Published online 23 February 2021 Nucleic Acids Research, 2021, Vol. 49, No. 5 **2583–2597**
doi: 10.1093/nar/gkab108

Conserved dual-mode gene regulation programs in higher eukaryotes

Jun-Yeong Lee[1,†], Jawon Song[2,†], Lucy LeBlanc[3], Ian Davis[1], Jonghwan Kim[3] and Samuel Beck[1,*]

[1]Davis Center for Regenerative Biology and Medicine, MDI Biological Laboratory, Bar Harbor, ME 04609, USA, [2]Texas Advanced Computing Center, The University of Texas at Austin, Austin, TX 78758, USA and [3]Department of Molecular Biosciences, The University of Texas at Austin, Austin, TX 78712, USA

Received September 14, 2020; Revised December 21, 2020; Editorial Decision February 06, 2021; Accepted February 08, 2021

ABSTRACTS

Recent genomic data analyses have revealed important underlying logics in eukaryotic gene regulation, such as CpG islands (CGIs)-dependent dual-mode gene regulation. In mammals, genes lacking CGIs at their promoters are generally regulated by interconversion between euchromatin and heterochromatin, while genes associated with CGIs constitutively remain as euchromatin. Whether a similar mode of gene regulation exists in non-mammalian species has been unknown. Here, through comparative epigenomic analyses, we demonstrate that the dual-mode gene regulation program is common in various eukaryotes, even in the species lacking CGIs. In cases of vertebrates or plants, we find that genes associated with high methylation level promoters are inactivated by forming heterochromatin and expressed in a context-dependent manner. In contrast, the genes with low methylation level promoters are broadly expressed and remain as euchromatin even when repressed by *Polycomb* proteins. Furthermore, we show that invertebrate animals lacking DNA methylation, such as fruit flies and nematodes, also have divergence in gene types: some genes are regulated by *Polycomb* proteins, while others are regulated by heterochromatin formation. Altogether, our study establishes gene type divergence and the resulting dual-mode gene regulation as fundamental features shared in a broad range of higher eukaryotic species.

INTRODUCTION

Eukaryotic genome exists either as condensed heterochromatin or loose euchromatin (1–4). The densely packed structure of heterochromatin physically blocks the access of transcriptional machineries and transcription factors (TFs) to the genome, thus the genes forming heterochromatin remain transcriptionally silent. Within the nucleus, heterochromatin constitutes a spatial compartment that is physically segregated from its open counterpart, euchromatin (5–8). Transitioning a gene from euchromatin into heterochromatin has been shown to silence gene expression (9), establishing heterochromatin formation as one of the primary eukaryotic gene inactivation mechanisms.

Another well-known gene inactivation mechanism in multi-cellular eukaryotes is mediated by *Polycomb* repressive complexes (PRCs) (10–13). *Polycomb* proteins are conserved in a broad range of eukaryotes, including vertebrates, plants, insects, nematodes, and even some yeasts (14). *Polycomb* proteins catalyze covalent modifications of histones: *Polycomb* repressive complex 1 (PRC1) forms mono-ubiquitination of lysine 119 of H2A (H2AK119ub1), while *Polycomb* repressive complex 2 (PRC2) forms the di- and tri-methylation at the lysine 27 of H3 (H3K27me2 and H3K27me3). Binding of *Polycomb* group proteins and resulting histone modifications repress gene expression by blocking the assembly of transcriptional machineries at gene promoters (13,15).

Interestingly, the relationships among these two gene inactivation mechanisms are not clearly defined in most eukaryotes; whether these two mechanisms work together to inactivate the same genes, or whether they inactivate distinct subsets of genes; if then, which classes of genes are inactivated by heterochromatin formation while other types of genes are repressed by *Polycomb* proteins. Clarification of this relationship is critical for the comprehensive understanding of the gene regulatory mechanisms in eukaryotic organisms.

As a part of answers to these questions, our recent study indicated that the presence (and absence) of CpG islands (CGIs) determines the mode of gene regulations in mammals (16,17). CGIs are mammalian long (minimum 200 bp, median 1 kb) DNA elements with unexpectedly high CpG

*To whom correspondence should be addressed. Tel: +1 207 288 9880, ext. 476; Fax: +1 207 288 2130; Email: sbeck@mdibl.org

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

(Cytosine-phosphate-Guanine) dinucleotides and high GC contents (17–20). Majority of cytosines at the CpG sites in mammalian genome are methylated (5-methylcytosine); however, CpG sites within CGIs remain largely unmethylated. CGIs are usually associated with gene promoters; ∼60% of genes in human and mouse are associated with CGIs in their promoter (CGI+ or CGI-containing genes) while the other 40% are not (CGI– or CGI-less genes). Through compendium analysis of genomic data sets, we revealed a mammalian dual-mode gene regulation program mediated by CGIs (Figure 1A): only CGI– genes are regulated by interconversion between euchromatin and heterochromatin (16). On the other hand, CGI+ genes remain as euchromatin regardless of their transcriptional activities, and are inactivated by *Polycomb* proteins (21,22).

Notably, while functional CGI elements are only observed in mammalian genome (Figure 1B) (19,23), the components of CGI-mediated dual-mode gene regulation, such as the distinction of eu-/hetero-chromatin, *Polycomb* proteins, or DNA methylation, are well conserved in a broad range of eukaryotes. Therefore, we question whether the eukaryotic species lacking CGI elements also have dual-mode gene regulation program. In this study, we performed a comparative/integrative analysis using published genome-scale datasets and present that the divergence of gene types and resulting dual-mode gene regulations are common in various eukaryotic model organisms.

MATERIALS AND METHODS

Genome version

For NextGen sequencing data analysis, the following genome assemblies were used. *Escherichia coli K12 MG1655*: 2001–10-15, *Schizosaccharomyces pombe*: EF2, *Saccharomyces cerevisiae*: sacCer3, *Magnaporthe oryzae*: MG8, *Neurospora crassa*: NC12, *Gossypium hirsutum*: AD1-NBIv1.1, *Zea mays*: AGPv3, *Glycine max*: Glyma1, *Solanum lycopersicum*: SL2.50, *Lotus japonicus*: Kazuza2009, *Oryza sativa japonica*: IRGSP-1.0, *Oryza sativa indica*: ASM465v1, *Arabidopsis thaliana*: TAIR10, *Caenorhabditis elegans*: ce10, *Drosophila melanogaster*: dm3, *Danio rerio*: danRer10, *Cynoglossus semilaevis*: V1.0, *Xenopus tropicalis*: XENTR9.1, *Macaca mulatta*: Mmul_8.0.1, *Homo sapiens*: hg19, *Macaca fascicularis*: macFas5, *Rattus norvegicus*: rn6, *Bos taurus*: bosTau8, *Mus musculus*: mm9, *Canis lupus familiaris*: CanFam3.1.

Definition of mammalian CGI+ and CGI– genes

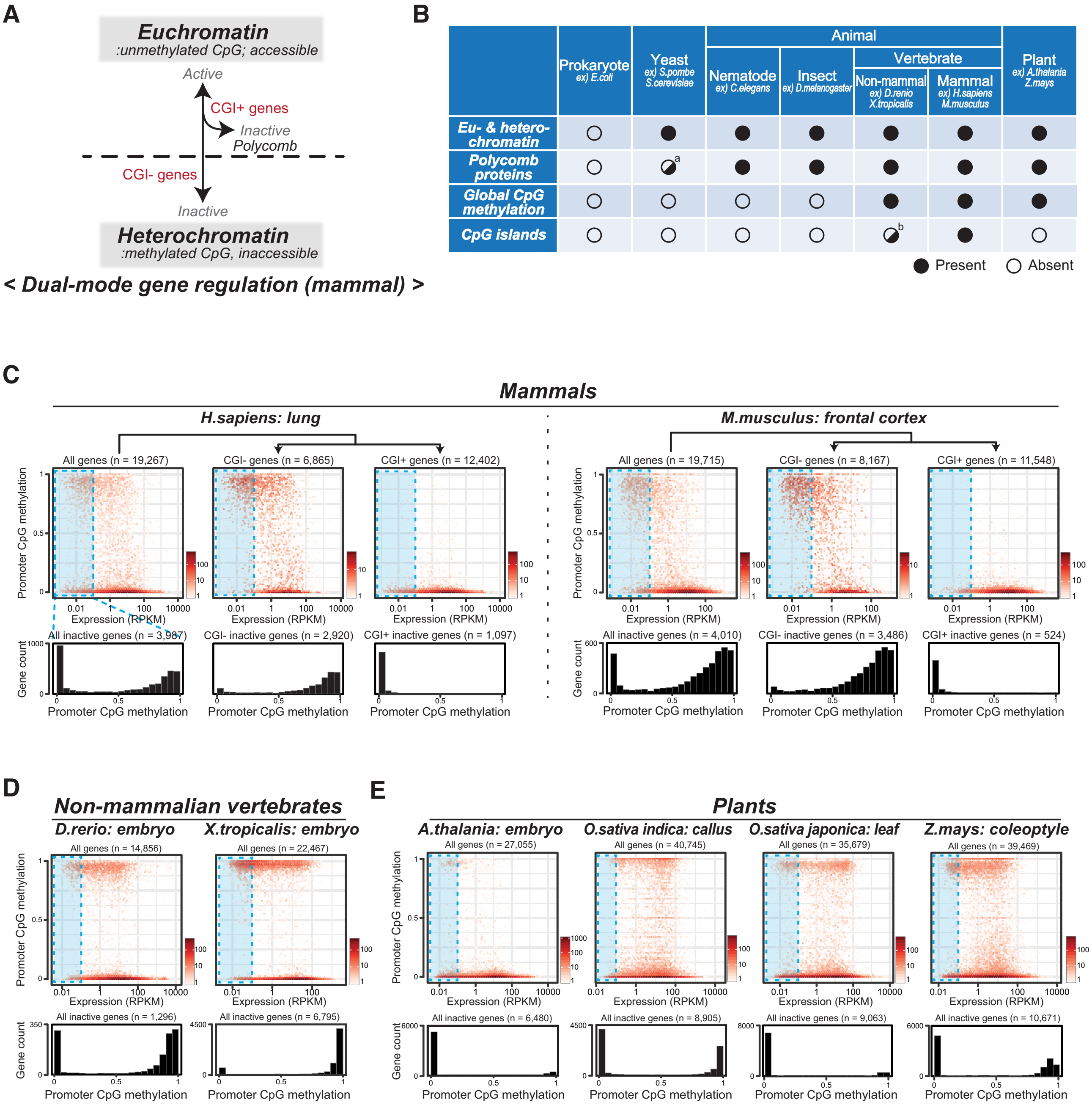
CGI-containing (CGI+) and CGI-less (CGI−) genes were defined as we did previously (16). Briefly, we used experimentally validated CGI elements identified by CxxC-affinity purification followed by parallel sequencing (CAP-seq) data (24). In detail, we listed CxxC-affinity purified regions in sperm, blood and cerebellum (both in mouse and human) with a general ChIP-seq data analysis pipeline (see ChIP-/ATAC-/DNase-seq analysis section in the Methods section), and identified non-tissue-specific consensus CxxC-domain binding regions. For gene classification, genes surrounded by consensus CxxC binding regions (within ±500 bp of the TSSs) were considered to be CGI+ genes, while genes without surrounding consensus CxxC binding regions were defined as CGI− genes (listed in Supplementary Table S3).

RNA-seq analysis

All available RNA-seq data deposited at Gene Expression Omnibus were initially listed on 9 April 2018, and further updated on 27August 2019 and 19 March 2020. RNA-seq data were downloaded from Sequence Read Archive (SRA) from the National Center for Biotechnology Information (NCBI) database. FASTQ files were extracted with the SRA Toolkit version 2.5.5, and then aligned to the genome using STAR version 2.4.2 (25). Gene expression was calculated as Read Per Kilobase per Million (RPKM) values using rpkmforgenes.py (26). Because the ranges of RPKM values span over three orders of magnitude and tend to give high random multiplicative error in high expression values, expression values were converted into $\log_{10}$ scale [$\log_{10}$(RPKM + 1)] for graphical summarization (Figures 2, 3, 5 and Supplementary Figures S3, S5 and S6). To test the relationships between gene expression and epigenetic modifications or motif occurrences (Figures 1C–E, 2, 3, 5, 6 and Supplementary Figures S1, S3, S5 and S6), only the protein-coding isoforms with the highest expression values among splicing variants (in RPKM) were used to minimize noise originating from rarely used alternative transcriptional start sites (TSSs), or poor annotation of TSSs of non-coding transcripts. All RNA-seq data used in this study are summarized in Supplementary Table S1.

Whole genome bisulfite sequencing (WGBS-seq) data analysis

All available WGBS-seq data deposited at Gene Expression Omnibus were initially listed on 9 April 2018, and the newly deposited data were added to our analysis on 27 August 2019 and 19 March 2020. WGBS-seq data analysis was done as we previously described (27). Raw data were downloaded from Sequence Read Archive (SRA) from the National Center for Biotechnology Information (NCBI) database. FASTQ files were extracted with the SRA Toolkit version 2.5.5, and then aligned to the genome using BSMAP (Bisulfite Sequence Mapping Program) 1.0 (28), and the methylation levels were calculated with BSMAPz (https://github.com/zyndagj/BSMAPz) for each cytosine within a given genome. Among three sequence contexts where cytosines can be methylated, CpG methylation showed the most distinct patterns between two promoter types (i.e. CGI+ versus CGI– & $\mathrm{Met}^{\mathrm{low}}$ versus $\mathrm{Met}^{\mathrm{high}}$; Supplementary Figures S1 and S5). Therefore, we monitored the coverage for CpG methylation in each WGBS-seq data, and selected and used only high-quality data for the subsequent analyses. To elaborate further, WGBS-seq data that detected at least 20% of all CpG sites in each genome were first selected, and then of these, only the data that detected at least 50% of promoters of total protein-coding genes were selected and were used for the further analyses. Promoter methylation levels (Figures 1C–E, 2A–B and Supplementary Figure 1) were calculated as the average methylation of all detected CpG sites within 200 bp-surrounding transcription start



	Prokaryote ex) E.coli	Yeast ex) S.pombe S.cerevisiae	Animal				Plant ex) A.thalania Z.mays
			Nematode ex) C.elegans	Insect ex) D.melanogaster	Vertebrate		
					Non-mammal ex) D.renio X.tropicalis	Mammal ex) H.sapiens M.musculus	
Eu- & hetero-chromatin	○	●	●	●	●	●	●
Polycomb proteins	○	◐[a]	●	●	●	●	●
Global CpG methylation	○	○	○	○	●	●	●
CpG islands	○	○	○	○	◐[b]	●	○

● Present ○ Absent

Figure 1. Overarching question of the study. (**A**) Dual-mode gene regulation in mammalian systems (16,17). Note that only CGI– genes are generally regulated by interconversion between eu- and hetero-chromatin, while CGI+ genes remain as euchromatin even when repressed by *Polycomb* proteins. (**B**) Conservation of components of dual-mode gene regulation in model eukaryotes. [a]Mostly absent except for rare cases (e.g. *Cryptococcus neoformans* (14)). [b]Non-mammalian vertebrates lack CpG islands with traditional definition, but have similar elements (i.e. non-methylated islands: NMIs) (44). (**C–E**) CpG methylation at the promoters of mammals (C), non-mammalian vertebrates (D), and plants (E). Upper 2D plots: Relationships between CpG methylation at the promoters (200 bp surrounding TSSs) of protein-coding genes (y-axis) versus expression of associated genes (x-axis). Lower bar plots: Distribution of promoter CpG methylation levels for transcriptionally inactive (RPKM < 0.1) genes.

site. In human data, WGBS-seq data generated from cancer cells/tissues were not used in measuring promoter methylation capacity, as some CGIs are often aberrantly methylated in these contexts (29). For the definition of Met^{high}/Met^{low} genes, promoter methylation level in each gene was calculated in each data/species, and then the methylation capacity of each promoter was calculated. To minimize noise from outliers, 98th percentile promoter methylation level was considered as max promoter methylation capacity (Figure 2A, B). Genes with max promoter methylation capacity higher than 0.7 and lower than 0.4 were considered as Met^{high} and Met^{low} genes, respectively. To measure total CpG methylation levels in each species (Supplementary Figure S2B), average methylation levels in all CpG sites were calculated in all available WGBS-seq data of each species and summarized as the boxplots. All WGBS-seq data used

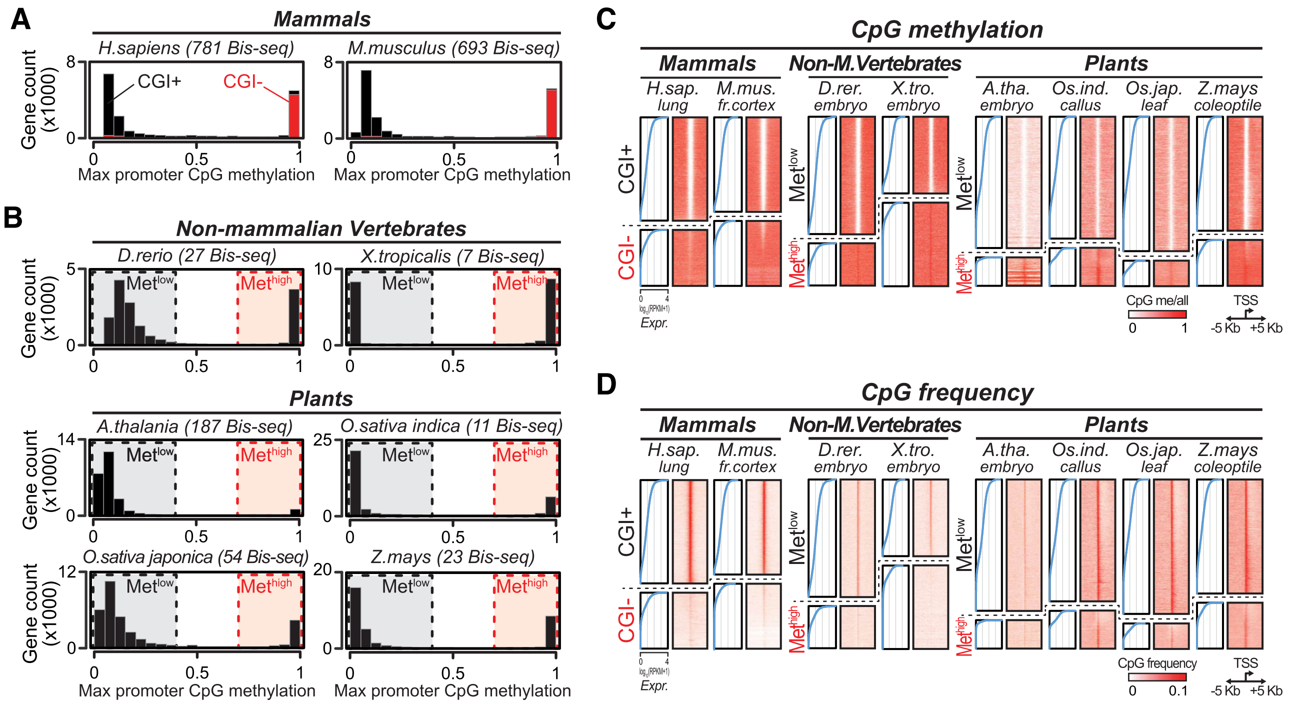


Figure 2. Conserved gene type divergence in vertebrates and plants. (**A**, **B**) Distributions of maximum CpG methylation levels at the promoters of protein-coding genes (200 bp surrounding TSSs) of mammals (A) and non-mammalian vertebrates and plants (B). In non-mammalian vertebrates and plants (B), Met^{high} (high methylation level) and Met^{low} (low methylation level) genes were defined as shown in dotted boxes (max promoter CpG methylation level > 0.7 and < 0.4, respectively). (**C**, **D**) CpG methylation (C) and frequencies (D) at the 10 kb-surrounding regions of TSSs of two types of genes are shown as heatmaps. Genes are sorted by their expression levels (left blue line plots) in the indicated cell/tissue types.

in this study are summarized in Supplementary Table S2. Met^{high} and Met^{low} genes defined in this study are listed in Supplementary Table S3.

ChIP-/ATAC-/DNase-seq analysis

ChIP-seq (Chromatin Immunoprecipitation followed by parallel-sequencing), ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) and DNaseI-seq (DNase I hypersensitive sites sequencing) data were listed (on 19 March 2020) and downloaded from Sequence Read Archive (SRA) from the National Center for Biotechnology Information (NCBI) database. FASTQ files were extracted with the SRA Toolkit version 2.5.5, and aligned using Bowtie 2.2.5 to the reference genomes (30). Signal based analyses were done using duplicate filtered read pileup bedGraph files made from Model-based Analysis of ChIP-seq (MACS) 1.4.2 (31). In order to summarize the ChIP signal enrichment over controls, the background subtracted bedGraph files with log likelihood ratios were made using MACS2 version 2.1.1 with 'bdgcmp -m logLR' command. As the heterochromatin ChIP-seq data often have high false positive signals due to the contamination of active genomic regions (32), we filtered H3K9me2/3 ChIP-seq data as we did previously (16) and only used data with high signal to noise ratio (SNR). For RNA-seq data shown in Figure 5, we first measured the median expression levels and selected the top 10% expressed genes, and defined them as generally active genes. Then we measured H3K9me2/3 ChIP-seq signals within 1 kb of TSSs of generally active genes and considered them as the background noise from active genomic regions. To measure the true heterochromatin signals, we monitored H3K9me2/3 signals at Giemsa positive regions (for mammals) or at the 1 kb-surrounding TSSs of generally inactive genes (bottom 10% of median expression; in non-mammalian species). If the average H3K9me2/3 signals from active genes (background noise) are higher than the signals in inactive regions, those data were excluded from our analysis. Unlike other species, within *C. elegans* genome, H3K9me3 is formed in H3K27me3 positive regions (which is formed by *Polycomb* proteins), while H3K9me2 occurs in heterochromatic regions similar to all other species (33,34). Therefore, only H3K9me2 was used as the heterochromatic mark in *C. elegans*. All ChIP-/ATAC-/DNase-seq data used in this study are summarized in Supplementary Table S4. Statistical test in Figure 3 and Supplementary Figure S6 were performed as follows. For chromatin accessibility data (Figure 3A and Supplementary Figure S6A), DNase-/ATAC-seq signals at 1Kb surrounding regions of TSSs and the expression of associated genes ($\log_{10}$ scale) were monitored separately in CGI+/Met^{low} and CGI–/Met^{high} gene groups, and the significance of their difference was measured with Chow test. For gene inactivation marks (Figure 3A, B and Supplementary Figure S6A, S6B), ChIP-seq signals at the surrounding 1 kb (*Polycomb*, H3K27me3) or 10 Kb (heterochromatin, H3K9me2/3) of the TSSs of inactive genes (RPKM < 0.1) were monitored separately in CGI+/Met^{low} and CGI–/Met^{high} gene groups, and the significance of their difference were measured with Wilcoxon signed-rank test.

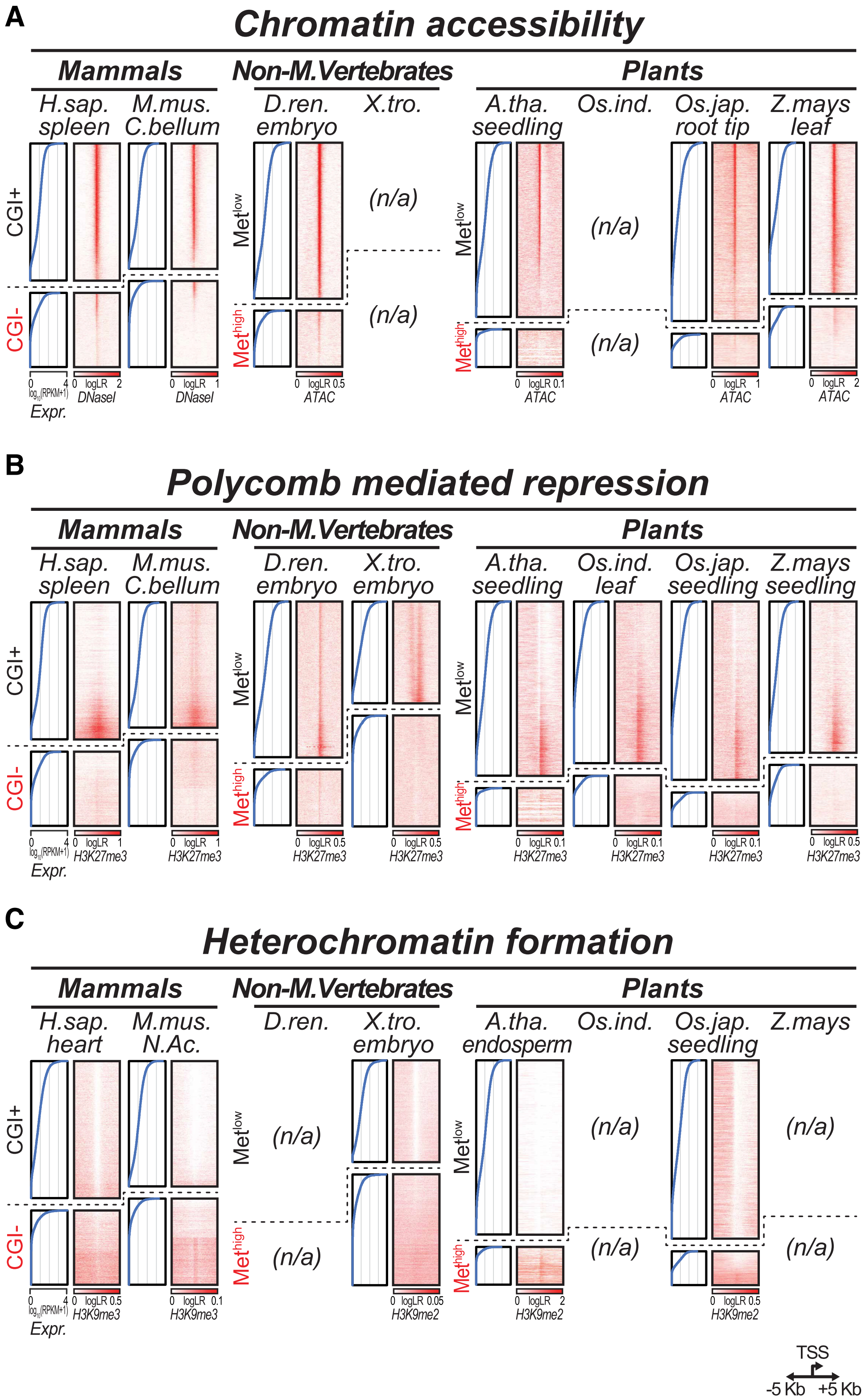


Figure 3. Distinct epigenetic gene regulations in two types of promoters. Chromatin accessibility (either ATAC-seq or DNase-seq, indicated below heatmaps) (**A**), *Polycomb*-mediated gene repression mark (H3K27me3) (**B**) and heterochromatin mark (H3K9me2/3) (**C**) signals at the 10 kb-surrounding regions of TSSs of two-types of genes are shown as heatmaps. Genes are sorted by their expression levels (left blue line plots) in the indicated cell/tissue types.

Order randomness test (Runs test)

To test whether the two types of genes we defined (CGI+/Met^{low} versus CGI–/Met^{high}) are linearly partitioned in the genome, we performed an order randomness test (Runs Test) (35) as we did previously (16). In detail, we simplified gene arrangement in each chromosome into binarized gene orders (CGI+ or CGI– in mammals, Met^{low} or Met^{high} in non-mammalian vertebrates and plants), and then counted the transition (runs) from one state to another in each chromosome (e.g. CGI+ to CGI– or CGI– to CGI+ in mammals; Met^{low} to Met^{high} or Met^{high} to Met^{low} in vertebrates and plants). These runs value were standardized into *Z*-score by comparing with mean (μ) and standard variation (σ) expected from randomized order (upper-left equations in Figure 4A, where n1 and n2 refer to the total number of CGI+/CGI− or Met^{low}/Met^{high} in each chromosome). As runs values (*Z*-score) form standardized normal distribution ($\mu = 0$, $\sigma = 1$) in completely randomized orders (35), *P*-values were calculated as the area-under-curve under the normal distribution.

Gene homology analysis (Basic Local Alignment Search Tool: BLAST)

To measure the level of homology between neighboring gene pairs in Figure 4B, we performed BLAST analysis using whole protein-coding genes in the genomes. Since the BLAST e-values are sensitive to the target database size, we fixed the *db* parameter as all protein-coding genes in the genome of each species, and performed all-by-all TBLASTX analysis using each single protein-coding gene as a *query*. Neighboring gene pairs with an *e*-value $<1 \times 10^{-10}$ were considered homologous pairs.

Hi-C data analysis

Hi-C data were listed (on 19 March 2020) and downloaded from Sequence Read Archive (SRA) from the National Center for Biotechnology Information (NCBI) database and Encyclopedia of DNA Elements (ENCODE). FASTQ files were extracted with the SRA Toolkit version 2.5.5, and aligned using Bowtie 2.3.2 to the reference genomes (30). For downstream analysis, hypergeometric optimization of motif enrichment (HOMER) 4.11.1 (36) was used with 50 kb resolution. Principal component analysis (PCA) was performed using runHiCpca.pl. All Hi-C data used in this study are summarized in Supplementary Table S4.

TF binding motif analysis

TF binding motifs that are enriched in one gene group (Met^{high}/CGI− or Met^{low}/CGI+ genes) over the other were identified in each species using analysis of motif enrichment (AME) 5.0.4 of the multiple expectation maximization for motif elicitation (MEME) Suite (37) with default parameters (Supplementary Table S5). Occurrence of these selected motifs at the 10 kb-surrounding regions of TSSs of two types of genes are shown as heatmaps (Figure 5D). The significances of the differences in the motif occurrences were tested using Wilcoxon signed-rank test. Known motif database was downloaded from HOMER Motif Database (‘HOMER Known Motifs’) (36).

Gene ontology analysis

Gene ontology data were downloaded from BioMart of Ensembl genome database (38). Gene overlaps between all Met^{high}/CGI− and Met^{low}/CGI+ genes of each species and gene ontology were first calculated using hypergeometric distribution as we previously described (16,17). Based on this result, we listed Gene Ontology that are enriched (*P*-values less than 1×10^{-15}) in any gene group/species. As the Gene Ontology annotation depths and qualities vary in various species, traditional overlap analyses (i.e. hypergeometric/binomial distribution, Fisher’s exact test) that are dependent to gene set size were not suitable for the fair evaluation of gene set overlaps. Therefore, we performed the permutation tests (100 times) and measured the overlaps between each gene group and Gene Ontology as the skewness from random expectation (*Z*-score).

RESULTS

Promoter CpG methylation analysis indicates conserved gene type divergence in vertebrates and plants

To test whether gene type divergence is common in non-mammalian eukaryotic species, we performed meta-analyses of genomic and epigenomic signatures. We first focused on the epigenetic modification mediated by cytosine methylation. Among four nucleotides constituting the eukaryotic genome, cytosine is the only nucleotide that can be methylated. Notably, the surrounding sequence context is important for cytosine methylation. For instance, except for rare cases (e.g. neuronal lineages (39,40)), only cytosines within CpG dinucleotides are globally methylated in mammals. In plants, virtually any cytosine in multiple different sequence contexts (usually classified as C(p)G, C(p)H(p)H, C(p)H(p)G, where H corresponds to A/T/C and (p) indicates phosphodiester bond between neighboring nucleotides) can be methylated (41,42).

We focused on the fact that mammalian CGI+ and CGI− genes exhibit distinct promoter CpG methylation patterns (Figure 1A) (16,17). As shown in Figure 1B, several non-mammalian eukaryotes (i.e. non-mammalian vertebrates & plants) also have global CpG methylation, but do not have CGIs in their genome. Therefore, we reasoned that promoter CpG methylation signatures in these species will allow us a glimpse into their gene regulatory mechanisms.

We specifically collected publicly available gene expression (mRNA-seq) and CpG methylation (WGBS-seq) data generated from same tissue and developmental stage of mammals, non-mammalian vertebrates & plants (Supplementary Tables S1 and S2), and monitored the relationships between promoter methylation and transcriptional activity of the associated genes. Two-dimensional plots in Figure 1C show the CpG methylation levels at human and mouse promoters as a function of the expression levels of associated genes. Among all genes of human and mouse (upper left plots in Figure 1C), the majority of highly active genes (i.e. RPKM > 100) are unmethylated in CpG

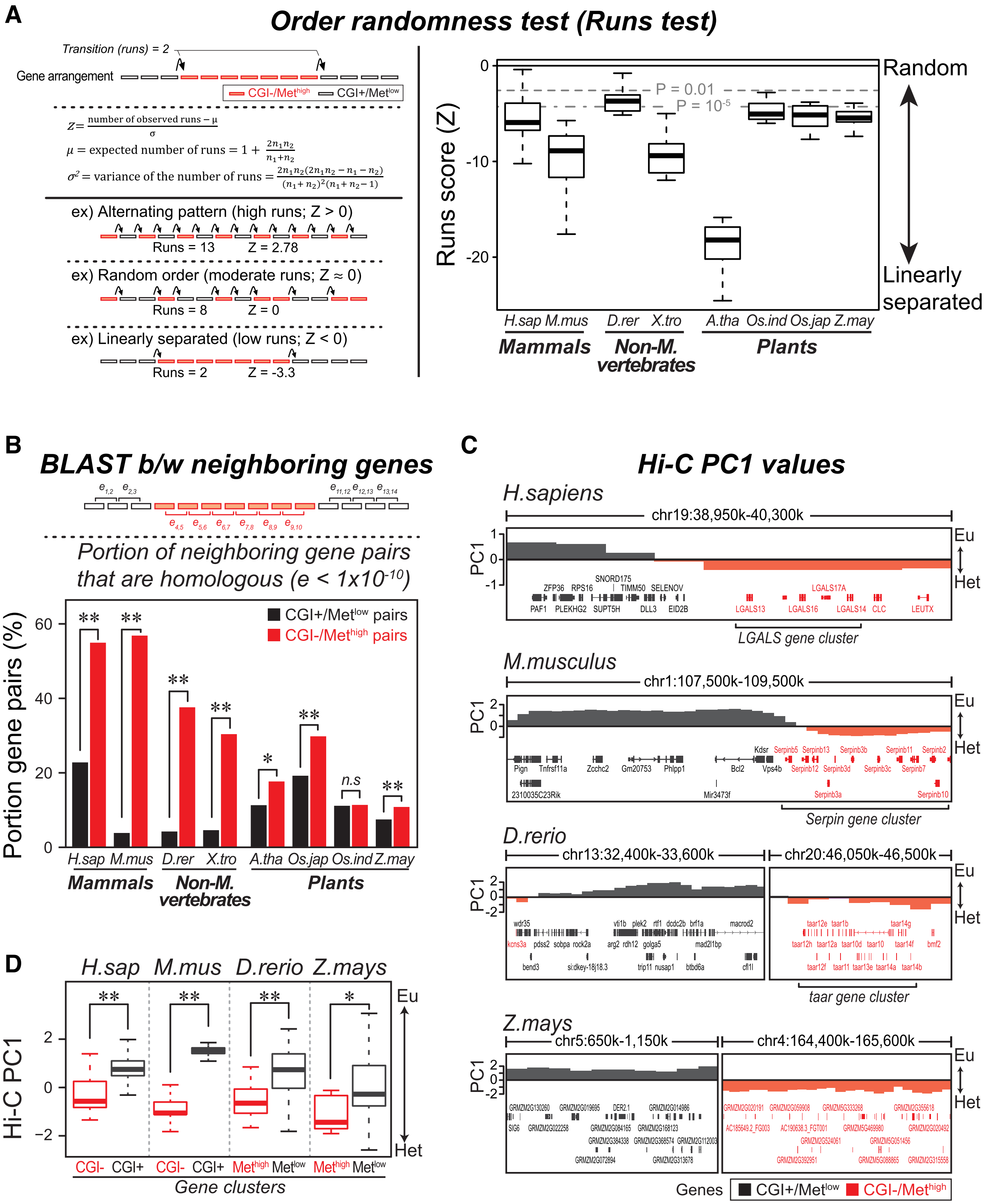


Figure 4. Genomic arrangement of the two types of promoters. (**A**) Order randomness test (Runs test) of CGI-/Met^{high} and CGI+/Met^{low} genes in each chromosome. Left panel shows a schematic representation of the Runs test (top) as well as examples of randomly arranged genes and well-organized genes (bottom). Runs test measures the occurrence of transitions from one gene type to another in linear order (e.g. from CGI–/Met^{high} to CGI+/Met^{low} or *vice versa*; recursive arrows in the left panel) that is further standardized (using the equation shown in the left panel, where *n*1 and *n*2 indicate the numbers of CGI–/Met^{high} and CGI+/Met^{low} genes, respectively; for the detail, see also Methods). Note that most chromosomes have negative runs score (Z), indicating a strong linear separation between CGI–/Met^{high} and CGI+/Met^{low}. (**B**) Homologous portion among all neighboring iso-type gene pairs (e.g. neighboring CGI–/Met^{high} gene pairs, or neighboring CGI+/Met^{low} gene pairs). Levels of homology among neighboring gene pairs were calculated as e-values from all-to-all protein-coding gene BLAST analyses (for detail, see Materials and Methods). Neighboring gene pairs with e-values less than 1 × 10^{-10} were considered as homologous gene pairs. (**C**) Genomic landscapes and Hi-C PC1 values in CGI−/Met^{high} (red) or CGI+/Met^{low} (black) gene clustered regions. (**D**) Hi-C PC1 values in CGI−/Met^{high} (red) or CGI+/Met^{low} (black) gene clustered regions. $^{**}P < 10^{-5}$, $^{*}P < 10^{-3}$, n.s: $P > 0.01$.

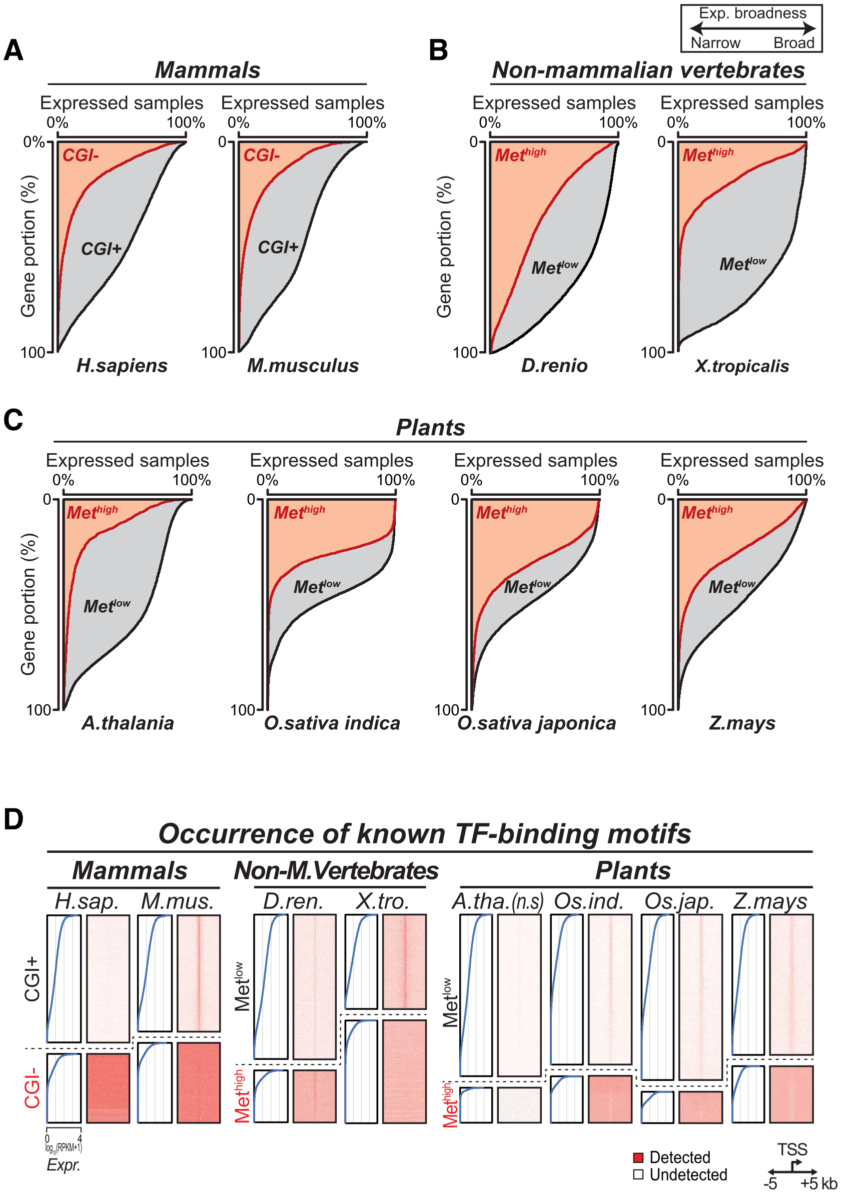


Figure 5. Distinct transcriptional regulation in two gene types. (**A–C**) Expression broadness measured as the portion of RNA-seq data expressing given genes over RPKM 0.5. Genes are sorted according to the expression broadness. (**D**) Occurrence for known TF binding motifs at the 10 kb-surrounding regions of TSSs of two gene types.

sites within their promoters. However, the promoters of inactive genes (i.e. RPKM < 0.1; blue boxes. See also the lower bar plots) show mixed methylation patterns; some are methylated, while the others are unmethylated. Interestingly, when these patterns are monitored separately, mammalian CGI+ and CGI– genes exhibit completely distinct promoter methylation patterns (center and right graphs in Figure 1C). As shown in our previous studies (11,16,17), CGI– promoters gain a high level of CpG methylation when transcriptionally inactive. On the other hand, CpG sites within CGI+ promoters remain unmethylated regardless of the transcriptional activities of associated genes. This shows that the mixed methylation levels in inactive promoters originate from the mixture of two types of promoters: with or without CGIs. These data also indicate that the promoter methylation analysis can be a valid approach in defining the types of gene promoters.

Based on these observations, we questioned whether the species lacking CGIs also have gene type divergence similar to mammals. Specifically, we investigated CpG methylation levels in the promoters of non-mammalian vertebrates and plants (Figure 1D, E). Interestingly, inactive promoters of both non-mammalian vertebrates (zebrafish and frog; Figure 1D) and plants (arabidopsis, two rice species, and maize; Figure 1E; for non-CpG methylation, see Supplementary Figure S1) can also be subdivided into two types, i.e. methylated and unmethylated. These data strongly suggest that the dual-mode of gene regulation we observed in mammals (Figure 1A) (16,17) may also exist in these non-mammalian eukaryotic species.

Gene type classification guided by integrative epigenome analyses.

Based on these observations, we defined the two types of promoters within the species lacking CGIs. For this, we again focused on the distinct CpG methylation patterns shown in mammalian two promoter types: i.e. CGI+ promoters that are constitutively unmethylated versus CGI– promoters changing the promoter methylation level during regulation. Unlike Figure 1C–E showing the methylation pattern in a single tissue/context, here, we collected all available methylation data (WGBS-seq) in each species generated from variable contexts (i.e. various cell/tissue types, developmental stages, diseases, and injuries, etc; Supplementary Table S2). Using this, we then measured each promoter's maximal CpG methylation levels in various conditions. As shown in Figure 2A, maximal CpG methylation levels of all promoters show clear bimodal distributions both in human and mouse. This bimodality clearly reflects the CGI-ness of the promoters; CGI– promoters can be methylated during their regulation (red; max CpG methylation ≈ 1), while most CGI+ genes are not engaged in methylation (black; max CpG methylation ≈ 0).

Figure 2B shows the cases in species that do not have CpG islands (i.e. non-mammalian vertebrates and plants). Interestingly, the maximal promoter CpG methylation levels in these species also show strong bimodalities: some promoters can be methylated (as in mammalian CGI– genes), while the others cannot (as in CGI+ genes). Here, we defined two classes of promoters within each of these species (i.e. Met^{high} (high methylation level, red) versus Met^{low} (low methylation level, black) in Figure 2B), and then further investigated the differences between them.

Heatmaps of Figure 2C show the global CpG methylation (5'me-CpG) ratio at the promoter surrounding regions. As shown in our previous study, mammalian CGI+ promoters are clearly unmethylated regardless of transcriptional activities of associated genes. On the other hand, CGI– promoters are preferentially methylated, and exhibit narrow demethylation only when they are transcriptionally active. As expected, our newly defined Met^{low} genes in CGI-lacking species are also unmethylated at the surrounding regions of the promoters at all level of transcriptional activities, which is similar to CGI+ genes. In contrast, promoters of Met^{high} genes are largely methylated both in non-mammalian vertebrates and plants (Figure 2C). This data shows that even for the species without clear CGIs, there exist two distinct types of genes that are regulated differently.

Sequence characteristics of Met^{low} and Met^{high} promoters.

Mammalian CpG islands are speculated to be the byproducts of cumulated CpG depletion within genome (20,43). CpG depletion, a process in which the genome loses CpG dinucleotides globally due to the irreversible deamination of methylated CpG, is often observed in species with global CpG methylation (Supplementary Figure S2). In human, cytosine and guanine each cover 20.4% of total nucleotides of the genome; thus, over 4% of total dinucleotides are expected to have CpG sequences (i.e. expected CpG frequency; $0.204 \times 0.204 = 0.0418$). However, the observed CpG frequency in human genome is only 0.98%, reflecting significant genome-wide depletion of CpG dinucleotides throughout generations (Supplementary Figure S3A). Interestingly, CpGs in CGIs constitutively remain unmethylated (Figure 2A, C), thus are protected from deamination of methylated CpG throughout generations and sustain relatively higher CpG frequencies compared to other parts of genome.

Here, we questioned whether our newly defined Met^{low} and Met^{high} genes also share similar sequence characteristics as their mammalian counterparts. Figure 2D shows the frequencies of CpG at the near surrounding regions of two types of promoters. As expected from the definition of CGIs (18,19), mammalian CGI+ promoters have dramatically higher CpG frequencies compared to surrounding regions (Figure 2D, left). On the other hand, CGI– promoters do not show significantly higher CpG frequencies compared to surrounding regions. These are well aligned with the CpG depletion level (i.e. observed/expected CpG frequency; Supplementary Figure S3B), indicating that the majority of mammalian genome, but not CGIs, have experienced serious CpG depletion (see also Supplementary Figures S2A and S3A, S3C).

Promoters of species lacking CGIs show strikingly different sequence characteristics. In non-mammalian vertebrates, only narrow surrounding regions of Met^{low} promoters show relatively milder increase of CpG sequences and lower CpG depletion levels compared to mammals (Figure 2D; see also Supplementary Figure S3B). This pattern is well aligned with the prior report demonstrating that CGI-like elements in non-mammalian vertebrates (non-methylated islands: NMIs) (44) tend to have less significant enrichment of CpG sequences compared to mammals. Plant genomes exhibit substantially less significant CpG depletion compared to vertebrates (Supplementary Figure S3B, lighter blue colors; see also Supplementary Figure S2A). Interestingly in plants, not only Met^{low} but also Met^{high} promoters have enriched CpG frequencies compared to surrounding regions (Figure 2D).

Our data imply several important facts underlying eukaryotic genome specializations and their gene regulatory mechanisms. Most of all, the unique sequence characteristics of CGIs could be formed in a specific environment that only mammalian genome has been exposed to (i.e. extensive CpG depletion along with high CpG methylation levels; Supplementary Figures S2 and S3). On the other hand, non-mammalian vertebrate and plant genomes have been much less CpG depletions; thus, even if they have DNA elements functioning like mammalian CGIs, they cannot have similar sequence characteristics as mammalian CGIs. Concomitantly, when DNA elements having the sequence characteristics of mammalian CGIs (i.e. GC content > 50%; length > 200 bp; $\text{CG}_{\text{observed/expected}} > 0.6$) are defined in non-mammalian species, they are generally not associated with gene promoters (Supplementary Figure S4). These data indicate that prior approaches to define CGI-like elements in plant genome based on sequence characteristics (45,46) were not suitable to detect the divergence of gene regulatory mechanisms.

Notably, most vertebrate cells have only one sequence context (CpG) where cytosines are globally methylated.

However, plants have three different methylatable sequence contexts (CG, CHG, CHH; Supplementary Figures S1 and S5) (41,42). Despite these specialized DNA sequence/methylation patterns distinct from mammals, plants still sustain two distinct types of promoters (Figure 2B). This implies that the divergence of promoter types had occurred before each species obtained specialized genomic/epigenomic characteristics, establishing dual-mode gene regulation as a fundamental feature of higher eukaryotic species.

Conserved dual-mode gene regulation in vertebrates and plants.

Based on these observations, we further investigated how those two types of promoters are regulated by distinct mechanisms. Figure 3A shows the chromatin accessibilities measured by DNase- or ATAC-seq. As we previously reported (16,17), mammalian CGI+ promoters sustain high chromatin accessibilities even when the associated genes are transcriptionally inactive (Figure 3A, left). On the other hand, CGI− promoters exhibit chromatin accessibilities only when the associated genes are actively expressed.

Met^{low} promoters both from non-mammalian vertebrates and plants also have higher level of chromatin accessibilities compared to Met^{high} promoters at all level of transcriptional activities (Figure 3A; $P < 0.001$ in all cases). Notably, this pattern is commonly observed in various cell/tissue types (Supplementary Figure S6A). High level of chromatin accessibility (Figure 3A) and unmethylated CpG (Figure 2A–C) are characteristics of euchromatin (3,4). Our data suggest that Met^{low} promoters would generally remain as euchromatin even when they are transcriptionally inactive, similar to their mammalian counterparts, CGI+ genes.

Our previous study showed that mammalian CGI+ genes are inactivated by *Polycomb* proteins, while CGI– genes are silenced by heterochromatin formation. To test whether Met^{low} and Met^{high} genes of non-mammalian vertebrates/plants are inactivated by similar mechanisms, we further analyzed published ChIP-seq datasets. As shown in Figure 3B, inactive Met^{low} promoters are enriched with H3K27me3, marks of *Polycomb*-mediated gene repression, while Met^{high} genes are depleted with H3K27me3 ($P < 2.2 \times 10^{-16}$ in all cases; see also Supplementary Figure S6B). These data imply that *Polycomb*-mediated gene repression is clearly limited to Met^{low} genes, but not to Met^{high} genes. On the other hand, H3K9me2/3, marks of heterochromatin, are largely enriched at the broad surrounding regions of inactive Met^{high} genes but not Met^{low} genes regardless of cell/tissue types, implying that heterochromatin formation mediated gene inactivation is limited to Met^{high} genes (Figure 3C and Supplementary Figure S6C; $P < 2.2 \times 10^{-16}$ in all cases).

Our comparative epigenome analyses clearly show an important logic in global gene regulation that is broadly conserved in higher eukaryotes with global DNA methylation. We demonstrate that all species we have tested in this study invariably have two types of genes: one group of genes is generally regulated by interconversion between euchromatin and heterochromatin [i.e. CGI– (mammals) and Met^{high} (non-mammalian vertebrates, plants) genes], while the other group preferentially forms euchromatin even when they are transcriptionally inactivated by *Polycomb* proteins (i.e. CGI+ and Met^{low} genes).

Genomic arrangement of Met^{low} and Met^{high} genes reflects their regulatory mechanisms.

An organism's current genome is the cumulative outcome of chromosomal mutation events that have occurred throughout generations (47), thus reflecting the intracellular environment that each gene has been exposed to (48). As our previous study demonstrated that the gene arrangement in mouse genome indeed reflects gene regulatory mechanisms (16), we further investigated how our newly defined Met^{low} and Met^{high} genes are arranged in the genome. Figure 4A shows the order randomness test (Runs test (35)) of Met^{low} and Met^{high} genes in each chromosome of the eight eukaryotic species. Our data demonstrate that Met^{low} and Met^{high} genes are linearly separated in most chromosomes; Met^{low} genes are clustered together in some part of chromosomes, while Met^{high} genes are also similarly concentrated in other part of chromosomes. CGI+ and CGI– genes in mammalian genomes also show similar pattern (Figure 4A) as we previously reported (16).

Figure 4B shows the neighboring gene homology analysis performed by BLAST (Basic Local Alignment Search Tool). In most eukaryotes tested, especially in mammals and vertebrates, neighboring Met^{high} (or CGI– in mammals) gene pairs tend to have a significantly higher chance to be homologous to each other when compared to Met^{low} (or CGI+) gene pairs. This shows that local gene duplication patterns are better sustained in CGI–/Met^{high} genes rather than CGI+/Met^{low} genes. Intriguingly, the differences in proportion of homologous gene pairs between Met^{low} and Met^{high} genes were much lower in the four plant species tested as well as being a lower proportion overall. This may reflect a lack of selective pressure to produce or preserve clusters of homologous genes in plants. Notably, the local gene duplication patterns are sequestered by chromosomal translocations (47,49), frequently occurring among spatially proximal regions (48,50). These data are well aligned with our observation that Met^{low} and Met^{high} (CGI+ and CGI– in mammals) genes are regulated by distinct mechanisms (Figure 3). Met^{low} (CGI+ in mammals) genes are generally placed at the interaction-prone environments (i.e. euchromatin), while Met^{high} (or CGI−) genes are generally located at interaction-depleted environments (i.e. heterochromatin).

We additionally tested whether this pattern is also supported by recent chromatin conformation analysis data. As Principal Component Analysis (PCA) of Hi-C data classifies the genome into euchromatin (high PC1) and heterochromatin (low PC1), we collected available Hi-C data and measured PC1 values in Met^{low} and Met^{high} (CGI+ and CGI– in mammals) gene clusters. As shown in Figure 4C and D, Met^{low} (or CGI+) gene clusters had significantly higher PC1 values compared to Met^{high} (or CGI−) gene clusters, demonstrating that Met^{low}/CGI+ genes generally form euchromatin while Met^{high}/CGI− genes form heterochromatin.

Distinct transcriptional regulation of Met$^{\text{low}}$ and Met$^{\text{high}}$ genes.

One of the most well characterized differences between mammalian CGI+ and CGI− genes is their expression pattern (17,19,20). CGI+ genes are broadly expressed throughout the body, while CGI– genes typically remain silent in most environments and are expressed in a context-dependent manner. In order to test whether their non-mammalian counterparts, Met$^{\text{low}}$ and Met$^{\text{high}}$ genes, also show similar expression patterns, we performed a large-scale gene expression analysis by integrating published mRNA-seq data generated from various tissues and developmental stages of each species (Supplementary Table S1).

Figure 5A shows gene expression broadness measured as the portion of samples expressing given genes (RPKM > 0.5). As discussed above, CGI+ genes exhibit significantly broader expression patterns than CGI– genes showing much narrower expression patterns ($P < 2.2 \times 10^{-16}$ in both human and mouse). In parallel to this pattern, Met$^{\text{low}}$ genes in non-mammalian vertebrates and plants also exhibit broader gene expression patterns compared to narrowly expressed Met$^{\text{high}}$ genes (Figure 5B, C; $P < 2.2 \times 10^{-16}$ in all cases).

Aligned with these distinct gene expression patterns, Met$^{\text{low}}$ and Met$^{\text{high}}$ genes also have distinct functions (GO (Gene Ontology) analysis; Supplementary Figure S7). Broadly-expressed Met$^{\text{low}}$ genes tend to encode proteins involved in general maintenance functions that are essential for every cell, such as the regulation of gene expression process (i.e. transcription, translation, modification, transport), cell cycle progression, and division (Supplementary Figure S7A, S7B), and this pattern is similar to mammalian CGI+ genes. On the other hand, genes involved in highly context-specific behaviors like cell-to-cell signaling and innate immune response are strongly enriched in Met$^{\text{high}}$ genes (Supplementary Figure S7A, S7B).

Figure 5D show the occurrence of known TF binding motifs at the promoter surrounding regions. Interestingly, mammalian CGI– promoters are largely enriched with TF binding sites compared to CGI+ genes. In parallel, Met$^{\text{high}}$ genes in most CGI-lacking species (except for *A. thaliana*) are also enriched with sequence-specific TF binding motifs compared to Met$^{\text{low}}$ genes ($P < 2.2 \times 10^{-16}$ in all cases). These data suggest that the context-dependent expression of Met$^{\text{high}}$ genes would be directly controlled by cell type- and stage-specific TF bindings. This is also supported by the data shown in Supplementary Figure S8 and Table S5, where the majority of cell/tissue type- and stage-specific TF binding motifs are commonly enriched in Met$^{\text{high}}$ (and mammalian CGI–) promoters. However, Met$^{\text{low}}$ promoters are depleted with TF binding motifs indicating that these genes are primarily regulated by the mechanisms beyond TF-mediated transcriptional regulations. These data are well-aligned with our prior observation made in mammals (16): traditional TF-mediated transcriptional regulation mechanism (i.e. local enhancer-loop model) explains the regulation of CGI− genes, not CGI+ genes, where the expressions are controlled by additional layer of regulations (i.e. long-range chromatin interactions (16) and PolII preloading/pausing/releasing (17,51)).

Interestingly, *A. thaliana* genome has unique characteristics that are distinct from other plants. Unlike other plant species, in *A. thaliana*, Met$^{\text{high}}$ genes cover only a very small portion of the entire genome (5.3%; compared to 16.1%/22.2% in *O. sativa japonica*/*indica* and 26.7% in *Zea mays*; see Figures 1E and 2B). Moreover, over half of these Met$^{\text{high}}$ genes are detected at centromeric/pericentromeric regions (i.e. 54.8% are within 3 Mb surrounding centromeres; the corresponding portion in Met$^{\text{low}}$ genes are only 9.1%; $P < 1 \times 10^{-10}$). This pattern is also observed as a more dramatic linear separation between Met$^{\text{low}}$ and Met$^{\text{high}}$ genes compared to other plants (Figure 4A). In addition, Met$^{\text{high}}$ genes in *A. thaliana* do not have a large number of TF binding motifs as in other plants (Figure 5D). These data indicate that the gene regulation program in *A. thaliana* has diverged from other plant species and specialized into the unique direction using *Polycomb*-mediated gene inactivation as a major gene inactivation mechanism. However, this does not exclude alternate hypotheses, such as the possibility that TFs in *A. thaliana* may have less binding selectivity for their canonical motifs. Despite these species-specific characteristics, all species we tested here still have common gene type divergence. These results emphasize that the dual-mode gene regulation program is a fundamental feature of eukaryotic gene regulation established long before the genome specialization of each species.

Conserved divergence in gene regulation mechanisms of *D. melanogaster* and *C. elegans*.

Other than the species tested thus far, *C. elegans* and *D. melanogaster* are broadly used non-vertebrate animal models and lack CGIs in their genome. Since they have heterochromatin/euchromatin distinction and have *Polycomb* machineries (Figure 1B), we further questioned whether the dual-mode gene regulation is also conserved in these non-vertebrate animals. Due to the lack of global DNA methylation, promoter classification by CpG methylation patterns (Figure 2A, B) were not applicable in these species. Therefore, we collected ChIP-seq data showing the signs of *Polycomb*-mediated gene repression (H3K27me3) and heterochromatin-mediated gene silencing (H3K9me2/3) marks generated from various tissues/contexts of *C. elegans* and *D. melanogaster* (Supplementary Table S4). We then measured the maximum *Polycomb* and heterochromatin signals in each promoter (Figure 6). Interestingly, our data show clear L-shaped distributions both in *C. elegans* and *D. melanogaster*. This is similar with the case in human (Figure 6A). These L-shaped distributions show an important underlying logic in the gene regulation within these species: *Polycomb*-mediated gene repression and heterochromatin-mediated gene silencing are mutually incompatible: some genes are programmed to be regulated by *Polycomb* proteins, while others are regulated by heterochromatin formation, and these distinctions are not interchangeable. These results show that the divergence of gene regulatory mechanisms that we observed in vertebrates and plants (Figures 1–5) also exist even in the animal models lacking global DNA methylation and CGI elements.

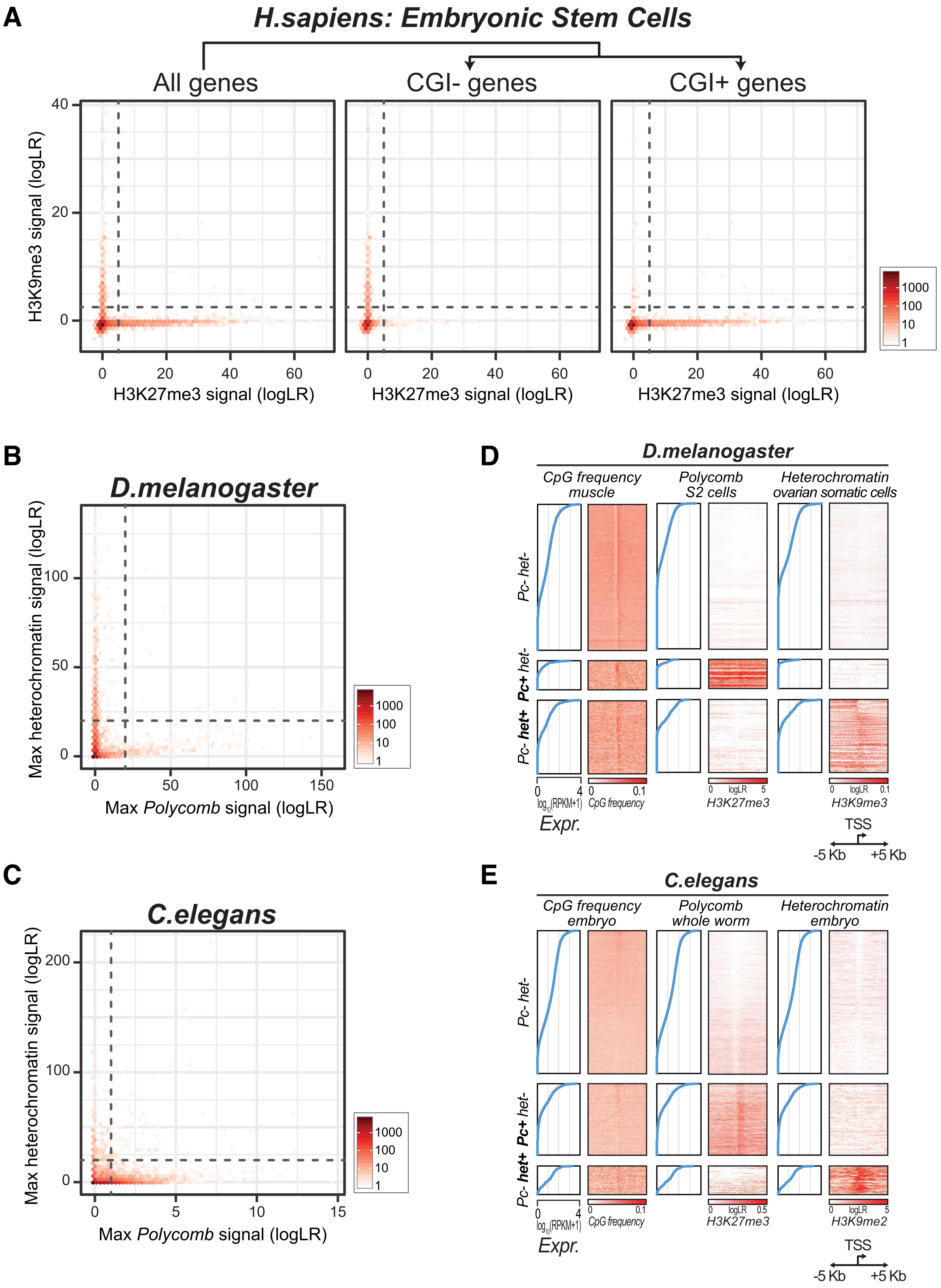


Figure 6. Mode of gene regulations in model animals lacking global CpG methylation. (**A**) Two-dimensional plots showing H3K27me3 and H3K9me3 signals at the promoters of protein-coding genes in human embryonic stem cells. Note that *Polycomb* (x-axis, H3K27me3) and heterochromatin (y-axis, H3K9me3) marks form L-shaped distribution, reflecting their mutual exclusiveness. (**B**, **C**) Two-dimensional plots showing maximum *Polycomb* (x-axis: H3K27me3) and heterochromatin (y-axis: H3K9me2/3) mark signals at the promoters of protein-coding genes in fruit fly (B) and nematode (C). (**D**, **E**) CpG frequencies, H3K27me3, and H3K9me2 signals at the 10 kb-surrounding regions of TSSs of genes are shown as heatmaps.

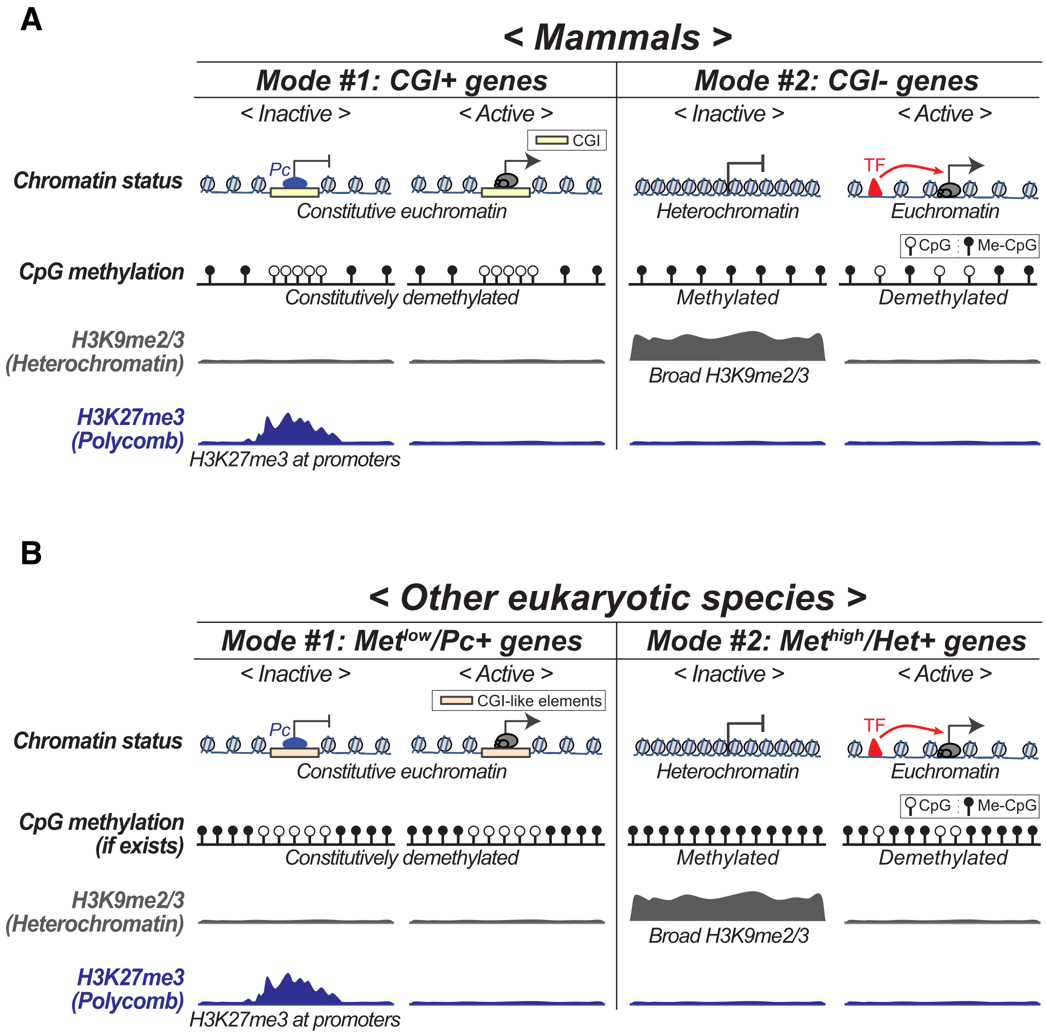


Figure 7. Models of dual-mode gene regulation in mammals (**A**) and non-mammalian eukaryotes (**B**). Note that only mammalian genomes have experienced strong CpG depletion, thus CGIs have unique sequence characteristics distinguishable from other part of genome. However, species lacking CGIs also have conserved dual-mode gene regulation programs.

DISCUSSION

As summarized in Figure 7, our study clearly demonstrates that the dual-mode gene regulation is a shared trait across different higher eukaryotes. One class of genes is regulated by inter-conversion between heterochromatin and euchromatin, and is expressed and functions in context-dependent manners (i.e. mammalian CGI– and vertebrate/plant Met^{high} genes). The other class of genes remains as euchromatin even when repressed by *Polycomb* proteins and exhibits broad expression patterns and have general functions essential for all cell/tissue types (CGI+ and Met^{low} genes).

We also show that some species have gained specialized characteristics distinct from others. In *C. elegans*, H3K9me3 has specialized function different from H3K9me2: H3K9me3 co-localizes with *Polycomb* mark (H3K27me3) rather than heterochromatin (H3K9me2) (33,34). Vertebrates have only one sequence context (CG) where cytosines are globally methylated, unlike plants with three different methylatable sequence contexts (CG, CHG, CHH; where H corresponds to A, T or C) (41,42). *A. thaliana* have unique TF binding motif occurrence patterns distinct from other plants and vertebrates (Figure 5D). Unlike vertebrates and plants, *C. elegans* and *D. melanogaster* do not have genome-wide global CpG methylations. Notably, all these eukaryotes with distinct characteristics consistently sustain dual-mode gene regulation programs. This indicates that this divergence of gene regulation mechanism is a fundamental and important characteristic of multicellular eukaryote systems that have been established long before each species to have specialized genome characteristics.

One question here is what would be of benefit to having two independent gene regulation programs within an organism. It is noteworthy here that the genes expressed in context-dependent manners (i.e. mammalian CGI– and vertebrate/plant Met^{high} genes in Figure 5) tend to fulfill tissue-specific functions rather than more universal biological processes. Our study shows that these silent genes generally form condensed heterochromatin (Figure 3C), which was often shown to be tethered to the nuclear periphery in various species (5,16,48,52–54). It is obvious that the spatial separation of these rarely expressed genes away from active nuclear compartments is beneficial for suppressing their misexpression, as well as for saving energy and resources required for their maintenance. Our data also indicate that linear gene arrangement is optimized for the efficient spatial segregation between distinct types of genes (i.e. co-clustering of same type of genes & linear separation between distinct gene types; Figure 4).

On the other hand, CGI+/Met^{low} genes are broadly expressed (Figure 5A) and are responsible for the general functions that are essential for every cell/tissue, or for the regulation of organism development (Supplementary Figure 7). Consistently, our previous study indicates that the strict regulation of CGI+ gene expression levels is critical for mammalian survival and development (17). Our data also demonstrated that the transcriptional activities of mammalian CGI+ genes are more tightly regulated by fine balancing of activators and repressors occupied on CGIs (e.g. MYC and PRC class proteins (16,17)). Similarly, activities of developmental genes in fruit fly (55–57) and plants (58,59) are also sharply controlled by balancing of activators and repressors (i.e. *Trithorax* and *Polycomb* group proteins) bound on TRE/PRE (*Trithorax*/*Polycomb* responsive elements).

Another question is what the determinant of the types of gene regulation is. One of the plausible culprits is ZF-CxxC (zinc finger-CxxC) domain-containing proteins (24,60–62). In mammals and non-mammalian vertebrates, CxxC domain-containing proteins constantly bind to unmethylated CGIs (or NMIs in non-mammalian vertebrates (44)) in tissue-/cell type-independent manners (44,60,61). Along with the binding of these proteins, CGIs/NMIs remain unmethylated even when repressed, which is a key characteristic of euchromatin. In addition, integration of artificial CGI-like DNA sequence into a gene desert in mouse embryonic stem cells successfully recapitulated the binding of CxxC domain proteins, as well as the formation of CGI-like epigenome (i.e. CpG demethylation and gain of H3K27me3) (63). These indicate that, in mammals, the mere presence of CGIs is enough for a gene to be bound by CxxC proteins that, in turn, enable a distinct gene regulation program from the genes lacking CGIs. Notably, CxxC domain-containing proteins are broadly conserved in multi-cellular eukaryotic organisms that were shown to have dual-mode gene regulation programs in this study (mammals (61), non-mammalian vertebrates (44), plants (42), *D. melanogaster* (64) and *C. elegans* (65,66)). Identification of CxxC-binding target regions in these species will allow a deeper understanding of shared eukaryotic gene regulatory mechanisms.

DATA AVAILABILITY

Accession numbers of the data used in this paper are summarized in Supplementary Tables S1, S2 and S4.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate Dr Gregory J. Zynda for sharing the codes for methylation data analysis. We thank the Texas Advanced Computing Center (TACC) for providing HPC resources.

FUNDING

National Institutes of Health [P20GM0103423, P20GM104318 to S.B., R01GM112722 to J.K.]. Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Henikoff,S. (2000) Heterochromatin function in complex genomes. *Biochim. Biophys. Acta - Rev. Cancer*, **1470**, O1–O8.
2. Hall,I.M., Shankaranarayana,G.D., Noma,K., Ayoub,N., Cohen,A. and Grewal,S.I.S. (2002) Establishment and maintenance of a heterochromatin domain. *Science*, **297**, 2232–2237.
3. Straub,T. (2003) Heterochromatin dynamics. *PLoS Biol.*, **1**, E14.
4. Grewal,S.I.S. and Jia,S. (2007) Heterochromatin revisited. *Nat. Rev. Genet.*, **8**, 35–46.
5. Padeken,J. and Heun,P. (2014) Nucleolus and nuclear periphery: Velcro for heterochromatin. *Curr. Opin. Cell Biol.*, **28**, 54–60.
6. Aebi,U., Cohn,J., Buhle,L. and Gerace,L. (1986) The nuclear lamina is a meshwork of intermediate-type filaments. *Nature*, **323**, 560–564.
7. van Steensel,B. and Belmont,A.S. (2017) Lamina-Associated Domains: Links with chromosome architecture, heterochromatin, and gene repression. *Cell*, **169**, 780–791.
8. Akhtar,A. and Gasser,S.M. (2007) The nuclear envelope and transcriptional control. *Nat. Rev. Genet.*, **8**, 507–517.
9. Muller,H.J. (1930) Types of visible variations induced by X-rays in Drosophila. *J. Genet.*, **22**, 299–334.
10. Beisel,C. and Paro,R. (2011) Silencing chromatin: Comparing modes and mechanisms. *Nat. Rev. Genet.*, **12**, 123–135.
11. Beck,S., Lee,B.K. and Kim,J. (2015) Multi-layered global gene regulation in mouse embryonic stem cells. *Cell. Mol. Life Sci.*, **72**, 199–216.
12. Boyer,L.A., Plath,K., Zeitlinger,J., Brambrink,T., Medeiros,L.A., Lee,T.I., Levine,S.S., Wernig,M., Tajonar,A., Ray,M.K. *et al.* (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349–353.
13. Simon,J.A. and Kingston,R.E. (2009) Mechanisms of Polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.*, **10**, 697–708.
14. Dumesic,P.A., Homer,C.M., Moresco,J.J., Pack,L.R., Shanle,E.K., Coyle,S.M., Strahl,B.D., Fujimori,D.G., Yates,J.R. and Madhani,H.D. (2015) Product binding enforces the genomic specificity of a yeast Polycomb repressive complex. *Cell*, **160**, 204–218.
15. Morey,L. and Helin,K. (2010) Polycomb group protein-mediated repression of transcription. *Trends Biochem. Sci.*, **35**, 323–332.
16. Beck,S., Rhee,C., Song,J., Lee,B.K., LeBlanc,L., Cannon,L. and Kim,J. (2018) Implications of CpG islands on chromosomal architectures and modes of global gene regulation. *Nucleic Acids Res.*, **46**, 4382–4391.
17. Beck,S., Lee,B.K., Rhee,C., Song,J., Woo,A.J. and Kim,J. (2014) CpG island-mediated global gene regulatory modes in mouse embryonic stem cells. *Nat. Commun.*, **5**, 5490.
18. Gardiner-Garden,M. and Frommer,M. (1987) CpG Islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
19. Bird,A., Taggart,M., Frommer,M., Miller,O.J. and Macleod,D. (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, **40**, 91–99.
20. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.
21. Mendenhall,E.M., Koche,R.P., Truong,T., Zhou,V.W., Issac,B., Chi,A.S., Ku,M. and Bernstein,B.E. (2010) GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet.*, **6**, e1001244.
22. Blackledge,N.P., Rose,N.R. and Klose,R.J. (2015) Targeting Polycomb systems to regulate gene expression: Modifications to a complex story. *Nat. Rev. Mol. Cell Biol.*, **16**, 643–649.

23. Craig,J.M. and Bickmore,W.A. (1994) The distribution of CpG islands in mammalian chromosomes. *Nat. Genet.*, **7**, 376–382.
24. Illingworth,R.S., Gruenewald-Schneider,U., Webb,S., Kerr,A.R.W., James,K.D., Turner,D.J., Smith,C., Harrison,D.J., Andrews,R. and Bird,A.P. (2010) Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. *PLoS Genet.*, **6**, e1001134.
25. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
26. Ramsköld,D., Wang,E.T., Burge,C.B. and Sandberg,R. (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, **5**, e1000598.
27. Song,J., Zynda,G., Beck,S., Springer,N.M. and Vaughn,M.W. (2016) Bisulfite sequence analyses using CyVerse discovery environment: from mapping to DMRs. *Curr. Protoc. Plant Biol.*, **1**, 510–529.
28. Xi,Y. and Li,W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 232.
29. Meissner,A., Mikkelsen,T.S., Gu,H., Wernig,M., Hanna,J., Sivachenko,A., Zhang,X., Bernstein,B.E., Nusbaum,C., Jaffe,D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.
30. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
31. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
32. Jain,D., Baldi,S., Zabel,A., Straub,T. and Becker,P.B. (2015) Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res.*, **43**, 6959–6968.
33. Bessler,J.B., Andersen,E.C. and Villeneuve,A.M. (2010) Differential localization and independent acquisition of the H3K9me2 and H3K9me3 chromatin modifications in the Caenorhabditis elegans adult germ line. *PLoS Genet.*, **6**, e1000830.
34. Camacho,J., Truong,L., Kurt,Z., Chen,Y.W., Morselli,M., Gutierrez,G., Pellegrini,M., Yang,X. and Allard,P. (2018) The memory of environmental chemical exposure in *C. elegans* is dependent on the Jumonji demethylases jmjd-2 and jmjd-3/utx-1. *Cell Rep.*, **23**, 2392–2404.
35. Wald,A. and Wolfowitz,J. (1940) On a test whether two samples are from the same population. *Ann. Math. Stat.*, **11**, 147–162.
36. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of Lineage-Determining transcription factors prime cis-Regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
37. McLeay,R.C. and Bailey,T.L. (2010) Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**, 165.
38. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
39. Lister,R., Mukamel,E.A., Nery,J.R., Urich,M., Puddifoot,C.A., Johnson,N.D., Lucero,J., Huang,Y., Dwork,A.J., Schultz,M.D. *et al.* (2013) Global epigenomic reconfiguration during mammalian brain development. *Science*, **341**, 1237905.
40. Varley,K.E., Gertz,J., Bowling,K.M., Parker,S.L., Reddy,T.E., Pauli-Behn,F., Cross,M.K., Williams,B.A., Stamatoyannopoulos,J.A., Crawford,G.E. *et al.* (2013) Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.*, **23**, 555–567.
41. Feng,S., Cokus,S.J., Zhang,X., Chen,P.Y., Bostick,M., Goll,M.G., Hetzel,J., Jain,J., Strauss,S.H., Halpern,M.E. *et al.* (2010) Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 8689–8694.
42. Iyer,L.M., Abhiman,S. and Aravind,L. (2011) Natural history of eukaryotic DNA methylation systems. *Progr. Mol. Biol. Transl. Sci.*, **101**, 25–104.
43. Bird,A.P. (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.*, **8**, 1499–1504.
44. Long,H.K., Sims,D., Heger,A., Blackledge,N.P., Kutter,C., Wright,M.L., Grützner,F., Odom,D.T., Patient,R., Ponting,C.P. *et al.* (2013) Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife*, **2013**, e00348.
45. Ashikawa,I. (2001) Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. *Plant J.*, **26**, 617–625.
46. Antequera,F. and Bird,A.P. (1988) Unmethylated CpG islands associated with genes in higher plant DNA. *EMBO J.*, **7**, 2295–2299.
47. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 11484–11489.
48. Bickmore,W.A. and Teague,P. (2002) Influences of chromosome size, gene density and nuclear position on the frequency of constitutional translocations in the human population. *Chromosom. Res.*, **10**, 707–715.
49. Tillier,E.R.M. and Collins,R.A. (2000) Genome rearrangement by replication-directed translocation. *Nat. Genet.*, **26**, 195–197.
50. Branco,M.R. and Pombo,A. (2006) Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.*, **4**, 780–788.
51. Kellner,W.A., Bell,J.S.K. and Vertino,P.M. (2015) GC skew defines distinct RNA polymerase pause sites in CpG island promoters. *Genome Res.*, **25**, 1600–1609.
52. Matheson,T.D. and Kaufman,P.D. (2016) Grabbing the genome by the NADs. *Chromosoma*, **125**, 361–371.
53. Cremer,T., Kurz,A., Zirbel,R., Dietzel,S., Rinke,B., Schrock,E., Speicher,M.R., Mathieu,U., Jauch,A., Emmerich,P. *et al.* (1993) Role of chromosome territories in the functional compartmentalization of the cell nucleus. *Cold Spring Harbor Symp. Quant. Biol.*, **58**, 777–792.
54. Cremer,T. and Cremer,C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292–301.
55. Ringrose,L. and Paro,R. (2007) Polycom/Trithorax response elements and epigenetic memory of cell identity. *Development*, **134**, 223–232.
56. Geisler,S.J. and Paro,R. (2015) Trithorax and polycomb group-dependent regulation: A tale of opposing activities. *Dev.*, **142**, 2876–2887.
57. Kassis,J.A., Kennison,J.A. and Tamkun,J.W. (2017) Polycomb and trithorax group genes in drosophila. *Genetics*, **206**, 1699–1725.
58. Alvarez-Venegas,R. (2010) Regulation by Polycomb and Trithorax Group Proteins in Arabidopsis. *Arab. B.*, **8**, e0128.
59. Deng,W., Buzas,D.M., Ying,H., Robertson,M., Taylor,J., Peacock,W.J., Dennis,E.S. and Helliwell,C. (2013) Arabidopsis polycomb repressive complex 2 binding sites contain putative GAGA factor binding motifs within coding regions of genes. *BMC Genomics*, **14**, 593.
60. Shin Voo,K., Carlone,D.L., Jacobsen,B.M., Flodin,A. and Skalnik,D.G. (2000) Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and Methyl-CpG binding domain protein 1. *Mol. Cell. Biol.*, **20**, 2108–2121.
61. Long,H.K., Blackledge,N.P. and Klose,R.J. (2013) ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochem. Soc. Trans.*, **41**, 727–740.
62. Blackledge,N.P., Thomson,J.P. and Skene,P.J. (2013) CpG island chromatin is shaped by recruitment of ZF-CxxC proteins. *Cold Spring Harb. Perspect. Biol.*, **5**, a018648.
63. Wachter,E., Quante,T., Merusi,C., Arczewska,A., Stewart,F., Webb,S. and Bird,A. (2014) Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *Elife*, **3**, e03397.
64. Ardehali,M.B., Mei,A., Zobeck,K.L., Caron,M., Lis,J.T. and Kusch,T. (2011) Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription. *EMBO J.*, **30**, 2817–2828.
65. Pokhrel,B., Chen,Y. and Biro,J.J. (2019) CFP-1 interacts with HDAC1/2 complexes in C. elegans development. *FEBS J.*, **286**, 2490–2504.
66. Beurton,F., Stempor,P., Caron,M., Appert,A., Dong,Y., Chen,R.A.J., Cluet,D., Couté,Y., Herbette,M., Huang,N. *et al.* (2019) Physical and functional interaction between SET1/COMPASS complex component CFP-1 and a Sin3S HDAC complex in *C. elegans*. *Nucleic Acids Res.*, **47**, 11164–11180.