In vertebrates, cytosine-guanine (CpG) dinucleotides are predominantly methylated, with \(\sim80\%\) 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 \(\sim80\%\) 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
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
CGIs, particularly when located in open chromatin [52]. G4 ChIP-seq data performed on human K562 chronic high stability G4 sequences associated with CGIs were hypomethylated compared with those found outside. Whole-genome bisulphite 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 cells showed that G4 quadruplexes 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
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 bisulphite 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

<table>
<thead>
<tr>
<th>Species</th>
<th>CpG sites (%)</th>
<th>sequence features</th>
<th>CGIs described</th>
<th>Identification method</th>
<th>Elevated CpG density at TSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sponge</td>
<td>0 25 50 75 100</td>
<td>No</td>
<td>N/A</td>
<td>BioCAP</td>
<td>Yes</td>
</tr>
<tr>
<td>Amphioxus</td>
<td></td>
<td>No</td>
<td>N/A</td>
<td>BioCAP</td>
<td>Yes</td>
</tr>
<tr>
<td>Sea vase</td>
<td></td>
<td>No</td>
<td>N/A</td>
<td>BioCAP</td>
<td>Yes</td>
</tr>
<tr>
<td>Oyster</td>
<td></td>
<td>No</td>
<td>N/A</td>
<td>BioCAP</td>
<td>Yes</td>
</tr>
<tr>
<td>Honeybee</td>
<td></td>
<td>No</td>
<td>N/A</td>
<td>BioCAP</td>
<td>Yes</td>
</tr>
<tr>
<td>Worm</td>
<td></td>
<td>No</td>
<td>N/A</td>
<td>BioCAP</td>
<td>Yes</td>
</tr>
<tr>
<td>Fruit fly</td>
<td></td>
<td>No</td>
<td>N/A</td>
<td>BioCAP</td>
<td>Yes</td>
</tr>
<tr>
<td>Frog</td>
<td></td>
<td>Yes</td>
<td>BioCAP</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td></td>
<td>Yes</td>
<td>Comp, CAP, BioCAP</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Zebrafish</td>
<td></td>
<td>Yes</td>
<td>BioCAP</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

**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 (*Crassostrea gigas*), honeybee (*Apis mellifera*), 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].
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 (*Crassostrea gigas*) [70] and in the sea slug *Aplysia* [71]. Interestingly, in *Caenorhabditis elegans*, enrichment of a CPF1 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.
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
87 Guenther, M.G., Jenner, R.G., Chevalier, B., Nakamura, T., Croce, C.M., Canaani, E. et al. (2005) Global and Hox-specific
Crassostrea gigas
1118 © 2021 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution License 4.0 (CC BY-NC-ND).