4382–4391 *Nucleic Acids Research, 2018, Vol. 46, No. 9* *Published online 26 February 2018*
doi: 10.1093/nar/gky147

Implications of CpG islands on chromosomal architectures and modes of global gene regulation

Samuel Beck[1,2,*], Catherine Rhee[1], Jawon Song[3], Bum-Kyu Lee[1], Lucy LeBlanc[1], Laurie Cannon[1] and Jonghwan Kim[1,4,5,*]

[1]Department of Molecular Biosciences, The University of Texas at Austin, Austin, Texas 78712, USA, [2]Kathryn W. Davis Center for Regenerative Biology and Medicine, MDI Biological Laboratory, Bar Harbor, Maine 04609, USA, [3]Texas Advanced Computing Center, The University of Texas at Austin, Austin, Texas 78712, USA, [4]Institute for Cellular and Molecular Biology, The University of Texas at Austin, Austin, Texas 78712, USA and [5]Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, Texas 78712, USA

Received November 2, 2017; Revised January 18, 2018; Editorial Decision February 15, 2018; Accepted February 15, 2018

ABSTRACT

CpG islands (CGIs) have long been implicated in the regulation of vertebrate gene expression. However, the involvement of CGIs in chromosomal architectures and associated gene expression regulations has not yet been thoroughly explored. By combining large-scale integrative data analyses and experimental validations, we show that CGIs clearly reconcile two competing models explaining nuclear gene localizations. We first identify CGI-containing (CGI+) and CGI-less (CGI−) genes are non-randomly clustered within the genome, which reflects CGI-dependent spatial gene segregation in the nucleus and corresponding gene regulatory modes. Regardless of their transcriptional activities, CGI+ genes are mainly located at the nuclear center and encounter frequent long-range chromosomal interactions. Meanwhile, nuclear peripheral CGI− genes forming heterochromatin are activated and internalized into the nuclear center by local enhancer–promoter interactions. Our findings demonstrate the crucial implications of CGIs on chromosomal architectures and gene positioning, linking the critical importance of CGIs in determining distinct mechanisms of global gene regulation in three-dimensional space in the nucleus.

INTRODUCTION

Three-dimensional (3D) chromosomal architecture plays critical roles in mammalian gene regulation. One model explaining nuclear gene localization is the chromosomal territory-extrusion model (***CT-extrusion model***) (1). In this model, each chromosome occupies its own territory, mainly located at the nuclear periphery and genes located in their own chromosome's territory form silent heterochromatin. Activation of genes occurs through the internalization of genes into the nuclear center in concert with the conversion to euchromatin. Alternatively, other studies introduced a concept of nuclear subcompartments with specialized functions (***nuclear subcompartment model***) (2,3). According to this model, the nucleus contains hundreds of transcription factories where active transcriptional machineries are focally enriched, while *Polycomb* repressive complex (PRC) proteins aggregate to form *Polycomb* bodies. It was shown that genes remain repressed when they are located within *Polycomb* bodies (3,4) while localizing to the transcription factories when activated (2,5,6). However, it has long been overlooked that not only transcription factories (2,7) but also *Polycomb* bodies (3,8,9) are mainly detected within the nuclear center, rather than the periphery. Therefore, in terms of location of inactive genes, the two models explaining chromosomal architectures, *CT-extrusion model* and *nuclear subcompartment model,* are mutually incompatible, and neither of them can explain the general localization behaviors of the all genes within the genome. In parallel, it is unclear which genes are inactivated by heterochromatin formation and which genes are repressed by PRC.

To resolve these contradictions, we performed a large-scale integrative data analysis, particularly focusing on the implications of CpG islands (CGIs) in 3D chromosomal architectures. CGIs, originally defined based on the sequence characteristics of high-GC contents and CpG-dinucleotide frequencies (10–12), have been recently recognized as hotspots for global gene regulation (13–16). In mammalian genome, ∼60% of genes have CGIs near their promoters (CGI+ genes) while the other 40% do not (CGI− genes) (15,16). However, how CGI+ and CGI− genes are organized within nucleus has not been well understood.

[*]To whom correspondence should be addressed. Tel: +1 207 288 9880 (Ext 476); Fax: +1 207 288 2130; Email: sbeck@mdibl.org
Correspondence may also be addressed to Jonghwan Kim. Tel: +1 512 232 8046; Fax: +1 512 471 1218; Email: jonghwankim@mail.utexas.edu

MATERIALS AND METHODS

Definition of CGI± and CGI− genes

CGI-containing (CGI+) and CGI-less (CGI−) genes were defined as follows. We used experimentally validated CGI elements identified by CxxC-affinity purification followed by parallel sequencing (CAP-seq) data (17). 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-seq, DamID-seq, MeDIP-seq and DNaseI-seq data analysis section in 'Material and Methods' section), and identified non-tissue specific consensus CxxC-domain binding regions (listed in Supplementary Table S3). 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 S4).

3D DNA-FISH (fluorescence *in situ* hybridization)

Fluorescence *in situ* hybridization (FISH) was performed as previously described using the Oligopaint technique (18) with modifications for the identification of exact 3D location of target loci. Fluorescence labeled FISH probe libraries were designed as either ssDNA 36mers (ATTO-550, Figure 2E and Supplementary Movie S1) or ssDNA 45mers (ATTO-550 and ATTO-488, Figure 2F) and synthesized by MYcroarray (Ann Arbor, MI, USA). Cells grown adherently were suspended by trypsinization and fixed using 4% PFA. To avoid loss of cells during the solution exchange and washing steps, suspended cells were stained with trypan blue whenever necessary. Single locus detection (Figure 2E) was done with 20 pmol probes in hybridization cocktail with 2× saline-sodium citrate buffer with 0.1% tween-20 (SSCT), 50% formamide, 10% (w/v) dextran sulfate and 10 μg RNase A. The target region for the CGI− gene cluster (Myosin heavy chain cluster in Figure 2E) was chr11:66,977,423–67,174,410 (n = 3,157), and the target region for the CGI+ cluster (Rbm24/Cap2 cluster in Figure 2E) was chr13:46,483,276–46,661,642 (n = 2,798). On the other hand, multiple loci detection (Figure 2F, ∼20 000 oligo library, target regions are listed in Supplementary Table S1) was done with 200 pmol probes in 2× SSCT, 50% formamide, 30% (w/v) dextran sulfate and 10 μg RNase A. After staining, cells were washed and resuspended in antifade mounting medium, and 10 μl of the cell resuspension was dropped onto a glass slide and gently covered by an 18 × 18mm coverslip. Nuclei were imaged with Zeiss LSM 710 Confocal Microscope with Z-stacks. To determine localization, loci detected within 10% of the longest diameter from nucleus periphery in reconstituted 3D-view images (for example, see Supplementary Movie S1) were considered as peripheral loci, while the rest were considered to be located at the nuclear center.

Selection of high-quality H3K9me2/3 ChIP-seq data

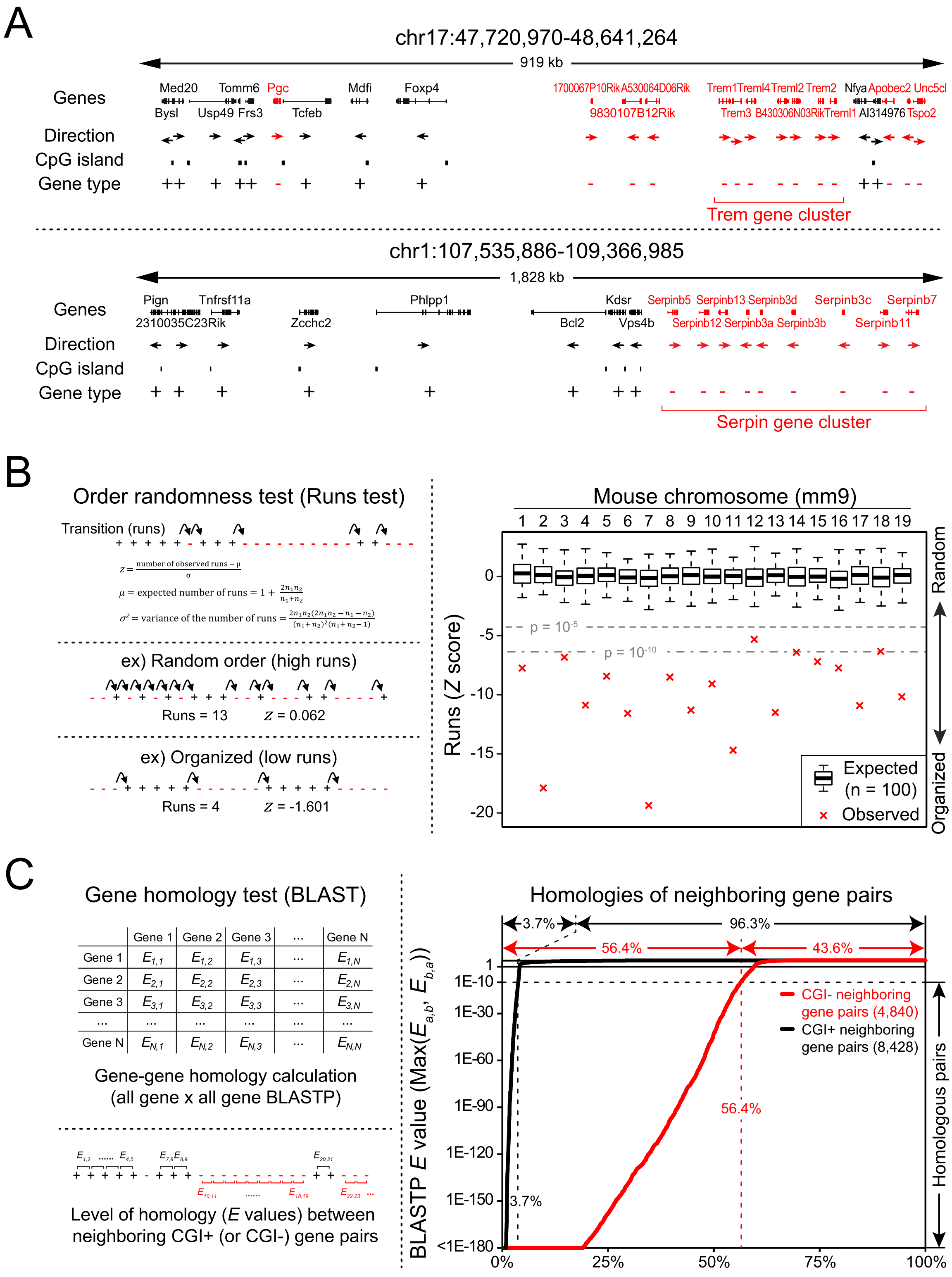
One of the most critical issues in ChIP-seq data analyses of heterochromatin regions is the contamination of active genomic regions, such as the promoters of housekeeping genes, during the ChIP process (19), which form high false positive signals. To circumvent this issue, we assessed H3K9me2/3 ChIP-seq data qualities in the published ChIP-seq data repository. Available mouse H3K9me2/3 ChIP-seq data deposited at the GEO website were listed and downloaded at 25 October 2015 and 29 September 2017 (total 287 experiments). Among them, the experiments done with in-house generated antibodies (12 experiments), experiments lacking antibody information (12 experiments), the experiments without author-provided control experiments (input or mock ChIP; 35 experiments) were not used for the further analyses. The quality of the remaining 228 experiments was assessed by two criteria: high signal to noise ratio (SNR) (16) and low false positive signal in active genomic regions. First, to assess SNR, H3K9me2/3 peaks of each experiment were called using histone modified region identification pipeline ('–nomodel –nolambda' option in MACS 1.4.2) using the High Performance Parallel Computing system (Texas Advanced Computing Center, the University of Texas at Austin). To monitor the background level, an SNR was calculated from duplicate read filtered bedGraph files generated by MACS 1.4.2 for each ChIP-seq data. After filtering out all high background data with a stringent filtering criterion of SNR 0.1, a total of 93 H3K9me2/3 ChIP-seq data were used for the further assessments. Second, to assess the false positive signal, the promoters of housekeeping genes were identified and used as the representative active genomic regions. For this, the 1524 pA+ RNA-seq profile (see RNA-seq data analysis in 'Materials and Methods' section; Supplementary Figure S6 and Table S5) was clustered with K-means clustering algorithm and 2442 genes that were active in all samples were identified, and the average H3K9me2/3 signal within the ± 500bp from the TSSs was calculated. For the reliable heterochromatin regions, Giemsa positive bands (gpos100 and gpos66) were used. As expected, a large portion of the experiments were contaminated with active genomic regions, and experiments whose average signal in Giemsa positive bands were at least threefold higher than the average signal in active regions were selected (39 experiments) and used for the further analyses shown in the examples of Figure 2B.

Gene frequency analyses in human G bands and isochores

For gene frequency analyses with regard to G bands (Supplementary Figure S2), human (hg19) cytoband lists were downloaded from the UCSC Genome Browser Database website (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/). Isochore lists were downloaded from the IsoFinder website (20) (http://bioinfo2.ugr.es:8080/isochores/human-isochores/).

Gene homology analysis (BLAST)

To measure the level of homologies between neighboring gene pairs in Figure 1C, we performed BLAST analysis using whole protein coding genes in the mouse genome. Since the BLAST *E*-values are sensitive to the target database size, we fixed the database to all proteins in the mouse genome and performed NxN BLASTP analysis using each single



	Gene 1	Gene 2	Gene 3	...	Gene N
Gene 1	$E_{1,1}$	$E_{1,2}$	$E_{1,3}$	...	$E_{1,N}$
Gene 2	$E_{2,1}$	$E_{2,2}$	$E_{2,3}$	...	$E_{2,N}$
Gene 3	$E_{3,1}$	$E_{3,2}$	$E_{3,3}$	...	$E_{3,N}$
...	...	...	...	...	...
Gene N	$E_{N,1}$	$E_{N,2}$	$E_{N,3}$	...	$E_{N,N}$

Figure 1

Figure 1. Non-random linear separation of CGI+ and CGI− genes in the genome and juxtaposition of CGI− homologs. (**A**) Examples of CGI+ (black) and CGI− (red) gene arrangements in the mouse genome. Homologous genes that co-cluster together are indicated by red brackets. (**B**) Gene order randomness test (Runs test) of CGI+/CGI− genes in the mouse genome (right). Box plots show expected transition from random gene shuffling. Left panel shows schematic representation of the Runs test (top), examples of randomly arranged genes and well-organized genes (bottom). (**C**) Neighboring gene homology test between all neighboring CGI+ (black) and CGI− (red) gene pairs. Gene pairs are sorted by E-value in ascending manner.

protein as query. Neighboring gene pairs with an *E*-value less than 1E-10 were considered homologous pairs.

Cell cultures

NIH-3T3 fibroblasts and C2C12 myoblasts were maintained in Dulbecco's Modified Eagle Medium (DMEM; GIBCO) containing 10% fetal bovine serum (GIBCO). For myotube differentiation of C2C12 myoblasts, cells were grown to high confluence and then the media was switched to DMEM supplemented with 2% donor equine serum (HyClone) and 1 μM insulin (Sigma-Aldrich).

ChIP-seq, DamID-seq, MeDIP-seq and DNaseI-seq data analysis

ChIP-seq, DamID-seq, MeDIP-seq and DNaseI-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 aligned using Bowtie 2.2.5 to the mouse genome (mm9, NCBI Build 37). To identify CTCF binding sites in mESC, CxxC binding regions for defining CGI+ and CGI− genes (see 'Definition of CGI+ and CGI− genes' in 'Material and Methods' section) and transcription factor (TF) binding sites (Figure 3E and F), model-based analysis for the ChIP-seq peak caller (MACS 1.4.2) (21) was used with a *P*-value cutoff of 1E-5. TF binding target genes were defined as genes with the TF binding within ±2 Kb of the TSS. Signal based analyses were done using duplicate filtered read pileup bedGraph files made from MACS 1.4.2. 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. For the data without control experiments (5me-C, DNaseI), the total area under the signal curve from bedGraph file was normalized to be one billion (1×10^9).

RNA-seq analysis

RNA-seq data were downloaded from SRA. FASTQ files were aligned to the mouse genome (mm9, NCBI Build 37) using STAR version 2.4.2 (22). Gene expression was calculated as RPKM values using rpkmforgenes.py (23). 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. For a unified gene expression profile of diverse tissues and cell lines shown in Supplementary Figure S6, all available mouse poly-A positive RNA-seq data (3818 samples) were summarized and downloaded on 5 May 2015. The measurement of gene expression of these samples was done using high-performance parallel computing system (Texas Advanced Computing Center, the University of Texas at Austin). Among the 3,818 samples, excluding single cell RNA-seq or experiments whose expression verified gene counts are small (less than 5,000 genes with RPKM 0.5 or higher), 1,524 high quality RNA-seq data were used. The gene expression profile was summarized as log10 scale ($\log_{10}$(RPKM+1)) and neighboring gene expression similarities were monitored by calculating Pearson Correlation Coefficients. Manually curated sample information was summarized in Supplementary Table S5, and the expression profile was deposited in the Gene Expression Omnibus (accession number GSE80797).

Microarray analysis

As a unified gene expression profile of diverse tissues and cell lines shown in Supplementary Figure S6, microarray data from GNF (Genomics Institute of the Novartis Research Foundation) Mouse Gene Atlas V3 (GSE10246) (24) were used. To precisely monitor expression values, raw data files (.cel files) were background corrected and normalized with GC Robust Multi-array Average expression measure using sequence information (GCRMA) (25) methods to minimize the background signal originating from probe sequence or high GC content. For genes with multiple probe sets, only probes with maximal signal were used for further analyses. For the expression data shown in Figure 3F, downloaded .cel files were normalized with the Robust Multi-array Average (RMA) (26) method.

ChIA-PET, Hi-C, Repli-chip data analysis

For ChIA-PET data analysis (Figure 3B and Supplementary Figure S5), the long-range chromatin interaction lists detected in HeLa, HCT116, K562, and MCF7 cells (27) were downloaded from ENCODE website. CGI+ and CGI− genes were sorted by expression level and binned into every 100 genes. Average detected count of long-range interactions within ±3 Kb from the TSSs was measured.

For inter-chromosomal interaction frequency analysis using Hi-C in Figure 3C, aligned read pairs were downloaded from Gene Expression Omnibus (GEO; GSE35156). Among the reads, the pairs matching with the CGI+ or CGI− gene clusters at both ends were filtered, and the inter-chromosomal interaction counts between gene clusters were normalized with restriction enzyme site count (HindIII, NcoI) within gene clusters. For Hi-C PCA analysis in Supplementary Figure S4A, 'runHiCpca.pl' was used in the HOMER (Hypergeometric Optimization of Motif EnRichment) program suite (28). For gene count analysis with regards to topologically associated domain (TAD) in Figure 3D, TAD lists determined in mouse embryonic stem (ES) cells were downloaded and used (29).

For replication timing analysis in Supplementary Figure S4B, Repli-chip wavelet-smoothed signal data (30) were downloaded from ENCODE website. For each CGI+ and CGI− gene clusters, average replication timing values ($\log_2$[early S phase signal/late S phase signal]) were calculated.

RESULTS

Non-random arrangement of CGI+ and CGI− genes

We first focused on the arrangements of CGI+ and CGI− genes in chromosomes. Since modern genome sequences are the cumulative outcome of a series of chromosomal rearrangement events throughout generations (31), we speculated that the current gene arrangements may expose important information about the environments to which each gene has been exposed (32). One apparent pattern we observed from the visual inspection of the mouse genome was the non-random linear separation of CGI+ and CGI− genes: multiple CGI+ genes are co-clustered together, while CGI− genes are also gathered separately in other regions (Figure 1A). To further test whether this pattern is prevalent in the mammalian genome, we performed an order randomness test (Runs Test) (33) by simplifying gene arrangements in each chromosome into binarized gene orders (CGI+ or CGI−; Figure 1B). As a result, we found significant linear separations of CGI+ or CGI− genes in all mouse and most human chromosomes (Figure 1B and Supplementary Figure S1; see also Supplementary Figure S2).

We additionally observed that homologous CGI− genes, presumably formed by local gene duplication, are often clustered together (Trem/Serpin gene clusters in Figure 1A), while CGI+ genes are not. By assessing the homology between neighboring gene pairs (Figure 1C), we found that more than half (56.4%) of all CGI− neighboring gene pairs are highly homologous (BLAST *E*-value < 1E-10), while only 3.7% of CGI+ gene pairs are homologous. Notably, the local gene duplication patterns (Supplementary Figure S3) can be sequestered by chromosomal translocations (31,34), frequently occurring among spatially proximal regions (32,35). Therefore, our findings imply that CGI+ and CGI− gene classes are located within totally different nuclear environments: CGI+ genes might be spatially proximate to each other and encounter relatively frequent chromosomal contacts leading to subsequent rearrangements, while CGI− genes might be spatially segregated from CGI+ genes and experience fewer chromosomal interactions.

Effects of CGIs on chromatin status

To understand the localization behaviors of CGI+ and CGI− genes, we thoroughly investigated chromatin status using datasets generated from mouse ES cells (16,36–38). As shown in Figure 2A, promoters with CGI elements are occupied by KDM2A, a CxxC domain-containing protein, regardless of associated gene activities. As we previously reported, the occupancy patterns of MYC or PRC class DNA-binding proteins are also confined to CGI+ genes (16). Accordingly, CGI+ promoters constantly remain unmethylated (5me-C) and sustain chromatin accessibility to some degree (DNaseI), even when genes are repressed by PRC factors. Since TF occupancy, CpG hypomethylation, and chromatin accessibility are hallmarks of euchromatin (39), these results evidently illustrate that CGI+ genes are generally associated with euchromatin.

On the other hand, the promoters of CGI− genes are generally methylated (5me-C) and inaccessible (DNaseI). To test whether these genes form heterochromatin, we additionally monitored H3K9me2/3 signals, which are representative heterochromatin signatures, and the occupancy of SUV39H1, an H3K9 specific methyltransferase, using systematically selected high-quality ChIP-seq data attained from multiple mouse samples (‘Materials and Methods’ section) (40–45). The results revealed that the surrounding regions of silent CGI− promoters are largely associated with heterochromatin signatures compared to active CGI− genes (Figure 2B). This implies that CGI− genes are associated with heterochromatin, and can be converted to euchromatin upon activation. Importantly, CGI+ genes show largely depleted H3K9me2/3 and SUV39H1 signals regardless of their activities. These data clarify that PRC suppresses CGI+ gene expression while CGI− genes are inactivated by heterochromatin formation.

Spatial segregation of CGI+ and CGI− genes in the nucleus

Based on our findings along with prior observations that the heterochromatin is located at the nuclear periphery while euchromatin is located at the core of the nucleus (46), we hypothesized that CGI+ and CGI− genes are spatially segregated within the nucleus and propose a revised model for nuclear gene localizations (Figure 2C). In our model, CGI+ genes stay within the nuclear center regardless of their activities, co-localizing with either transcription factories or *Polycomb* bodies depending on their activities as explained in the *nuclear subcompartment model*. On the other hand, generally silent CGI− genes are positioned at the nuclear periphery, forming heterochromatin. Upon activation, they extrude into the nuclear center and form euchromatin, which fits into the *CT-extrusion model*.

To test our revised model, we examined chromosomal association with nuclear peripheral lamina (47) by assessing the association of lamin B1 (LMNB1) at the surrounding regions of CGI+ and CGI− promoters using published DamID-seq data (48). LMNB1 preferentially occupies the surrounding regions of silent CGI− promoters, suggesting that silent CGI− genes are located at the lamina-associated nuclear periphery (Figure 2D). The results also imply that active CGI− genes without LMNB1 association are located distantly from the nuclear periphery, and this fits into the *CT-extrusion model*. Notably, LMNB1 association is clearly depleted near almost all CGI+ gene promoters regardless of gene activities, implying their consistent localization at the nuclear center area as explained in the *nuclear subcompartment model*.

We further experimentally validated our model by performing FISH, mapping the nuclear position of CGI+ and CGI− gene clusters that are inactive in fibroblasts, but active in myotubes. As shown in Figure 2E (for 3D view, Supplementary Movie S1), the silent CGI− gene cluster was mainly detected at the nuclear periphery of fibroblasts, but at the nuclear center in myotubes. On the other hand, the CGI+ gene cluster was mostly found at the nuclear center in both fibroblasts and myotubes. We then monitored multiple gene clusters collectively using Oligopaint techniques (18). As the most important differences expected are localization behaviors between inactive CGI+ and CGI− genes (Figure 2C; center and periphery, respectively), we designed FISH probes targeting multiple CGI+ and CGI− gene clusters that are inactive in fibroblasts (eight clusters each; Supplementary Table S1). As shown in Figure 2F, we confirmed that silent CGI− gene clusters are generally located at the nuclear periphery, while inactive CGI+ gene clusters reside mainly at the nuclear center, clearly supporting our model

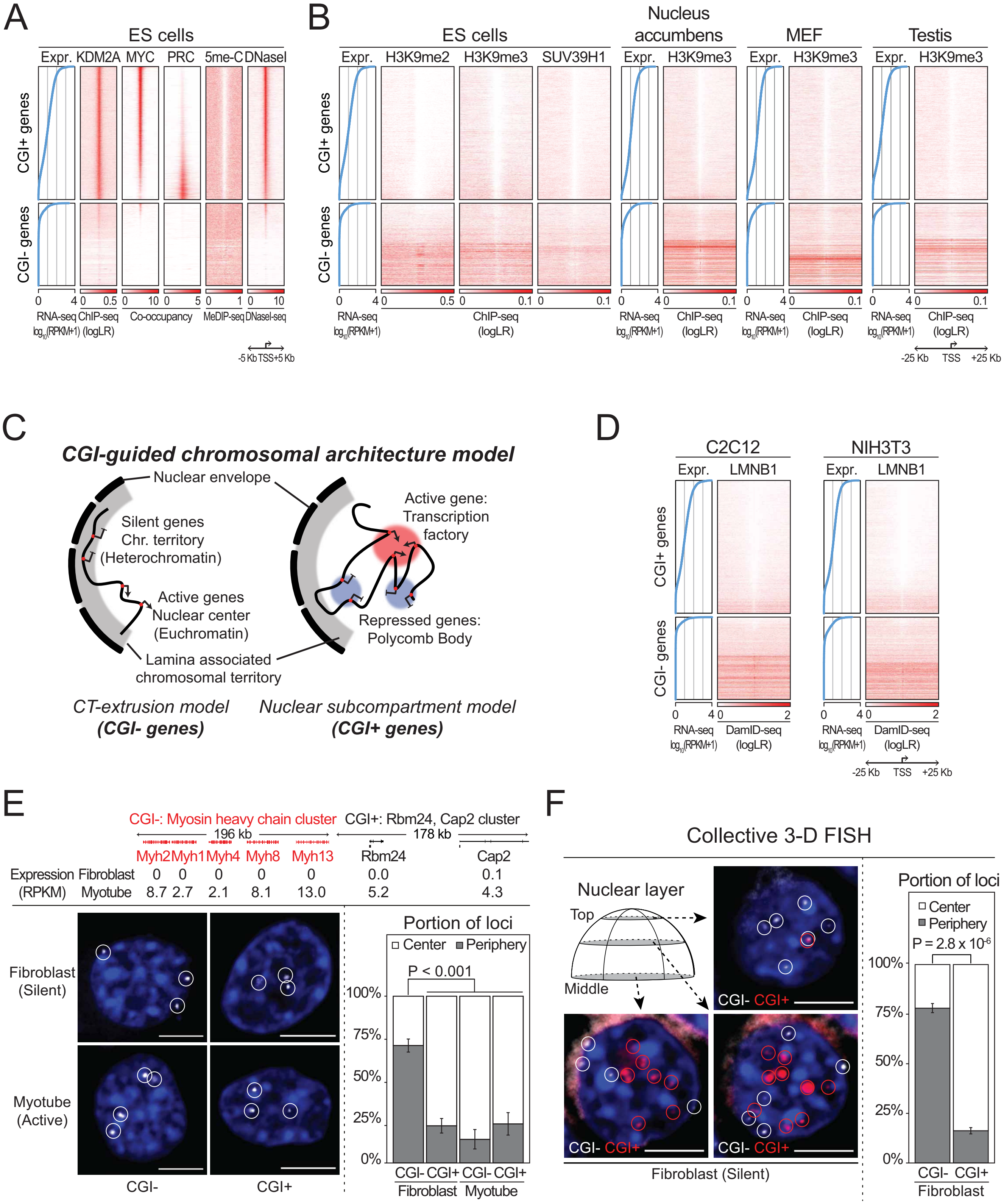


Figure 2. Spatial segregation of CGI+ and CGI− genes. (**A**) CGI-centered regulation mechanisms in CGI+ genes, and (**B**) heterochromatin features in CGI− genes. All protein coding genes are sorted by their expression values (left blue line plot), and the genomic landscape of promoter surrounding regions are shown. (**C**) CGI-guided chromosomal architecture model. (**D**) Association of peripheral nuclear lamina (LMNB1) with silent CGI− genes, but not CGI+ genes. (**E**) Localization of single CGI+ (or CGI−) gene cluster in fibroblasts and myotubes, and (**F**) multiple silent CGI+/CGI− gene clusters in fibroblasts (eight clusters each) determined by 3D FISH. Percentages of gene loci located in the nuclear center compared to the periphery in each cell are shown in bottom right. *P*-values: Wilcoxon signed-rank tests. Scale bar: 5 μm. Error bars show S.E.M.

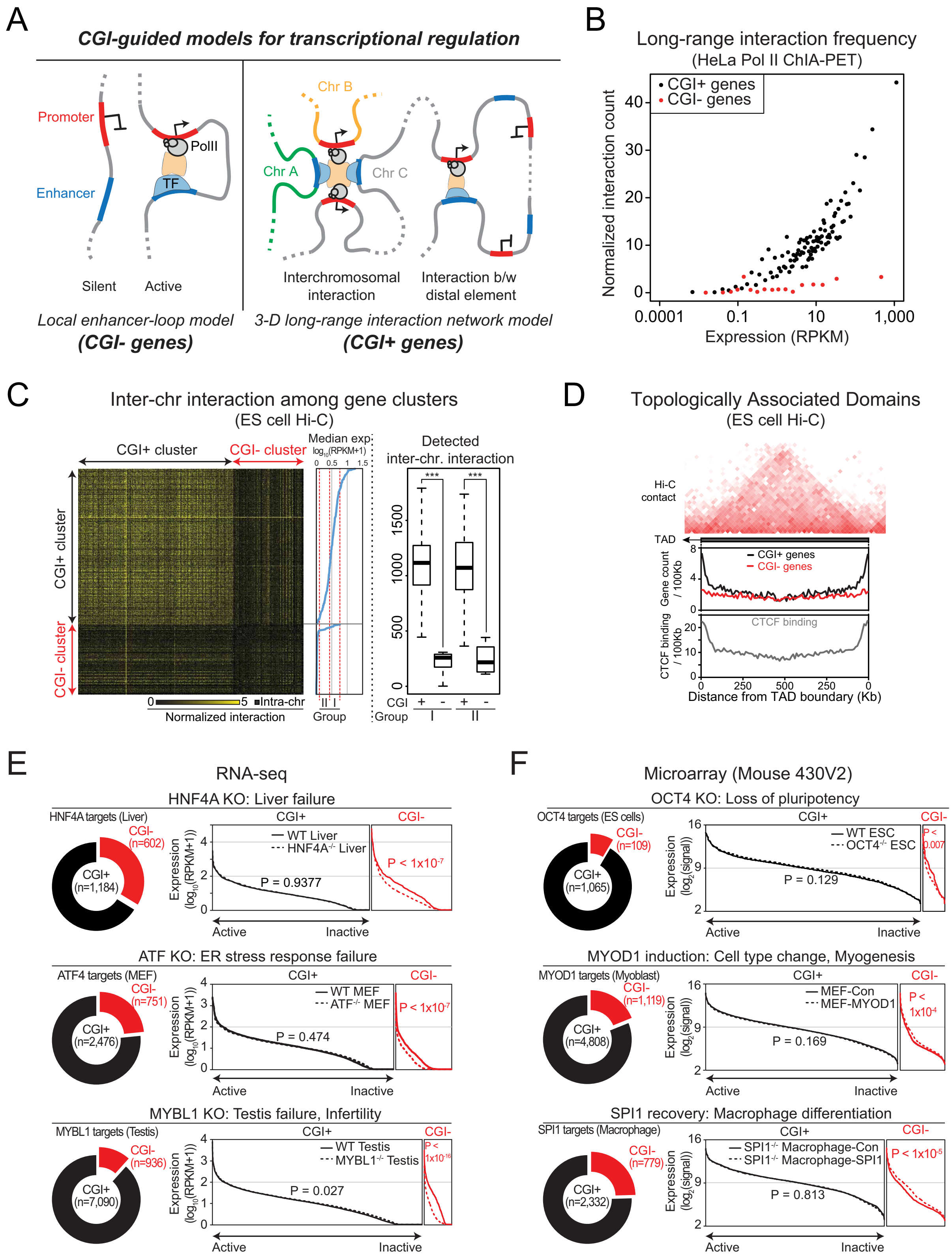


Figure 3. Transcriptional regulation mechanisms of CGI+ and CGI− genes. (**A**) CGI-guided models for mammalian transcriptional regulation mechanisms. (**B**) Higher frequency of Pol II-mediated chromosomal long-range interactions in CGI+ genes compared to CGI− genes. Genes were sorted by expression level and binned into every 100 genes, and average expression and normalized long-range interaction count of bins are shown. (**C**) Frequency of inter-chromosomal interactions in CGI+ and CGI− gene clusters (Supplementary Table S2; sorted by their median expression) determined by mouse ES cell Hi-C data. Normalized total detected inter-chromosomal interaction counts in cluster groups are shown in right. $***P < 0.0001$ from Wilcoxon signed-rank test. (**D**) CGI+ and CGI− gene frequencies as well as CTCF binding frequencies with regards to TAD structures. (**E** and **F**) TF binding target gene expression analysis upon TF perturbation. In each line plot, target genes are sorted by expression level. *P*-values are calculated from Kolmogorov–Smirnov tests.

of CGI-dependent spatial gene segregation (Figure 2C; see also Supplementary Figure S4).

CGI and transcriptional regulation mechanisms

We further questioned whether the different localization patterns of CGI+ and CGI− genes would resolve the discrepancy among the proposed models explaining transcriptional regulation mechanisms (Figure 3A). The most widely accepted transcriptional regulation mechanisms is the ***local enhancer-loop model*** proposed almost three decades ago (49), where enhancers occupied by TFs communicate with nearby promoters via chromosomal looping for gene activation. In this model, the expression levels of the target genes associated with a specific TF are expected to be directly affected by perturbation of the TF. However, TF target genes often do not show expected expression changes upon TF perturbations (50). Alternatively, recent high-throughput chromatin interaction analyses have revealed most enhancers are associated with promoters beyond the nearest ones, and multiple enhancers and promoters form complex 3D long-range interaction networks (***3D long-range interaction network model***) (51). This new model explains the inconsistencies between TF targets and expression changes upon perturbations of the TFs (50). Nevertheless, considering that the conventional *local enhancer-loop model* has been also supported by multiple examples (52–54), each model seems to explain the regulation of only limited sets of genes.

To test whether our finding of CGI-mediated gene segregation would explain the discrepancy between models, we measured the long-range interaction frequencies from surrounding regions of the CGI+ and CGI− gene promoters using Pol II ChIA-PET data (27). As shown in Figure 3B, CGI+ genes encounter more frequent long-range interactions than CGI− genes with similar activities (see also Supplementary Figure S5), indicating that nuclear central CGI+ genes are more strongly involved in Pol II-mediated chromosomal long-range interactions compared to CGI− genes, which are generally condensed as heterochromatin. We additionally monitored the relative frequencies of inter-chromosomal interactions using Hi-C data obtained from mouse ES cells (29). As shown in Figure 3C, CGI+ gene clusters show significantly higher frequencies of inter-chromosomal interactions than CGI− gene clusters. Surprisingly, the gene expression levels of the CGI+ gene clusters did not show any strong correlation with the inter-chromosomal interaction frequencies (Figure 3C, right box plot), suggesting that not only active, but also inactive CGI+ genes encounter frequent chromosomal long-range interactions, presumably mediated by *Polycomb* bodies (3,4).

We then questioned whether our observations of frequent long-range interactions among CGI+ genes agree with the structures formed by local genome entanglement: TADs defined from Hi-C data analysis (29). To delineate the relationship between local topological structures and CGIs, we investigated the location of CGI+ and CGI− genes with regards to the TADs. As shown in Figure 3D, CGI+ genes, but not CGI− genes, are largely enriched at each end of TADs (6.2-fold more than at the center of TADs). As a result, about 23.2% of total CGI+ genes are detected at the inter-TAD regions or the end of TADs (<20 Kb from the ends of TADs; 13.9% for CGI− genes). This is similar with the binding frequency of CTCF (3.5-fold more than at the center of TADs), which is also known to be involved in the chromosomal long-range interactions and responsible for the physical separation of chromosomal domains (29,55). These data show that the chromosomal long-range interactions of CGI+ genes strongly influence local genome structures, similarly to CTCF bindings.

Since CGI− genes are not highly involved in long-range chromosomal interactions (Figure 3B), we speculated that CGI− genes would be more dependent on the local enhancer-mediated regulations. We compared the ChIP-seq and the target gene expression data upon TF removal or induction. Although the majority of targets for each TF are CGI+ genes, not CGI− genes, the depletion or induction of each TF did not significantly affect the global expression levels of the CGI+ target genes (Figure 3E and F). On the other hand, CGI− target genes, which constitute only a minor portion of the total targets of the TF, showed significant changes in gene expression upon depletion or induction of the TF. Thus, CGI− genes respond more strongly to TF perturbation than CGI+ genes. Altogether, our findings strongly suggest that *3D long-range interaction network model* explains the regulatory mode of CGI+ genes, while the regulation of CGI− genes fit squarely into *local enhancer-loop model* (Figure 3A).

DISCUSSION

In this study, we reveal CGI+ and CGI− genes are organized as separated clusters within the genome, and further show CGI-dependent chromatin architectures and CGI-associated global gene regulatory modes. Notably, our findings can explain the prior discrepancies among multiple models describing chromosomal architectures and transcriptional regulation as summarized in Table 1. These distinctions highlight the critical implications of CGIs on gene regulation which has never been clearly elucidated before.

Genome-wide linear organization of CGI+ and CGI− genes seems to be optimal for the cost-effective gene regulation. It is reasonable to speculate that the co-clustering patterns of CGI+ or CGI− genes (Figure 1) are beneficial for efficient spatial segregation (Figure 2), which in turn allows differential regulation without mutual interferences (Figure 3). The co-clustering patterns (Figure 1), as well as the similar directionality (Supplementary Figure S3) of the neighboring CGI− genes also seem advantageous for simultaneous tissue or stage-specific activation (Supplementary Figure S6) of the multiple CGI− genes with similar function (Figure 1C).

Interestingly, the key components of CGI-mediated dual mode gene regulation, PRC and heterochromatin-mediated repression mechanisms, are universally observed in a broad range of eukaryotes even in the species without CGIs (15,39,56,57). Therefore, it will be imperative to test whether there are distinct modes of gene regulation mediated by specific DNA elements corresponding to the mammalian CGIs in other eukaryotes.

Table 1. Summary of CGI-dependent chromosomal architectures and gene regulations

	CGI− genes	CGI+ genes
Gene regulation	*Local enhancer-loop model*	*3D long-range interaction network model*
Chromatin status	Heterochromatin—Euchromatin	Euchromatin
Chromosomal architectures	*Chromosomal territory (CT)-extrusion model*	*Nuclear subcompartment model*
Default gene location in nucleus	Nuclear periphery (extruded into the center when activated)	Nuclear center (regardless of the gene activity)
Dimension of regulation[a]	Linear local regulation	3D spatial regulation
Nuclear landscape model[b]	*Chromosomal territory-interchromatin compartment (CT-IC) model*	*Interchromatin network (ICN) model*

[a] See also Supplementary Figure S6.
[b] Discussed in Supplementary Figure S7.

DATA AVAILABILITY

The data reported in this paper are deposited at the Gene Expression Omnibus (accession number GSE80797).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate Drs Edward Marcotte, Vishwanath Iyer (University of Texat at Austin), Jian Xu (University of Texas Southwestern) and Andrew J. Woo (University of Western Australia) for their critical comments on the manuscript, and Drs Ting Wu and Brian J. Beliveau (Harvard Medical School) for technical advice on Oligopaint technique. 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 [P20GM104318].

Conflict of interest statement. None declared.

REFERENCES

1. Chubb,J.R. and Bickmore,W.A. (2003) Considering nuclear compartmentalization in the light of nuclear dynamics. *Cell*, **112**, 403–406.
2. Sutherland,H. and Bickmore,W.A. (2009) Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, **10**, 457–466.
3. Pirrotta,V. and Li,H.B. (2012) A view of nuclear Polycomb bodies. *Curr. Opin. Genet. Dev.*, **22**, 101–109.
4. Bantignies,F., Roure,V., Comet,I., Leblanc,B., Schuettengruber,B., Bonnet,J., Tixier,V., Mas,A. and Cavalli,G. (2011) Polycomb-dependent regulatory contacts between distant hox loci in drosophila. *Cell*, **144**, 214–226.
5. Brown,J.M., Green,J., das Neves,R.P., Wallace,H.A.C., Smith,A.J.H., Hughes,J., Gray,N., Taylor,S., Wood,W.G., Higgs,D.R. *et al.* (2008) Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. *J. Cell Biol.*, **182**, 1083–1097.
6. Buckley,M.S. and Lis,J.T. (2014) Imaging RNA Polymerase II transcription sites in living cells. *Curr. Opin. Genet. Dev.*, **25**, 126–130.
7. Iborra,F.J., Pombo,A., Jackson,D.A. and Cook,P.R. (1996) Active RNA polymerases are localized within discrete transcription 'factories' in human nuclei. *J. Cell Sci.*, **109**, 1427–1436.
8. Hernández-Muñoz,I., Taghavi,P., Kuijl,C., Neefjes,J. and van Lohuizen,M. (2005) Association of BMI1 with polycomb bodies is dynamic and requires PRC2/EZH2 and the maintenance DNA methyltransferase DNMT1. *Mol. Cell. Biol.*, **25**, 11047–11058.
9. van der Stoop,P., Boutsma,E.A., Hulsman,D., Noback,S., Heimerikx,M., Kerkhoven,R.M., Voncken,J.W., Wessels,L.F.A. and van Lohuizen,M. (2008) Ubiquitin E3 ligase ring1b/Rnf2 of polycomb repressive complex 1 contributes to stable maintenance of mouse embryonic stem cells. *PLoS One*, **3**, e2235.
10. 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.
11. Gardiner-Garden,M. and Frommer,M. (1987) CpG Islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
12. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
13. Fouse,S.D., Shen,Y., Pellegrini,M., Cole,S., Meissner,A., Van Neste,L., Jaenisch,R. and Fan,G. (2008) Promoter CpG Methylation Contributes to ES Cell Gene Regulation in Parallel with Oct4/Nanog, PcG Complex, and Histone H3 K4/K27 Trimethylation. *Cell Stem Cell*, **2**, 160–169.
14. 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.
15. Deaton,A.M. and Bird,A. (2011) CpG islands and the regulation of transcription. *Genes Dev.*, **25**, 1010–1022.
16. 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.
17. 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.
18. Beliveau,B.J., Joyce,E.F., Apostolopoulos,N., Yilmaz,F., Fonseka,C.Y., McCole,R.B., Chang,Y., Li,J.B., Senaratne,T.N., Williams,B.R. *et al.* (2012) Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 21301–21306.
19. 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.
20. Oliver,J.L., Carpena,P., Hackenberg,M. and Bernaola-Galván,P. (2004) IsoFinder: Computational prediction of isochores in genome sequences. *Nucleic Acids Res.*, **32**, W287–W292.
21. 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.
22. 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.
23. 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.
24. Lattin,J.E., Schroder,K., Su,A.I., Walker,J.R., Zhang,J., Wiltshire,T., Saijo,K., Glass,C.K., Hume,D.A., Kellie,S. *et al.* (2008) Expression analysis of G protein-coupled receptors in mouse macrophages. *Immun. Res.*, **4**, 5.

25. Wu,Z., Irizarry,R.A., Gentleman,R., Martinez-Murillo,F. and Spencer,F. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.
26. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
27. Li,G., Ruan,X., Auerbach,R.K., Sandhu,K.S., Zheng,M., Wang,P., Poh,H.M., Goh,Y., Lim,J., Zhang,J. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.
28. 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.
29. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
30. Hiratani,I., Ryba,T., Itoh,M., Rathjen,J., Kulik,M., Papp,B., Fussner,E., Bazett-Jones,D.P., Plath,K., Dalton,S. *et al.* (2010) Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. *Genome Res.*, **20**, 155–169.
31. 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.
32. 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. *Chromosome Res.*, **10**, 707–715.
33. Wald,A., Wolfowitz,J. and Wald2,A. (1940) On a test whether two samples are from the same population. *Ann. Math. Stat.*, **11**, 147–162.
34. Tillier,E.R. and Collins,R.A. (2000) Genome rearrangement by replication-directed translocation. *Nat. Genet.*, **26**, 195–197.
35. 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.
36. Blackledge,N.P., Zhou,J.C., Tolstorukov,M.Y., Farcas,A.M., Park,P.J. and Klose,R.J. (2010) CpG islands recruit a histone H3 lysine 36 demethylase. *Mol. Cell*, **38**, 179–190.
37. Shen,L., Wu,H., Diep,D., Yamaguchi,S., D'Alessio,A.C., Fung,H.L., Zhang,K. and Zhang,Y. (2013) Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell*, **153**, 692–706.
38. Yue,F., Cheng,Y., Breschi,A., Vierstra,J., Wu,W., Ryba,T., Sandstrom,R., Ma,Z., Davis,C., Pope,B.D. *et al.* (2014) A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, **515**, 355–364.
39. Henikoff,S. (2000) Heterochromatin function in complex genomes. *Biochim. Biophys. Acta*, **1470**, O1–O8.
40. Maze,I., Feng,J., Wilkinson,M.B., Sun,H., Shen,L. and Nestler,E.J. (2011) Cocaine dynamically regulates heterochromatin and repetitive element unsilencing in nucleus accumbens. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 3035–3040.
41. Bulut-Karslioglu,A., DeLaRosa-Velázquez,I.A., Ramirez,F., Barenboim,M., Onishi-Seebacher,M., Arand,J., Galán,C., Winter,G.E., Engist,B., Gerle,B. *et al.* (2014) Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. *Mol. Cell*, **55**, 277–290.
42. Pedersen,M.T., Agger,K., Laugesen,A., Johansen,J. V, Cloos,P.A.C., Christensen,J. and Helin,K. (2014) The demethylase JMJD2C localizes to H3K4me3 positive transcription start sites and is dispensable for embryonic development. *Mol. Cell. Biol.*, **34**, 1031–1045.
43. Pezic,D., Manakov,S.A., Sachidanandam,R. and Aravin,A.A. (2014) piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. *Genes Dev.*, **28**, 1410–1428.
44. Elsässer,S.J., Noh,K.M., Diaz,N., Allis,C.D. and Banaszynski,L.A. (2015) Histone H3.3 is required for endogenous retroviral element silencing in embryonic stem cells. *Nature*, **522**, 240–244.
45. von Meyenn,F., Iurlaro,M., Habibi,E., Liu,N.Q., Salehzadeh-Yazdi,A., Santos,F., Petrini,E., Milagre,I., Yu,M., Xie,Z. *et al.* (2016) Impairment of DNA methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Mol. Cell*, **62**, 848–861.
46. 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 Harb. Symp. Quant. Biol.*, **58**, 777–792.
47. Guelen,L., Pagie,L., Brasset,E., Meuleman,W., Faza,M.B., Talhout,W., Eussen,B.H., de Klein,A., Wessels,L., de Laat,W. *et al.* (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**, 948–951.
48. Wu,F. and Yao,J. (2013) Spatial compartmentalization at the nuclear periphery characterized by genome-wide mapping. *BMC Genomics*, **14**, 591.
49. Wang,J.C. and Giaever,G.N. (1988) Action at a distance along a DNA. *Science*, **240**, 300–304.
50. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
51. Zhang,Y., Wong,C.-H.C.-H., Birnbaum,R.Y., Li,G., Favaro,R., Ngan,C.Y., Lim,J., Tai,E., Poh,H.M., Wong,E. *et al.* (2013) Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, **504**, 306–310.
52. Tolhuis,B., Palstra,R.J., Splinter,E., Grosveld,F. and De Laat,W. (2002) Looping and interaction between hypersensitive sites in the active β-globin locus. *Mol. Cell*, **10**, 1453–1465.
53. Drissen,R., Palstra,R.J., Gillemans,N., Splinter,E., Grosveld,F., Philipsen,S. and De Laat,W. (2004) The active spatial organization of the β-globin locus requires the transcription factor EKLF. *Genes Dev.*, **18**, 2485–2490.
54. Rada-Iglesias,A., Wallerman,O., Koch,C., Ameur,A., Enroth,S., Clelland,G., Wester,K., Wilcox,S., Dovey,O.M., Ellis,P.D. *et al.* (2005) Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Hum. Mol. Genet.*, **14**, 3435–3447.
55. Lee,B.K. and Iyer,V.R. (2012) Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. *J. Biol. Chem.*, **287**, 30906–30913.
56. Simon,J.A. and Kingston,R.E. (2009) Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat. Rev. Mol. Cell Biol.*, **10**, 697–708.
57. 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.