Biochemical Society Transactions (2021) **49** 1109–1119
https://doi.org/10.1042/BST20200695

PORTLAND PRESS

Review Article

Sequence determinants, function, and evolution of CpG islands

Allegra Angeloni[1,2] **and Ozren Bogdanovic**[1,2]

[1]Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, Australia; [2]School of Biotechnology and Biomolecular Sciences, Faculty of Science, UNSW, Sydney, Australia

Correspondence: Ozren Bogdanovic (o.bogdanovic@garvan.org.au)

In vertebrates, cytosine-guanine (CpG) dinucleotides are predominantly methylated, with ~80% of all CpG sites containing 5-methylcytosine (5mC), a repressive mark associated with long-term gene silencing. The exceptions to such a globally hypermethylated state are CpG-rich DNA sequences called CpG islands (CGIs), which are mostly hypomethylated relative to the bulk genome. CGIs overlap promoters from the earliest vertebrates to humans, indicating a concerted evolutionary drive compatible with CGI retention. CGIs are characterised by DNA sequence features that include DNA hypomethylation, elevated CpG and GC content and the presence of transcription factor binding sites. These sequence characteristics are congruous with the recruitment of transcription factors and chromatin modifying enzymes, and transcriptional activation in general. CGIs colocalize with sites of transcriptional initiation in hypermethylated vertebrate genomes, however, a growing body of evidence indicates that CGIs might exert their gene regulatory function in other genomic contexts. In this review, we discuss the diverse regulatory features of CGIs, their functional readout, and the evolutionary implications associated with CGI retention in vertebrates and possibly in invertebrates.

Introduction

CpG islands (CGIs) represent a pervasive DNA sequence class frequently associated with vertebrate gene promoters [1,2], where their sequence features adapt them for transcriptional activity [3]. CGIs can be identified according to DNA sequence and chromatin determinants, which include elevated CpG and GC content, lack of DNA methylation (5-methylcytosine, 5mC), presence of trimethylation at lysine 4 of histone H3 (H3K4me3), and enrichment in transcription factor binding sites (TFBS) (Figure 1). Approximately 50–70% of all annotated vertebrate gene promoters are found associated with a CGI, including the majority of housekeeping genes as well as a subset of tissue-specific genes [2,4]. While CGIs are most commonly studied within the context of vertebrate gene promoters, approximately half of all identified CGIs, classed 'orphan' CGIs (oCGIs), are located in inter- and intragenic regions. A number of emerging studies have proposed that oCGIs, while distinct from promoter-associated CGIs, can also contribute to transcriptional regulation [2,5–9].

CGIs constitute conserved features of gene regulatory elements in highly divergent vertebrate species. Vertebrate genomes are heavily methylated, with ~80% of all CpG dyads containing 5mC [10–12]. 5mC is particularly susceptible to spontaneous deamination to thymidine, thus vertebrate genomes are CpG poor [13–15]. A major defining feature of CGIs is that they are mostly refractory to 5mC targeting, which may partly explain the retention of CpG density at these genomic locations [16]. Conversely, most invertebrate genomes are sparsely methylated and are characterised by CpG density at the expected frequency [17,18]. The possibility of invertebrate genomes containing CGIs has therefore not been greatly considered. However, a number of studies have identified CGI-like features in invertebrates ranging from sponges to cephalochordates [19–21]. Furthermore, a family of

Received: 23 January 2021
Revised: 25 May 2021
Accepted: 26 May 2021

Version of Record published:
22 June 2021

Biochemical Society Transactions (2021) 49 1109–1119
https://doi.org/10.1042/BST20200695

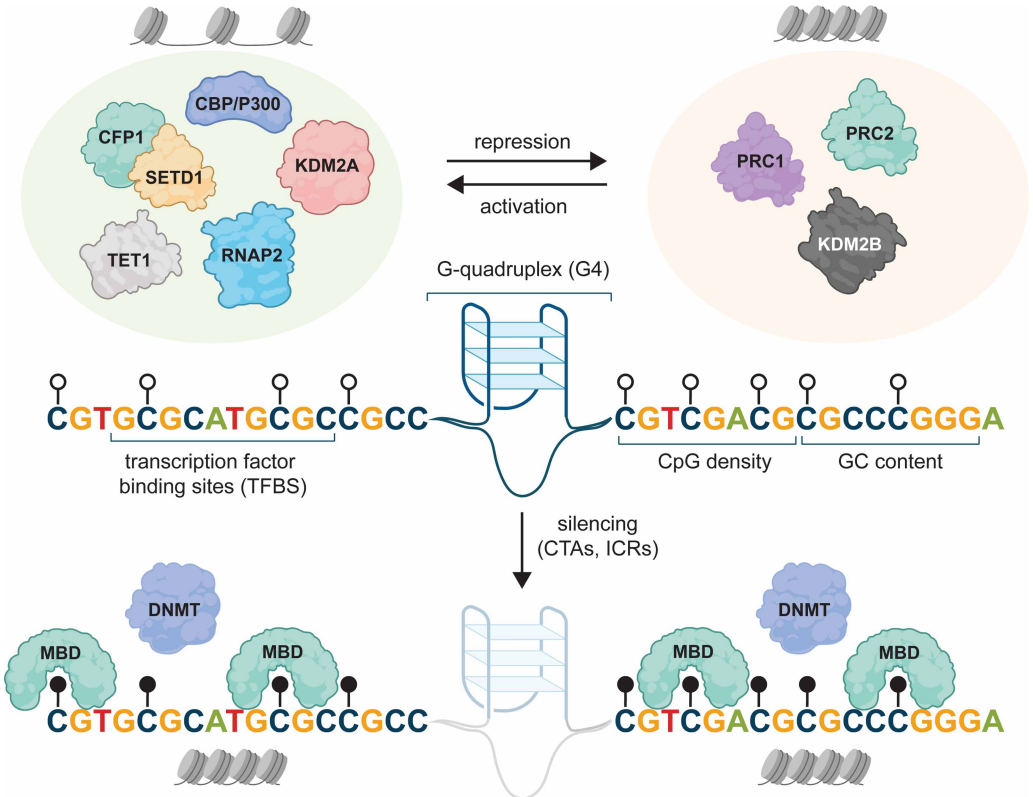

Figure 1. Sequence and chromatin features of CGIs.

CGIs are characterised by elevated CpG density and GC content [10,11,14,74], transcription factor binding sites (TFBS) [42–46], and G quadruplex (G4) DNA sequences [47–49,52,53]. CGIs overlap key gene regulatory elements, such as promoters [2,4,5,27,113,114] and enhancers [6,7,9] and can thus switch between active/poised and repressive chromatin states, depending on the activity of the gene which they are regulating. These states are influenced by the complement of so called 'reader' proteins targeted to CGIs, which include transcriptional activators (CBP/P300, SETD1, CFP1, TET1, KDM2A, RNAP2) and repressors (PRC1, PRC2, KDM2B) [6,9,23,28,30,38,39,81–84,110–112]. In exceptional cases, such as in imprinted control regions (ICRs), or cancer testis antigen gene (CTA) promoters, CGIs can be stably silenced through DNA methylation (5mC) and methyl-CpG binding proteins (MBDs), a state which is reinforced by constant targeting of DNA methyltrasferases (DNMTs) [3,24,25,56,61,62]. It is not yet clear how continuous presence of 5mC within CGIs influences G4 sequences [52].

proteins that specifically recognise non-methylated CpGs and that contain a zinc finger CXXC (ZF-CXXC) domain are deeply conserved in metazoans [21–23]. The preservation of CpG-rich sequences at metazoan gene promoters underscores the important role these features play in gene regulation.

Sequence features and chromatin signatures of CGIs

Early studies performed in mammalian cells identified correlations between diverse sequence features of CGIs and their functional readout. The occurrence of non-methylated CGIs specifically at the 5′ end of genes was suggestive of a potential relationship between CpG-richness and DNA hypomethylation related to gene regulatory function. This hypothesis was verified through transfection assays in cell lines, where it was demonstrated that artificial methylation of CGI promoters was inhibitory to transcription [24,25]. Furthermore, restriction enzyme digests performed using HeLa cell bulk chromatin found that non-nucleosomal regions are associated with regions of high CpG and GC density and low 5mC [26]. These assays indicated that the DNA sequence and the chromatin state of CGIs prime them for transcriptional activity. The discovery that 5mC was a mutation 'hotspot' in the *lacI* gene in *E. coli* led to the hypothesis that elevated CpG concentration in CGIs is maintained in the genome as non-methylated CpGs are refractory to rapid mutability [13]. Following extended evolutionary periods, CpG sites become underrepresented in the genome, such as in humans where CpG

dinucleotides occur at ∼20% the expected frequency [27]. However, the exact evolutionary forces that act on CGIs and the diverse regulatory features within them (CpG density, GC content, TFBS) are far from being completely understood.

One study developed mathematical models that aimed to describe evolutionary regimes in primate species that drive CGI maintenance in distinct genomic contexts. This work revealed multiple major classes of CGI-like sequences [16]. Those include: (i) canonical unmethylated CGIs, characterised by low deamination rates and variable CpG and GC content, (ii) exonic CGIs exhibiting variable 5mC levels and low CpG divergence rates, (iii) biased gene conversion islands, displaying high 5mC levels and rapid deamination rates, and (iv) pseudo-CGIs, characterised by significant CpG loss. Importantly, in each regime described, the CpG density was largely dependent on the interplay between 5mC levels and deamination rates, with little evidence for purifying selection acting on CpG density itself. Nevertheless, CpG density of CGIs is an important regulatory feature that contributes to the formation of histone signatures associated with transcription. For example, H3K4me3 is universally associated with CpG-rich gene promoters and is compatible with gene expression [28–31]. H3K4me3 is deposited by the deeply conserved COMPASS complex [32], which is implicated in transcriptional activation through association with proteins such as the Spt-Ada-Gcn5 histone acetyltransferase (SAGA) [33,34]. The presence of H3K4me3 at transcriptional start sites is conserved in eukaryotes [35]. H3K4me3 and 5mC are mutually exclusive, thus it has been suggested that H3K4me3 excludes 5mC from CGIs through an antagonistic relationship with the ADD domain of the *de novo* DNA methyltransferase 3L (DNMT3L) [36,37]. It has also been demonstrated that non-methylated CGIs are enriched in H3K4me3 and CXXC finger protein 1 (CFP1) [38,39]. CFP1 is known to associate with the H3K4 methyltransferase SETD1 [40] to selectively bind non-methylated CGIs. An exogenous CpG-rich sequence inserted at loci that typically lack H3K4me3 in mouse embryonic stem cells (mESCs) recruited Cfp1 and gained H3K4me3, indicating that increased CpG density facilitates recruitment of chromatin-modifying enzymes that enable a transcriptionally permissive state. Further to this, the inserted sequence did not gain 5mC, suggestive of the contribution of CpG density to DNA hypomethylation at CGIs [38].

However, elegant functional experiments have demonstrated that H3K4me3 recruitment is not dependent on CpG density alone [41]. In mESCs, some CGIs at developmental genes are maintained in a poised configuration, adopting a bivalent chromatin state that includes both H3K4me3 and the repressive, Polycomb-mediated H3K27me3 mark. Insertion of a 1000 bp GC-rich, CpG-poor DNA sequence in a human gene desert in mESCs established that high GC content alone was insufficient to create a bivalent chromatin domain. Similarly, AT- and CG-rich sequences inserted into gene deserts became methylated without gaining H3K4me3 or H3K27me3. Conversely, GC-rich CGIs were refractory to *de novo* 5mC deposition, suggestive of the importance of both GC content and CpG density for the formation of permissive chromatin at CGIs. Many promoter-associated mammalian TFBS such as general transcription factor SP1, nuclear respiratory factor 1 (NRF1), and E2F [42–45] exhibit high CpG density and elevated GC-content. CpG-rich sequences derived from *E. coli* that lack mammalian TFBS become methylated when inserted in mESCs, indicating that CpG-richness alone is likely insufficient to retain CGIs in a hypomethylated state [45]. Mutation of TFBS in the hypomethylated *Gtf2a1l* CGI promoter such as motifs for SP1, CCCTC-binding factor (CTCF) and members of the RFX winged-helix family result in increased 5mC. Similarly, a study that aimed to model the relative contribution of individual determinants to CGI hypomethylation performed parallel insertion and methylation profiling of thousands of DNA fragments in mESCs [46]. Mutation of mammalian TF binding motifs in mouse DNA fragments resulted in alterations to 5mC levels, while insertion of the RE1-Silencing Transcription factor (REST) binding motif in a fully methylated *E. coli* fragment resulted in loss of 5mC. In line with previous results, this study also provided further support for the overall negative correlation between CpG density and 5mC, by assessing the 5mC state of multiple integrated fragments of varying CpG frequency. It is therefore evident that CpG density, GC content and TFBS are each significant determinants in maintaining the chromatin state necessary for the functional readout of CGIs.

Recent work suggests that G-quadruplex (G4) DNA sequences contribute to the maintenance of hypomethylation at CGIs [47,48] (Figure 1). G4 sequences are guanine-rich four-stranded DNA secondary structures containing stacked planar guanine-tetrads. *In silico* and experimental identification of G4 sequences performed predominantly in human cell lines have revealed enrichment of G4 sequences at transcriptional start sites [49–51]. Whole-genome bisulfite sequencing of DNA extracted from human embryonic stem cells (hESCs) revealed that high stability G4 sequences associated with CGIs were hypomethylated compared with those found outside CGIs, particularly when located in open chromatin [52]. G4 ChIP-seq data performed on human K562 chronic

Biochemical Society Transactions (2021) **49** 1109–1119
https://doi.org/10.1042/BST20200695

myelogenous leukemia cells integrated with DNMT1 binding sites found DNMT1 to be localised to and inhibited at G4 structures, suggesting that CGIs evade 5mC targeting through sequestering of DNMT1 at G4 sequences [53]. *In silico* G4 profiling performed in 37 eukaryotic species encompassing fungi, protozoa and a diverse range of metazoan species found G4 sequences to be conserved at some gene promoters [54]. In this study, the relationship between 5mC and G4 sequences was explored through comparison of G4 sequences at promoters in the highly methylated *Sus scrofa domesticus* (pig) genome and the sparsely methylated *Bombyx mori* (silkworm) genome. This analysis revealed in both species that G4 sequences had low 5mC levels relative to the bulk genome, indicating an antagonistic and evolutionarily conserved relationship between 5mC and G4 sequences. However, further research is required to elucidate the potential for cross-talk between G4s, CGIs and DNMTs as well as other chromatin remodelling factors.

Orphan CGIs (oCGIs)

CGIs are most commonly studied in the context of promoters; however, multiple reports have indicated that CGIs can exert gene regulatory functions in a variety of genomic contexts. For example, orphan CGIs (oCGIs) coincide with developmental enhancers in zebrafish, frog and mouse embryos that are linked to key developmental pathways. These enhancers become developmentally activated during the vertebrate phylotypic period, when they undergo active DNA demethylation mediated by Ten-eleven translocation (TET) enzymes, while gaining classic enhancer chromatin marks such as H3K4me1 and H3K27ac [6]. oCGIs have also been described as conserved features of broadly expressed enhancers in placental mammals, containing canonical H3K4me1 and H3K27ac chromatin marks and TFBS [7]. A recent study put forward an exciting possibility that oCGIs might act as enhancer boosters by increasing physical and functional communication between poised enhancers and CpG-rich gene promoters at developmental genes in mouse anterior neural progenitor cells [9]. When poised for activation in mESCs, these enhancer oCGIs are enriched in H3K27me3, H3K4me1 and are bound by Polycomb-group proteins and CBP/p300. Apart from transcriptional enhancers, distal CGIs are also known to be associated with non-coding RNA promoters, and unannotated transcripts [2]. CGIs therefore exhibit a flexible repertoire of regulatory functions in the genome, some of which appear to have been retained through millions of years of divergent evolution.

5mC and transcriptional repression at CGIs

The presence of CpG-rich DNA sequences in vertebrate genomes was first identified through methylation-sensitive restriction enzyme digest assays, which unravelled an inverse correlation between CpG density and 5mC [11,14]. This led to the hypothesis that the emergence and evolution of CGIs might be causally related to 5mC. The advent of massively parallel sequencing alongside the development of sodium bisulfite treatment for 5mC identification enabled base-resolution analyses of CGIs and the precise quantification of their 5mC state [12,55]. Global 5mC assessment in mouse and human ESCs found prevalent hypomethylation at promoter-associated CGIs, independently of gene activity [12,55]. The exception to this widespread hypomethylated state are CGI promoters of cancer testis antigen (CTA) genes, which are targeted by 5mC during embryogenesis in mouse, human and zebrafish [56]. This results in organism-wide CTA silencing (Figure 1) that is relieved only during germline development or oncogenic processes [57]. Nevertheless, such examples are extremely limited, and it yet needs to be determined whether 5mC is a major determinant of CTA silencing.

The relationship between CGIs and 5mC has also been explored through studies of imprinted genes. Monoallelic expression of imprinted genes occurs through parental-specific 5mC states at discrete genetic elements termed imprinting control regions (ICRs) (Figure 1). Among the best studied examples are murine maternally expressed *Igf2r*, *Slc22a2*, and *Slc22a3* genes and the paternally expressed long non-coding (lncRNA) *Airn*. Each parentally-derived allele is distinguishable by the presence of differentially methylated CGIs; the paternal allele contains a methylated CGI promoter in *Igf2r* while the maternal allele contains a methylated CGI in intron 2 of *Igf2r* that is co-localised with the *Airn* promoter [58–60]. Early studies induced demethylation at the *Igf2r* locus with the potent demethylating agent 5-azacytidine (5-aza-C) in cultured human and mouse astrocyte cells [61] and in newborn mice [62]. In these studies, 5-aza-C treatment induced global DNA demethylation and biallelic gene expression of *Igf2r*. However, later studies revealed *Airn* to be the primary *cis*-acting silencer of *Igf2r, Slc22a2* and *Slc22a3* on the paternal allele [63]. Among the three genes silenced by *Airn*, only *Igf2r* gains methylation on the paternal allele [59]. Intriguingly, *Airn* expression is sufficient to silence *Igf2r* in the absence of 5mC, suggesting that promoter 5mC presence is not necessary for gene silencing [64]. The inefficacy of 5mC to act as a dominant repressive mechanism is further supported by *in vivo*

experiments in *Xenopus* embryos, which demonstrated that methylated CpG-rich promoter-reporter gene constructs are robustly expressed at late-blastula and gastrula stages [65]. Two different studies, which employed precise epigenome editing to target the catalytic domain of DNMT3A to CpG-rich genomic locations via a zinc finger effector, came to different conclusions related to the repressive potential of 5mC at CGIs [66,67]. While one study observed efficient 5mC-mediated gene repression [66], the other revealed varying effects including the compatibility of 5mC, H3K4me3 and RNA polymerase II at numerous genomic loci [67]. Notably, these two studies were not carried out in the same cell line. It is therefore evident that the repressive role traditionally attributed to 5mC at CGIs is not as straightforward as suggested by early studies; rather, the relationship between 5mC, CGIs and gene expression might largely depend on the biological context.

DNA methylation and the evolutionary maintenance of CGIs

Although mechanisms that describe how individual sequence and chromatin features of CGIs facilitate the maintenance of hypomethylation have been proposed, it remains elusive how CGIs have remained refractory to 5mC targeting throughout evolution. Furthermore, it is unclear to what degree genome hypermethylation contributed to the formation and maintenance of CGIs. Intriguingly, analysis of human chromosome 21 inserted into a mouse genome found that hypomethylated regions marked by H3K4me3 present on human chromosome 21 were appropriately recapitulated in the transchromosomic mouse model, indicating that DNA sequence is largely sufficient to prevent 5mC accumulation at CGIs irrespective of the host species [68]. A similar result was observed following insertion of bacterial artificial chromosomes (BACs) containing mouse-derived genomic sequences into zebrafish zygotes, where it was seen that promoter-associated mouse hypomethylated regions were again appropriately specified.

Unlike vertebrates, invertebrates contain variable genomic 5mC levels, ranging from 0% (such as in *Drosophila melanogaster* and *Caenorhabditis elegans*) to 80% (such as in sponge *Amphimedon queenslandica*) (Figure 2). In invertebrates that display mosaic 5mC patterns, targeting is mostly limited to gene bodies, where 5mC is thought to prevent spurious transcriptional initiation by RNA polymerase II [69]. In sparsely methylated invertebrate genomes, the possibility of CGI presence has thus not been greatly considered, however the presence of CGI-like sequences has already been described in several species (Figure 2). Perhaps the most

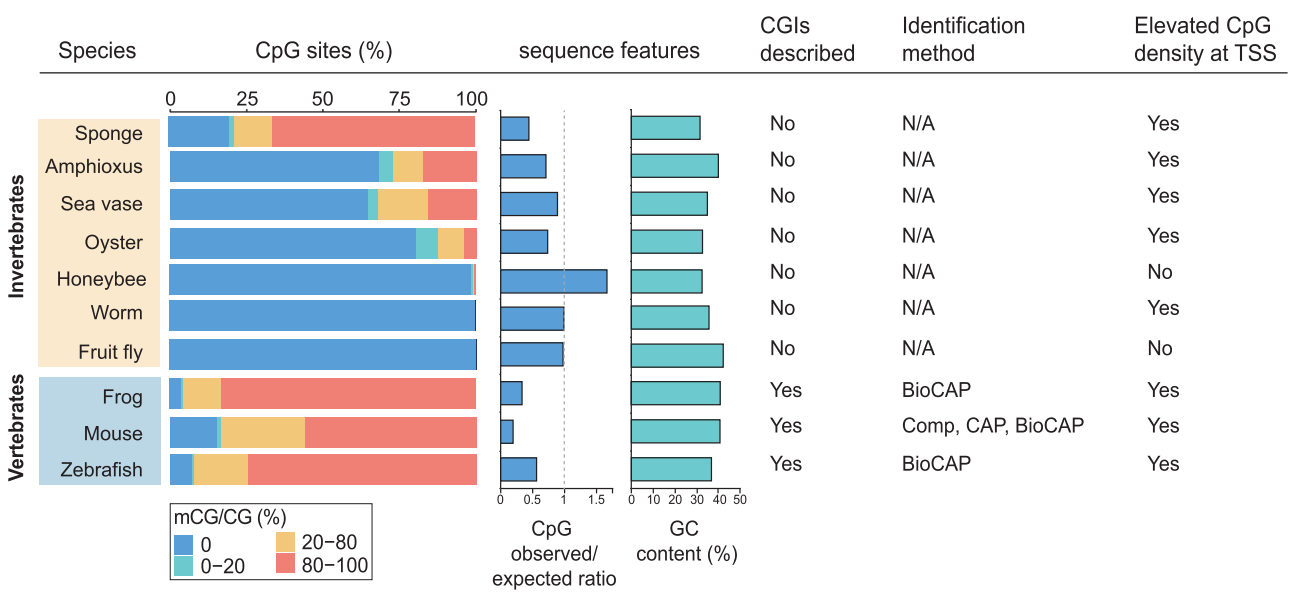

Figure 2. DNA methylation status, sequence features, and CGI presence in ten metazoan genomes.

Global DNA methylation levels obtained from whole-genome bisulfite sequencing datasets of the following species: sponge (*Amphimedon Queenslandica*), lancelet (*Branchiostoma lanceolatum*), sea vase (*Ciona intestinalis*), pacific oyster (*Crassotrea gigas*), honeybee (*Apis melifera*), worm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), frog (*Xenopus tropicalis*), mouse (*Mus musculus*), and zebrafish (*Danio rerio*) [6,20,21,73,115]. Presence of CGIs in the genome has been previously described in the following studies [2,5,116], whereas elevated CpG density at TSS was previously discussed here [2,19–21,70,72].

Biochemical Society Transactions (2021) **49** 1109–1119
https://doi.org/10.1042/BST20200695

striking example comes from the demosponge *Amphimedon queenslandica*, which displays a fully hypermethylated genome as well as unmethylated regions of elevated CpG content that overlap transcription start sites (TSS). Furthermore, such *Amphimedon* promoters contain DNA binding motifs for methyl-sensitive transcription factors such as NRF1, Ying Yang 1 (YY1), early growth response protein (EGR) and GL1 [21]. Sea vase *Ciona intestinalis* exhibits a mosaic DNA methylome, with sharp transitions between roughly comparable amounts of methylated and unmethylated DNA co-localising with transcription units. Bisulfite sequencing analysis revealed the presence of unmethylated CpG-rich domains, with a CpG density similar to that of vertebrate CGI promoters [19]. CpG-dense regions surrounding TSS have also been described in the European amphioxus (*Branchiostoma lanceolatum*) [20], as well as in the pacific oyster (*Crassotrea gigas*) [70] and in the sea slug *Aplysia* [71]. Interestingly, in *Caenorhabditis elegans*, enrichment of a CFP1 orthologue has been reported at nucleosome-depleted CpG-rich gene promoters marked by H3K4me3 [72]. While CGIs are most extensively characterised in hypermethylated vertebrate genomes, it remains elusive whether they are a vertebrate-specific innovation, or rather a deeply conserved feature of metazoan gene regulatory elements. Understanding how CGIs emerged and evolved to have functional significance in gene regulatory elements will require further genomic and epigenomic studies involving diverse metazoan species [73].

Computational and biochemical methods for CGI identification

Historically, the sequence features of CGIs have been extensively used for genome-wide prediction of CGI locations [74–77]. However, these algorithms were largely based on the sequence composition of CGIs in mouse and human (i.e GC content >50%, CpG O/E >0.6, length >200 bp). Consequently, while successful in mammals, such algorithms gave mixed results in non-mammalian vertebrates such as zebrafish [2]. This issue was overcome through the development of biochemical methods to identify CGIs. CXXC affinity purification (CAP) exploits a purified CXXC3 protein domain from mouse Mbd1 that captures unmethylated CGIs specifically [78]. CAP revealed a similar number of CGIs in mouse and human (23 000 and 25 500 respectively) with the same proportion of CGIs found at annotated TSS in both mouse and human (60% and 59%, respectively) [5]. A later study employed profiling of non-methylated CGIs in seven divergent vertebrate species through BioCAP [2], a modified CAP protocol that captures CGIs using human KDM2B ZF-CXXC protein domain immobilised on an avidin-based support [79]. Overall, CAP-based approaches provide an unbiased methodology for the identification of CGIs from purified genomic DNA of vertebrate and potentially invertebrate DNA.

CGI reader proteins

Concordant with the functional conservation of CGIs, protein domains that specifically recognise and interact with CGIs are evolutionarily conserved. Many CGI reader proteins contain a ZF-CXXC protein domain that recognises clusters of unmethylated CpG-rich sequences. This protein family is found in complexes that nucleate specifically at CGIs and may play roles in protecting CGIs from 5mC deposition and inducing context-dependent chromatin states [23]. The ZF-CXXC domain contains two conserved cysteine-rich clusters that coordinate two zinc ions in a tetrahedral structure, intervened by a linker sequence that provides rigidity to the domain structure. Binding is mediated by a DNA-binding loop that forms specific side-chain and backbone interactions with the CpG site on double stranded DNA. The DNA binding loop is in such close proximity to the cytosine that the presence of a methyl group would create a severe steric clash [80]. Examples of proteins enriched at CGIs and containing a ZF-CXXC domain include the histone lysine-specific demethylases KDM2A/B that contribute to the depletion of H3K36me2 at promoters [81–83], the histone lysine methyltransferase CFP1 that deposits H3K4me3 [38,84], and the histone lysine methyltransferases MLL1/2 [85–88].

A major conserved protein family associated with CGIs are the Polycomb repressive complexes 1 and 2 (PRC1/2) that are critical regulators of gene expression during development [89–92]. PRC1 is an E3 ubiquitin ligase that targets the C-terminal tail histone H2A whereas PRC2 is a histone H3 lysine 27 methyltransferase. Although PRC1 and PRC2 play distinct roles in H3K27me3 establishment, they ultimately function to establish and maintain repressive chromatin states (Figure 1). PRC1 and PRC2 function almost exclusively at CGIs. A well-studied target is the deeply conserved Hox gene cluster consisting of a conserved group of related genes responsible for establishing animal body plans [93–97]. Hox genes closely resemble CGIs in vertebrates, being rich in CpG and GC content and lacking 5mC. Intriguingly, the canonical protein structure of PRC1/2 does not contain a sequence-specific DNA binding domain. Studies performed in cancer cell lines and mESCs have

indicated a co-occupancy of a variant PRC1 complex and KDM2B, suggesting that a variant PRC1 complex associates with KDM2B that recruits PRC1 to its genomic targets [98–101].

Ten-eleven translocation (TET) dioxygenase enzymes are a protein family involved in 5mC removal [102] (Figure 1). TET proteins actively mediate iterative demethylation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [103–105]. 5fC and 5caC are recognised and cleaved by thymine-DNA glycosylase (TDG), followed by excision and replacement with unmethylated cytosine through base excision repair pathways [106,107]. In mammals, TET1 and TET3 contain a ZF-CXXC protein domain while the ancestral TET2 ZF-CXXC domain is present in the TET2-interacting protein IDAX/CXX4 [108]. Three TET protein copies (TET1/2/3) are found in mammals and some vertebrates such as zebrafish [109]. TET orthologues containing a conserved ZF-CXXC domain have been described in invertebrates [20,21]. Enrichment of 5hmC and TET1 at CpG-rich gene promoters has been reported in mESCs in numerous studies, indicating a potential functional role of TET1 in maintaining CGIs in a hypomethylated state [110–112]. Altogether, CGIs are associated with highly diverse readers including components of COMPASS and Polycomb complexes as well as the TET dioxygenase enzymes.

Perspectives

- CGIs are essential components of vertebrate gene regulatory elements such as promoters and enhancers. CGIs and their reader protein complexes are deeply conserved in the vertebrate lineage. Unravelling how CGIs evolved is fundamental to understanding the mechanisms by which these key regulatory sequences exert functional readout.

- Although significant efforts have been made to elucidate the evolution of CGIs, the possibility of CGIs being present in metazoans beyond vertebrates (where they are most extensively characterised) remains understudied. Future research on CGI evolution should employ CAP-based profiling of diverse vertebrate and invertebrate genomes with the aim of understanding better which features (i.e CpG density, GC content, TFBS, G4 sequences) are conserved within which lineage.

- Besides canonical promoter CGIs, orphan CGIs (oCGIs), which are found in intergenic regions and associated with enhancer activity, have recently been extensively characterised. oCGI display remarkable functional conservation in vertebrate genomes and appear to be required for regulation of key developmental genes. Understanding the molecular mechanisms that allow for the establishment of developmental stage- and tissue-specific 5mC patterns and enhancer (H3K4me1/H3K27ac) signatures at these regions will be a major focus of future studies.

Competing Interests
The authors declare that there are no competing interests associated with the manuscript.

Funding
No particular funding has been received for this work.

Author Contributions
A.A. and O.B. conceived the study, prepared the figures, and wrote the manuscript.

Abbreviations
CAP, CXXC affinity purification; CFP1, CXXC finger protein 1; CGIs, CpG islands; CTA, cancer testis antigen; ICRs, imprinting control regions; NRF1, nuclear respiratory factor 1; TET, Ten-eleven translocation; TFBS, transcription factor binding sites; TSS, transcription start sites; ZF-CXXC, zinc finger CXXC.

Biochemical Society Transactions (2021) **49** 1109–1119
https://doi.org/10.1042/BST20200695

References

1 Bird, A.P. (1987) Cpg islands as gene markers in the vertebrate nucleus. *Trends Genet.* **3**, 342–347 https://doi.org/10.1016/0168-9525(87)90294-0
2 Long, H.K., Sims, D., Heger, A., Blackledge, N.P., Kutter, C., Wright, M.L. et al. (2013) Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife* **2**, e00348 https://doi.org/10.7554/eLife.00348
3 Blackledge, N.P. and Klose, R. (2011) Cpg island chromatin: a platform for gene regulation. *Epigenetics* **6**, 147–152 https://doi.org/10.4161/epi.6.2.13640
4 Saxonov, S., Berg, P. and Brutlag, D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. U.S.A.* **103**, 1412–1417 https://doi.org/10.1073/pnas.0510310103
5 Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J. et al. (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* **6**, e1001134 https://doi.org/10.1371/journal.pgen.1001134
6 Bogdanović, O., Smits, A.H., de la Calle Mustienes, E., Tena, J.J., Ford, E., Williams, R. et al. (2016) Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat. Genet.* **48**, 417–426 https://doi.org/10.1038/ng.3522
7 Bell, J.S.K. and Vertino, P.M. (2017) Orphan CpG islands define a novel class of highly active enhancers. *Epigenetics* **12**, 449–464 https://doi.org/10.1080/15592294.2017.1297910
8 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 https://doi.org/10.1038/nature09165
9 Pachano, T., Sánchez-Gaya, V., Mariner-Faulí, M., Ealo, T., Asenjo, H.G., Respuela, P. et al. (2020) Orphan CpG islands boost the regulatory activity of poised enhancers and dictate the responsiveness of their target genes. *biorxiv* https://doi.org/10.1101/2020.08.05.237768
10 Cooper, D.N., Taggart, M.H. and Bird, A.P. (1983) Unmethylated domains in vertebrate DNA. *Nucleic Acids Res.* **11**, 647–658 https://doi.org/10.1093/nar/11.3.647
11 Bird, A., Taggart, M., Frommer, M., Miller, O.J. and Macleod, D. (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91–99 https://doi.org/10.1016/0092-8674(85)90312-5
12 Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J. et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 https://doi.org/10.1038/nature08514
13 Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**, 775–780 https://doi.org/10.1038/274775a0
14 Bird, A.P. (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 https://doi.org/10.1093/nar/8.7.1499
15 Shen, J.C., Rideout, III, W.M. and Jones, P.A. (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.* **22**, 972–976 https://doi.org/10.1093/nar/22.6.972
16 Cohen, N.M., Kenigsberg, E. and Tanay, A. (2011) Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* **145**, 773–786 https://doi.org/10.1016/j.cell.2011.04.024
17 Bird, A.P., Taggart, M.H. and Smith, B.A. (1979) Methylated and unmethylated DNA compartments in the sea urchin genome. *Cell* **17**, 889–901 https://doi.org/10.1016/0092-8674(79)90329-5
18 Tweedie, S., Charlton, J., Clark, V. and Bird, A. (1997) Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol. Cell. Biol.* **17**, 1469–1475 https://doi.org/10.1128/MCB.17.3.1469
19 Suzuki, M.M., Kerr, A.R.W., De Sousa, D. and Bird, A. (2007) Cpg methylation is targeted to transcription units in an invertebrate genome. *Genome Res.* **17**, 625–631 https://doi.org/10.1101/gr.6163007
20 Marlétaz, F., Firbas, P.N., Maeso, I., Tena, J.J., Bogdanovic, O., Perry, M. et al. (2018) Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 https://doi.org/10.1038/s41586-018-0734-6
21 de Mendoza, A., Hatleberg, W.L., Pang, K., Leininger, S., Bogdanovic, O., Pflueger, J. et al. (2019) Convergent evolution of a vertebrate-like methylome in a marine sponge. *Nat. Ecol. Evol.* **3**, 1464–1473 https://doi.org/10.1038/s41559-019-0983-2
22 Lee, J.H., Voo, K.S. and Skalnik, D.G. (2001) Identification and characterization of the DNA binding domain of CpG-binding protein. *J. Biol. Chem.* **276**, 44669–44676 https://doi.org/10.1074/jbc.M107179200
23 Long, H.K., Blackledge, N.P. and Klose, R.J. (2013) ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochem. Soc. Trans.* **41**, 727–740 https://doi.org/10.1042/BST20130028
24 Stein, R., Razin, A. and Cedar, H. (1982) In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc. Natl Acad. Sci. U.S.A.* **79**, 3418–3422 https://doi.org/10.1073/pnas.79.11.3418
25 Busslinger, M., Hurst, J. and Flavell, R.A. (1983) DNA methylation and the regulation of globin gene expression. *Cell* **34**, 197–206 https://doi.org/10.1016/0092-8674(83)90150-2
26 Tazi, J. and Bird, A. (1990) Alternative chromatin structure at CpG islands. *Cell* **60**, 909–920 https://doi.org/10.1016/0092-8674(90)90339-G
27 Consortium, I.H.G.S. (2001) International human genome sequencing consortium. initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 https://doi.org/10.1038/35057062
28 Schneider, R., Bannister, A.J., Myers, F.A., Thorne, A.W., Crane-Robinson, C. and Kouzarides, T. (2004) Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat. Cell Biol.* **6**, 73–77 https://doi.org/10.1038/ncb1076
29 Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J. et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181 https://doi.org/10.1016/j.cell.2005.01.001
30 Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D. et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 https://doi.org/10.1038/ng1966
31 Hughes, A.L., Kelley, J.R. and Klose, R.J. (2020) Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation. *Biochim. Biophys. Acta Gene Regul. Mech.* **1863**, 194567 https://doi.org/10.1016/j.bbagrm.2020.194567
32 Shilatifard, A. (2012) The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu. Rev. Biochem.* **81**, 65–95 https://doi.org/10.1146/annurev-biochem-051710-134100
33 Vermeulen, M., Eberl, H.C., Matarese, F., Marks, H., Denissov, S., Butter, F. et al. (2010) Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142**, 967–980 https://doi.org/10.1016/j.cell.2010.08.020

34 Bian, C., Xu, C., Ruan, J., Lee, K.K., Burke, T.L., Tempel, W. et al. (2011) Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation. *EMBO J.* **30**, 2829–2842 https://doi.org/10.1038/emboj.2011.193
35 Soares, L.M., He, P.C., Chun, Y., Suh, H., Kim, T. and Buratowski, S. (2017) Determinants of histone H3K4 methylation patterns. *Mol. Cell* **68**, 773–85. e6 https://doi.org/10.1016/j.molcel.2017.10.013
36 Ooi, S.K.T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z. et al. (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714–717 https://doi.org/10.1038/nature05987
37 Otani, J., Nankumo, T., Arita, K., Inamoto, S., Ariyoshi, M. and Shirakawa, M. (2009) Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. *EMBO Rep.* **10**, 1235–1241 https://doi.org/10.1038/embor.2009.218
38 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 https://doi.org/10.1038/nature08924
39 Brown, D.A., Di Cerbo, V., Feldmann, A., Ahn, J., Ito, S., Blackledge, N.P. et al. (2017) The SET1 complex selects actively transcribed target genes via multivalent interaction with CpG island chromatin. *Cell Rep.* **20**, 2313–2327 https://doi.org/10.1016/j.celrep.2017.08.030
40 Lee, J.-H., Tate, C.M., You, J.-S. and Skalnik, D.G. (2007) Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. *J. Biol. Chem.* **282**, 13419–13428 https://doi.org/10.1074/jbc.M609809200
41 Wachter, E., Quante, T., Merusi, C., Arczewska, A., Stewart, F., Webb, S. et al. (2014) Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *eLife* **3**, e03397 https://doi.org/10.7554/eLife.03397
42 Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A. et al. (1994 Sep 29) Sp1 elements protect a CpG island from de novo methylation. *Nature* **371**, 435–438 https://doi.org/10.1038/371435a0
43 Macleod, D., Charlton, J., Mullins, J. and Bird, A.P. (1994) Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev.* **8**, 2282–2292 https://doi.org/10.1101/gad.8.19.2282
44 Gebhard, C., Benner, C., Ehrich, M., Schwarzfischer, L., Schilling, E., Klug, M. et al. (2010) General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells. *Cancer Res.* **70**, 1398–1407 https://doi.org/10.1158/0008-5472.CAN-09-3406
45 Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F. and Schübeler, D. (2011) Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.* **43**, 1091–1097 https://doi.org/10.1038/ng.946
46 Krebs, A.R., Dessus-Babus, S., Burger, L. and Schübeler, D. (2014) High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *eLife* **3**, e04094 https://doi.org/10.7554/eLife.04094
47 Varizhuk, A., Isaakova, E. and Pozmogova, G. (2019) DNA G-Quadruplexes (G4s) modulate epigenetic (Re)Programming and chromatin remodeling. *Bioessays* **41**, 1900091 https://doi.org/10.1002/bies.201900091
48 Mukherjee, A.K., Sharma, S. and Chowdhury, S. (2019) Non-duplex G-Quadruplex structures emerge as mediators of epigenetic modifications. *Trends Genet.* **35**, 129–144 https://doi.org/10.1016/j.tig.2018.11.001
49 Huppert, J.L. and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **35**, 406–413 https://doi.org/10.1093/nar/gkl1057
50 Zhao, Y., Du, Z. and Li, N. (2007) Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. *FEBS Lett.* **581**, 1951–1956 https://doi.org/10.1016/j.febslet.2007.04.017
51 Hänsel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A. et al. (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* **48**, 1267–1272 https://doi.org/10.1038/ng.3662
52 Jara-Espejo, M. and Line, S.R. (2020) DNA G-quadruplex stability, position and chromatin accessibility are associated with CpG island methylation. *FEBS J.* **287**, 483–495 https://doi.org/10.1111/febs.15065
53 Mao, S.-Q., Ghanbarian, A.T., Spiegel, J., Martínez Cuesta, S., Beraldi, D., Di Antonio, M. et al. (2018) DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol.* **25**, 951–957 https://doi.org/10.1038/s41594-018-0131-8
54 Wu, F., Niu, K., Cui, Y., Li, C., Lyu, M., Ren, Y. et al. (2021) Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. *Commun. Biol.* **4**, 98 https://doi.org/10.1038/s42003-020-01643-4
55 Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A. et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 https://doi.org/10.1038/nature10716
56 Skvortsova, K., Tarbashevich, K., Stehling, M., Lister, R., Irimia, M., Raz, E. et al. (2019) Retention of paternal DNA methylome in the developing zebrafish germline. *Nat. Commun.* **10**, 3054 https://doi.org/10.1038/s41467-019-10895-6
57 Fratta, E., Coral, S., Covre, A., Parisi, G., Colizzi, F., Danielli, R. et al. (2011) The biology of cancer testis antigens: putative function, regulation and therapeutic potential. *Mol. Oncol.* **5**, 164–182 https://doi.org/10.1016/j.molonc.2011.02.001
58 Wutz, A. and Barlow, D.P. (1998) Imprinting of the mouse Igf2r gene depends on an intronic CpG island. *Mol. Cell. Endocrinol.* **140**, 9–14 https://doi.org/10.1016/s0303-7207(98)00022-7
59 Zwart, R., Sleutels, F., Wutz, A., Schinkel, A.H. and Barlow, D.P. (2001) Bidirectional action of the Igf2r imprint control element on upstream and downstream imprinted genes. *Genes Dev.* **15**, 2361–2366 https://doi.org/10.1101/gad.206201
60 Sleutels, F. and Barlow, D.P. (2001) Investigation of elements sufficient to imprint the mouse Air promoter. *Mol. Cell. Biol.* **21**, 5008–5017 https://doi.org/10.1128/MCB.21.15.5008-5017.2001
61 Hu, J.F., Vu, T.H. and Hoffman, A.R. (1996) Promoter-specific modulation of insulin-like growth factor II genomic imprinting by inhibitors of DNA methylation. *J. Biol. Chem.* **271**, 18253–18262 https://doi.org/10.1074/jbc.271.30.18253
62 Hu, J.F., Oruganti, H., Vu, T.H. and Hoffman, A.R. (1998) Tissue-specific imprinting of the mouse insulin-like growth factor II receptor gene correlates with differential allele-specific DNA methylation. *Mol. Endocrinol.* **12**, 220–232 https://doi.org/10.1210/mend.12.2.0062
63 Latos, P.A., Pauler, F.M., Koerner, M.V., Senergin, H.B., Hudson, Q.J., Stocsits, R.R. et al. (2012) Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* **338**, 1469–1472 https://doi.org/10.1126/science.1228110
64 Santoro, F., Mayer, D., Klement, R.M., Warczok, K.E., Stukalov, A., Barlow, D.P. et al. (2013) Imprinted Igf2r silencing depends on continuous Airn lncRNA expression and is not restricted to a developmental window. *Development* **140**, 1184–1195 https://doi.org/10.1242/dev.088849

Biochemical Society Transactions (2021) **49** 1109–1119
https://doi.org/10.1042/BST20200695

65 Bogdanovic, O., Long, S.W., van Heeringen, S.J., Brinkman, A.B., Gómez-Skarmeta, J.L., Stunnenberg, H.G. et al. (2011) Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis. *Genome Res.* **21**, 1313–1327 https://doi.org/10.1101/gr.114843.110

66 Broche, J., Kungulovski, G., Bashtrykov, P., Rathert, P. and Jeltsch, A. (2021) Genome-wide investigation of the dynamic changes of epigenome modifications after global DNA methylation editing. *Nucleic Acids Res.* **49**, 158–176 https://doi.org/10.1093/nar/gkaa1169

67 Ford, E., Grimmer, M.R., Stolzenburg, S., Bogdanovic, O., de Mendoza, A., Farnham, P.J. et al. (2020) Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation. *biorxiv* https://doi.org/10.1101/170506

68 Long, H.K., King, H.W., Patient, R.K., Odom, D.T. and Klose, R.J. (2016) Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. *Nucleic Acids Res.* **44**, 6693–6706 https://doi.org/10.1093/nar/gkw258

69 de Mendoza, A., Lister, R. and Bogdanovic, O. (2019) Evolution of DNA methylome diversity in eukaryotes. *J. Mol. Biol.* **432**, 1687–1705 https://doi.org/10.1016/j.jmb.2019.11.003

70 Rivière, G. (2014) Epigenetic features in the oyster *Crassostrea gigas* suggestive of functionally relevant promoter DNA methylation in invertebrates. *Front. Physiol.* **5**, 129 https://doi.org/10.3389/fphys.2014.00129

71 Rajasethupathy, P., Antonov, I., Sheridan, R., Frey, S., Sander, C., Tuschl, T. et al. (2012) A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. *Cell* **149**, 693–707 https://doi.org/10.1016/j.cell.2012.02.057

72 Chen, R.A.-J., Stempor, P., Down, T.A., Zeiser, E., Feuer, S.K. and Ahringer, J. (2014) Extreme HOT regions are CpG-dense promoters in *C. elegans* and humans. *Genome Res.* **24**, 1138–1146 https://doi.org/10.1101/gr.161992.113

73 Zemach, A., McDaniel, I.E. Silva, P. and Zilberman, D. (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 https://doi.org/10.1126/science.1186366

74 Gardiner-Garden, M. and Frommer, M. (1987) Cpg islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 https://doi.org/10.1016/0022-2836(87)90689-9

75 Takai, D. and Jones, P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. U.S.A.* **99**, 3740–3745 https://doi.org/10.1073/pnas.052410099

76 Ponger, L. and Mouchiroud, D. (2002) CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* **18**, 631–633 https://doi.org/10.1093/bioinformatics/18.4.631

77 Hackenberg, M., Previti, C., Luque-Escamilla, P.L., Carpena, P., Martínez-Aroza, J. and Oliver, J.L. (2006) Cpgcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics* **7**, 446 https://doi.org/10.1186/1471-2105-7-446

78 Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, 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 https://doi.org/10.1371/journal.pbio.0060022

79 Blackledge, N.P., Long, H.K., Zhou, J.C., Kriaucionis, S., Patient, R. and Klose, R.J. (2012) Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic Acids Res.* **40**, e32 https://doi.org/10.1093/nar/gkr1207

80 Xu, C., Bian, C., Lam, R., Dong, A. and Min, J. (2011) The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. *Nat. Commun.* **2**, 227 https://doi.org/10.1038/ncomms1237

81 Blackledge, N.P., Zhou, J.C., Tolstorukov, M.Y., Farcas, A.M., Park, P.J. and Klose, R.J. (2010) Cpg islands recruit a histone H3 lysine 36 demethylase. *Mol. Cell* **38**, 179–190 https://doi.org/10.1016/j.molcel.2010.04.009

82 Farcas, A.M., Blackledge, N.P., Sudbery, I., Long, H.K., McGouran, J.F., Rose, N.R. et al. (2012) KDM2B links the polycomb repressive complex 1 (PRC1) to recognition of CpG islands. *eLife* **1**, e00205 https://doi.org/10.7554/eLife.00205

83 Wu, X., Johansen, J.V. and Helin, K. (2013) Fbxl10/Kdm2b recruits polycomb repressive complex 1 to CpG islands and regulates H2A ubiquitylation. *Mol. Cell* **49**, 1134–1146 https://doi.org/10.1016/j.molcel.2013.01.016

84 Carlone, D.L. and Skalnik, D.G. (2001) Cpg binding protein is crucial for early embryonic development. *Mol. Cell. Biol.* **21**, 7601–7606 https://doi.org/10.1128/MCB.21.22.7601-7606.2001

85 Birke, M. (2002) The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation. *Nucleic Acids Res.* **30**, 958–965 https://doi.org/10.1093/nar/30.4.958

86 Bach, C., Mueller, D., Buhl, S., Garcia-Cuellar, M.P. and Slany, R.K. (2009) Alterations of the CxxC domain preclude oncogenic activation of mixed-lineage leukemia 2. *Oncogene* **28**, 815–823 https://doi.org/10.1038/onc.2008.443

87 Guenther, M.G., Jenner, R.G., Chevalier, B., Nakamura, T., Croce, C.M., Canaani, E. et al. (2005) Global and Hox-specific roles for the MLL1 methyltransferase. *Proc. Natl Acad. Sci. U.S.A.* **102**, 8603–8608 https://doi.org/10.1073/pnas.0503072102

88 Milne, T.A., Dou, Y., Martin, M.E., Brock, H.W., Roeder, R.G. and Hess, J.L. (2005) MLL associates specifically with a subset of transcriptionally active target genes. *Proc. Natl Acad. Sci. U.S.A.* **102**, 14765–14770 https://doi.org/10.1073/pnas.0503630102

89 Whitcomb, S.J., Basu, A., Allis, C.D. and Bernstein, E. (2007) Polycomb group proteins: an evolutionary perspective. *Trends Genet.* **23**, 494–502 https://doi.org/10.1016/j.tig.2007.08.006

90 Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G. et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 https://doi.org/10.1038/nature06008

91 Simon, J.A. and Kingston, R.E. (2009) Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.* **10**, 697–708 https://doi.org/10.1038/nrm2763

92 Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S. et al. (2010) GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet.* **6**, e1001244 https://doi.org/10.1371/journal.pgen.1001244

93 Santini, S., Boore, J.L. and Meyer, A. (2003) Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. *Genome Res.* **13**, 1111–1122 https://doi.org/10.1101/gr.700503

94 Lee, A.P., Koh, E.G.L., Tay, A., Brenner, S. and Venkatesh, B. (2006) Highly conserved syntenic blocks at the vertebrate Hox loci and conserved regulatory elements within and outside Hox gene clusters. *Proc. Natl Acad. Sci. U.S.A.* **103**, 6994–6999 https://doi.org/10.1073/pnas.0601492103

95 de Rosa, R., Grenier, J.K., Andreeva, T., Cook, C.E., Adoutte, A., Akam, M. et al. (1999) Hox genes in brachiopods and priapulids and protostome evolution. *Nature* **399**, 772–776 https://doi.org/10.1038/21631

96 Ryan, J.F., Mazza, M.E., Pang, K., Matus, D.Q., Baxevanis, A.D., Martindale, M.Q. et al. (2007) Pre-bilaterian origins of the Hox cluster and the Hox code: evidence from the sea anemone, *Nematostella vectensis*. *PLoS ONE* **2**, e153 https://doi.org/10.1371/journal.pone.0000153

97 Krumlauf, R. (1994) Hox genes in vertebrate development. *Cell* **78**, 191–201 https://doi.org/10.1016/0092-8674(94)90290-9
98 Gearhart, M.D., Corcoran, C.M., Wamstad, J.A. and Bardwell, V.J. (2006) Polycomb group and SCF ubiquitin ligases are found in a novel BCOR complex that is recruited to BCL6 targets. *Mol. Cell. Biol.* **26**, 6880–6889 https://doi.org/10.1128/mcb.00630-06
99 Sánchez, C., Sánchez, I., Demmers, J.A.A., Rodriguez, P., Strouboulis, J. and Vidal, M. (2007) Proteomics analysis of Ring1B/Rnf2 interactors identifies a novel complex with the Fbxl10/Jhdm1B histone demethylase and the Bcl6 interacting corepressor. *Mol. Cell Proteomics* **6**, 820–834 https://doi.org/10.1074/mcp.M600275-MCP200
100 Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F. et al. (2012) PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Mol. Cell* **45**, 344–356 https://doi.org/10.1016/j.molcel.2012.01.002
101 Boulard, M., Edwards, J.R. and Bestor, T.H. (2015) FBXL10 protects polycomb-bound genes from hypermethylation. *Nat. Genet.* **47**, 479–485 https://doi.org/10.1038/ng.3272
102 Ross, S.E. and Bogdanovic, O. (2019) TET enzymes, DNA demethylation and pluripotency. *Biochem. Soc. Trans.* **47**, 875–885 https://doi.org/10.1042/BST20180606
103 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 https://doi.org/10.1126/science.1170116
104 Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C. and Zhang, Y. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 https://doi.org/10.1038/nature09303
105 Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A. et al. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 https://doi.org/10.1126/science.1210597
106 Maiti, A. and Drohat, A.C. (2011) Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286**, 35334–8 https://doi.org/10.1074/jbc.C111.284620
107 He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q. et al. (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 https://doi.org/10.1126/science.1210944
108 Akahori, H., Guindon, S., Yoshizaki, S. and Muto, Y. (2015) Molecular evolution of the TET gene family in mammals. *Int. J. Mol. Sci.* **16**, 28472–28485 https://doi.org/10.3390/ijms161226110
109 Almeida, R.D., Loose, M., Sottile, V., Matsa, E., Denning, C., Young, L. et al. (2012) 5-hydroxymethyl-cytosine enrichment of non-committed cells is not a universal feature of vertebrate development. *Epigenetics* **7**, 383–389 https://doi.org/10.4161/epi.19375
110 Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A.C., Rappsilber, J. et al. (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343–348 https://doi.org/10.1038/nature10066
111 Xu, Y., Wu, F., Tan, L., Kong, L., Xiong, L., Deng, J. et al. (2011) Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Mol. Cell* **42**, 451–464 https://doi.org/10.1016/j.molcel.2011.04.005
112 Wu, H., D'Alessio, A.C., Ito, S., Xia, K., Wang, Z., Cui, K. et al. (2011) Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**, 389–393 https://doi.org/10.1038/nature09934
113 Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) Cpg islands as gene markers in the human genome. *Genomics* **13**, 1095–1107 https://doi.org/10.1016/0888-7543(92)90024-m
114 Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M. et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 https://doi.org/10.1038/ng1990
115 Wang, X., Li, Q., Lian, J., Li, L., Jin, L., Cai, H. et al. (2014) Genome-wide and single-base resolution DNA methylomes of the pacific oyster *Crassostrea gigas* provide insight into the evolution of invertebrate CpG methylation. *BMC Genomics* **15**, 1119 https://doi.org/10.1186/1471-2164-15-1119
116 Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 https://doi.org/10.1038/nature01262