CGGBP1-regulated heterogeneous C–T transition rates relate with G-quadruplex potential of terrestrial vertebrate genomes

Praveen Kumar and Umashankar Singh†
Department of Biological Sciences and Engineering, Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat. PIN: 382055.
Praveen Kumar: kumarpraveen@iitgn.ac.in
ORCID ID: 0000-0001-8485-5396
† Address for correspondence: Umashankar Singh: usingh@iitgn.ac.in
ORCID ID: 0000-0001-8578-8201

Highlights

- Evolution of vertebrate genome G4-forming potential mirrors that of CGGBP1.
- By preventing GC-loss, CGGBP1 preserves the genome's ability to form G4 structures.
- CGGBP1 curtails C-to-T transitions at potential G4-sites, with its impact shaped by evolutionary context.
- The G4-forming regions most responsive to CGGBP1 exemplify this methylation-linked defense against C-to-T substitutions, distinguishing homeotherms from poikilotherms.

Abstract

The formation of G-quadruplexes is fundamentally linked to the DNA sequence, particularly regions rich in guanine. Unlike other DNA features, evolution has not favored specific agents to stabilize G-quadruplex structures; their formation depends strictly on inherent sequence properties. Vertebrate promoters are notable for their consistently high GC-content and pronounced G/C strand asymmetry. The evolutionary acquisition of CGGBP1 in vertebrates has functionally constrained methylation-driven cytosine-to-thymine transition mutations at GC-rich transcription factor binding sites within promoters, thereby preserving GC-content in these regulatory elements across amniote genomes. CGGBP1 target sequences are associated with GC-rich DNA and display G/C asymmetry, while also modulating G-quadruplex formation. This underscores the importance of investigating G4-forming potential retention across vertebrates through CGGBP1's role in protecting GC-content. In this study, we reveal the extensive potential for G-quadruplex formation in the promoters and genomes of over 100 vertebrate species. The evolutionary conservation of these GC-rich regions arises from selective mechanisms that protect both the G-rich strand, capable of forming the G4 structures, and the unused C-rich strand by suppressing C-to-T transitions. We demonstrate that G4-forming regions under nuclear constraints are governed by this restrictive process, mediated by CGGBP1. Furthermore, sequences most influenced by CGGBP1 levels also conform to this protective evolutionary strategy. Our data suggest that CGGBP1 acts not only to prevent excessive secondary structure formation but also to sustain genome-wide GC-richness by regulating methylation-associated C-to-T transitions in these sequences. Collectively, these findings underscore the evolutionary emergence of CGGBP1 as a guardian of genomic GC-richness.

Keywords

CGGBP1, meC-T transition, Genome, G-quadruplex, Evolution, Vertebrates

Abbreviations

Ac: *Anolis carolinensis*

DCRs: Differentially Captured Regions
Ev: Empty vector
G4s: G-quadruplexes
Gg: *Gallus gallus*
Hs: *Homo sapiens*
Lc: *Latimeria chalumane*
meC-T: methylated cytosine to thymine
MAP-MAP: Methylation-associated point mutation assessment pipeline
MeDIP: Methylated DNA Immuno-precipitation
TFBSs: Transcription Factor Binding Sites
dsDNA: double-stranded DNA
ssDNA: single-stranded DNA

Introduction

G-quadruplexes are well characterized structures formed by nucleic acids with Hoogsteen hydrogen bonding between tandemly repeating units of guanine residues (Varshney et al., 2020). Guanine-rich DNA and RNA sequences possess the intrinsic capacity to fold into G-quadruplex structures, which are stabilized by coordination between four guanines and monovalent cations to form distinctive tetrad geometries (Balasubramanian et al., 2011). The prevalence of G-quadruplex-forming sequences throughout the genome, particularly within regulatory regions such as promoters and telomeres, underscores their biological significance in controlling fundamental cellular processes including gene expression and chromosome stability (Huppert & Balasubramanian, 2005; Varshney et al., 2020). Formation of DNA quadruplexes is definitely dependent on the sequence properties, most importantly the occurrence of interrupted stretches of G triplets, yet unpredictable just from the DNA sequence alone. A series of reports have shown that formation of G-quadruplexes is under regulation of different proteins as well as hitherto uncharacterized mechanisms (Balasubramanian et al., 2011; Datta et al., 2024; Huppert & Balasubramanian, 2005; Varshney et al., 2020). Independent of the rather transient formation of G-quadruplexes, the capacity of specific DNA sequences to form G-quadruplexes can be viewed as their G4-forming potential. The collective G4-forming potential of a genome can hence be used to study evolution of genomic G4-forming potentials.

Whether G4 structures materialize *in vivo* remains variable and context-dependent, similar to other epigenetic traits. G4 formation is fundamentally determined by DNA sequence properties and regulated through two distinct molecular mechanisms: in the first, specialized helicases such as BLM, WRN, and DHX36 actively disrupt G4 structures while preserving the underlying G4-forming potential of the DNA sequence (Mendoza et al., 2016; Varshney et al., 2020; W. Wu et al., 2018) in the second, the intrinsic sequence properties permitting G4 formation are conserved, allowing the DNA to retain its capacity for G4 structure formation (Du et al., 2008). This sequence conservation has been demonstrated to be under evolutionary selection within gene regulatory regions, indicating the functional importance of these structures (Du et al., 2008). However, if this sequence conservation is disrupted by mutagenic processes, particularly methylation-associated C-to-T transitions, the G4-forming potential is compromised or lost entirely (Mao et al., 2018; Neupane et al., 2023). Such mutations can occur both in G-rich runs critical for tetrads (GGG regions) and in loop regions between runs, with transitions accounting for approximately 22% of G4 structure-destabilizing mutations (Neupane et al., 2023). Furthermore, the relationship between DNA methylation and G4 formation is bidirectional. While methylation can destabilize certain G4 structures or prevent their formation, G4 structures themselves protect CpG islands from methylation by sequestering and inhibiting DNA methyltransferase 1 (Mao et al., 2018). This interplay suggests that G4 formation and the preservation of G4-forming sequence potential represent a critical regulatory interface between genomic sequence architecture and epigenomic regulation (Mao et al., 2018).

Cytosine methylation is one of the most well-conserved epigenetic mechanisms known to occur in almost all organisms with double-stranded DNA as their genetic material. Traditionally, it has been studied as a chemical modification that affects DNA-protein interactions, chromatin compaction, gene expression, and a molecular memory for genome function over and above the DNA sequence (Baubec & Schübeler, 2014; Jones, 2012; Moore et al., 2013). More recently, cytosine methylation has been studied in the context of non-canonical DNA structures, most importantly the G-quadruplexes (G4s) (Niu et al., 2025; Wang et al., 2020). The formation of G4s primarily depends on short and spaced repeats of G on the same DNA strand. Although cross-strand annealings of G-rich segments do yield inter-strand quadruplex-like structures, it only affects the length of G-rich segments on a particular strand of DNA without negating the necessity of skewed G-richness on the strands participating in G4 formation. The G4-forming region inherently has a local GC-skew with the C-rich strand offering substrates for non-CpG cytosine methylation. In the event of G4 folding of the G-rich strand, the C-rich strand, which is left single-stranded, runs the risk of a higher mutation load, which is exacerbated by cytosine methylation. The meC-T transitions occur spontaneously at a rate much lower than possible otherwise due to the double-stranded nature of the DNA. While most of the U (generated by spontaneous deamination of C) is rectified back to C guided by G:::C pairing on dsDNA, the T (generated by spontaneous deamination of 5meC) poses a G-A mutational load on the complementary strand. GC-richness combined with GC-skew is rich in transcription factor binding sites (TFBSs), often concentrated in the *cis* promoters and enhancers genome-wide (Deaton & Bird, 2011; Hartono et al., 2015).

TFBSs in vertebrate genomes are abundantly GC-rich, with cytosine methylation functioning as a mitigator of TFBS-DNA binding (Yin et al., 2017). The primarily unmethylated status of CpG dinucleotides in CpG islands is a classic demonstration of the cooperation between TFBS activity, gene expression, and the unmethylated status of cytosines (Yin et al., 2017). The palindromic nature of the CpG context ensures symmetrical cytosine content on complementary strands, as well as the maintenance of methylation across replication and repair. The CpG dinucleotides thus abrogate any inter-strand GC-skew and associated secondary structures such as G4, which are potentially formed at regions with non-CpG cytosines (Agarwal et al., 2015; Datta et al., 2024; Patel et al., 2019). There is evidence that vertebrate TFBSs have an abundance of non-CpG GC-contents (Khuu et al., 2007), which are prone to strand-asymmetric cytosine methylation, loss of C to T, and eventual loss of TFBS GC-content. The concentration of non-CpG TFBSs in the *cis* promoters, as well as the “dark matter” of the genome, increases the chances of GC-skew and resultant G4 formation potential. Since the sparse non-CpG cytosines tend to be predominantly methylated, due to an ancient mechanism of guarding against non-self genetic material, it combines two distinct processes: a chemical mark that is methylation of cytosines and a physical structure that is the G quadruplex. Biochemical studies have shown that G4 stability is dually affected by cytosine methylation, with evidence that cytosine methylation generally enhances G4 stability (Lin et al., 2013). Since cytosines are not involved in the G4 Hoogsteen bonding directly, their influence on G4 stability is likely to be indirect and as yet poorly understood.

Given the above premise, there is a strong case for the loss of GC-content in the course of evolution through the meC-T mutation. However, against all odds, the course of terrestrial vertebrate evolution has not only preserved the GC-content, but also enriched it (Kumar et al., 2025; Morbia et al., 2024). As compared to the coelacanths, the sarcopterygian lineage ancestral to the terrestrial vertebrates, and non-amniotes, the avian and mammalian genomes generally have higher GC-content (Kumar et al., 2025). The various lineages of the divergent reptilian groups show highly divergent GC-contents. Such a retention of GC-content, especially in the non-CpG contexts, stands at odds with the default nature of non-CpG island cytosines. Such a retention of GC-content, with implicit conserving effects on TFBS-DNA interactions, could not have occurred without an active mechanism. The transition of vertebrates from aquatic to terrestrial habitats has been accompanied by the evolution of CGGBP1, a gene that seems to fill this functional void. CGGBP1 binds to GC-rich DNA and, amongst a host of other effects, manifests functions pertinent to our discussion: mitigation of cytosine methylation in non-CpG contexts, and prevention of G4 formation at regions of GC-skew, and counteraction of R-loop accumulation and transcription-replication conflicts at short CGG-repeat tracts prone to R-loop formation (Agarwal et al. 2015; Datta et al. 2024; Ummethum et al. 2025). Recent reports have shown that the TFBS landscapes have retained/gained GC in higher amniotes (which are also endothermic homeotherms), whereas the poikilothermic lower amniotes have progressively lost TFBS GC-contents under the influence of CGGBP1 (Kumar et al., 2025). Bernardi and colleagues demonstrated that DNA methylation levels are inversely correlated with body temperature across vertebrate species, with poikilothermic (cold-blooded) organisms exhibiting significantly higher DNA methylation levels compared to homeothermic (warm-blooded) organisms (Varriale & Bernardi, 2006). This temperature-dependent methylation pattern, reflecting evolutionary adaptations to organismal physiology, appears to be fundamentally encoded within the CGGBP1 protein itself. Recent experimental evidence demonstrates that simply expressing CGGBP1 orthologs derived from poikilothermic species (*Latimeria chalumnae* and *Anolis carolinensis*) versus homeothermic species (*Homo sapiens* and *Gallus gallus*) recapitulates the ancestral methylation signatures associated with their organismal origin (Kumar et al., 2025) in human somatic cell cultures. Remarkably, when poikilothermic CGGBP1 orthologs are overexpressed, the resulting genome-wide methylation profile exhibits elevated methylation levels characteristic of cold-adapted species, whereas homeothermic CGGBP1 orthologs impose methylation restriction patterns typical of warm-blooded organisms (Kumar et al., 2025). We assume and argue that a similar cytosine methylation regulatory function of CGGBP1 would have operated in the germline giving rise to methylation associated with mutational biases and genomic base compositional changes. To the extent this holds true, comparisons of genome-wide G4-forming potentials and CGGBP1-driven point mutation rates can inform us about how cytosine methylation has affected G4-forming potentials in the course of vertebrate evolution.

The GC-richness of higher amniotic genomes, especially the promoters, inherently increases their G4-forming capacity. The evolutionary expansion and conservation of G-quadruplex motifs suggest they play crucial roles in gene regulation and genome stability (F. Wu et al., 2021). Particularly, the progressive enrichment of G4 motifs in regulatory regions and their increasing density and diversity during species evolution indicates that higher organisms have evolved G-quadruplexes into an elaborate transcriptional regulatory mechanism to meet complex physiological and behavioral demands (F. Wu et al., 2021). Apparently, the evolution of terrestrial vertebrates has the following underlying coincidental genomic feature enrichment, which is likely to be more concentrated at promoters: an increase in G4-forming potential reflected as GC-skew, GC-richness, and an enhanced property of cytosine methylation mitigation. Evidently, the evolution of G4-forming potential of genomes has an epigenetic basis in the form of cytosine methylation, at the TFBS-rich proximal promoters, as well as the genomic dark matter and gene expression. A systematic assessment of the evolution of the G4-forming potential of genomes and its relationship with cytosine methylation is required to understand this mechanism. Through a simple yet rigorous meta-analysis of 105 genomes, we demonstrate an evolutionary path of G4-forming potential in taxon groups ranging from coelacanths to eutherians. We provide evidence that the evolution of the cytosine methylation mitigator protein CGGBP1 correlates with this enhanced G4-forming potential of mammalian genomes. We show that CGGBP1 expression specifically restricts meC-T transition mutations (not other types of transitions or transversions) strongly at genome-wide regions with GC-skew and high G4-forming potential, which are concentrated at *cis*-proximal promoters. Our findings are based on the predictable G4-forming potential, actual GC-skew, and a base-level analysis of the point mutation landscape. These results explain the pattern of evolution of the G4-forming potential of terrestrial vertebrate genomes and its evolution alongside the progressive emergence of amniotes, homeotherms, and mammals. The results presented here provide an epigenetic basis for the lineage-specific differences in G4-forming potentials, present C deamination as a key molecular mechanism that indirectly drives the evolution of G4-forming potential and posit CGGBP1 as a key arbiter of G4-forming potential by determining the mutability of G4-complementary C-rich strand.

Results

GC-content of potential G4-forming regions genome-wide and 1kb promoters has different selective pressure on GC retention in vertebrate genomes

DNA can form various secondary structures beyond the familiar double helix; one such structure is the G-quadruplex (G4), which arises from guanine-rich sequence patterns. These signatures are useful for predicting potential G4-forming regions genome-wide. In this study, we used pqsfinder to identify potential G4-forming regions in vertebrate genomes (Hon et al., 2017). This tool detects these regions based on sequence signatures characterized by four consecutive (though sometimes imperfect) guanine runs, separated by variable-length loops. The classical G4-forming motif typically fits the pattern G≥3N1-7G≥3N1-7G≥3N1-7G≥3, where 'G' denotes guanine tracts of at least three residues, and 'N' denotes intervening loop sequences of 1-7 nucleotides (Hon et al., 2017). Unlike earlier methods, pqsfinder can also detect non-canonical G4s that include imperfections, such as bulges or longer loops, broadening the range of identifiable G4 structures (Labudová et al., 2020). Analyses of potential G4-forming regions genome-wide across 105 vertebrate species (57 mammals, 27 aves, 17 reptiles, and 4 non-amniotes (summary in Supplementary Table S1 and details in methods) show that G4-forming potential per megabase (Mb) is positively correlated with the GC-content across three classes of vertebrate genomes significantly (Mammals, aves, and reptiles as lower amniotes and non-amniotes in one class) (Fig. 1a; Supplementary Table S2). However, the GC-content of these potential G4-forming regions is correlated with the GC-content of the genomes only in mammals (Fig. 1b; Supplementary Table S3). On the other hand, non-mammalian genomes (Aves, reptiles, and non-amniotes) GC-content is uncorrelated with the potential G4s' GC-content (Fig. 1b; Supplementary Table S3). Avian genomes' GC-content is tightly clustered in a range with varying GC-content of potential G4-forming regions (Fig. 1b; Supplementary Table S3). Reptiles (lower amniotes) and non-amniotes, being the most divergent group, have lineage-specific genome GC-content divergence with varying potential G4s GC-content (Fig. 1b). Further, compositional analysis of these potential G4s' length against their GC-content reveals GC-poor regions in non-mammalian vertebrates, and mammals significantly preserve the GC-content of these potential G4-forming regions (Fig. 1c; Supplementary Table S4). This indicates that overall G4-forming potential of vertebrate genomes is maintained such that it can form G4, but the internal GC-richness of these G4s is susceptible to loss. Repeating the same analyses using all annotated *cis* promoters of vertebrate genomes (summary in Supplementary Table S5 and for details see methods) revealed that the G4-forming potential of 1 kb *cis* promoters, from non-amniotes to mammals, is significantly higher than the genome-wide G4-forming potential(Fig. 1d; Supplementary Table S6).

Furthermore, this G4-forming potential is positively correlated with promoter GC-content across all taxonomic groups (Fig. 1d). Unlike the genome-wide pattern, the GC-content of potential G4s within 1 kb *cis* promoters in vertebrates from non-amniotes to mammals is correlated with the promoters' GC-content (Fig. 1e; Supplementary Table S7). Homeothermic vertebrates (Aves and mammals) 1 kb *cis* promoters exhibit GC retention in potential G4-forming regions, as indicated by a positive correlation between G4 length and GC-content except in non-amniotes and reptiles (poikilotherms), which show a weak correlation (Fig. 1f; Supplementary Table S8). These results indicate two key points. First, the selective pressures on G4-forming potential differ between genome-wide regions and promoters. Second, GC retention in potential G4-forming regions varies across taxonomic groups, both genome-wide and within promoters. The observed loss of GC-content is linked to the identification of these G4-forming sequences; if the GC were not selectively lost, these regions might not be detected by the prediction program. This suggests that these regions experience selective GC loss that does not disrupt G4 formation but reduces the overall GC-content in G4-forming regions. Regions not identified as potential G4s may have also undergone such GC loss, although this cannot be confirmed here.

CGGBP1, a GC-rich binding protein, is known to prevent GC loss genome-wide and specifically within transcription factor binding sites in 1 kb *cis* promoters of certain target genes. The GC retention function of CGGBP1 is associated with its role in preventing the meC-T transition. Additionally, CGGBP1 preferentially binds regions with GC asymmetry that have the potential to form G4 structures. Notably, depletion of CGGBP1 leads to increased G4 formation, indicating that CGGBP1 acts to suppress G4 formation. Therefore, potential G4-forming regions in the human genome likely exhibit these features and show dependence on CGGBP1 activity.

Human CGGBP1 prevents GC loss in potential G4-forming regions by limiting C-to-T transitions genome-wide

We next focused on potential G4-forming regions identified by pqsfinder in the human genome to investigate whether evolutionary forms of CGGBP1 influence methylation-associated GC-loss in these regions. We used recently published data (Kumar et al., 2025), to assess the evolutionary trajectory of CGGBP1-mediated regulation of genome-wide methylation patterns, vertebrate-specific CGGBP1 orthologs representing key evolutionary lineages non-amniotes (*Latimeria chalumnae*), reptiles (*Anolis carolinensis*), aves (*Gallus gallus*), and mammals (*Homo sapiens*) were ectopically expressed in HEK293T cells. Methylation-associated point mutations affecting genome-wide cytosine methylation were subsequently assessed using the Methylation Associated Point-Mutation Assessment Pipeline (MAP-MAP), a computational framework designed to profile mutation dynamics in methylation patterns across the experimental samples, we observed an overall increase in mutation rates (summary in Supplementary Table S9), with the highest rates seen in *Homo sapiens* (Hs) CGGBP1 (Fig. S1, more details in Supplementary Table S10). CGGBP1 is known to restrict cytosine methylation and meC-T transition rates genome-wide. Analysis of meC-T transitions within potential G4-forming regions revealed that endogenous CGGBP1 (Ev) limits these transitions, thereby reducing GC-loss in these sequences (Fig. 2a). Evolutionary forms of CGGBP1 differ in this capacity: evolutionary forms of CGGBP1 exhibit differing abilities in this regard: *Latimeria chalumnae* (Lc) and *Anolis carolinensis* (Ac) forms were unable to prevent meC-T transitions in human cells, *Gallus gallus* (Gg) showed partial inhibition, and *Homo sapiens* (Hs) displayed the strongest restriction, minimizing these transitions to the greatest extent (Fig. 2a). Notably, over-expression of Hs markedly minimized all mutations leading to GC-loss (Fig. S1). These findings suggest that human CGGBP1 prevents genome-wide GC-loss, including in potential G4-forming regions, by restricting meC-T transitions, a function likely acquired through evolution. This mechanism may contribute to the preservation of both G4-forming potential and the GC-content of G4 sequences in vertebrate genomes. Our earlier findings showed that CGGBP1 preserves GC content at transcription factor binding sites in repressed promoters by limiting meC-T transitions, a property developed over evolutionary time. This underscores the importance of examining promoter regions, their G4-forming potential, and the impact of CGGBP1 on GC retention both in G4 structures and the broader promoter context.

Higher meC-T transition rates in promoter G4-forming sequences compared to background promoters are shaped by evolutionary differences in CGGBP1

We further focused on all annotated 1 kb promoters in the human genome and their potential G4-forming regions. The predicted G4 sequences within these promoters exhibited a CGGBP1 evolutionary form-dependent restriction of meC-T transition rates (Fig. 2b). This pattern mirrored the genome-wide trend, with Hs CGGBP1 acting as the strongest suppressor of cytosine methylation-induced transitions in predicted G4-forming sequences. The endogenous CGGBP1 represented by empty vector (Ev) showed moderate restriction compared to input controls, Gg displayed a weaker effect, while Lc and Ac failed to restrict these transitions effectively (Fig. 2b). These differences correspond to previously reported distinctions between homeothermic and poikilothermic vertebrates in GC retention mediated by CGGBP1's evolutionary forms, particularly through the restriction of meC-T transitions in transcription factor binding sites. GC asymmetry is a characteristic sequence feature associated with CGGBP1-regulated regions. GC asymmetry measured as interstrand GC-skew is known to overcome the ds of DNA and form G4 (Datta et al., 2024). To explore this further, we examined GC-skew in potential G4-forming regions genome-wide, within 1 kb cis promoter G4-forming regions, and in regions lacking potential G4 formation. Our analysis revealed that potential G4-forming regions exhibit higher GC-skew overall. Specifically, genome-wide predicted G4-forming regions display greater GC-skew than those within 1 kb promoter G4-forming regions (Fig. S2). In contrast, regions without potential G4 formation showed no significant enrichment of GC-skew (Fig. S2). Mutation-free regions across MeDIP samples exhibit varying G4 formation potential depending on the over-expressed CGGBP1 variant (Fig. S3). Hs maintains the highest G4-forming potential among these, accompanied by detectable GC-skew in these regions, although the skew is not centrally enriched (Fig. S3). In contrast, methylation-enriched, mutation-free regions lack GC-skew, with only Hs showing GC-skew comparable to the genomic background. For Lc, Ac, Gg, and Ev, G4-forming potential decreases significantly alongside a marked reduction in GC-skew (Fig. S3). This means that mutation-free regions enriched in MeDIP DNA in different samples in Hs show the highest G4 formation capability with retention of GC-skew as that of the genomic background.

Next, we compared meC-T transition rates in potential G4-forming regions to those in background promoter sequences by examining overall point mutation rates across entire 1 kb promoter regions, as well as within G4-forming segments located in promoters (see Supplementary Table S11 and Table S12). In both analyses, we observed that homeothermic samples (Hs and Gg) displayed increased point mutation rates that favored GC retention, distinguishing them from the poikilotherms (Lc and Ac) (Supplementary Fig. S4a and Fig. S4b). This trend—preferential GC retention—was evident both at the level of whole promoters and within their potential G4-forming regions. All observed point mutations, expressed as fractions of all mutations within each sample, were statistically significant for both promoters and predicted G4s in promoters (Supplementary Table S13 for promoters, Table S14 for G4s within promoters). In all samples, meC-T transition rates differed significantly in both promoter regions and their potential G4s (Supplementary Table S15 and Table S16, respectively).

To further discern whether meC-T transitions are more prevalent in promoters as a background or specifically within G4s, we compared meC-T to all point mutations in both contexts. Results showed that in homeotherms (Hs and Gg), G4-forming regions were more prone to meC-T transitions than the broader promoter background, whereas in poikilotherms (Lc and Ac), the reverse was true. This difference reached significance only for Hs among homeotherms and for Lc among poikilotherms (Supplementary Tables S17–S21). After adjusting for input controls, potential G4-forming sequences maintained higher meC-T transition rates compared to background regions, in line with the evolutionary dependence on CGGBP1, except for Lc (Fig. 2c). Overexpression of Hs CGGBP1 resulted in the strongest restriction of meC-T transitions in both promoters and predicted G4 regions, followed by the endogenous CGGBP1 level and then Gg, while Lc and Ac showed the highest mutation rates in both regions (Fig. 2c). This elevated mutation rate in G4 regions is expected, as G4 formation requires the local unwinding and separation of the two DNA strands, exposing single-stranded DNA (ssDNA). ssDNA is more vulnerable to various mutagenic processes, including spontaneous deamination of methylated cytosines, which leads to increased meC-T transitions. Thus, the structural dynamics essential for G4 formation inherently increase the susceptibility of these regions to methylation-associated mutations. After confirming the impact of CGGBP1 on GC preservation via limiting meC-T transitions in potential G4-forming promoters and genome-wide regions, it is essential to realize that these sites harbor G4-forming potential but may not always form G4 structures. Consequently, to study methylation effects on G4s, experiments must account for both chromatin-dependent environments influenced by CGGBP1 and settings devoid of nuclear constraints.

G4-forming sequences retain GC under chromatin constraint imposed by CGGBP1

G4 formation reflects the intrinsic propensity of sequences to adopt these alternative structures beyond the thermodynamic constraints imposed by the nuclear environment, which generally favors double-stranded DNA (dsDNA). Examining regions capable of forming G4s under native nuclear conditions provides an accurate representation of GC retention at experimentally validated G4-forming sites. We utilized previously identified regions characterized under two conditions: G4-IP-nat, representing G4s formed under true native nuclear constraints, and G4-IP-Denat, representing heat-denatured DNA capable of forming G4s *in vitro* without nuclear or chromatin constraints, as well as dsDNA (Datta et al., 2024).

Since CGGBP1 functions exclusively within the nuclear environment where chromatin structure and nuclear constraints influence G4 formation our focus was to understand CGGBP1’s role at G4-forming sites that persist under native conditions. Analysis of meC-T transition rates in regions forming G4s under nuclear constraints revealed a strong restriction imposed by over-expressed Hs CGGBP1 (Fig. 3a). This restrictive ability was characteristic of homeothermic CGGBP1 variants (Hs and Gg) but absent in poikilothermic variants (Lc and Ac) (Fig. 3a). Interestingly, the heat-denatured DNA (G4-IP-Denat) capable of forming G4s *in vitro* did not exhibit C-to-T restriction by any CGGBP1 form, consistent with the notion that these sequences do not form G4s under native nuclear conditions (Fig. 3b). This suggests that CGGBP1’s regulation under nuclear constraints is essential to link potential G4-forming regions with cytosine methylation-driven deamination.

These findings demonstrate that CGGBP1 acquisition in vertebrate evolution functionally constrains meC-T transitions in sequences capable of forming G4s within the chromatin context. Further examination of sequence features showed that, under nuclear constraints, G4-forming regions are enriched with high GC-skew, surrounded by broader GC-skew-poor areas. In contrast, G4-IP-Denat regions show only background genomic levels of GC-skew without central enrichment (Fig. S5). This supports the idea that the double-stranded nature of DNA restricts G4 formation, with localized regions (~0.1 Mb flanks) of pronounced GC-skew overcoming the constraints of strand complementarity and Watson-Crick pairing to enable G4 formation genome-wide (Fig. S5). Moreover, the increased frequency of G4 formation at regions with low and narrow GC-skew upon heat denaturation underscores that disrupting dsDNA structure facilitates G4 formation where it is otherwise limited.

Further, we focused on regions showing the most significant quantitative differences in G4 capture when CGGBP1 levels were altered, defined as Differentially Captured Regions (DCRs) as reported in earlier studies (Datta et al., 2024). DCRs represent genomic loci demonstrating significant variation in G-quadruplex formation when comparing CGGBP1-sufficient and CGGBP1-depleted cellular environments. These regions, identified through comparative analysis of G4 signal intensities across 0.2 kb genomic bins using the AbC G4-ChIP methodology, predominantly exhibit elevated G4 signal intensity upon CGGBP1 removal, indicating that this protein exerts suppressive effects on quadruplex materialization at these genomic locations (Datta et al., 2024). The genomic architecture of DCRs is characterized by substantial guanine-cytosine richness and pronounced interstrand G/C-skew asymmetry, features that correlate with CGGBP1-mediated restraint of G4 formation (Datta et al., 2024). Analysis of C-to-T transition rates in these regions showed that homeothermic CGGBP1 forms (Hs and Gg) restrict these transitions, whereas poikilothermic forms (Lc and Ac) do not (Fig. 3c and Supplementary Table S22-S23). Over-expression of Hs CGGBP1 strongly suppresses C-to-T transitions, and even endogenous CGGBP1 restricts these transitions compared to input (Fig. 3c). These regions are highly sensitive to CGGBP1 levels and have the potential to form G4s. The data suggest that GC-content retention in potential G4-forming regions is largely due to Hs CGGBP1 and, to some extent, avian CGGBP1. We propose a mechanism of co-evolution of CGGBP1, heterogeneous C-to-T transition rates and G4-forming potential in the vertebrate genomes (Fig. 4). This appears to be an evolutionary trait acquired by CGGBP1. Since G4 formation is a molecular byproduct driven by DNA sequence chemistry, CGGBP1 has evolved to limit C-to-T transitions and preserve GC-content in vertebrate genomes, where G4s are part of this GC-rich DNA. This offers a new evolutionary perspective on the cellular function of CGGBP1.

Discussion

CGGBP1 is reported or predicted in all terrestrial vertebrate classes, with indications of gene-loss in most amphibian groups and retention only in amniotes (Kumar et al., 2025; Morbia et al., 2024; Singh & Westermark, 2015). Within the amniotes, the evolutionary conservation pattern of CGGBP1 follows the endo-homeothermic versus poikilothermic divide. The G4-forming potential, its GC-richness, and GC-skew of G4-forming regions also follow the same pattern. The best bifurcation of the G4-forming potential genome-wide is achieved only when we cluster the 105 species through a simple homeothermic-poikilothermic divide, not otherwise. These findings establish that with the evolution of the terrestrial vertebrates, the epigenome (in the form of cytosine methylation landscape) has affected the genome, its GC-skew, and G4-forming potential under the influence of CGGBP1. This enrichment of G4-forming potential in homeotherms coincides with a G4-preventing function of CGGBP1. The increase in G4-forming potential has been under an evolutionary check by the same agent that facilitates it, the homeothermic CGGBP1, best represented by its highly conserved mammalian form.

Our analysis establishes that the G4 GC-contents have increased and correlated with genomic GC-contents during vertebrate evolution. The evolution of CGGBP1 and its C-retentive effects appears to be causal to this phenomenon. Interestingly, of all the GC-relevant point mutations analyzed, all C-retentive (secondarily G-retentive) point mutations are antagonized by CGGBP1. This function of CGGBP1 ascends from non-amniotes to mammals, mimicking the GC-content increase in G4s as shown in Figure 1. A direct test of the potentially causal role of CGGBP1 in G4 GC-retention is to test C-retention rates in actual G4-formed regions *in cellulo* for which a regulatory role of CGGBP1 is established. We discover that the acute expression of different CGGBP1 forms inside HEK293T cells affects C-retention differently very strongly at CGGBP1-mitigated G4s. When compared against the G4-forming potential of the genomes, not under chromatin constraints, we observed a distinction between G4 structures in actuality, rather than the potential of DNA sequences to form G4 reflected as DNA sequence properties. Specific meC-T transition restriction by CGGBP1 is observed only at regions that can overcome the constraints on G4 formation by double-strandedness. These regions have strong G4 folding properties such that, given a loss of chromatin constraints, *in vitro*, they would form G4. In contrast, the regions with potential for G4 formation, which can be realized only when the double-strandedness is physically overcome, are not targets for meC-T restriction by CGGBP1. *In vivo*, the G4-forming potential realization into actual G4s is impeded not only by double double-strandedness of the DNA but also by chromatin constraints. Nucleosomes, DNA-binding proteins, and epigenetic features render the G4-forming potential into actual G4 folding depending on a combination of the following: innate G4-forming potential, default state of double-strandedness, chromatin constraints, transitory single-stranded state due to processes like replication, transcription, repair, and recombination (Johnson, 2020; Maizels, 2012). Hs CGGBP1 has been shown to be required for exerting the chromatin constraints on G4 formation (Datta et al., 2024). *In vitro*, recombinant CGGBP1 can impede G4 folding of ssDNA by physically associating with it through mechanisms that are not completely understood. CGGBP1 could be collaborating with additional factors in mitigating G4 folding; however, it remains clear that just by itself, CGGBP1 can interfere with G4 formation (Datta et al., 2024; Ummethum et al., 2025). The current work sheds light on the biochemical nature of this function of CGGBP1 and the effect of its evolution on the G4 landscape of higher vertebrates.

It is interesting to note that through the evolution of higher vertebrates, there has been no selection against G4-forming potential. To the contrary, as our results demonstrate, genomic features that favour G4 formation seem to have become enriched. One such primary feature, apart from GC-richness, is that of inter-strand GC compositional asymmetry. Without invoking the role of CGGBP1 in G4 biology, it remains unclear what the true evolutionary force underlies this G4-forming potential enrichment in higher vertebrates. It could either be the actual G4 formation, or perhaps the GC-skew offers other functional roles, such as TFBSs and DNA biophysical properties compatible with replication and transcription bubble formation. The latter would imply that G4 folding of dsDNA is an unavoidable outcome of the DNA composition actually selected for functions requiring DNA structural turnovers. The absence of any known G4-stabilizing proteins or domains and an abundance of helicases and G4-resolving proteins lend support to this possibility. The role of CGGBP1 provides an additional layer of regulation on the G4-forming potential: it interferes with the G4 folding and, at the same time, protects and helps retain G4-forming potential. Interestingly, the mechanism underlying this dual effect of CGGBP1 on G4-forming regions also favours GC-retention at TFBSs in target gene promoters (Kumar et al., 2025). Biochemically, CGGBP1 interferes with cytosine methylation, restricts meC-T transitions, and thereby affects GC-contents locally (Agarwal et al., 2015; Kumar et al., 2025; Morbia et al., 2024; Singh & Westermark, 2015). In terms of G4-forming potential retention and regions with high GC-skew, this attributes a unique function to CGGBP1: It is not the G-rich G4-forming strand, rather the C-rich complementary strand, which is under the influence of cytosine methylation and a higher load of meC-T transitions. A primary loss of C on the C-rich strand would secondarily lead to G-loss on the G4-forming strand. Such a mechanism of G4 regulation is not reported and is unique to CGGBP1. Through cytosine methylation restriction, CGGBP1 ensures that C and eventually G are retained, which is reflected as G4-forming potential.

Since this effect of CGGBP1 is directly proportional to the GC-content of the genomes, one would expect that the increasing correlations between GC-contents of G4s and respective genomes from non-amniotes and lower amniotes to the higher amniotes are due to a parallel evolution of CGGBP1. Our findings here show that representative forms of CGGBP1 from these groups mimic the biochemical effects of CGGBP1 on meC-T transitions in acute overexpression experiments. Remarkably, these acute biochemical effects of CGGBP1 on the cytosine methylome reflect the evolutionary changes in genomic GC-content and G4-forming potential associated sequence features.

Together, these findings show that CGGBP1 mediates the evolution of G4-forming potential of the genomes through two mechanisms. One is by exerting a chromatin constraint and disallowing G4 folding, and the second, by minimizing the meC-T loss on the C-rich non-G4 strand through a cytosine methylation restriction. Apparently, this function of CGGBP1 is enriched in vertebrate evolution and observed as concentrated only at regions with a strong tendency to form G4s, not at all locations genome-wide, with a varying range of G4-forming potentials. The emergence of higher vertebrates, which are homeothermic, coincides with genomic GC-richness, favouring TFBS richness against a G4-forming potential which can affect TF-binding and genome function. The evolutionary pressures on establishing a balance between these two opposing features apparently operate not through regulation of G-richness, but rather through maintenance of the complementary C-richness. The evolution of CGGBP1 has played an important role in reconciling non-CpG cytosine methylation, GC-skew, GC-richness, and G4-forming potential with functional outcomes of the genome compatible with the emergence of higher amniotes and homeothermy.

Conclusions

- Our findings position CGGBP1 as a key evolutionary safeguard of GC-rich, methylation-vulnerable regions, where it limits C→T transitions and stabilizes GC content at promoters, TFBSs, and G4-forming elements, particularly in mammals and birds.
- By integrating vertebrate comparative genomics with base-level mutation profiling, we show that CGGBP1 activity relates with G4-forming potential and GC stability, with human CGGBP1 displaying a uniquely strong capacity to suppress methylation-driven GC loss at structurally complex regulatory loci.
- Overall, this work supports a model in which CGGBP1 protects G4-prone, GC-dense regulatory DNA from methylation-associated mutagenesis, thereby contributing to the maintenance of genomic integrity and regulatory complexity in higher vertebrates.

Methods

Mapping the G4 landscape in vertebrate genomes and their 1 kb cis promoters via pqsfinder

To assess G4-forming potential across vertebrate genomes, we analyzed 105 reference genomes, including 57 mammals, 27 birds, 17 reptiles, and 4 non-amniotes, sourced from Ensembl as more details in Supplementary Table S17 (Kumar et al., 2025). Genome FASTA files were systematically processed using pqsfinder, a G4 prediction tool uniquely capable of identifying not only canonical G-quadruplex-forming sequences but also those with imperfections, such as bulges, mismatches, and atypically long loops (Hon et al., 2017; Labudová et al., 2020). In contrast to previous algorithms that enforce the strict [G3N1-7]4 motif, pqsfinder accommodates the greater sequence diversity observed in experimentally validated G4 structures, including longer loops and non-standard spacers. Its high computational efficiency and parameterization using high-throughput G4-seq experimental datasets make it particularly suited for large-scale genomic analysis. We applied a recommended score threshold of 52 to prioritize reliability and specificity, effectively reducing false positives and aligning predictions with experimentally validated G4 motifs (Hon et al., 2017; Labudová et al., 2020). We applied the same analysis to all annotated promoters across vertebrate genomes. We supplied 1kb *cis* promoter sequence FASTA files for each species were retrieved using coordinates previously extracted from genome annotation files (GTF), as described elsewhere (Kumar et al., 2025).

Genome-wide sequence feature characterization of predicted G-quadruplex potential across vertebrate genomes

In addition to obtaining the raw output generated by pqsfinder—which includes coordinates, sequences, and scores of potential G-quadruplex (G4)-forming regions—we developed a customized pipeline to calculate comprehensive summary statistics and associated features for both the predicted G4-forming sequences and their respective input sequences. These metrics encompass genome size or promoter region size (defined as the total nucleotide length of all annotated promoters combined), GC-content of the entire genome and promoter regions, the total number of predicted G4 motifs within genomes and promoters, and the proportion of the genome or promoter sequences covered by these G4-forming regions. Furthermore, we computed the density of G4 motifs per megabase, average G4 motif length, average G4 prediction scores, average GC-content of the predicted G4 sequences, and the number of chromosomes or scaffolds per species. This statistical characterization was carried out both genome-wide and across all annotated promoters for each species analyzed. These summary statistics formed the basis for downstream analyses and enabled us to draw biologically meaningful conclusions supported by statistical significance testing (details in Supplementary Table S1 and Table S5).

Characterization of point mutations within potential G4-Forming sites in human genomic and promoter regions

Leveraging our recently published methylation-enriched DNA point mutation analysis pipeline (MAP-MAP) (see additional file 11 (Kumar et al., 2025)), we delved deep into the mutation landscape of potential G4-forming regions within the human genome. Our study focused on methylation-associated point mutation analysis in MeDIP DNA and the role of different CGGBP1 forms in modulating the mutability of potential G4-forming sequences. The primary goal was to determine whether these G4 motifs exhibit high mutation rates that vary according to CGGBP1 variants. We retrieved mutation profiles for predicted G4 regions across the genome, including entire promoters and potential G4 sequences therein. We also examined two experimental conditions—G4-IP-native (GSM6122749) and G4-IP-denatured (GSM6122750)—that reflect G4 formation in the presence and absence of nuclear constraints, respectively, with the denatured condition representing DNA disrupted by heat treatment (Datta et al., 2024). Mutation patterns in these regions were thoroughly plotted and analyzed. Further, we concentrated on differentially captured regions (DCRs) in GSE202456 (Datta et al., 2024), which are genomic loci strongly influenced by CGGBP1 levels and possess potential G4-forming capability. For these DCRs, we calculated mutation profiles with a specific focus on methylation-dependent C-to-T transitions and assessed how different CGGBP1 forms impact these mutational processes.

Heatmap and plotProfile generation

All heatmaps and profile plots were generated using deepTools. To visualize C-to-T transitions across the genome, and specifically within promoter regions and potential G4-forming regions of the human genome, we first produced normalized bigWig files for each MeDIP sample (Lc, Ac, Gg, Hs, Ev, and Input). These files represent the distribution of C-to-T transitions genome-wide, with each bigWig encoding the signal as the fraction of all transitions identified in each sample, normalized by subtracting the sample mean to account for global differences. These scaled bigWig files were subsequently used to plot C-to-T transition signals for each sample in 1kb windows upstream and downstream of predicted G4-forming regions in a bin size of 10 bp. Similarly, we plotted these scaled signals in relation to G4-IP-native (GSM6122749) and G4-IP-denatured (GSM6122750) samples, examining their distributions within 0.1Mb flanking intervals in a bin size of 1000 bp from each peak center. To further investigate the distribution of potential G4 signals in regions devoid of mutations, we generated additional scaled bigWig tracks in 10bp bins for mutation-free sites genome-wide, using the predicted G4 bed coordinates for each sample. For inter-regional comparisons, the number of mutation-free regions per million bases was normalized for each flank by dividing each value by the highest observed value in the dataset (after subtracting the Input signal), resulting in relative values ranging from 0 (minimum) to 1 (maximum). In parallel, we created normalized bigWig tracks for GC-skew (calculated as the modulus of GC skew) in 25bp bins genome-wide and plotted these signals in relation to mutation-free regions for each sample. The GC-skew signal was also assessed for enrichment by plotting the 25bp bin GC-skew modulus bigWig track in 0.1Mb flanking intervals in a bin size of 1000 bp around G4-IP-native and G4-IP-denatured peak centers. Finally, to study sequence features linked to potential G4-forming regions identified by pqsfinder, we plotted GC-skew signals in 1kb intervals both upstream and downstream of these coordinates. As a control, GC-skew signals were also plotted for G4-free regions, defined as genome-wide regions where potential G4-forming motifs were absent.

Statistical analysis

The proportions of transition and transversion mutations, calculated as fractions of total point mutation entries in each MeDIP sample within potential G4-forming regions, were found to be highly significant (P-value < 1×10□□) based on chi-square tests. In parallel, for each sample, methylated C-to-T (meC-T) transitions—expressed as a fraction of total transitions—were also highly significant (P-value < 1×10□□) upon application of the same statistical test. These statistical comparisons were performed across three genomic contexts within the human genome: potential G4-forming regions genome-wide, predicted G4 motifs located within 1kb promoter regions, and the full 1kb promoters themselves. This was possible because MeDIP data corresponding to various CGGBP1 forms were specifically available for the human genome.

To assess C-to-T transition rates genome-wide within potential G4-forming regions, we calculated the proportion of C-to-T transitions among all transition events for each sample and subtracted the average value, representing deviations from the mean. To illustrate differences in point mutation rates across samples, each point mutation type was expressed as a percentage of total point mutations within the respective sample; this percentage was then normalized by subtracting the corresponding value obtained from the Input sample. This approach allowed a clear demonstration of methylation-associated point mutation rates in potential G4-forming regions located within 1kb annotated promoter regions of the human genome.

Finally, analysis revealed that the percentage of point mutations occurring in potential G4-forming regions within 1kb promoter intervals, relative to the total point mutation count per sample, varied significantly among samples (P-value < 1×10□□, chi-square test), further supporting the specificity and reproducibility of these findings.

Code availability

Custom scripts and code utilized in this study are publicly accessible at the following GitHub repository: https://github.com/kb-praveen/MAP-MAP-and-G4-Landscape.

Limitations of the study

1. A direct cause-effect relationship between G4 formation and evolutionary forms of CGGBP1 needs targeted experiments.
2. The time-series experiments, through which cytosine methylation and meC-T transition restrictions effects of various CGGBP1 forms at high G4-forming potential regions have been deduced, shall be extended to *in-cellulo* G4 landscape characterization experimentally.

Author contributions

PK performed all the analyses, generated the data, and participated in manuscript writing. US provided supervision and participated in writing the manuscript.

Statements and Declarations

The authors declare no conflict of interest.

Acknowledgments

This work was carried out with funds from ANRF (erstwhile SERB) [CRG/2021/000375] to US with additional support from IIT Gandhinagar. PK was supported by the MHRD fellowship**.**

Figure legends:

Figure 1: Selective GC-retention bias differs between genomic and promoter G4-forming sequences in vertebrates. (a) G4 density per megabase (Mb) of vertebrate genomes plotted against genomic GC-content shows a significant positive correlation, as revealed by simple linear regression ($P < 0.01$; see Supplementary Table S2 for full statistics). (b) The GC-content of genome-wide potential G4-forming sequences increases with genomic GC-content only in mammals ($P < 0.01$; Supplementary Table S3), but not in aves, reptiles, or non-amniotes. This is depicted by linear regression of genome GC-content against the mean GC-content of all predicted G4s in each species, grouped by taxon. (c) Length of potential G4-forming regions genome-wide is positively correlated with G4 GC-content in mammals, but shows no significant relationship in non-mammalian groups, as indicated by simple linear regression ($P < 0.01$; Supplementary Table S4). (d) G4 density (per Mb) in 1 kb promoters of vertebrate genomes positively correlates with promoter GC-content across all vertebrate taxonomic groups, according to simple linear regression ($P < 0.01$; Supplementary Table S6). (e) The GC-content of potential G4-forming sequences within 1 kb promoters increases significantly with promoter GC-content in all taxon groups (non-amniotes, reptiles, aves, and mammals), as shown by linear regression ($P < 0.01$; Supplementary Table S7). (f) In 1 kb promoters, the GC-content of predicted G4-forming sequences increases significantly with G4 length in aves and mammals (homeotherms), but not in reptiles or non-amniotes (poikilotherms), based on linear regression analysis ($P < 0.01$; Supplementary Table S8). For each taxon group, G4 density (per Mb), G4 GC-content, and G4 length are represented as the mean for each species. Non-amniotes and reptiles (green circles and green regression line) are plotted as one cluster (poikilotherms, n=21), while aves (purple diamonds and line; n=27) and mammals (orange squares and line; n=57) are shown as separate groups (homeotherms). All regressions include means with error bars representing 95% confidence intervals.

Figure 2: CGGBP1 evolutionary forms regulate the restriction of meC-T transition in methylation-enriched, G4-forming genomic regions, distinguishing between poikilothermic and homeothermic species. (a): Genome-wide analysis of potential G4-forming regions, defined by 1 kb flanks around G4 coordinates, reveals that meC-T transition rates are restricted in a CGGBP1 form-dependent manner. Samples expressing human (Hs), avian (Gg) forms of CGGBP1, and empty vector (Ev)

representing endogenous CGGBP1, as well as the input control, exhibit lower meC-T transition rates, with Hs showing the strongest restriction (lowest transition rate). In contrast, poikilothermic forms (Lc and Ac) display positive meC-T transition rates, indicating less restriction. The plotted data show genome-wide meC-T transition signals for all samples (Lc, Ac, Gg, Hs, Ev, and Input) across potential G4-forming regions. (b): Potential G4-forming motifs within 1 kb promoters of the human genome also exhibit CGGBP1 form-dependent restriction of meC-T transitions. Human CGGBP1 is the most effective at limiting methyl-cytosine to thymine mutation rates, mirroring the genome-wide pattern, whereas Lc and Ac CGGBP1 forms do not achieve similar restriction. Gg and Ev reduce transition rates to a degree, but both still maintain a higher mutation rate than the input sample, which reflects the baseline background mutation rate in non-methylated DNA. (c): meC-T transition rates as a fraction of all point mutations in 1 kb gene promoters and in predicted G4-forming regions (details in Supplementary Table S15) show significant differences depending upon evolutionary forms of CGGBP1, as indicated by chi-square testing ($P < 0.01$; full details in Supplementary Table S16). G4-forming regions within promoters display differential meC-T transition rates as compared to background promoters, and both endogenous and overexpressed human CGGBP1 demonstrate strong restriction of meC-T transitions relative to other forms. Gg and Ev moderately restrict mutation rates, while poikilothermic forms (Lc and Ac) fail to do so. Hs CGGBP1 maintains the lowest meC-T rates in both G4-forming regions and background promoters, though meC-T rates are still marginally but significantly higher in G4-forming regions (see Supplementary Table S17), whereas Lc shows higher meC-T in promoters versus corresponding G4-forming regions (see Supplementary Table S18). Ac, Gg, and Ev do not show statistically significant differences in promoters, though for Gg and Ev, meC-T appears higher in G4-forming regions, opposite to Ac (see Supplementary Tables S19-S21). To highlight methylation-specific changes, input as background was subtracted from both promoters and G4-forming regions, and these normalized meC-T transition rates are shown as bar graphs: Lc (Blue), Ac (Green), Gg (Red), Hs (Yellow), and Ev (Grey), with solid fill for promoters and lined fill for potential G4-forming regions.

Figure 3: CGGBP1 maintains GC-content in DNA regions known to form G-quadruplexes by restricting methylation-associated C-to-T transitions. (a): Regions capable of forming G-quadruplexes within the chromatin-constrained nuclear environment (G4-IP-Nat; GSM6122749) show sensitivity to CGGBP1-dependent methylation-associated C-to-T transitions. Human CGGBP1 (Hs) strongly restricts these meC-T transitions, while empty vector control (Ev) and avian (Gg) forms exhibit slightly higher transition rates compared to input (background). Poikilothermic forms (Lc and Ac) display elevated meC-T transition rates in these verified G4-forming regions. (b): Potential G-quadruplex regions identified from heat-denatured genomic DNA, which lacks chromatin constraints and double-stranded structure (G4-IP-Denat; GSM6122750), do not associate with methylation patterns. No central enrichment or CGGBP1-dependent differences in meC-T transitions are observed in these regions. C-to-T transition signals, calculated as the fraction of all transition mutations with average background subtraction, are plotted ±0.1 Mb around peak centers for both G4-IP-Nat and G4-IP-Denat samples. (c): Differentially captured regions (DCRs), experimentally validated G4-forming loci highly sensitive to CGGBP1 levels (GSE202456), display species-specific meC-T transition patterns distinguishing poikilotherms (Lc, Ac) from homeotherms (Hs, Gg, Ev). Poikilotherms show poor restriction of meC-T transitions, whereas homeotherms, particularly Hs CGGBP1, strongly restrict these transitions. The empty vector control (Ev) represents endogenous CGGBP1 and displays intermediate restriction, with Gg showing moderate restriction relative to background input. Transition rates were calculated as the fraction of C-to-T mutations among all transitions, with averages subtracted per sample. Differences between groups are statistically significant (Chi-square test, $P < 0.01$; Supplementary Table S23).

Figure 4: A schematic of the mechanism of G4-forming potential regulation by CGGBP1 in vertebrate genomes. CGGBP1 restricts meC-to-T transition on the G-poor and C-rich strand. Through replicative mechanisms this C-retention leads to G-retention on the complementary G4-forming strand. The C-retention caused by cytosine methylation restriction imposed by CGGBP1 from higher vertebrates leads to higher G4-forming potential. This G4-forming potential enhancing effect of CGGBP1 has evolved in higher amniotes alongside its function of mitigating G4-formation. Higher amniotic CGGBP1 thus links cytosine methylation, GC-retention and G4-forming potential without increasing the risk of inordinate G4 formation.

Supplementary Figure legends:

Figure S1: Genome-wide point mutation rates in potential G4-forming regions show evolution-dependent enhancement favoring GC retention by CGGBP1. Point mutation rates were determined as the percentage of each mutation type out of the total mutations for each sample, adjusted by subtracting the input control background. Different mutation types are illustrated as bar graphs. The analysis reveals an increased mutation rate in Hs, reflecting a genome-wide bias toward GC-content preservation in potential G4-forming regions. Bars are color-coded by sample: Lc (blue), Ac (green), Gg (red), Hs (yellow), and Ev (grey). All mutation type proportions relative to total mutations are statistically significant (Chi-square test, $P < 0.01$; Supplementary Table S10).

Figure S2: Potential G4-forming regions are distinguished by their notable GC asymmetry. (a): G4-free regions lack central enrichment of GC-skew within their ±1 kb flanks. (b): Predicted G4s in promoters exhibit pronounced central GC-skew enrichment. (c): Potential G4-forming regions show strong G/C asymmetry enrichment. GC-skew signals were plotted across ±1 kb flanks around the center of G4-free regions, promoter-associated predicted G4s, and genome-wide potential G4-forming regions. Genome-wide potential G4-forming regions display higher GC-skew than those found specifically in promoters.

Figure S3: Human CGGBP1 maintains elevated G/C asymmetry and G4-forming potential in mutation-free regions. Scaled and normalized counts of potential G4-forming regions were plotted ±1 kb around one million randomly selected mutation-free regions across samples. Hs exhibits higher genome-wide G4-forming potential compared to Lc, Ac, Gg, and Ev. Similarly, GC-skew signals, also scaled and normalized by the number of mutation-free regions per sample, were plotted around the same set of regions. Regions with higher associated G4-forming potential display greater G/C asymmetry, with Hs maintaining significantly higher genome-wide G/C asymmetry than the other samples.

Figure S4: CGGBP1 drives evolutionary-dependent GC bias in point mutations within promoters and associated potential G4-forming regions. (a) & (b): Point mutation rates were calculated as the percentage of each mutation type relative to the total mutations within each sample, with background correction applied by subtracting the input control. Different mutation types are shown as bar graphs. Overall, the analysis reveals an elevated mutation rate in Hs, reflecting a bias toward preserving GC content in both potential G4-forming (b) regions and background promoters (a). Bars are color-coded by sample: Lc (blue), Ac (green), Gg (red), Hs (yellow), and Ev (grey). All mutation type proportions relative to total point mutations are statistically significant (Chi-square test, $P < 0.01$; Supplementary Table S13 for promoters and Table S14 for potential G4s in promoters).

Figure S5: Chromatin-constrained G4-forming regions are defined by pronounced G/C asymmetry. G4-forming regions within the chromatin-constrained nuclear environment (G4-IP-Nat, GSM6122749) exhibit pronounced GC-skew enrichment, flanked by broader areas of low GC-skew. In contrast, G4-IP-Denat regions (GSM6122750) display only background genomic GC-skew levels without any central enrichment. These patterns are illustrated by heat maps generated from ±0.1 Mb flanks around the peak centers of each region.

Supplementary Table legends:

Table S1: Summary of G4-forming potential and associated features of 105 vertebrate genomes
Table S2: Simple linear regression of Genomic GC-content vs G4 per Mb between different vertebrate taxa
Table S3: Simple linear regression of Genomic GC-content vs G4 GC-content between different vertebrate taxa
Table S4: Simple linear regression of Genomic G4 length vs G4 GC-content between different vertebrate taxa
Table S5: Summary of G4-forming potential and associated features of 105 vertebrate genomes 1kb *cis* promoters
Table S6: Simple linear regression of 1kb *cis* promoters GC-content vs G4 per Mb in between different vertebrate taxa
Table S7: Simple linear regression of 1kb *cis* promoters GC-content vs G4 GC-content between different vertebrate taxa
Table S8: Simple linear regression of 1kb *cis* promoters G4 length vs G4 GC-content between different vertebrate taxa
Table S9: Summary of point mutation in potential G4-forming regions genome-wide in MeDIP data after over-expression of evolutionary forms of CGGBP1
Table S10: Summary of Chi-square test of individual point mutations in potential G4-forming regions genomewide, where each mutation is calculated as a fraction of all point mutations occurring in G4s
Table S11: Summary of point mutation in potential G4-forming regions found in all annotated 1kb *cis* promoters in MeDIP data after over-expression of evolutionary forms of CGGBP1
Table S12: Summary of point mutation in all annotated 1kb *cis* promoters in MeDIP data after over-expression of evolutionary forms of CGGBP1
Table S13: Summary of Chi-square test of individual point mutations in 1kb *cis* promoters, where each mutation is calculated as a fraction of all point mutations occurring in 1kb *cis* promoters overall
Table S14: Summary of Chi-square test of individual point mutations in potential G4-forming regions in 1kb *cis* promoters, where each mutation is calculated as a fraction of all point mutations occurring in G4s
Table S15: Summary of meC-T in 1kb *cis* promoters and potential G4s found in them
Table S16: Summary of Chi-square test for meC-T in 1kb *cis* promoters and G4s found in them
Table S17: Summary of Chi-square test for meC-T in promoters and G4s in promoters after over-expression of Hs CGGBP1
Table S18: Summary of Chi-square test for meC-T in promoters and G4s in promoters after over-expression of Lc CGGBP
Table S19: Summary of Chi-square test for meC-T in promoters and G4s in promoters after over-expression of Ac CGGBP1
Table S20: Summary of Chi-square test for meC-T in promoters and G4s in promoters after over-expression of Gg CGGBP1

Table S21: Summary of Chi-square test for meC-T in promoters and G4s in promoters after endogenous level of CGGBP1 represented by Ev
Table S22: Summary of meC-T transition and overall transition in Differentially Captured Regions in MeDIP samples
Table S23: Summary of Chi-square for C-T transition vs all transitions found in DCRs in MeDIP samples

References:

Agarwal, P., Collier, P., Fritz, M. H.-Y., Benes, V., Wiklund, H. J., Westermark, B., & Singh, U. (2015). CGGBP1 mitigates cytosine methylation at repetitive DNA sequences. *BMC Genomics*, *16*(1), 390.

Balasubramanian, S., Hurley, L. H., & Neidle, S. (2011). Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nature Reviews. Drug Discovery*, *10*(4), 261–275.

Baubec, T., & Schübeler, D. (2014). Genomic patterns and context specific interpretation of DNA methylation. *Current Opinion in Genetics & Development*, *25*, 85–92.

Datta, S., Patel, M., Sathyaseelan, C., Ghosh, C., Mudgal, A., Patel, D., Rathinavelan, T., & Singh, U. (2024). G-quadruplex landscape and its regulation revealed by a new antibody capture method. *Oncotarget*, *15*, 175–198.

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, *25*(10), 1010–1022.

Du, Z., Zhao, Y., & Li, N. (2008). Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. *Genome Research*, *18*(2), 233–241.

Hartono, S. R., Korf, I. F., & Chédin, F. (2015). GC skew is a conserved property of unmethylated CpG island promoters across vertebrates. *Nucleic Acids Research*, *43*(20), 9729–9741.

Hon, J., Martínek, T., Zendulka, J., & Lexa, M. (2017). pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics (Oxford, England)*, *33*(21), 3373–3379.

Huppert, J. L., & Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. *Nucleic Acids Research*, *33*(9), 2908–2916.

Johnson, F. B. (2020). Fundamentals of G-quadruplex biology. *Annual Reports in Medicinal*

Chemistry, *54*, 3–44.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews. Genetics*, *13*(7), 484–492.

Khuu, P., Sandor, M., DeYoung, J., & Ho, P. S. (2007). Phylogenomic analysis of the emergence of GC-rich transcription elements. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(42), 16528–16533.

Kumar, P., Morbia, I., Satish, A. L., Datta, S., & Singh, U. (2025). CGGBP1 from higher amniotes restricts cytosine methylation and drives a GC-bias in transcription factor-binding sites at repressed promoters. *Transcription*, 1–36.

Labudová, D., Hon, J., & Lexa, M. (2020). pqsfinder web: G-quadruplex prediction using optimized pqsfinder algorithm. *Bioinformatics (Oxford, England)*, *36*(8), 2584–2586.

Lin, J., Hou, J.-Q., Xiang, H.-D., Yan, Y.-Y., Gu, Y.-C., Tan, J.-H., Li, D., Gu, L.-Q., Ou, T.-M., & Huang, Z.-S. (2013). Stabilization of G-quadruplex DNA by C-5-methyl-cytosine in bcl-2 promoter: implications for epigenetic regulation. *Biochemical and Biophysical Research Communications*, *433*(4), 368–373.

Maizels, N. (2012). G4 motifs in human genes. *Annals of the New York Academy of Sciences*, *1267*, 53–60.

Mao, S.-Q., Ghanbarian, A. T., Spiegel, J., Martínez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hänsel-Hertsch, R., Tannahill, D., & Balasubramanian, S. (2018). DNA G-quadruplex structures mold the DNA methylome. *Nature Structural & Molecular Biology*, *25*(10), 951–957.

Mendoza, O., Bourdoncle, A., Boulé, J.-B., Brosh, R. M., Jr, & Mergny, J.-L. (2016). G-quadruplexes and helicases. *Nucleic Acids Research*, *44*(5), 1989–2006.

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology : Official Publication of the American College of Neuropsychopharmacology*, *38*(1), 23–38.

Morbia, I., Kumar, P., Satish, A. L., Mudgal, A., Datta, S., & Singh, U. (2024). The amniote-conserved DNA-binding domain of CGGBP1 restricts cytosine methylation of transcription factor binding sites in proximal promoters to regulate gene expression. *BMC Genomic Data*, *25*(1), 98.

Neupane, A., Chariker, J. H., & Rouchka, E. C. (2023). Analysis of Nucleotide Variations in Human G-Quadruplex Forming Regions Associated with Disease States. *Genes*, *14*(12). https://doi.org/10.3390/genes14122125

Niu, K., Xiang, L., Zhang, X., Li, X., Yao, T., Li, J., Zhang, C., Liu, J., Peng, Y., Xu, G., Xiang, H., Wang, H., Song, Q., & Feng, Q. (2025). DNA 5mC methylation inhibits the formation of G-quadruplex structures in the genome. *Genome Biology*, *26*(1), 202.

Patel, D., Patel, M., Datta, S., & Singh, U. (2019). CGGBP1 regulates CTCF occupancy at repeats. *Epigenetics & Chromatin*, *12*(1), 57.

Singh, U., & Westermark, B. (2015). CGGBP1--an indispensable protein with ubiquitous cytoprotective functions. *Upsala Journal of Medical Sciences*, *120*(4), 219–232.

Ummethum, H., Murriello, A. C., Werner, M., Márquez-Gómez, E., König, A.-C., Kruse, E., Lalonde, M., Trauner, M., Chanou, A., Weiβ, M., Lee, C. S. K., Ettinger, A., Erhard, F., Hauck, S. M., & Hamperl, S. (2025). The CGG triplet repeat binding protein 1 counteracts R-loop induced transcription-replication stress. *EMBO Reports*. https://doi.org/10.1038/s44319-025-00550-1

Varriale, A., & Bernardi, G. (2006). DNA methylation and body temperature in fishes. *Gene*, *385*, 111–121.

Varshney, D., Spiegel, J., Zyner, K., Tannahill, D., & Balasubramanian, S. (2020). The regulation and functions of DNA and RNA G-quadruplexes. *Nature Reviews. Molecular Cell Biology*, *21*(8), 459–474.

Wang, Z.-F., Li, M.-H., Chu, I.-T., Winnerdy, F. R., Phan, A. T., & Chang, T.-C. (2020). Cytosine epigenetic modification modulates the formation of an unprecedented G4 structure in the WNT1 promoter. *Nucleic Acids Research*, *48*(3), 1120–1130.

Wu, F., Niu, K., Cui, Y., Li, C., Lyu, M., Ren, Y., Chen, Y., Deng, H., Huang, L., Zheng, S., Liu, L., Wang, J., Song, Q., Xiang, H., & Feng, Q. (2021). Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. *Communications*

Biology, *4*(1), 98.

Wu, W., Rokutanda, N., Takeuchi, J., Lai, Y., Maruyama, R., Togashi, Y., Nishikawa, H., Arai, N., Miyoshi, Y., Suzuki, N., Saeki, Y., Tanaka, K., & Ohta, T. (2018). HERC2 Facilitates BLM and WRN Helicase Complex Interaction with RPA to Suppress G-Quadruplex DNA. *Cancer Research*, *78*(22), 6371–6385.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C., & Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (New York, N.Y.)*, *356*(6337). https://doi.org/10.1126/science.aaj2239

Figure 1

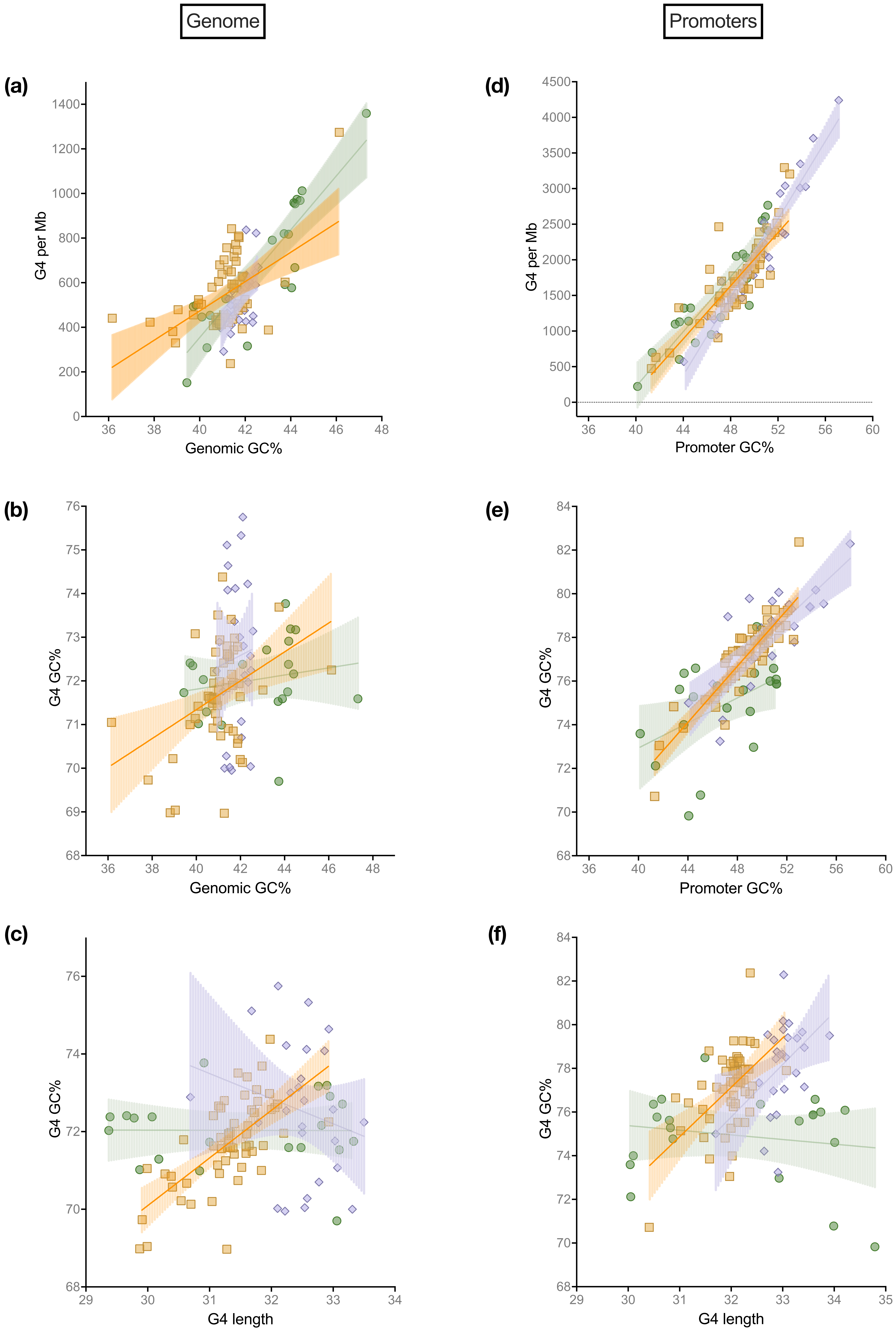

Genome
Promoters
(a)
G4 per Mb
0
200
400
600
800
1000
1200
1400
36
38
40
42
44
46
48
Genomic GC%
(d)
G4 per Mb
0
500
1000
1500
2000
2500
3000
3500
4000
4500
36
40
44
48
52
56
60
Promoter GC%
(b)
G4 GC%
68
69
70
71
72
73
74
75
76
36
38
40
42
44
46
48
Genomic GC%
(e)
G4 GC%
68
70
72
74
76
78
80
82
84
36
40
44
48
52
56
60
Promoter GC%
(c)
G4 GC%
68
70
72
74
76
29
30
31
32
33
34
G4 length
(f)
G4 GC%
68
70
72
74
76
78
80
82
84
29
30
31
32
33
34
35
G4 length

Figure 2

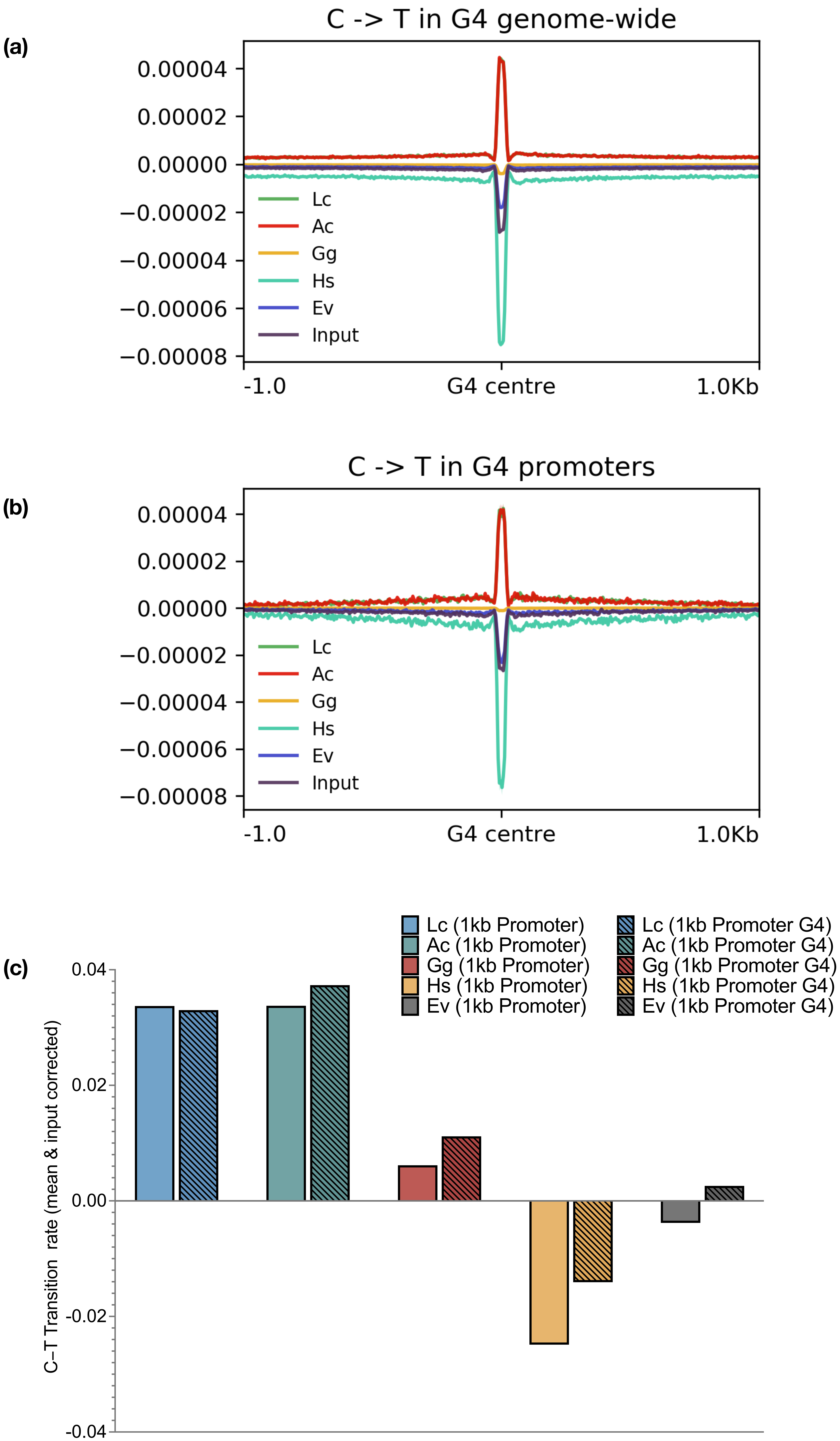
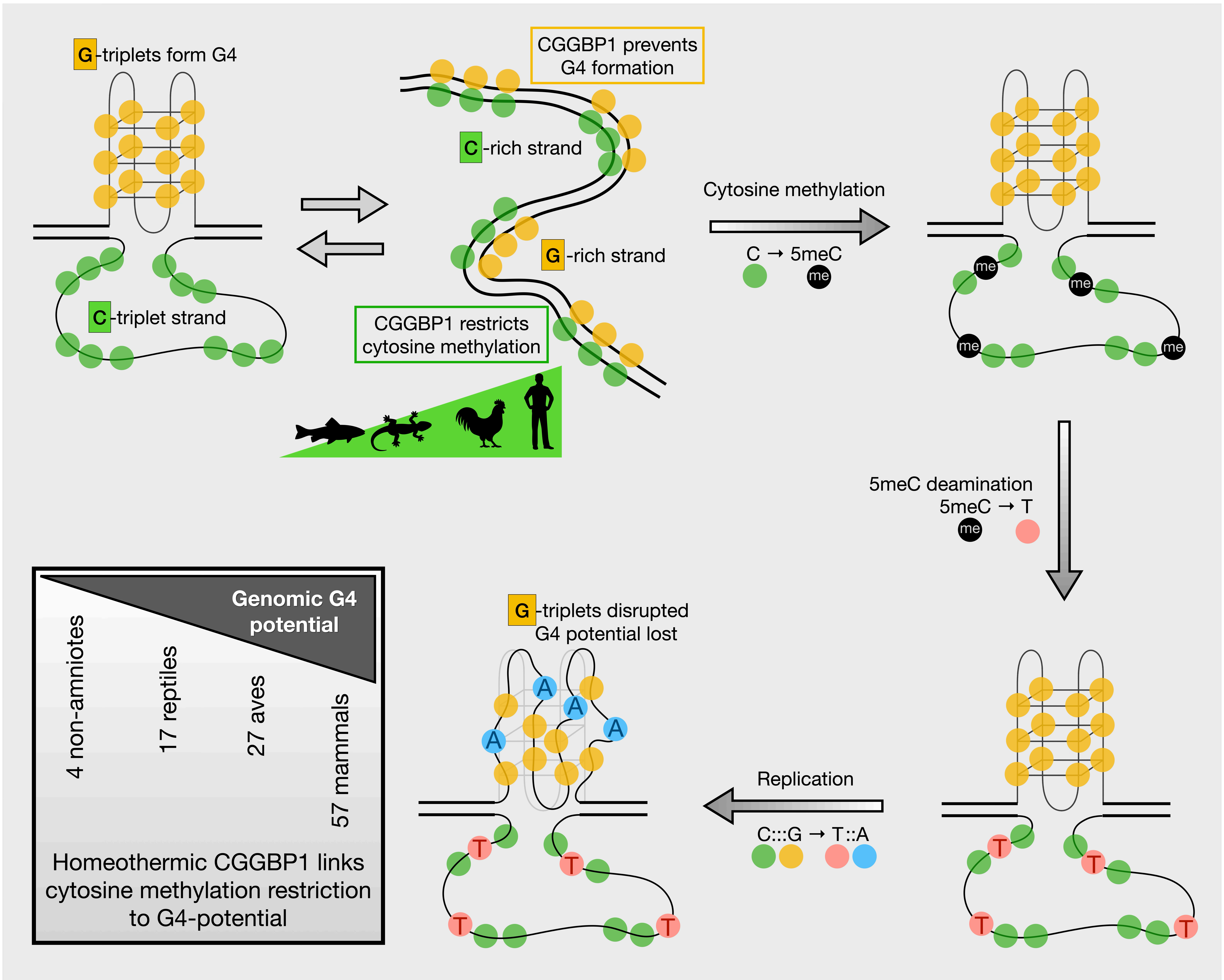

(a)
C -> T in G4 genome-wide
0.00004
0.00002
0.00000
−0.00002
−0.00004
−0.00006
−0.00008
Lc
Ac
Gg
Hs
Ev
Input
-1.0
G4 centre
1.0Kb
(b)
C -> T in G4 promoters
0.00004
0.00002
0.00000
−0.00002
−0.00004
−0.00006
−0.00008
Lc
Ac
Gg
Hs
Ev
Input
-1.0
G4 centre
1.0Kb
(c)
Lc (1kb Promoter)
Ac (1kb Promoter)
Gg (1kb Promoter)
Hs (1kb Promoter)
Ev (1kb Promoter)
Lc (1kb Promoter G4)
Ac (1kb Promoter G4)
Gg (1kb Promoter G4)
Hs (1kb Promoter G4)
Ev (1kb Promoter G4)
C–T Transition rate (mean & input corrected)
0.04
0.02
0.00
-0.02
-0.04

Figure 3

(a)
G4 IP native
0.000015
0.000010
0.000005
0.000000
−0.000005
−0.000010
−0.000015
-0.1
Peak centre
0.1Mb
Lc
Ac
Gg
Hs
Ev
Input
(b)
G4 IP Denaturation
0.000015
0.000010
0.000005
0.000000
−0.000005
−0.000010
−0.000015
-0.1
Peak centre
0.1Mb
Lc
Ac
Gg
Hs
Ev
Input
(c)
DCRs C -> T
C–T Transition rate (sample-mean)
0.04
0.02
0.00
-0.02
-0.04
0.0298
0.0290
-0.0025
-0.0317
-0.0124
-0.0122
Lc
Ac
Gg
Hs
Ev
Input

Figure 4


G-triplets form G4
C-triplet strand
CGGBP1 prevents G4 formation
C-rich strand
G-rich strand
CGGBP1 restricts cytosine methylation
Cytosine methylation
C → 5meC
me
5meC deamination
5meC → T
me
T
Genomic G4 potential
4 non-amniotes
17 reptiles
27 aves
57 mammals
Homeothermic CGGBP1 links cytosine methylation restriction to G4-potential
G-triplets disrupted
G4 potential lost
A
Replication
C:::G → T::A