already silent in ES cells (Mohn et al. 2008), providing further evidence that silencing precedes DNA methylation. The genes affected are often expressed only in the germline, such as the MAGE family of testis-specific antigens (De Smet et al. 1999). Differences in gene-associated CGI methylation between one somatic tissue and another have also been reported, although these are relatively rare compared with differences between germline and somatic CGIs (Meissner et al. 2008).

In contrast to the rarity of methylated CGIs at the promoters of annotated genes, orphan CGIs are methylated much more frequently. About 17% of orphan CGIs have been found so far in a methylated state, compared with ~3% of CGIs at annotated gene promoters (illlengworth et al. 2008, 2010; Rauch et al. 2009; Maunakea et al. 2010). By further separating orphan CGIs into intragenic and intergenic categories, it becomes apparent that intragenic CGIs are especially prone to methylation (~20%–34%) (illlengworth et al. 2010; Maunakea et al. 2010). Accordingly, CGIs located within gene bodies show the greatest number of DNA methylation differences between different somatic cells and tissues (Maunakea et al. 2010).
In addition to its occurrence during normal development, CGI methylation has been well documented in cancer. The CGIs of several tumor suppressor genes acquire cancer-specific methylation, and many genes involved in familial forms of cancer undergo DNA methylation-associated silencing in sporadic cancers (for review, see Jones and Baylin 2002, 2007). These changes are thought to contribute to uncontrolled proliferation and thus tumor development. However, whether DNA methylation is the initiating event in gene silencing or is acquired at already silenced genes remains unclear. Genome-wide studies have demonstrated that many of the CGIs that acquire aberrant methylation in cancer are not associated with tumor suppressor genes [Weber et al. 2005; Keshet et al. 2006; Illingworth et al. 2010; Ruiken et al. 2010]. A key issue regarding cancer-specific CGI methylation is whether it mirrors normal developmental methylation. Profiling DNA methylation in the flanks of CGIs suggested that cancer-specific methylation patterns resemble those occurring in normal tissues [Irizarry et al. 2009]. A recent study, however, suggested that cancer-specific CGI methylation can be distinguished from that in normal tissues. Unlike normal tissues, where orphan CGIs are preferentially methyleated, in colorectal cancers, the proportion of TSSs, intragenic CGIs, and intergenic CGIs acquiring tumor-specific methylation was approximately equal [Illingworth et al. 2010].

CGIs and polycomb-mediated silencing

In addition to DNA methylation, CGI promoters can be silenced by polycomb group proteins (PcG). There are two PcG complexes in mammals: polycomb-repressive complex 1 (PRC1) and PRC2. PRC2 mediates H3K27me3, and this mark is recognized by PRC1, which is thought to inhibit transcriptional elongation by a mechanism involving H2A ubiquitylation [Stock et al. 2007; Zhou et al. 2008] and/or chromatin compaction [Escalante et al. 2010]. Differentiation of ES cells into neurons is accompanied by losses and gains of H3K27me3 at many promoters at various stages of differentiation. In contrast, DNA methylation is gained at a relatively small number of promoters only during the early stages of differentiation, remaining relatively stable thereafter [Mohn et al. 2008]. This indicates that polycomb is a more dynamic repression system than DNA methylation, at least during later developmental stages. In ES cells, CGIs silenced by polycomb possess the “active” mark H3K4me3 as well as H3K27me3 [Azuara et al. 2006; Bernstein et al. 2006; Mikkelsen et al. 2007]. These “bivalent” CGI promoters are poised between two alternative states: either active transcription or stable repression. Upon differentiation, they can lose H3K27me3 and become active or be subject to more stable transcriptional repression. Bivalent CGI promoters account for approximately one-fifth of CGI promoters in ES cells [Ku et al. 2008] but are also found in other cell types, although to a lesser extent [Roh et al. 2006; Mikkelsen et al. 2007].

Acquisition of CGI methylation in cancer has been found to occur preferentially at genes marked by H3K27me3 in ES cells [Ohm et al. 2007; Widschwendter et al. 2007; Illingworth et al. 2010]. By mirroring the pattern of polycomb-dependent gene silencing established in ES cells, DNA methylation may hypothetically facilitate maintenance of a “pseudo-pluripotent state,” thereby favoring unrestrained proliferation of cancer cells [Keshet et al. 2006; Ohm et al. 2007; Schlesinger et al. 2007]. It has also been found that promoters that are positive for H3K27me3 in ES cells are more than four times more likely to acquire DNA methylation, supporting the idea that there is a relationship of some sort between the two systems [Mohn et al. 2008]. One report claimed that the PcG protein Ezh2, which catalyzes H3K27me3, can recruit DNA methyltransferases directly [Vire et al. 2006], but this mechanistic link remains unproven. In many respects, DNA methylation and polycomb behave as alternative silencing mechanisms in ES cells [Fouse et al. 2008; Mohn et al. 2008].

Interestingly, polycomb tends to be targeted to CGI-containing regions of the genome [Mikkelsen et al. 2007; Ku et al. 2008], whereas non-CGI promoters do not often associate with H3K27me3 [Mikkelsen et al. 2007; Mohn et al. 2008]. The observation that some CGIs are specifically targeted by polycomb led to the hypothesis that these DNA sequences are prone to polycomb recruitment [Fig. 4C]. Although H3K27me3 often tends to mark large genomic domains, binding of Ezh2 [the catalytic component of PRC2] in human ES cells strongly correlates with CGIs and, in particular, those lacking motifs for activating transcription factors [such as Ets1 and 2, NFY, YY1, and c-Myc] [Ku et al. 2008]. Intriguingly, a polycomb recruitment element identified in human cells is a nucleosome-deficient sequence that includes a CGI [Woo et al. 2010]. A recent study provided functional evidence of a role for CGIs in polycomb recruitment. The CGI of the bivalent Zfp12 locus was found to be sufficient to recruit H3K27me3, H3K4me3, and PRC2 to a human gene desert region inserted into mouse ES cells [Mendenhall et al. 2010]. Interestingly, GC-rich DNA from a bacterial source also had the ability to recruit PRC2, suggesting that base composition plays an important role.

As the majority of CGIs do not normally recruit polycomb, base composition per se is very unlikely to be the only factor involved in attracting or excluding polycomb. To explain this selectivity, it was proposed that the presence of transcription factor-binding motifs within a CGI might...
be sufficient to protect it from PRC2 recruitment [Ku et al. 2008]. Consistent with this hypothesis, deletion of activating motifs at a CGI promoter that normally lacks H3K27me3 in ES cells resulted in PRC2 recruitment and modification of H3K27 [Mendenhall et al. 2010]. This suggests that transcription factor binding or productive transcription may be sufficient to protect against silencing by polycomb. A corollary of this hypothesis is that polycomb-mediated silencing, like DNA methylation, is secondary to gene silencing by other mechanisms. This fits with early work on Drosophila showing that polycomb group proteins do not affect the establishment of developmental gene expression patterns, but are essential to maintain those patterns through time. Indeed, recent experiments indicate that polycomb is preferentially targeted to stalled promoters of coding and noncoding transcripts [Beisel and Paro 2011; Enderle et al. 2011] or to forms of RNA polymerase that are not competent to couple RNA synthesis with cotranscriptional modification [Brookes and Pombo 2009]. Transcriptionally silent CGIs may not always be sufficient to invite polycomb, however, as PRC2-recruitment is proposed to depend on transcription of ncRNAs from CGIs [Rinn et al. 2007; Zhao et al. 2008]. Short transcripts produced from bivalent CGIs in ES and T cells are reported to form stem–loop structures that may be involved in recruiting polycomb to these CGIs [Kanhere et al. 2010]. In addition, the JmjC domain-containing protein Jarid2 has been identified as a novel component of PRC2 and has been implicated in polycomb targeting in mammals [for review, see Herz and Shilatifard 2010]. We have clues about the mechanisms behind polycomb recruitment to specific sites in the mammalian genome and the role played by CGIs in this process, but much remains to be learned.

What protects CGIs from DNA methylation?

CGIs are generally unmethylated CpG-enriched domains that occur against a backdrop of genome-wide DNA methylation and consequent CpG depletion. How are they protected from the waves of de novo DNA methylation that occur during development? A simple explanation would be that CGIs are intrinsically refractory to the action of DNA methyltransferases, but this is very unlikely. Some CGIs do become methylated during normal development; for example, hundreds of CGIs are heavily methylated on one female X chromosome, but are nonmethylated on the other. An alternative possibility is that methylation at CGIs is actively removed by a DNA demethylase [Wu and Zhang 2010], but uncertainty concerning the identity of the DNA demethylating activities in animals has made this hypothesis difficult to test. The discovery that 5-methylcytosine can be converted to 5-hydroxymethylcytosine (hmC) poses the tantalizing possibility that hmC is an intermediate in the demethylation process [Ito et al. 2010; Wu and Zhang 2010]. The demethylating enzyme Tet1 possesses a domain related to the CpG-binding CXXC domain found in Cpf1 [Tahiliani et al. 2009], and could in theory be targeted to CGIs. Compatible with this hypothesis, recent studies locate Tet1 preferentially at CGIs in mouse ES cells [Ficz et al. 2011; Wu et al. 2011a,b], and depletion of Tet1 results in increased CpG methylation at CGIs [Wu et al. 2011a]. An attractive scenario is that CGIs are subject to sporadic de novo methylation, but are continually swept clean by a mechanism involving oxidation of 5-methylcytosine. Defects in such a system may predispose to de novo CGI methylation, as seen in many cancers. Indeed, mutations in the TET2 gene are frequent in leukemias and compromise the hydroxylation reaction [Ko et al. 2010].

Regardless of the detailed molecular mechanism, there is evidence that the methylation-free state of CGIs is causally related to their function as promoters. Deletion or mutation of Sp1 transcription factor-binding sites in the mouse Aprt promoter results in a failure to maintain the unmethylated state of the Aprt CGI in transgenic experiments [Brandeis et al. 1994; Macleod et al. 1994]. This suggests that the binding of transcription factors, or the act of transcription itself, during early development is required for establishment of the DNA methylation-free state. A prediction of this hypothesis is that all CGI promoters should be active during the waves of de novo methylation that occur at the blastocyst stage and in developing germ cells of the embryo [Morgan et al. 2005]. Indeed, studies of a small number of CGI promoters for highly tissue-specific CGI-associated genes showed that several are expressed in early embryonic cells [Daniels et al. 1997]. Large-scale analyses have strengthened this relationship, showing that 90% of genes with CGI promoters are expressed in the early embryo or testis [Sequeira-Mendes et al. 2009]. The idea that transcription is antagonistic to CGI methylation also fits well with the observation that the presence of RNApII at CGIs, irrespective of gene activity, is associated with resistance to DNA methylation in cancer [Takeshima et al. 2009]. DNA sequence-specific transcription factors also have been implicated in preventing DNA methylation, as cooperative binding of the transcription factors Sp1, Nrf-1, and YY1 in normal monocytes correlates with protection from CGI methylation in leukemia cells [Gebhard et al. 2010]. In this case, CGI promoters bound by these factors also had the highest expression levels. A survey of human CGIs reinforced the correlation with transcription factor binding [Straussman et al. 2009]. Subsets of CGIs that become methylated or remain nonmethylated during development were screened for DNA sequence motifs that might influence this decision. An algorithm based on these motifs, several of which were transcription factor-binding sites, was able to predict which CGIs would be immune from DNA methylation, again suggesting that the switch is dependent on cis-acting factors.

There is evidently a close relationship between transcription in the early embryo and lack of CGI methylation, but mechanisms that relate the two are unknown. It has been proposed that origins of DNA replication [ORIs] are the missing link. Based on evidence that CGIs often colocalize with ORIs [Delgado et al. 1998], it was speculated that intermediates in the process of replication initiation lead to local exclusion of DNA methylation and, over time, an altered base composition [Antequera
and Bird 1999]. A large fraction of CGIs (83%) colocalizes with ORIs in ES cells (Sequeira-Mendes et al. 2009), which may reflect the situation in the early embryo. A causal relationship between ORI function and CGIs has yet to be established, however, and therefore other kinds of DNA-based metabolism might be responsible for excluding DNA methylation from these regions. For example, CGI promoters are typically loaded with polymerases that create short abortive transcripts [Core et al. 2008] even when the associated gene is inactive (Kanhere et al. 2010). This “futile” transcription cycle may somehow protect CGIs from the action of DNA methyltransferases, allowing these “silent” promoters to exclude DNA methylation.

Another possible explanation for the immunity of most CGIs to DNA methylation is that their signature chromatin mark, H3K4me3, interferes with DNA methyltransferase activity. Chromatin binding of Dnmt3L, a partner protein of the de novo DNA methyltransferases Dnmt3a and Dnmt3b, is inhibited by H3K4me3 [Ooi et al. 2007]. Importantly, the ADD domains of both Dnmt3a and Dnmt3b also fail to interact with H3K4me3 and are catalytically less active in vitro on chromatin containing this modification compared with unmodified or H3K9me3-modified chromatin [Zhang et al. 2010]. Multiple potential mechanisms for preventing CGI methylation, including those discussed above, are not mutually exclusive, but may act in concert. It is noteworthy that a promoterless CpG-rich sequence marked by Cpf1 binding and H3K4me3 was only partially immune to DNA methylation, suggesting that this chromatin modification by itself is insufficient [Thomson et al. 2010]. It may be that a combination of factors—including, perhaps, initiation of transcription—is required to exclude DNA methylation from CGIs.

Concluding remarks regarding CGI function

CGIs represent a dispersed but coherent DNA sequence class whose members function as genomic platforms for regulating transcription at their associated promoters. These properties depend on the shared features of their DNA sequence; notably, CpG richness and a higher than average G+C content. Paradoxically, CpG richness by itself attracts protein complexes that promote H3K4me3 [Thomson et al. 2010], but G+C-rich DNA also recruits H3K27me3 [Mendenhall et al. 2010]. An equivocal activity state due to the coexistence of these contradictory tendencies prevails at bivalent CGIs, particularly in ES cells, where transcription decisions at many promoters are pending (Azuara et al. 2006; Bernstein et al. 2006). Of course, CpG is also the substrate for DNA methyltransferases, providing another opportunity for regulation. Why should CGI DNA be poised to set up both active and inactive chromatin structures? A possible reason is to create tension between these opposing states and thus facilitate decisive switching. A nonbiological analogy is the spring in an electrical switch, which ensures that a light can be flicked easily between on and off states that are subsequently stable. A key part of such a “spring” model is that features of the silent or active condition should be self-reinforcing once a decision has been made. Biochemically, this can be achieved by positive feedback on pathways that emphasize the chosen state, or by inhibition of opposing pathways. For example, H3K4me3 can attract the ING4 zinc finger protein, which indirectly augments acetylation of histone H3 tails [Hung et al. 2009], thereby emphasizing the active chromatin configuration. At the same time, this mark inhibits the action of Dnmts, which might otherwise impose transcriptional silence. It is likely that a cascade of related positive and negative feedback loops ensures spring-like stability. Indeed, the more we learn about the marking of chromatin, the more loops of this kind are uncovered, suggesting that consolidation of transcriptional states could be a major role for these epigenetic systems. How is the decision between activity or silence made? The evidence suggests that DNA sequence-specific transcriptional regulators are often ultimately responsible, and that their influence, and that of the proteins they recruit, is dominant over the tendencies imposed by CGI DNA. It is notable that the ability of GC-rich DNA to recruit H3K4me3 and H3K27me3 was detected using promoterless exogenous DNA domains [Mendenhall et al. 2010; Thomson et al. 2010]. Thus, factors that repress or activate transcription appear to exaggerate the chromatin changes to which CGIs are already prone. The construction of abrupt switches based on biochemical processes that are often continuously variable is a challenge in many biological processes. CGIs may provide one solution at the level of transcriptional regulation.

Acknowledgments

We thank Shaun Webb for analyzing the base composition of transcription factor-binding sites. We are also grateful to Thomas Clouaire, Jim Selfridge, and Elisabeth Wachter for helpful comments on the manuscript. The Bird laboratory is funded by The Wellcome Trust, the Medical Research Council, and Cancer Research UK.

References

CpG islands and transcription


Deaton and Bird


