



Roles of transposable elements and DNA methylation in the formation of CpG islands and CpG-depleted regulatory elements

Jueng Soo You^{a,1} , Steve Pierce^{b,1}, Gangning Liang^c , and Peter A. Jones^{b,2}

Affiliations are included on p. 8.

Contributed by Peter A. Jones; received February 7, 2025; accepted September 18, 2025; reviewed by Bradley E. Bernstein and Ting Wang

The origins of CpG islands (CGIs) are not known. They are relatively short GC-rich regions of DNA with a higher-than-expected occurrence of CpG dinucleotides compared to most of the genome. They constitute less than 1% of the human genome but harbor approximately 40% of all transcription start sites (TSSs). CGIs are usually modulated by histone modifications in somatic cells or, in a minority of cases, permanently silenced by CpG methylation. Those that do not have TSSs are called “orphan CGIs”. Here, we show that CGIs containing TSSs almost never contain any of three major classes of transposable elements (TEs) and orphan CGIs rarely do. We hypothesize that CGIs persist across evolutionary time due to counterselection against TE insertion in the germ line. The 99% of the vertebrate genome, which is not CG rich, contains 60% of TSSs and putative enhancers. We postulate that conversion of an ancestral CpG-rich genome into the current CpG-depleted version present in vertebrates may also have allowed reversible DNA methylation to function in complex and dynamic gene control circuits. Therefore, we propose an evolutionary model in which vertebrate TEs are indirectly responsible for the existence of CGIs, and the formation of regulatory elements such as TSSs and enhancers that can potentially utilize dynamic DNA methylation for gene control.

CpG islands (CGIs) | transposable elements (TEs) | transcription start sites (TSSs) | DNA methylation

In a previous study of 53 eukaryotic species (1), we found a direct correlation between genome size and TE content and an inverse correlation with the Observed/Expected (O/E) CpG content. We argued that the original evolutionary benefit of DNA methylation was the silencing of TEs, allowing for genome expansion. This, in turn, led to depletion of the CpG sequence due to the spreading of DNA methylation into host DNA in the germ line (1). Consistent with this hypothesis is the fact that methylated CpG dinucleotides show an increased rate of C-T transition mutations relative to unmethylated sequences (2). Also, organisms with compact genomes, such as *Caenorhabditis elegans* and *Drosophila melanogaster*, have a lower proportion of TEs, little or no measurable DNA methylation, less CpG suppression, and few if any CGIs. Among other processes, developmental programs in these organisms rely on histone variants and modifications, such as those applied by the Polycomb Repressive Complex (PRC), to enable epigenetic regulation (3).

In contrast, vertebrate genomes are characterized by their large sizes, the occurrence of widespread DNA cytosine methylation, mostly in the CpG dinucleotide sequence, and the presence of a substantial proportion of TEs (1, 4, 5). In mammals, TEs are mainly represented by long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs) (6, 7). As will be discussed later, components of evolutionary young versions of these elements fulfill the algorithmic criteria used to identify CGIs, complicating the distinction between host and TE.

TEs exert a wide range of functions; they can be sources of genetic variation or be actively involved in the regulation of the host genome, through processes that can occur in combination with epigenetic mechanisms (8–11). The most abundant SINEs in the human genome are *Alu*S elements, which are primate-specific and account for about 10% of the genome. TEs have been shown to have the ability to mediate interactions between transcription factors (TFs) and regulatory elements (8, 12, 13). However, active transpositions or genomic rearrangements mediated by TEs can be inherently deleterious. For instance, TE insertions are associated with tumor evolution in somatic cells as well as with potentially lethal mutations in the germline (14, 15). Consequently, epigenetic mechanisms to suppress TE activity are essential (16–19).

DNA methylation silences TEs in the germline, protecting the complex host gene expression programs during embryogenesis (20, 21). Additionally, as mentioned previously, the deamination of 5-methylcytosine (5mC) to thymine (T) leads to C-to-T mutations, further contributing to eventual loss of their transcriptional potentials (1, 22–24). While

Significance

Transposable elements (TEs) constitute about half of the vertebrate genome. Previously, we provided evidence that TE silencing by CpG methylation in the germ line and subsequent deamination resulted in strong CpG depletion not only in the TEs but also in the host genome. Present day vertebrate genomes are CpG poor except for the 1% of the genome represented by CpG islands (CGIs). We propose a theory that TEs are indirectly responsible for the existence of CGIs through selection bias. Additionally, the dilution of the CpG sequence in the majority of host DNA allowed for the evolution of regulatory elements (REs). These REs utilize reversible DNA methylation for gene control and include tissue-specific transcription start sites and enhancers.

Author contributions: J.S.Y., G.L., and P.A.J. designed research; J.S.Y. and S.P. analyzed data and performed experiments; and J.S.Y., S.P., and P.A.J. wrote the paper.

Reviewers: B.E.B., Harvard Medical School; and T.W., Washington University.

Competing interest statement: P.A.J. is consultant for Zymo Research Corporation and Gangning Liang is consultant for Pangea Laboratory. No competing interests were disclosed by the other authors.

Copyright © 2025 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

¹J.S.Y. and S.P. contributed equally to this work.

²To whom correspondence may be addressed. Email: Peter.Jones@vai.org.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2502963122/-/DCSupplemental>.

Published October 24, 2025.

recent findings suggest that replication errors, potentially involving polymerase ϵ (Pol ϵ), may also contribute to the generation of C-to-T transition mutations (25), 5mC deamination appears to be the primary cause of such mutations in the human genome (26).

CpG is the only dinucleotide strongly depleted in the vertebrate genome with the notable exception being the CpG islands (CGIs). These short CpG enriched regions represent less than 1% of the total genomic DNA (27) but are associated with the transcription start sites (TSSs), as a critical landmark for the promoters, of about 40% of all vertebrate genes (28–30). However, little is known about their origins, particularly because the interplay between CGIs and GC-rich TEs is difficult to disentangle. Here, we hypothesize that CGIs persist because they resist or avoid TE insertion across evolutionary time, thereby remaining unmethylated in the germ line. Their conservation is favored because they often function as promoters of essential genes (31–34).

While CGIs have remained unmethylated in germlines, their roles in somatic cells are frequently misunderstood. Somatic methylation of CGIs is not frequent and not involved in the reversible control of gene expression. Rather, it is used as a long-term silencing strategy for TEs and a minority of cellular genes such as those on the inactive X-chromosome, germ cell-specific, and imprinted genes (35). Reversible regulation of CGIs is instead achieved by chromatin modifications analogous to the PRC mechanisms in organisms lacking either DNA methylation or recognizable CGIs (3). This suggests that chromatin modifications other than DNA methylation evolved early in evolution.

In this study, we expand on our earlier work (1) by also examining the effects of CpG depletion on the remaining 99% of the genome which contains most regulatory elements (REs) including 60% of the TSSs and putative enhancers. We further explore the interrelationships between TEs and CpG-poor TSSs. These TSSs tend to have TEs in closer proximity than those located in CGIs. They are also more frequently associated with tissue-specific genes (33, 36, 37), and they more often contain TEs at the TSS. An important outcome of the generation of CpG poor functional elements may therefore be the addition of an additional reversible chromatin mark not available to organisms with more compact genomes and no cytosine methylation. Overall, we propose a model in which the occurrence of TEs in vertebrate DNA has resulted in the preservation of CGIs derived from the ancestral genome. Invasion of TEs into nonessential regions may also have led to the evolution of new REs including TSSs and enhancers. The erosion of high densities of the CpG sequence might therefore have unlocked the potential for DNA methylation to participate in gene modulation rather than simply participating in permanent silencing for CGIs and present-day TEs.

Results

CGIs Are Enriched for TSSs and Depleted of TEs. The definition of CGIs was first formulated by Gardiner-Garden and Frommer (31) before accurate mapping of TEs was possible (31). They defined CGIs as being at least 200 bp in length with a G + C content >50% and an (O/E) CpG ratio of greater or equal to 0.6. This definition is often still used, but it includes evolutionary young SINEs, and the 5'UTRs of LINES and ERVs (LTRs) which are about 300 to 500 bp in length and easily fulfill these algorithmic criteria (38) (*SI Appendix, Table S1*). Earlier, we proposed a more stringent definition to help exclude young TEs from the data given that we are interested in the host DNA and its ability to act as a TSS for cellular genes (33) and TE can initiate CpG loss in host DNA (1). In the current paper, we used a cutoff for length of 400 bp rather

than 200 bp, given that the interval of 200 to 400 bp is largely made up of TEs rather than host TSSs (*SI Appendix, Fig. S1*). We also assessed directly the percentages of TE-derived DNA fragments in the CGIs we examined, to make sure we were not excluding small remnants of ancient TEs in our analysis.

We first examined the colocalization of human CGIs and TSSs and found, as expected, that TSSs are preferentially located within CGIs (Fig. 1A). Given that neither CGIs nor TSSs are randomly distributed across the genome, we used a permutation test to randomly select equal-sized intervals within 1 Mb of actual CGI locations. We identified 21,586 CGIs using these criteria. TSSs are enriched in CGIs, as almost 60% of CGIs have TSSs, compared to the 3.4% expected if distributions were random (Fig. 1B). We then examined the co-occurrence of TEs with CGIs and found that TE presence within CGIs is far less abundant than in the surrounding DNA (Fig. 1C). Note that the content of TEs in CGIs is less than it appears visually due to the fact that the pixels are proportionally larger than the actual densities. Quantitation of the actual contributions of each TE-derived DNA to the total DNA content in the CGIs revealed that SINEs account for 2.5%, LINES 0.9%, and LTRs 1.3% of the sequences in these regions (Fig. 1D). This is significantly less than that found in the surrounding 100 kb of DNA on both sides of the CGIs (Fig. 1D). Fig. 1D also shows that SINEs are more prevalent than expected in the vicinity of these actively transcribed genes compared to the total SINE content of human DNA. The reason for the enrichment of SINEs in euchromatin is not known but was first recognized some time ago (39).

Next, we examined the mouse genome and found 12,399 CGIs with an even higher enrichment of TSSs (*SI Appendix, Fig. S2A*)—almost 78% of CGIs have TSSs (*SI Appendix, Fig. S2B*). We also observed that both SINEs and LTRs were depleted in CGIs (*SI Appendix, Fig. S2 C and D*). Surprisingly, LINES were enriched in CGIs relative to surrounding DNA (*SI Appendix, Fig. S2D*). These observations align with previous reports in which LINE elements were shown to be more active in mice than in humans, where they play a prominent role in shaping the genome (40–42). Therefore, in both mouse and human, TEs are generally but not always depleted from CGIs, suggesting a conserved evolutionary constraint for TE insertions into CGIs.

CGIs Differ Depending on TSS Presence. The exclusion of TEs from human and, to a lesser extent, mouse CGIs could have at least two explanations. The higher prevalence of CpG could be inherently resistant to TE insertion. Alternatively, the insertion of TEs in the germline might be negatively selected against, leading to a survivorship bias in the examined genomes. To begin distinguishing between these possibilities, we categorized CGIs based on the presence (TSS_CGIs) or absence of a TSS within the boundaries of the island region. The latter have been named “orphan CGIs” (43), which despite lacking TSSs, can still have regulatory functions, such as being enhancers (44–46). Interestingly, it has recently been shown that the turnover of orphan CGIs between species contributes to regulatory innovation (47).

We found that TSS_CGIs are larger, have more CpG dinucleotides, and have a higher O/E CpG ratio than orphan CGIs (Fig. 2A) and are more depleted of TEs than orphan CGIs (Fig. 2B and C). In detail, TSS_CGIs, SINEs, LINES, and LTRs accounted for only 1.8, 0.3, and 0.2%, respectively, of the total nucleotides (Fig. 2C). These TEs were also found to be 11- to 57-fold less abundant in TSS_CGIs than in the surrounding 100 kb regions, which is higher than that observed for orphan CGIs (Fig. 2D), clearly illustrating the preferential exclusion of TEs from TSS_CGIs over evolutionary time. TE depletion in orphan

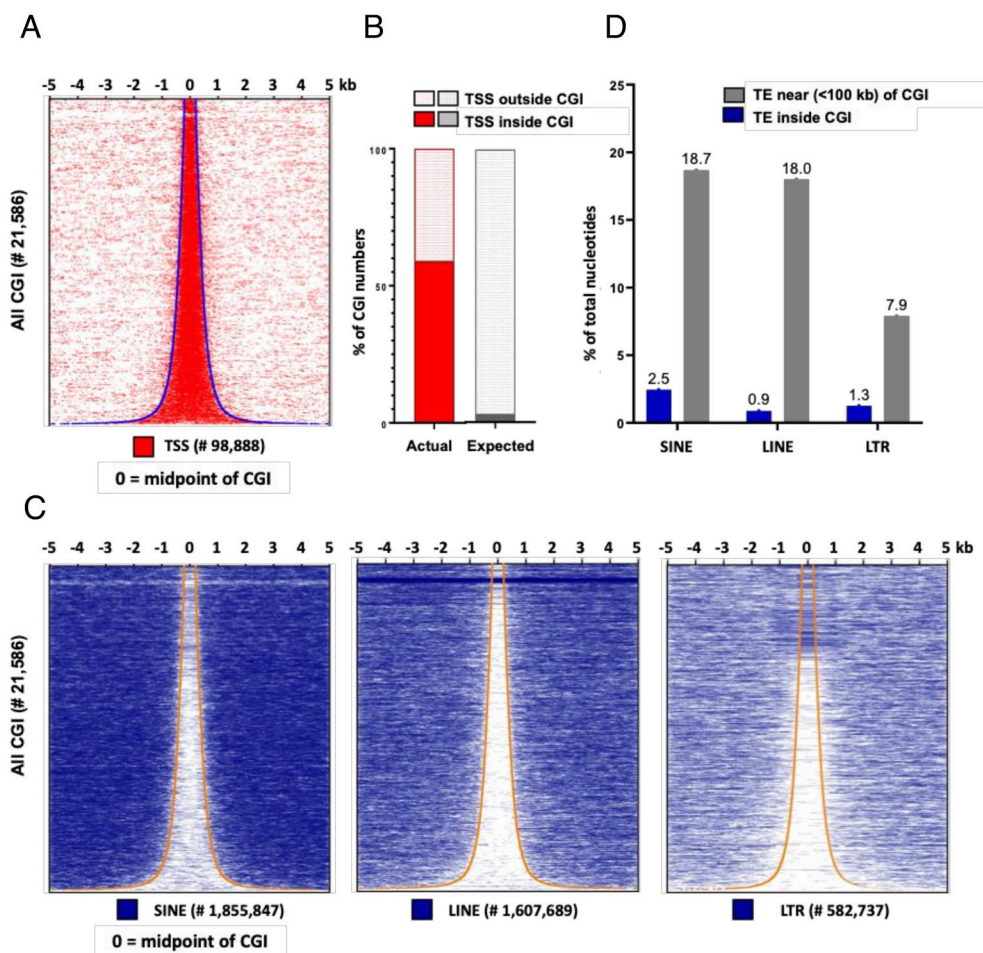


Fig. 1. TSSs are frequently located within CGIs, while TEs are often depleted from these regions. (A) CGIs were plotted according to width and aligned by midpoint, and TSSs were placed relative to each CGI. (B) Percentage of coincidence with TSSs for actual CGIs (Actual) or random intervals of identical lengths and similar locations to CGIs (Expected). (C) SINEs, LINEs, and LTRs are plotted relative to CGIs. Note that the content of TEs in CGIs is less than it appears visually because the pixels are proportionally larger than the actual densities. (D) Percentage of total nucleotides corresponding to TE according to the three main categories SINE, LINE, and LTR. These were measured inside of CGIs and 50 kb flanking regions.

CGIs may reflect the fact that they have functional roles, such as acting as enhancers, although the degree of depletion is less pronounced compared to TSS_CGIs (Fig. 2 B–D).

While the mouse genome showed slightly smaller sizes of CGIs, most of the features of TSS_CGIs were well conserved between mouse and human (SI Appendix, Fig. S3A). By discriminating between the two types of CGIs, we observed that LINE insertions in CGIs were mostly located in orphan CGIs (SI Appendix, Fig. S3B); 50% in orphan CGIs versus 1% in TSS_CGIs (SI Appendix, Fig. S3C and 3D). Interestingly, we found that most LINEs located in CGIs belonged to relatively young LINE subfamilies, known to have transposition potential, in both human and mouse (SI Appendix, Fig. S3D). As discussed earlier, LINE 5'UTRs fulfill the numerical criteria of a CGI (SI Appendix, Table S1) so that 50% of these orphan CGIs are probably not derived from the host DNA but represent the creation of new "CGIs" by TEs which may not have acquired a biological function. Altogether, these data suggest that there has been selection pressure against TE invasion in TSS_CGIs, while orphan-CGIs are apparently more likely to contain TEs, and these are conserved features across species.

TE Insertions Are More Common in CpG Poor than CGI_TSSs. Building upon our prior findings of strong TE exclusion at all annotated TSSs regardless of CGI status (1), we investigated

this phenomenon in more detail by analyzing the distribution of TEs around TSSs categorized as either CGI_TSSs or CpG poor (non-CGI_TSS) (Fig. 3). Almost none of the three TE classes were found at CGI_TSSs, which is consistent with our previous analysis (Fig. 1). In contrast, they were more often observed at non-CGI_TSS (Fig. 3B). A small portion of TSSs, all of which belong to non-CGI_TSSs, had TEs exactly overlapping the TSS. This observation aligns with the notion that these elements could have been repurposed as regulatory elements, such as alternative promoters, throughout evolution (8).

Interestingly, as observed previously (1), there was a pronounced decrease in the densities of all three TE classes downstream of the TSSs (Fig. 3B). One explanation for this observation is that TEs may interfere with the initial phases of transcriptional elongation, leading to evolutionary selective pressure to reduce the presence of TEs downstream of the TSS. In general, the mouse genome showed similar features (SI Appendix, Fig. S4).

Previously, we reported that all three classes of TEs were more frequently centered at enhancers than TSSs (1). Importantly, the frequencies of their occurrence were markedly age-dependent and evolutionary older versions, which had lost CpG sites were more likely to be present at these regulatory elements (REs). This loss of CpG sites was accompanied by changes in the binding sites for

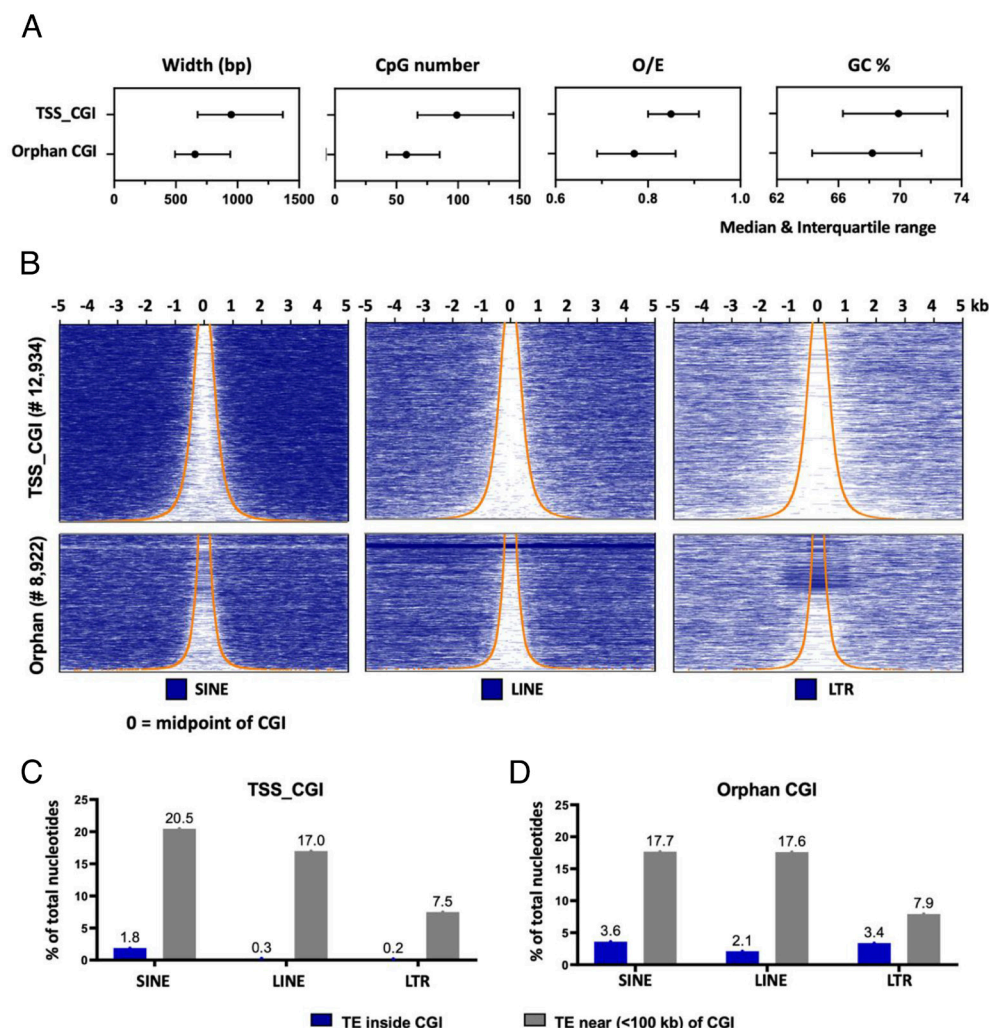


Fig. 2. TE insertions are less likely to occur in TSS_CGIs. (A) CGIs width, CpG number, Observed/Expected (O/E) ratio, and GC% were averaged separately in TSS_CGIs (TSS_CGIs comprise about 60% of all CGIs) and orphan CGIs. (B) SINEs, LINEs, and LTRs are plotted relative to each category of CGI. (C and D) Percentage of total nucleotides corresponding to TEs according to the three main categories SINE, LINE, and LTR in TSS_CGI or orphan CGI. These were measured in each CGI category and in 50 kb flanking regions.

TFs associated with the aged TEs (1). Thus, CpG depletion in TEs can result in the generation of new REs.

TEs Near Non-CGI TSSs Have Limited Effects on Gene Expression.

It is possible that the TE sequence per se has a functional role in the regulation of genes in which the promoter, as opposed to the actual TSS, co-occurs with a TE. To test this possibility, we selected two well-characterized “stemness” genes, *POU5F1* (also known as *OCT4*), and *NANOG*, contain *Alu* elements, in their promoters 291 bp and 213 bp upstream of the TSS, respectively (Fig. 4). An evolutionarily old *Alu* (*AluJ*) is located in the promoter of *OCT4* (Fig. 4A), likely inserted into the genome about 65 Mya (12). In the promoter of *NANOG*, a copy of the most abundant *Alu* element in the human genome, an intermediate-age *AluS* (~37 Mys old), is present (13) (Fig. 4B). It is important to note that the *Alus* in question are not the most studied ones in stemness (48), which are located about 6 kb upstream.

To test the effect of these *Alu* insertions on transcriptional activity, we generated luciferase constructs in which the candidate *Alu* was either included or absent from the canonical promoter. The *OCT4* luciferase construct (2.6 kb region upstream of TSS) and *NANOG* construct (1.8 kb region upstream of TSS) are not at CGI and contain both the proximal enhancers and

promoters, respectively. We transfected them into NCCIT cells, a surrogate for human embryonic stem cells, in which both *OCT4* and *NANOG* are highly expressed as all the necessary TFs for their expression are also present (49). The *OCT4* transcriptional activity was slightly, but significantly, stimulated in the absence of *AluJ* (Fig. 4A). In contrast, *NANOG*’s transcriptional activity was only slightly decreased with the ablation of the *AluS* (Fig. 4B). In neither case did the presence or absence of the *Alu* markedly alter transcription when upstream or downstream host sequences were either present or not (Fig. 4A and B). These *Alus* therefore have modest effects, if any, on the promoter functions.

Next, we tested whether in vitro methylation of the CpGs included in these constructs with the bacterial methylase *SssI* would downregulate luciferase expression by either promoter (Fig. 5A and B). Once again, the presence or absence of the *Alu* in either promoter had little discernible effect on the ability of DNA methylation to inhibit luciferase expression. We had earlier investigated the temporal acquisition of DNA methylation on these promoters during differentiation (50). The *Alus* in *OCT4* and *NANOG* were largely bereft of endogenous cytosine methylation when the genes were expressed but became more methylated after retinoic acid treatment, in concert with the other upstream

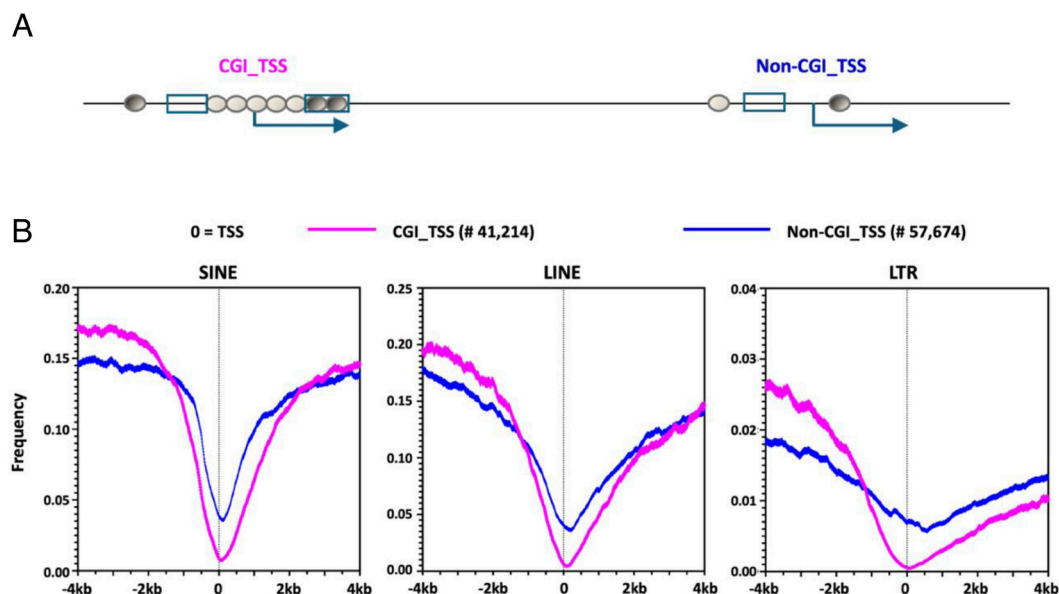


Fig. 3. CGI_TSSs are less likely to contain TE insertions than non-CGI_TSSs. (A) Graphical representation of CGI_TSS and non-CGI_TSS. CpGs are represented as beads, dark ones represent hypothetical methylated CpGs. We estimated that CGI_TSSs constitute 42% of all TSSs, which we use to calculate the frequencies in panel B. (B) Plots displaying the differences in the frequencies (TE proportional counts per 100 bp) of SINE, LINE, and LTR elements between CGI_TSSs and non-CGI_TSSs in the human genome. TSSs within CGIs (pink) show very little immediate proximity to TEs but starting around 1.5 kb upstream show a higher frequency of TE positioning compared to non-CGI_TSSs (blue).

regions. Also, it has been reported that these *Alus* become demethylated in heterokaryons during reprogramming (51). Together, these observations suggest that the methylation state of TEs mirrors the flanking host promoter sequences with limited effects on transcription.

Discussion

The results of our current analysis, which are an extension of our previous work (1), reveal remarkable differences between the composition of promoters that contain CGIs and those that do not.

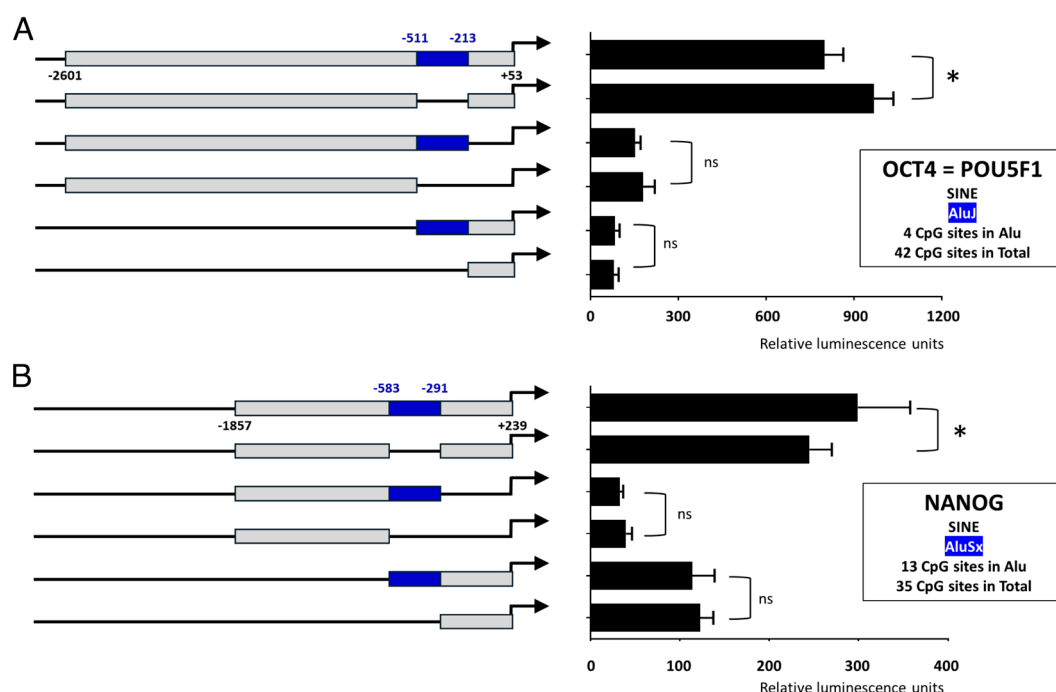


Fig. 4. SINEs located near non-CGI_TSSs have limited effects on transcriptional activity. Schematic diagrams (Left) illustrate the luciferase reporter constructs containing the regulatory regions of *OCT4* [(A) -2601 to +53] and *NANOG* [(B) -1857 to +239], with or without nearby SINE elements (*Alu*) or *AluXs*, shown in blue. The numbers indicate positions relative to the TSS, with +1 representing the TSS; negative values indicate upstream regions, while positive values indicate downstream regions. The corresponding bar graphs (Right) show transcriptional activity measured by dual-luciferase reporter assays in NCCIT cells. Data represent mean $\pm$ SEM from three independent experiments ($n = 3$). Graphs show mean and $\pm$ SEM ($n = 3$, * $P < 0.05$, ns: not statistically significant).

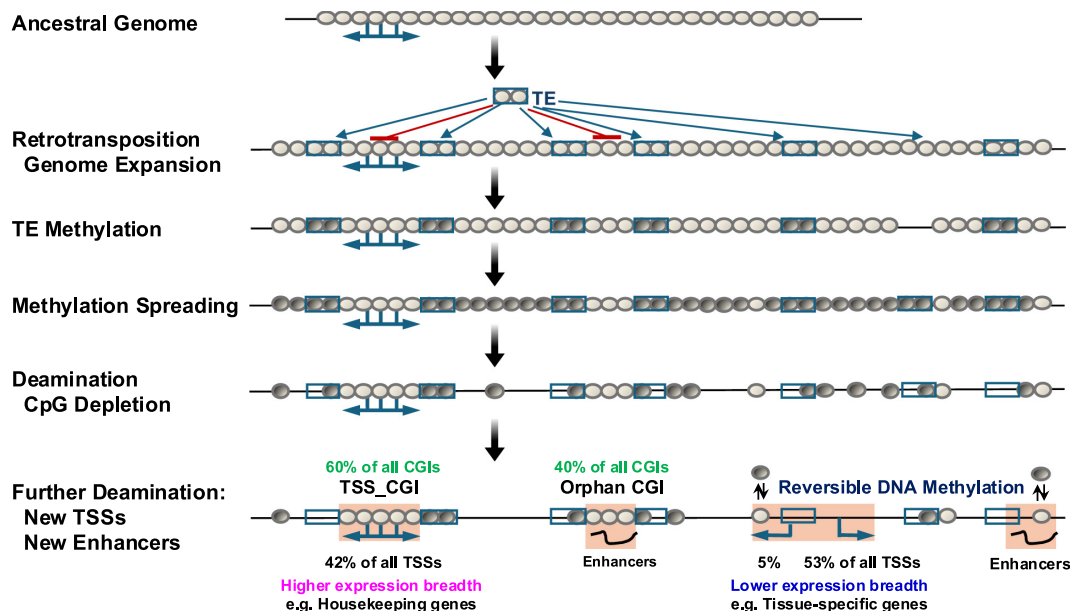


Fig. 6. Roles of TEs and DNA methylation in shaping the genome. We propose an updated version of the roles of TEs and DNA methylation as a follow up to our earlier work (1). In this model, the ancestral genome was relatively compact and rich in CpGs, resembling the genome of organisms that lack DNA methylation. The subsequent coevolution of TEs and DNA methylation machinery led to significant genome expansion. After TE insertion, the methylation of TEs and neighboring sequences initiated a gradual decline in CpG content within these regions. This process, coupled with the natural rate of C-to-T mutations, resulted in the loss of CpG sites over evolutionary time. CGIs have escaped large-scale invasion by TEs, likely due to their essential role in controlling TSSs (CGIs_TSS) and enhancers (orphan CGIs) during development. TSSs are distributed across the genome as follows: 42% reside within CGIs, 5% are located within TEs, and the remaining 53% are found outside of both CGIs and TEs. In this context, CGIs have evolved more like “mesas” or “buttes”, rather than “islands”, evading the evolutionary depletion of CpG sites due to their critical role in housekeeping functions. TEs could also have provided opportunities by depletion of CpG sites for the generation of TFs binding sequences (Non-CGI TSS) and REs evolution, such as enhancers. Additionally, TEs act as substrates for extensive genome rearrangement and recombination, which is not represented in this figure. We propose that the interplay between TE insertion and DNA methylation facilitated the evolution of more specialized, tissue-specific gene transcription. It also allowed for dynamic DNA methylation to participate in multiple-layer epigenetic controls. Image credit: adapted from ref. 1.

expression control (67–71), which allows for dynamic DNA methylation to participate in multiple-layer epigenetic controls.

In summary, the use of the term “junk DNA” fails to capture the profound symbiosis between the host genome and the TEs. Reduction of the CpG frequency in both the TE and the host DNA has most likely led to the existence of genomic landmarks (CGIs or “Mesas”) which can easily be recognized by TFs and serve as TSSs for 40% of all genes, many of which are essential for survival or organismal development. TEs can also function as part of REs (1, 8), but their most consequential contribution is probably the reduction of the CpG content of the remaining 99% of the genome, allowing for DNA methylation to function as a “tunable” but heritable epigenetic mark. Also, as the CpG content of the TEs themselves decreases during evolution, their presence near non-CGI_TSSs does not appear to interfere with gene control, as we demonstrated for *OCT4* and *NANOG*. The expansion of the genome by TEs might therefore have given rise to more options for gene control not available to organisms with compact genomes.

Materials and Methods

Genome Data and Statistics. All sequence analyses were carried out using the hg38 human and mm10 mouse genome assemblies. TE annotations were downloaded from RepeatMasker: hg38 – Dec. 2013 – open-4.0.5 – Repeat Library 20140131, and mm10 – Dec. 2011 – open-4.0.5 – Repeat Library 20140131. Human and mouse TSSs were obtained from refTSS 4.0.0 (72). CGI locations were derived from UCSC based on an unmasked genome. There are 51,402 CGI in unmasked hg38 based on criteria from Gardiner-Garden & Frommer criteria (31), which is (G + C) content of 0.50 or greater and an observed to expected CpG dinucleotide (O/E) ratio of 0.60 or greater, within a sequence window of 200 bp or greater. This number dropped to 21,856 when a threshold of 400 bp or larger was used. Typically,

researchers use a set of CGIs which have been derived from a masked genome in order to exclude repetitive DNA. For example, SINES-*ALU*, which, in the active form, is about 300 bp long, is 57% GC, and has a CpG OE = 0.95 and so would be considered a CGIs spuriously. We found that unmasked_CGI annotated regions, smaller than 400 bp, were more likely to be DNA that was mostly or only repetitive and TE DNA, whereas unmasked_CGIs larger than 400 bp were more likely to be masked_CGIs plus repetitive and TE DNA. In terms of numbers, unmasked_CGIs consist of 29,586 small (200 to 400 bp) intervals and 21,816 large intervals, while the (normally used) masked_CGIs consist of 9,411 small intervals and 18,538 large intervals.

Interval positioning and overlap measurements were computed with bedtools V2.31 and/or R implemented GenomicRanges V1.58.0. By default, Granges are 1-based sequences while bedtools and UCSC files are 0-based, meaning that some discrepancies for 1 bp overlapping intervals are possible.

Permutation testing was used in Fig. 1B to calculate the expected background TSS/interval overlap rate. Because TSS and TEs are not randomly distributed across the genome, an appropriate comparison for the expected association should exclude empty regions or other portions of the genome where TSS, TEs, and CGIs are not located. The same number and size distribution (as for actual CGI intervals) of randomized intervals was selected from sequence within 1 Mb of actual CGI positions. The random sets were generated 1,000 times and compared to TSS locations to determine the expected TSS-CpG association frequency.

Luciferase-Based Transcription Assays. Luciferase vectors were cloned into the pCpGL basic vector (73). For methylated plasmid, Luciferase vectors were subjected to in vitro methylation by methyltransferase M.SssI in the presence of methyl group donor S-adenosylmethionine (SAM). NCCIT cells were plated in 6-well plates, and 24 h later, luciferase constructs were transfected into the cells using Lipofectamine LTX (Invitrogen) according to the manufacturer’s protocol. Cells were lysed 24 h later, and luciferase assays were performed using a Dual-Luciferase® Reporter Assay System (Promega, Madison, WI) on a single automatic injection Mithras (Berthold technologies, Bad Wildbad, Germany) luminometer following the manufacturer’s instructions. Luciferase readings were taken as

singlets. Ratios of Renilla- to Firefly-luciferase readings were recorded for each experiment, and triplicates were then averaged.

Cell Culture. Embryonic carcinoma NCCIT cells (American Type Culture Collection) were cultured in RPMI 1640 (Life Technologies) supplemented with 10% fetal bovine serum.

Data, Materials, and Software Availability. No new data was generated for this research; all genomic sequence and annotation data came from publicly available sources, as referenced. Primary data used consists of; RepeatMasker: hg38 – Dec. 2013 – open-4.0.5 – Repeat Library 20140131, and mm10 – Dec. 2011 – open-4.0.5 – Repeat Library 20140131, at <https://www.repeatmasker.org> (74). Human and mouse TSSs were obtained from refTSS 4.0 0, at <https://refTSS.riken.jp> (72). Genomic interval and statistical analysis were done using standard

R code and/or bash scripts, which are available on Github at https://github.com/sepvai/CPGI_TE_overlap (75).

ACKNOWLEDGMENTS. This work was supported by the Van Andel Institute, the National Cancer Institute (R35CA209859; P.A.J. and G.L.), as well as the Konkuk University's research support program for faculty on sabbatical in 2024 (to J.S.Y). We thank Martha L. Escobar Galvis, PhD, for her help with editing the manuscript.

Author affiliations: ^aDepartment of Biochemistry, School of Medicine, Konkuk University, Chungju-si 27478, Korea; ^bDepartment of Epigenetics, Van Andel Institute, Grand Rapids, MI 49503; and ^cDepartment of Urology, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA 90089

- W. Zhou, G. Liang, P. L. Molloy, P. A. Jones, DNA methylation enables transposable element-driven genome expansion. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 19359–19366 (2020).
- A. P. Bird, DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
- B. E. Bernstein *et al.*, A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
- S. Nurk *et al.*, The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
- M. V. Almeida, G. Vernaz, A. L. K. Putman, E. A. Miska, Taming transposable elements in vertebrates: from epigenetic silencing to domestication. *Trends Genet.* **38**, 529–553 (2022).
- O. Deniz, J. M. Frost, M. R. Branco, Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* **20**, 417–431 (2019).
- G. Bourque *et al.*, Ten things you should know about transposable elements. *Genome Biol.* **19**, 199 (2018).
- A. Y. Du, J. D. Chobirko, X. Zhuo, C. Feschotte, T. Wang, Regulatory transposable elements in the encyclopedia of DNA elements. *Nat. Commun.* **15**, 7594 (2024).
- R. Fuego, J. Judd, C. Feschotte, J. Wysocka, Roles of transposable elements in the regulation of mammalian transcription. *Nat. Rev. Mol. Cell Biol.* **23**, 481–497 (2022).
- H. A. Lawson, Y. Liang, T. Wang, Transposable elements in mammalian chromatin organization. *Nat. Rev. Genet.* **24**, 712–723 (2023).
- E. C. Pehrsson, M. N. K. Choudhary, V. Sundaram, T. Wang, The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat. Commun.* **10**, 5640 (2019).
- P. Deininger, Alu elements: Know the SINEs. *Genome Biol.* **12**, 236 (2011).
- S. Kim, C. S. Cho, K. Han, J. Lee, Structural variation of Alu element and human disease. *Genomics Inform.* **14**, 70–77 (2016).
- M. A. Lodato, J. Vijg, Editorial: Somatic mutations, genome mosaicism and aging. *Front Aging* **3**, 1115408 (2022).
- L. M. Payer, K. H. Burns, Transposable elements in human genetic disease. *Nat. Rev. Genet.* **20**, 760–772 (2019).
- N. S. Catlin, E. B. Josephs, The important contribution of transposable elements to phenotypic variation and evolution. *Curr. Opin. Plant Biol.* **65**, 102140 (2022).
- C. Sun, A. Zhang, J. Chen, S. Schaack, "Junk" that matters: The role of transposable elements in bumblebee genome evolution. *Curr. Opin. Insect Sci.* **59**, 101103 (2023).
- R. N. Platt 2nd, M. W. Vandeweghe, D. A. Ray, Mammalian transposable elements and their impacts on genome evolution. *Chromosom. Res.* **26**, 25–43 (2018).
- A. D. Senft, T. S. Macfarlan, Transposable elements shape the evolution of mammalian development. *Nat. Rev. Genet.* **22**, 691–711 (2021).
- A. Zemach, I. E. McDaniel, P. Silva, D. Zilberman, Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).
- E. E. Grundy, N. Diab, K. B. Chiappinelli, Transposable element regulation and expression in cancer. *FEBS J* **289**, 1160–1179 (2022).
- X. Cheng, R. M. Blumenthal, Mediating and maintaining methylation while minimizing mutation: Recent advances on mammalian DNA methyltransferases. *Curr. Opin. Struct. Biol.* **75**, 102433 (2022).
- R. Prasad, T. J. Yen, A. Bellacosa, Active DNA demethylation—The epigenetic gatekeeper of development, immunity, and cancer. *Adv. Genet.* **2**, e10033 (2021).
- H. Ohtani, M. Liu, W. Zhou, G. Liang, P. A. Jones, Switching roles for DNA and histone methylation depend on evolutionary ages of human endogenous retroviruses. *Genome Res.* **28**, 1147–1157 (2018).
- M. Tomkova *et al.*, Human DNA polymerase epsilon is a source of C>T mutations at CpG dinucleotides. *Nat. Genet.* **56**, 2506–2516 (2024).
- A. B. Silveira *et al.*, Base-excision repair pathway shapes 5-methylcytosine deamination signatures in pan-cancer genomes. *Nat. Commun.* **15**, 9864 (2024).
- M. McClelland, R. Ivarie, Asymmetrical distribution of CpG in an "average" mammalian gene. *Nucleic Acids Res.* **10**, 7865–7877 (1982).
- N. P. Blackledge, R. Klose, CpG island chromatin: A platform for gene regulation. *Epigenetics* **6**, 147–152 (2011).
- A. M. Deaton, A. Bird, CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).
- A. Nishiyama, M. Nakanishi, Navigating the DNA methylation landscape of cancer. *Trends Genet.* **37**, 1012–1027 (2021).
- M. Gardiner-Garden, M. Frommer, CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
- L. Zhang, Z. Dai, J. Yu, M. Xiao, CpG-island-based annotation and analysis of human housekeeping genes. *Brief Bioinform.* **22**, 515–525 (2021).
- D. Takai, P. A. Jones, Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 3740–3745 (2002).
- J. Zhu, F. He, S. Hu, J. Yu, On the nature of human housekeeping genes. *Trends Genet.* **24**, 481–484 (2008).
- P. A. Jones, Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
- G. Nagae *et al.*, Tissue-specific demethylation in CpG-poor promoters during cellular differentiation. *Hum. Mol. Genet.* **20**, 2710–2721 (2011).
- B. Miao *et al.*, Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol.* **21**, 255 (2020).
- F. C. Grandi *et al.*, Retrotransposition creates sloping shores: A graded influence of hypomethylated CpG islands on flanking CpG sites. *Genome Res.* **25**, 1135–1146 (2015).
- J. R. Korenberg, M. C. Rykowski, Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**, 391–400 (1988).
- E. M. Ostertag, H. H. Kazazian Jr., Biology of mammalian L1 retrotransposons. *Ann. Rev. Genet.* **35**, 501–538 (2001).
- R. Rebollo, Y. Zhang, D. L. Mager, Transposable elements: Not as quiet as a mouse. *Genome Biol.* **13**, 159 (2012).
- M. Bodak, J. Yu, C. Claudio, Regulation of LINE-1 in mammals. *Biomol. Concepts* **5**, 409–428 (2014).
- R. S. Illingworth *et al.*, Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* **6**, e1001134 (2010).
- J. Yao *et al.*, Integration of multi-omics data revealed the orphan CpG islands and enhancer-dominated cis-regulatory network in glioma. *iScience* **27**, 110946 (2024).
- T. Pachano *et al.*, Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. *Nat. Genet.* **53**, 1036–1049 (2021).
- J. S. K. Bell, P. M. Vertino, Orphan CpG islands define a novel class of highly active enhancers. *Epigenetics* **12**, 449–464 (2017).
- A. A. Kocher *et al.*, CpG island turnover events predict evolutionary changes in enhancer activity. *Genome Biol.* **25**, 156 (2024).
- A. Morales-Hernandez *et al.*, Alu retrotransposons promote differentiation of human carcinoma cells through the aryl hydrocarbon receptor. *Nucleic Acids Res.* **44**, 4665–4683 (2016).
- J. S. You *et al.*, Depletion of embryonic stem cell signature by histone deacetylase inhibitor in NCCIT cells: Involvement of nanog suppression. *Cancer Res.* **69**, 5716–5725 (2009).
- J. S. You *et al.*, OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 14497–14502 (2011).
- N. Bhutani *et al.*, Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature* **463**, 1042–1047 (2010).
- F. Antequera, A. Bird, Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 11995–11999 (1993).
- F. Antequera, A. Bird, CpG islands. *EXS* **64**, 169–185 (1993).
- P. Lavia, D. Macleod, A. Bird, Coincident start sites for divergent transcripts at a randomly selected CpG-rich island of mouse. *EMBO J.* **6**, 2773–2779 (1987).
- N. Adachi, M. R. Lieber, Bidirectional gene organization: A common architectural feature of the human genome. *Cell* **109**, 807–809 (2002).
- D. Takai, P. A. Jones, Origins of bidirectional promoters: Computational analyses of intergenic distance in the human genome. *Mol. Biol. Evol.* **21**, 463–467 (2004).
- J. C. Lin *et al.*, Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island. *Cancer Cell* **12**, 432–444 (2007).
- C. J. Barger *et al.*, Co-regulation and function of FOXM1/RHNO1 bidirectional genes in cancer. *elife* **10**, e55070 (2021).
- G. Elgar *et al.*, Small is beautiful: Comparative genomics with the pufferfish (*Fugu rubripes*). *Trends Genet.* **12**, 145–150 (1996).
- R. A. Irizarry *et al.*, The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
- M. C. Frith *et al.*, Evolutionary turnover of mammalian transcription start sites. *Genome Res.* **16**, 713–722 (2006).
- J. C. Klein, A. Keith, V. Agarwal, T. Durham, J. Shendure, Functional characterization of enhancer evolution in the primate lineage. *Genome Biol.* **19**, 99 (2018).
- M. A. Zabidi *et al.*, Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
- A. Curina *et al.*, High constitutive activity of a broad panel of housekeeping and tissue-specific cis-regulatory elements depends on a subset of ETS proteins. *Genes Dev.* **31**, 399–412 (2017).

65. G. Nagy, D. Bojcsuk, P. Tzerpos, T. Cseh, L. Nagy, Lineage-determining transcription factor-driven promoters regulate cell type-specific macrophage gene expression. *Nucleic Acids Res.* **52**, 4234–4256 (2024).
66. M. Dejosez *et al.*, Regulatory architecture of housekeeping genes is driven by promoter assemblies. *Cell Rep.* **42**, 112505 (2023).
67. Y. Yin *et al.*, Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
68. H. Han *et al.*, DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Hum. Mol. Genet.* **20**, 4299–4310 (2011).
69. E. Kreibich, R. Kleinendorst, G. Barzaghi, S. Kaspar, A. R. Krebs, Single-molecule footprinting identifies context-dependent regulation of enhancers by DNA methylation. *Mol. Cell* **83**, 787–802. e789 (2023).
70. S. L. Thomas *et al.*, DNA strand asymmetry generated by CpG hemimethylation has opposing effects on CTCF binding. *Nucleic Acids Res.* **51**, 5997–6005 (2023).
71. W. A. Flavahan *et al.*, Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).
72. I. Abugessaisa *et al.*, refTSS: A reference data set for human and mouse transcription start sites. *J. Mol. Biol.* **431**, 2407–2422 (2019).
73. M. Klug, M. Rehli, Functional analysis of promoter CpG methylation using a CpG-free luciferase reporter vector. *Epigenetics* **1**, 127–130 (2006).
74. A. F. A. Smit, R. Hubley, P. Green, Data from "RepeatMasker Open-4.0." RepeatMasker. <http://www.repeatmasker.org>. Accessed 6 March 2024.
75. J. S. You, S. Pierce, G. Liang, P. A. Jones, Role of transposable elements and DNA methylation in the formation of CpG islands and CpG depleted regulatory elements. Github. https://github.com/sepvai/CPGI_TE_overlap. Deposited 8 October 2025.