Biophysical Journal
Article

DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells

Ling Zhang,[1,2] Wen Jun Xie,[1] Sirui Liu,[1] Luming Meng,[1] Chan Gu,[1,2] and Yi Qin Gao[1,2,*]
[1]Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering and [2]Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, Peking University, Beijing, China

ABSTRACT The relationship between DNA methylation and chromatin structure is still largely unknown. By analyzing a large set of published sequencing data, we observed a long-range power law correlation of DNA methylation with cell class-specific scaling exponents in the range of tens of kilobases. We showed that such cell class-specific scaling exponents are caused by different patchiness of DNA methylation in different cells. By modeling the chromatin structure using high-resolution chromosome conformation capture data and mapping the methylation level onto the modeled structure, we demonstrated that the patchiness of DNA methylation is related to chromatin structure. The scaling exponents of the power law correlation are thus a display of the spatial organization of chromatin. Besides the long-range correlation, we also showed that the local correlation of DNA methylation is associated with nucleosome positioning. The local correlation of partially methylated domains is different from that of nonpartially methylated domains, suggesting that their chromatin structures differ at the scale of several hundred base pairs (covering a few nucleosomes). Our study provides a novel, to our knowledge, view of the spatial organization of chromatin structure from a perspective of DNA methylation, in which both long-range and local correlations of DNA methylation along the genome reflect the spatial organization of chromatin.

INTRODUCTION

Composed of DNA and histones, chromatin has a three-dimensional (3D) structure at different hierarchical levels (1). The spatial organization of chromatin plays an essential role in many genomic functions, including gene expression, DNA replication, and cell mitosis (2–6). Several lines of evidence show that epigenetics can remodel chromatin structure at different levels (7–12). Super-resolution imaging recently showed that chromatin folding varies for different epigenetic states (9).

DNA methylation, as the most abundant epigenetic modification in eukaryotic chromosomes, is also thought to influence chromatin structure (10). DNA methylation has a close relationship with nucleosome positioning (11), and the binding of CCCTC-binding factor can be partly influenced by DNA methylation and thus changes chromatin structure (12). Recently, DNA methylation was also used to reconstruct A/B compartments of chromatin revealed by high-resolution chromosome conformation capture (Hi-C) experiments (13). Nevertheless, how DNA methylation relates to chromatin structure remains largely unknown.

On the other hand, the distribution of DNA methylation in chromatin, and thus the correlation of DNA methylation levels between different genomic segments, may provide hints on the spatial organization of chromatin. Here, we investigate long-range and local correlations in the DNA methylation landscape using published whole-genome bisulfite sequencing (WGBS) data, which we expect to reflect the packing of DNA in the 3D space, and try to obtain information on the underlying chromatin structure. DNA methylation possesses long-range power law correlation with a cell class-specific scaling exponent. In addition, the scaling exponent can be used to discern cell classes. We find that the degree of DNA methylation patchiness is cell-specific and that this patched methylation pattern contributes to the different scaling exponents in different cells. Using polymer modeling with Hi-C data, we show that the partially methylated domains (PMDs) spatially segregate from the non-PMDs (genomic regions that are not classified as PMDs) in the IMR90 cell line, leading to it having patchiness of DNA methylation that differs from that of the h1 cell line. In this way, the cell class-specific exponents for the long-range DNA methylation correlation reflect the spatial organization of chromatin. We also demonstrate that the local DNA

Submitted May 8, 2017, and accepted for publication August 11, 2017.

*Correspondence: gaoyq@pku.edu.cn

Ling Zhang and Wen Jun Xie contributed equally to this work.

Editor: Anatoly Kolomeisky.
http://dx.doi.org/10.1016/j.bpj.2017.08.019

methylation correlation is related to nucleosome occupancy, and suggest that there are different chromatin structures of PMDs and non-PMDs at nucleosome level. Therefore, both long-range and local DNA methylation correlations can reflect the spatial organization of chromatin.

MATERIALS AND METHODS

Sources of WGBS data

In this work, we used WGBS data for different cells, including 36 somatic cells, 49 cancer cells and the corresponding normal cells, 8 human brain cells, 1 mouse brain cell, 12 embryonic stem cell lines and related cells, and 6 cells with neurodegenerative diseases (NDDs) (105 in total). All the methylomes were summarized (Tables S1–S5) including the references, URLs, and sample details. The methylomes of cancer samples were downloaded from The Cancer Genome Atlas (TCGA) project. Of all the samples in TCGA, nine types of cancer samples have WGBS data and these nine samples were used.

The Hg18 reference genome was used for human brain cells and human embryonic stem cells (ESCs). The other cells used Hg19 as the reference genome. We determined that the reference genome used had little effect on the methylation correlation found here (Fig. S9 *A*).

Identification of PMDs, non-PMDs, and PMD-like regions

PMDs were identified genome-wide in tumor samples using a sliding window approach and the parameters in (14) were adopted in this work. The window size was set as 10 kb. A region was identified as a PMD if there were at least 10 methylated (β value > 0) CpG dinucleotides within it, of which the average methylation level was < 0.7. The contiguous PMD windows were then merged into a longer PMD. Only PMDs with lengths longer than 100 kb are used in the following analysis. Non-PMDs are identified as the complementary set of the PMDs and only non-PMDs whose lengths are $>$ 100 kb are used. PMD-like regions were defined in noncancer samples as the corresponding genomic regions of cancer PMDs. Thus, PMD-like regions are defined only for the noncancer samples that have corresponding cancer samples.

Calculation of scaling exponents

The scaling exponents of the long-range power law correlations were calculated as the maximum slope of the fitted double-log correlation data in the genomic region of tens of kilobases. To systematically identify those chromosomes whose slopes of double-log plot of methylation correlation are not well-defined in the concerned (tens of kilobases) region, we calculated the SD of the first-order derivatives for each fitted double-log plot. Low SDs indicate linear behavior with small slope fluctuation for methylation correlation, whereas high SDs indicate large fluctuation.

Fast Fourier transform of the local correlation

Fast Fourier transform (FFT) was performed on the local correlation of the two genomic regions (PMDs and non-PMDs) of imr90, respectively. To avoid the finite length effect and influences of length distribution of genomic regions, we used PMDs and non-PMDs with a genomic length $>$ 0.1 Mb.

Detrended fluctuation analysis

A brief explanation of Detrended Fluctuation Analysis (DFA) was given here and the details can also be found in the Supporting Material: Detrended Fluctuation Analysis for different cell classes. In DFA, the root mean-square fluctuation $F(r)$ of DNA methylation as a function of genomic distance was defined. For purely uncorrelated random sequences, $F(r) \sim r^{1/2}$, corresponding to a ~0.5 slope in double-log plot. If the correlation of a sequence decays exponentially, indicating a finite-range correlation, the fluctuation scaling exponent will also be 0.5. Only when a long-range correlation with an infinite characteristic length is expected, will the scaling exponent deviate from 0.5 and thus may be described by a power law.

Gene expression analysis

The level 3 RNA-seq by expectation maximization data from TCGA RNAseq version2 was downloaded from https://portal.gdc.cancer.gov. The RNA-seq by expectation maximization data were then converted to transcripts per million by multiplying by 10^6. To compare the differentially expressed genes between tumor and normal samples, we chose the tumor-normal sample pairs that were taken from the same patient and gene expression data of both tumor and normal samples that were available. We finally obtained four tumor-normal pairs, namely, brca_t5-brca_n5, coad_t2-coad_n2, luad_t5-luad_n5, and ucec_t5-ucec_n5. We compared gene expression for these four tumor-normal pairs.

Genes with intragenic regions intersecting with tumor PMDs (or PMD-like regions in corresponding normal samples) were identified. Then gene expression fold change was calculated as $\mathrm{TPM}_{tumor}/\mathrm{TPM}_{normal}$ for each gene. These genes were divided into four categories: activated genes (fold change ≥ 2), repressed genes (fold change ≤ 0.5), specifically expressed in tumor sample ($\mathrm{TPM}_{normal} = 0$ and $\mathrm{TPM}_{tumor} \neq 0$), and specifically expressed in normal sample ($\mathrm{TPM}_{normal} \neq 0$ and $\mathrm{TPM}_{tumor} = 0$). We also defined the gene density of the genome and the specific genomic regions like PMDs as the number of genes per million base pairs. Gene functional classification was carried out using The Database for Annotation, Visualization and Integrated Discovery (15) (16). The housekeeping genes list can be downloaded from https://www.tau.ac.il/~elieis/HKG/ (17).

Structural modeling using Hi-C data

We developed a restraint-based method to construct an ensemble of 3D chromosome models (18). The method was verified by the reproduction of experimental Hi-C contact frequencies. In our method, chromosome was coarse-grained as a polymer chain consisting of a string of beads. The Hi-C data for the IMR90 and h1 cell lines were obtained from Rao et al. (19) and Dixon et al. (20), respectively. According to the resolution of Hi-C data, in our modeling for IMR90 and h1, each bead represents a 50 or 40 kb genomic region, respectively. The polymer structure was optimized according to distance restraints derived from Hi-C data. To achieve this, we first converted the contact frequency matrix measured by the Hi-C experiment to a distance matrix that provides the spatial restraints for the coarse-grained beads. Then, we performed MD simulations starting from randomly generated initial conformations using biased potentials to generate an ensemble of conformations based on the restraint distance matrix. Further modeling details and validation were presented in Xie et al. (18).

RESULTS

DNA methylation shows long-range power law correlation

We compared the Pearson correlation coefficients of DNA methylation levels (β values) within the methylome across a wide range of human cells, including normal somatic cells, cancer cells, brain cells, gland cells, and stem cells. In calculating the long-range correlation, the methylation level was first averaged using a 200 bp window. The sources of relative WGBS data were summarized in Tables S1–S5.

Taking chromosome 1 as representative, all the methylation correlations strikingly present a long-range power law decay as the genomic distance increases (Fig. 1). The power law correlation implies a scale-free property of DNA methylation and the scale-invariant genomic segment lies in the tens of kilobases scale. Power law scaling is of general interest (21) and is often noticed in evolving systems that may be produced by hierarchical structure of several length-scales (22). The scale-invariant genomic scale (tens of kilobases) also involves the sizes of genes and chromatin domains, and can be important for a variety of genomic functions (19,23).

The correlation coefficients still have finite values in the order of 0.01–0.1 even for the 1 Mb genomic separation (Fig. 1). To verify the statistical significance of the power law decay found here, we also calculated the correlation of a randomly methylated DNA sequence for comparison. Specifically, we generated a randomized methylation pattern by randomly assigning the methylation level of each CpG following the overall distribution of the original sample. The correlation coefficient for the random sample immediately drops to zero and the power law decay disappears (Fig. S1). This comparison clearly shows the nonrandom nature of DNA methylation in the cells and that the methylation level of CpGs separated by a very long genomic distance is indeed significantly correlated.

Interestingly, the scaling exponents of long-range DNA methylation correlation differ substantially between normal somatic cells and cancer cells, and the respective values are −0.26 ± 0.02 and −0.06 ± 0.02. The value for cancer cells is significantly smaller than that for normal somatic cells. Small SDs show that the scaling exponents are conserved among either normal somatic cells or cancer cells (Fig. 1, *A* and *B*), although the methylation levels of individual CpGs (and even the average values among all CpGs) vary greatly (24). Similarly, the differences between normal somatic cells and brain or gland cells are substantial but consistent within each cell class (Fig. 1, *C* and *D*), suggesting that cellular differentiation causes systematic variations of DNA methylation landscape. It was found that the scaling exponents for chromosome 1 of normal somatic cells in three different individuals are conserved (Fig. 2 *A*; Figs. S2 *A* and S3 *A*) and that the power law scaling is also present in mouse brain cells (Fig. S2 *B*).

In addition, DFA (25) was also performed that again show the long-range correlation in the DNA methylome (Fig. S4). DFA was used previously to describe the long-range correlation in DNA sequences (25), which is more robust than direct correlation calculation when determining the average behavior of a long-range effect. The average scaling exponents of 0.76 ± 0.01 and 0.92 ± 0.02 are observed separately for normal somatic cells (Fig. S4 *A*) and cancer cells (Fig. S4 *B*). Their deviation from 0.5 and small variances indicate a uniform power law decay within certain cell states among different types of tissues. Cancer cells hold an obviously higher scaling exponent, in

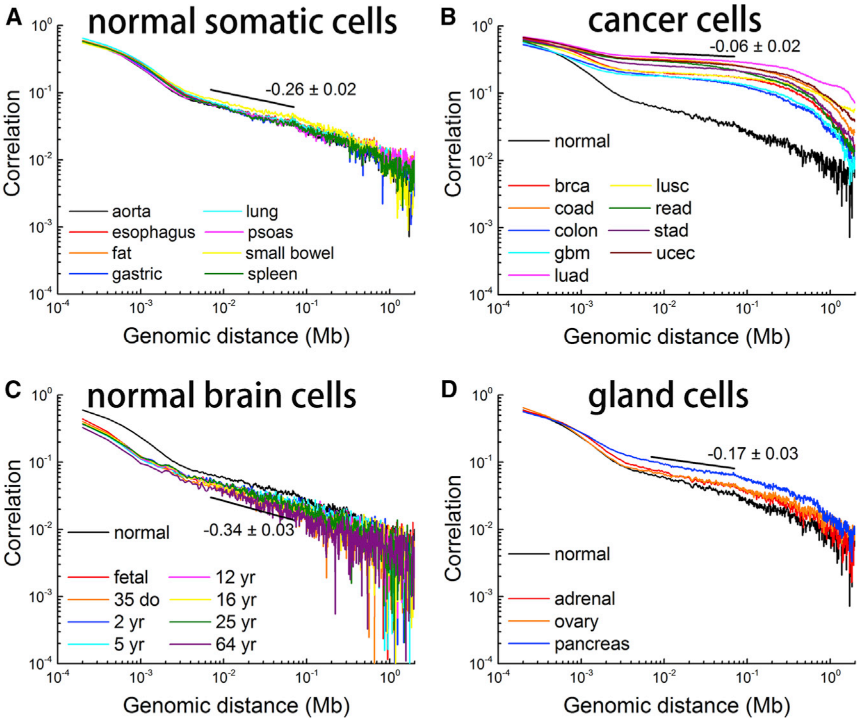

FIGURE 1 Long-range correlations in DNA methylation are distinct among different cell classes. The Pearson correlation coefficients for chromosome 1 of different cells are shown in log-log plots. The average scaling exponents are annotated in the figure. (*A*) Eight different somatic cells: aorta, esophagus, fat, gastric, lung, psoas, small bowel, and spleen. (*B*) Nine different cancer cells: bladder urothelial carcinoma (blca), breast invasive carcinoma (brca), colon adenocarcinoma (coad), colorectal cancer (colon), lung adenocarcinoma (luad), lung squamous cell carcinoma (lusc), rectum adenocarcinoma (read), stomach adenocarcinoma (stad), and uterine corpus endometrial carcinoma (ucec). The cells are labeled after TCGA except for colorectal cancer (colon). (*C*) Normal brain cells of different ages (fetal, 35 days old, and 2–64 years old). (*D*) Three different gland cells (adrenal, ovary, and pancreas). Correlation for normal aorta cells (normal) is also plotted for comparison in (*B*), (*C*), and (*D*). To see this figure in color, go online.

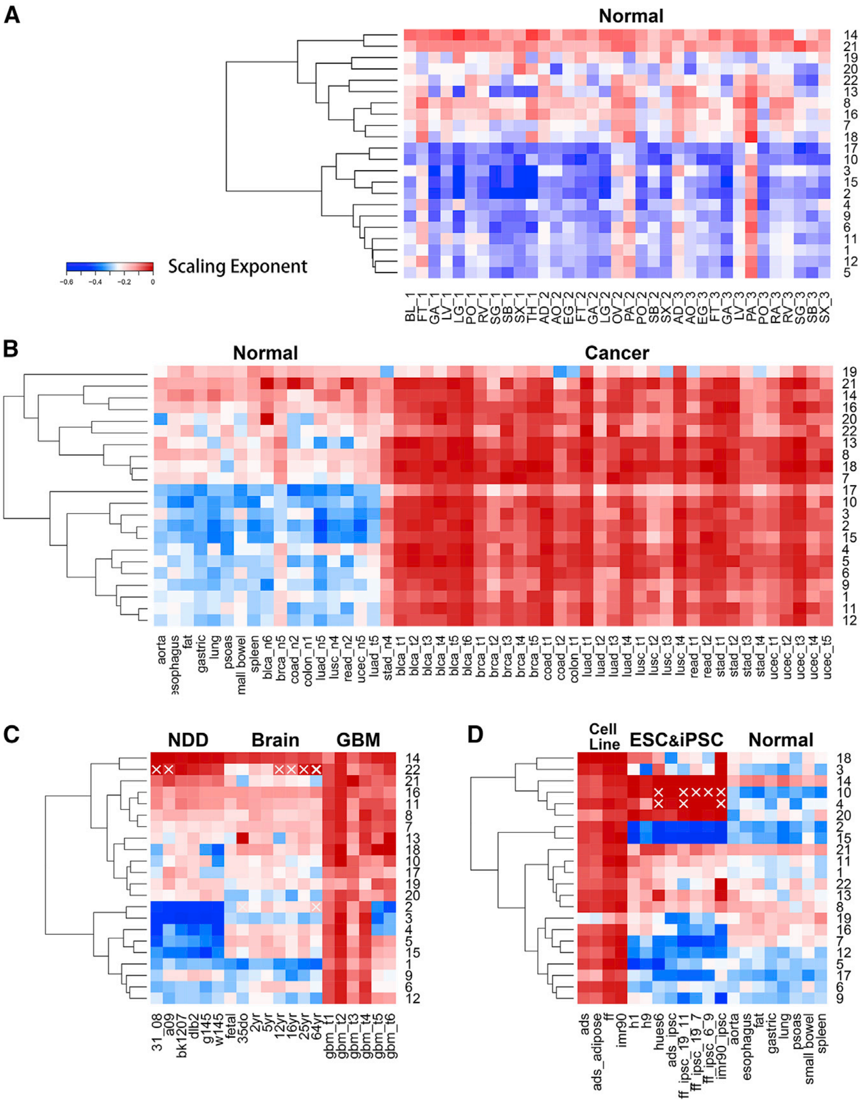

FIGURE 2 Heatmap clustering for the scaling exponents of all autosomal chromosomes shows differences between different cell classes. Sample labels are summarized in Supporting Material. (*A*) The scaling exponents of normal somatic cells are robust among three different individuals and the gland cells segregate from other cells. (*B*) Normal somatic cells (*left to right*: aorta to ucec_n5) segregate from cancer cells (brca_t1 to ucec_t5). luad_t5 and stad_n4 are further analyzed in Supporting Material. (*C*) Normal brain cells (fetal to 64 year) segregate from glioblastoma (GBM, gbm_t1 to gbm_t6) or NDDs (31_08 to w145). (*D*) ESCs and induced pluripotent stem cells (h1 to imr90_ipsc) segregate from cell lines including adult stem cell lines (ads) and somatic cell lines (ads_adipose, ff, and imr90). Normal somatic cells (aorta to spleen) were shown again for comparison. The chromosomes with correlations deviating from power law distributions are labeled with white forks (see Supporting Material for more analyses). To see this figure in color, go online.

accordance with their flatter double-log correlation curves (Fig. 1 *B*). The DFA of gland cells, brain cells, and ESCs are presented in Fig. S4, *C–E*, respectively.

The significant difference between different cell classes demonstrates that the long-range correlations in the DNA methylome cannot simply originate from the DNA sequence (25,26). DNA methylation was previously demonstrated to have long-range correlations by establishing a firm link with the A/B compartment (13), suggesting the scale-free property found here for DNA methylation to originate from chromatin structure, which is discussed later in more detail.

Clustering on the scaling exponents of chromosomes can be used to discern different cell types

The power law scaling behavior is observed in almost all chromosomes across a large variety of samples (Fig. S2 *C*). Hierarchical clustering for scaling exponents on all autosomal chromosomes demonstrates that most chromosomes behave similarly within each cell class, whereas chromosomes 14 and 21 tend to always have a higher scaling exponent (Fig. 2). When all chromosomes are compared, it can be clearly seen that cancer cells are distinguished from normal somatic cells (Fig. 2 *B*), consistent with the clustering on cell types (Fig. S3 *B*). Systematic differences are also clearly seen among normal brain cells, glioblastoma, and NDDs (Fig. 2 *C*). Different types of NDDs have similar scaling exponents whilst behaving significantly differently from glioblastoma, possibly highlighting their different pathogenesis (Fig. 2 *C*). In addition, the scaling exponent also clearly distinguishes ESCs and induced pluripotent stem cells from somatic cell lines and adult stem cell lines (Fig. 2 *D*).

When compared to the normal brain cells, all NDD samples analyzed here possess more negative scaling exponents for chromosome 2, 3, 5, and 15, suggesting their common roles associated with the neural diseases. In contrast, chromosome 19 shows little variation among all samples.

A small number of chromosomes possess correlations that deviates from a simple power law scaling (Fig. S2 *D*). We systematically identify such chromosomes (see Materials and Methods) which are, interestingly, mainly found in certain cells and particular chromosomes, namely chromosome 22 of the brain samples and chromosomes 4 and 10 of ESCs and induced pluripotent stem cells. Although the atypical power law behavior of these chromosomes may reflect the large fluctuation of the original methylation data, the clustered behavior could also suggest that these particular chromosomes have peculiar structures and functions that call for further studies. For example, it is known that genes in chromosome 22 are dense and that genetic disorder in chromosome 22 is associated with brain abnormalities (27).

Patchiness of DNA methylation is found along the genome and contributes to the power law scaling

Extensive changes in DNA methylation take place during tumorigenesis (28,29). In cancer cells, a large amount of long-range DNA hypomethylation was identified, distinct from the DNA methylation of normal cells (28). A domain with long-range DNA hypomethylation is termed a PMD (14). The DNA methylation profile illustrating the PMD formation in cancer cells can also be seen in Fig. 6 *C*. The IMR90 cell line also has such DNA hypomethylation character (14,28,30).

For cancer cells or the IMR90 cell line, the whole chromosome can be viewed as composed of alternating low-methylation-level domains (i.e., PMDs) and high-methylation-level domains (i.e., non-PMDs) in contrast to other cells. That is, the patchiness of DNA methylation for cancer cells or the IMR90 cell line is more apparent than in normal somatic and stem cells. The scaling exponents for IMR90 and cancer cells are similar to each other (Fig. 2). Such a coincidence promoted us to investigate whether the patchiness of DNA methylation contributes to the different scaling exponents of long-range DNA methylation in different cell classes.

Chromosome 1 in the IMR90 and h1 cell lines is taken as an example. There are 34% PMDs in IMR90, whereas h1 lacks PMDs. Namely, IMR90 and h1 have different degrees of DNA methylation patchiness. To understand how such patchiness is generated in IMR90 but not h1 cells, their Hi-C data, which are available, are used in the next section for structural modeling (19,20).

Here, we show that the high-low alternative pattern of DNA methylation is enough to mathematically reproduce the slow-decaying correlation in IMR90. We discretized the DNA methylation level of IMR90 and h1 into 1 and 0 with the methylation average as a reference value. Specifically, for chromosome 1 of each cell type, we assign a value of 1 to every 200 bp unit with a methylation level greater than that of the chromosome average, and 0 to those with a methylation level smaller than average. The correlations of the two discrete model series were calculated and shown in Fig. 3 *B*. The corresponding correlations of experimental DNA methylation level are also plotted in Fig. 3 *A* for comparison. The discrete model series also possesses the power law scaling behavior at the tens of kilobases scale (Fig. 3 *B*). The comparison between Fig. 3, *A* and *B* shows that the discrete model is able to reproduce the different scaling exponents in IMR90 and h1 cell lines, proving that the difference mainly comes from the different patchiness of their DNA methylation patterns. That is, the alternation of low and high methylation alternation along the genome in IMR90 results in the lower power law scaling exponent compared to the h1 cell line.

In addition, we calculated the correlations of the discrete model series for all the samples used in Fig. 1. The results are shown in Fig. S5. The scaling exponents of long-range DNA methylation for normal somatic cells, cancer cells, normal brain cells, and gland cells are −0.18 ± 0.02, −0.06 ± 0.02, −0.26 ± 0.03, and −0.13 ± 0.03, respectively. The scaling exponents using the discrete model series are the same as the experimental DNA methylation for cancer cells. For the other three cell classes, these two values differ, but only slightly. The order of the scaling

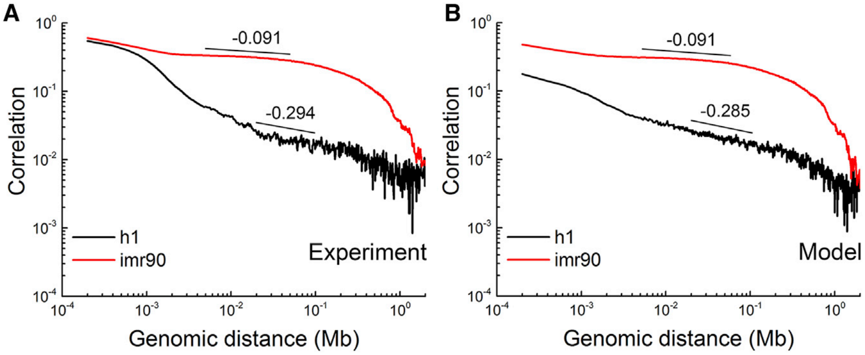

FIGURE 3 The comparison between the experimental and discrete model DNA methylation correlation indicates that the power law scaling mainly originates from the patchiness of DNA methylation. The long-range correlation of DNA methylation for chromosome 1 in IMR90 and h1 cell lines from: (*A*) experimental methylation data and (*B*) the discrete model. To see this figure in color, go online.

exponents among the four cell classes is maintained after discretization, again demonstrating that the high-low alternative pattern of DNA methylation accounts for the cell class-specific scaling exponents.

Patchiness of DNA methylation in IMR90 is related to chromatin structure

Next, we show the patchiness of DNA methylation reflects the packing of DNA in the 3D chromatin structure by mapping methylation level onto the modeled chromatin structure using Hi-C data.

We have developed a polymer modeling strategy using Hi-C data to construct the chromatin structure (see Materials and Methods and (18)). Hi-C data provide the frequency of physical interactions between any different genomic loci (31), and the frequencies can be further related with spatial distances (32). We used structural optimization to obtain the coarse-grained chromatin conformations meeting the distance constraints derived from Hi-C data.

We modeled the structures of chromosome 1 from IMR90 and h1 cells and mapped their DNA methylation levels onto the structures, which are respectively shown in Fig. 4, *A* and *B*. The two chromatin structures have obviously different organizations. Chromosome 1 of IMR90 shows a somewhat spherical appearance (Fig. 4 *A*), whereas the h1 chromosome adopts a scissor-like conformation (Fig. 4 *B*), suggesting structural changes during cellular differentiation. The mapping of DNA methylation level might provide a clue of how the different patchiness of DNA methylation in the two cell lines happens. It is interesting to observe that genomic regions with low methylation levels (colored *blue* in Fig. 4 *A*) are largely located close to each other in the chromatin model reconstructed based on the Hi-C data in IMR90. In contrast, the segregation of DNA methylation is not obvious in the h1 cell line (Fig. 4 *B*).

In our previous work, we have shown that the segregated low methylation regions (PMDs) in IMR90 are related to lamina-associated domains and chromatin compartment B, as well as other genome features (18), showing that the formation of PMDs may be caused by the improper function of DNA methyltransferase in chromatin compartment B and may be the origin of different patchiness in IMR90 compared to h1. Nearly all of the PMDs locate in chromatin compartment B and segregate from other genomic regions (chromatin compartment A) (18). This confirms the spatial segregation of the DNA methylation level in IMR90, qualitatively seen from the rendered chromosome structure (Fig. 4 *A*). These results suggest that the different patchiness of DNA methylation in different cells is related to their different chromatin structures. Thus, the long-range power law correlation for DNA methylation can reflect the spatial organization of chromatin, which in itself is hierarchical.

Local methylation correlations suggest the different chromatin structure in PMDs and non-PMDs

In the previous section, we have shown that the long-range correlation of DNA methylation reflects the global packing of DNA in chromatin. Next, we show that the local methylation correlations in PMDs and non-PMDs reflect their different structures. IMR90 cells, whose DNA methylation and nucleosome occupancy were obtained together using the nucleosome occupancy and methylome sequencing technique, is used as an example (33).

We compared the local correlation of CpG methylation in PMDs and non-PMDs in IMR90 cells (Fig. 5 *A*). Consistent with previous studies on the IMR90 cell line (34), the decay of PMD correlation clearly shows an obvious periodic behavior at base-resolution. The non-PMD regions, in contrast, show very weak periodic behavior.

The periodicities in different genomic regions were then quantified using FFTs of their local correlations. For PMD, the FFT of its local correlation shows a strong peak at 181 bp. At a similar position, a much weaker peak was found for non-PMD regions (Fig. 5 *B*). The period of 181 bp is consistent with the nucleosome repeat length

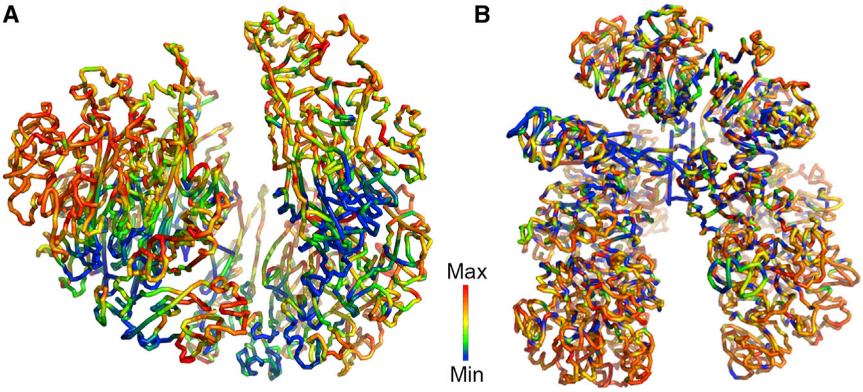

FIGURE 4 Patchiness of DNA methylation is related to chromatin structure. Modeled chromatin structures of chromosome 1 with mapped methylation level in (*A*) IMR90 and (*B*) h1 cell lines. Blue and red colors represent low and high DNA methylation levels, respectively. The DNA methylation data were obtained from Lister et al. (14). To see this figure in color, go online.

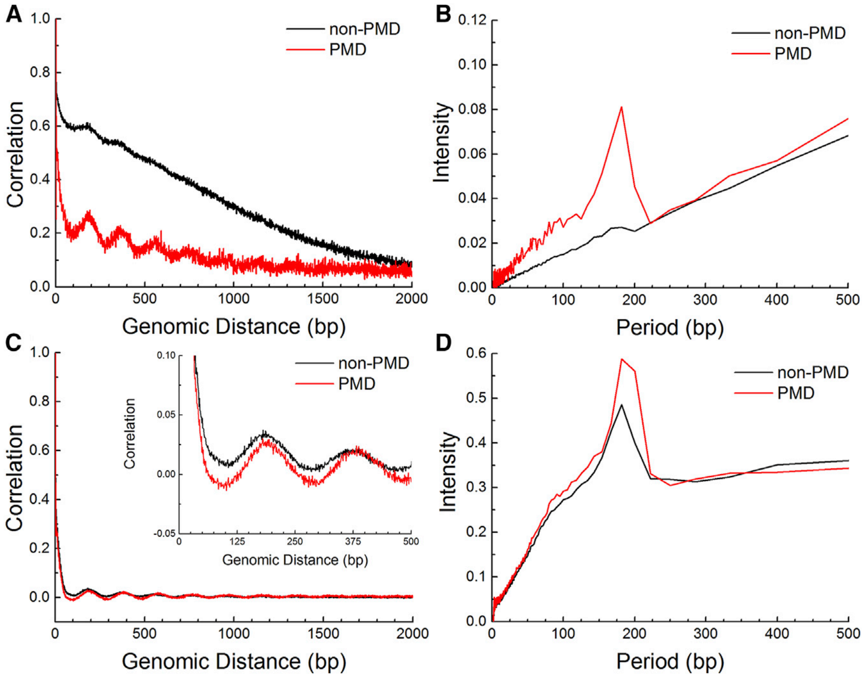

FIGURE 5 The local correlations of DNA methylation and nucleosome occupancy in PMD and non-PMD genomic regions of IMR90 cells. (*A* and *B*) Local correlations of DNA methylation and FFT analysis. (*C* and *D*) Local correlations of nucleosome occupancy and FFT analysis. The genomic distances below 500 bp are enlarged and shown in the inset in (*C*). To see this figure in color, go online.

(NRL), suggesting that this periodicity may come from the regular organization of nucleosomes in PMDs and that the nucleosomes in non-PMDs are relatively irregularly spaced.

We further analyzed the nucleosome occupancy and methylome sequencing data of nucleosome occupancy (33). The local correlation of nucleosome occupancy in PMD and non-PMD regions is shown in Fig. 5 *C* and their FFTs are plotted in Fig. 5 *D*. The local correlation of nucleosome occupancy has a 182 bp period in both PMD and non-PMD genomic regions, as seen from the FFTs of the local correlations (Fig. 5 *D*). Interestingly, the periodicity for PMDs is stronger than that in non-PMDs (Fig. 5 *C*), similarly to their differences in methylation patterns but to a lesser extent. Such a result is consistent with the possibility that the local DNA methylation correlates with nucleosome organization. The regularity of nucleosome arrangement can be weakened by nucleosome depletion or different NRLs. Both these factors severely affect chromatin organization and compaction. Nucleosome depletion massively influence chromatin flexibility (26,35,36). With different NRLs, the nucleosomes can form 30 nm higher order chromatin structure or other chromatin fibers (37,38). Therefore, the local correlation of DNA methylation suggests that the chromatin structure of PMDs and non-PMDs is different at the kilobase genomic scale.

Such a conclusion is also consistent with our previous analysis of Hi-C data (18). We found that the Hi-C patterns for PMDs and non-PMDs are obviously different, which again shows that these two domains have different spatial organization. From Hi-C data, it is easy to see that all the PMDs have uniform physical contact within its interior, whereas the majority of non-PMDs contain localized interaction domains.

Gene expression in PMDs and PMD-like regions are repressed

To understand how the patchiness of DNA methylation is related to biological functions, we analyzed the gene expression in PMDs and non-PMDs. As explained in the Materials and Methods, we analyzed the four tumor-normal sample pairs in TCGA. It was previously found that the PMDs in IMR90 correlate with repressive and anticorrelate with active histone marks (28). In addition, the CGI promoters are hypermethylated in PMDs (28).

Consistent with earlier studies (24,30,39), we find that genes within PMDs in cancer samples tend to be transcriptionally repressed (Fig. S6; Tables S6 and S7) and, interestingly, these genes are related to specific functions. Genes within cancer PMDs mainly relate to Gene Ontology terms such as cell membrane, glycoprotein, disulfide bond, olfaction, cadherin, and receptor (Table S8), which suggests that some intra-PMD genes regulating cell communication tend to be repressed. In addition, almost all housekeeping genes (3794 of 3796) are located outside PMDs, consistent with their essential role in fundamental cellular function (Table S8).

Taking the brca_t5 tumor sample as an example, there are 473 genes intersecting with PMDs, of which 305 are located within the PMD body and, in particular, 156 are in the PMD center (defined as the central 60% of the PMD), indicating that most genes embed in the PMD body. In addition, among the 473 genes intersecting with PMDs, 57.7% of them have non-CGI promoters (the definition of non-CGI and CGI promoter is from (40)) and this ratio is significantly higher than that of all genes (34.2%, 8420 of the total 24,630 genes), which indicates that genes with non-CGI promoters are enriched in PMDs (Table S8).

Besides the repressed gene expression level, we also find that the repression degree correlates with the PMD lengths in the four tumor-normal sample pairs. Fig. 6 *A* shows the correlation between gene repression levels and PMD lengths in brca and the results in the two other tumor-normal sample pairs (colon adenocarcinoma (coad) and uterine corpus endometrial carcinoma (ucec)) are presented in Fig. S7. With the increasing length of PMDs or PMD-like regions, the percentage of repressed genes increases, which may result from these genomic regions being buried in the compact chromatin regions in 3D space and probably also impedes the binding with transcription factors, RNA polymerase, or other regulators. The gene density of PMDs (2.737 in the brca_t5 sample) is also much lower than that of non-PMDs (6.526 in the brca_t5 sample), indicating the gene sparsity in PMDs.

We plotted the local correlation of different genomic regions of breast cells in Fig. 6 *B*. The decay of DNA correlation in the PMDs of breast cancer cells clearly shows a periodic behavior at base-resolution, just like the IMR90 cell line. The corresponding genomic regions of cancer PMDs in normal cells are defined as PMD-like regions. We found that PMD-like regions of breast cells have an average methylation level higher than PMDs and lower than non-PMDs (Fig. S8), and a less obvious methylation correlation periodicity (Fig. 6 *B*). Furthermore, the average expression level of genes in PMD-like regions is lower than that in non-PMDs and higher than that in PMDs, which is consistent with PMD-like regions' intermediate behavior in the DNA methylation level (Fig. S8) and periodicity of local methylation correlation (Fig. 6 *B*).

The intermediate properties and the similar genomic locations of PMD-like regions to the PMDs imply the role of PMD-like regions in tumor development. PMD-like regions may be the precursor of tumor PMDs in which the genes regulating cell communications are further repressed. Interestingly, PMD or PMD-like domains tend to lie in genomic regions with lower CpG density (Fig. 6 *C*), suggesting that they belong to different isochores (41). In this analysis, the Fisher's exact test between the expression level of PMDs and PMD-like regions and that between PMDs and non-PMDs is shown in Table S7. The average expression level in PMDs, PMD-like regions, and non-PMDs is shown in Table S6. The comparison of expression level in PMD and PMD-like regions in the four tumor-normal sample pairs is shown in Fig. S6.

DISCUSSION

Power law scaling in cancer cells is not caused by the lower average methylation level or copy number variation (CNV). One difference between cancer and normal somatic cells in DNA methylation is that the former appears to be demethylated in PMDs compared to the latter. To show that the more sustained correlation of cancer cell DNA methylation is not caused by this overall demethylation, we checked the scaling exponents of methylation correlations among cells with large variations in methylation levels. We calculated the methylation correlations of human inner cell mass (42) and primordial germ cells (43). The average methylation levels of these cells are both significantly lower than normal somatic cells, as has been found for cancer cells (Fig. S9 *B*). However, the scaling exponents for inner cell mass and primordial germ cells are nearly the same as normal cells and much lower than those of cancer cells (Fig. S9 *C*). Thus, it

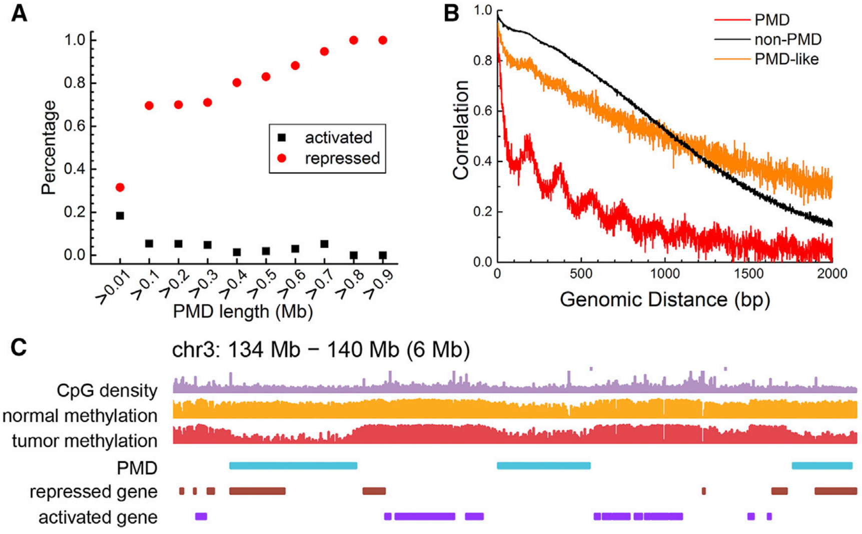

FIGURE 6 Gene expression and sequence property of PMD and non-PMD in breast cells. (*A*) Percentage of genes that are transcriptionally repressed or activated in oncogenesis as a function of PMD length in the brca_t5-brca_n5 sample pair. The comparisons in coad_t2-coad_n2 and ucec_t5-ucec_n5 sample pairs are shown in Fig. S6. (*B*) The local correlations of DNA methylation in PMD, non-PMD, and PMD-like genomic regions in breast cells. (*C*), Plot of a representative region showing the relationship between CpG density, methylation levels before and after oncogenesis, PMD, and repressed and activated genes in oncogenesis. In (*A*), (*B*), and (*C*), the data from brca cells were used. To see this figure in color, go online.

can be concluded that average methylation level does not determine the different scaling exponents across cell classes.

Since tumor samples are enriched in CNVs, we showed that the DNA methylation correlation is little affected by CNVs in cancer cells. For example, the long-range methylation correlation for chromosome 1 behaves similarly whether the CpG sites within CNVs of the brca_t5 tumor sample are included or excluded (Fig. S9 *D*) in the correlation calculations. We also checked the single-cell WGBS sequencing data and found that the long-range correlation pattern was quite well-conserved (Fig S9 *E*). Therefore, DNA methylation correlation found in this work is also conserved among different individual cells.

CONCLUSION

Through exploiting the chromatin structure modeled based on Hi-C data and the underlying long-range and local correlations of the DNA methylome, our study provides a comprehensive view of the flow of genetic information, connecting DNA sequence, CpG methylation, local and long-range chromatin structure, and gene expression. In normal somatic cells, DNA sequences with low CpG density correlate with low methylation levels and low expression levels (PMD-like). The development of cancers is associated with further decreases of the average methylation level in PMD-like regions, some of which turn into PMDs containing further suppressed genes. The correlation of methylation shows consistent differences among different classes of cells, including normal somatic cells, cancer cells, brain cells, gland cells, and stem cells, that are highly conserved within each class. The clear cell class dependence of the long-range power law scaling in methylation correlation shows that it can serve as a simple measure to discriminate cells at normal and pathological states. Such a finding points to a new direction, to our knowledge, in the analysis of the development of different diseases, such as cancers and NDDs, at the chromatin level.

SUPPORTING MATERIAL

Supporting Materials and Methods, nine figures, and eight tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)30911-6.

AUTHOR CONTRIBUTIONS

Y.Q.G designed the research. L.Z., W.J.X., S.L., and L.M. performed research. W.J.X., L.Z., S.L., and L.M. wrote the manuscript. C.G. contributed to the analysis of the data.

ACKNOWLEDGMENTS

We thank Fuchou Tang, Chengqi Yi, and Xiaoliang S. Xie for helpful comments on the manuscript. The results shown here are partly based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. We are grateful to Manel Esteller for providing the methylome of neurodegenerative diseases. This work was supported by the National Natural Science Foundation of China (grants 21573006, 21233002, and 91427304 to Y.Q.G.).

REFERENCES

1. Gibcus, J. H., and J. Dekker. 2013. The hierarchy of the 3D genome. *Mol. Cell.* 49:773–782.
2. Levine, M., C. Cattoglio, and R. Tjian. 2014. Looping back to leap forward: transcription enters a new era. *Cell.* 157:13–25.
3. Galupa, R., and E. Heard. 2015. X-chromosome inactivation: new insights into cis and trans regulation. *Curr. Opin. Genet. Dev.* 31:57–66.
4. Naumova, N., M. Imakaev, …, J. Dekker. 2013. Organization of the mitotic chromosome. *Science.* 342:948–953.
5. Pombo, A., and N. Dillon. 2015. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* 16:245–257.
6. Dekker, J. 2008. Gene regulation in the third dimension. *Science.* 319:1793–1794.
7. Jenuwein, T., and C. D. Allis. 2001. Translating the histone code. *Science.* 293:1074–1080.
8. Aranda, S., G. Mas, and L. Di Croce. 2015. Regulation of gene transcription by Polycomb proteins. *Sci. Adv.* 1:e1500737.
9. Boettiger, A. N., B. Bintu, …, X. Zhuang. 2016. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature.* 529:418–422.
10. Cedar, H., and Y. Bergman. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* 10:295–304.
11. Chodavarapu, R. K., S. Feng, …, M. Pellegrini. 2010. Relationship between nucleosome positioning and DNA methylation. *Nature.* 466:388–392.
12. Bell, A. C., and G. Felsenfeld. 2000. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature.* 405:482–485.
13. Fortin, J. P., and K. D. Hansen. 2015. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.* 16:180.
14. Lister, R., M. Pelizzola, …, J. R. Ecker. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature.* 462:315–322.
15. Huang, W., B. T. Sherman, and R. A. Lempicki. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37:1–13.
16. Huang, W., B. T. Sherman, and R. A. Lempicki. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4:44–57.
17. Eisenberg, E., and E. Y. Levanon. 2013. Human housekeeping genes, revisited. *Trends Genet.* 29:569–574.
18. Xie, W. J., L. Meng, …, Y. Q. Gao. 2017. Structural modeling of chromatin integrates genome features and reveals chromosome folding principle. *Sci. Rep.* 7:2818.
19. Rao, S. S., M. H. Huntley, …, E. L. Aiden. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* 159:1665–1680.
20. Dixon, J. R., S. Selvaraj, …, B. Ren. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature.* 485:376–380.
21. Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Rev.* 51:661–703.
22. Eugene, V. K., I. W. Yuri, and P. K. Georgy. 2006. Power Laws, Scale-Free Networks and Genome Biology. Springer, New York.

23. Schneider, R., and R. Grosschedl. 2007. Dynamics and interplay of nuclear architecture, genome organization, and gene expression. *Genes Dev.* 21:3027–3043.

24. Schultz, M. D., Y. He, …, J. R. Ecker. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature.* 523:212–216.

25. Peng, C. K., S. V. Buldyrev, …, H. E. Stanley. 1992. Long-range correlations in nucleotide sequences. *Nature.* 356:168–170.

26. Arneodo, A., C. Vaillant, …, C. Thermes. 2011. Multi-scale coding of genomic information: from DNA sequence to genome structure and function. *Phys. Rep.* 498:45–188.

27. McDermid, H. E., and B. E. Morrow. 2002. Genomic disorders on 22q11. *Am. J. Hum. Genet.* 70:1077–1088.

28. Berman, B. P., D. J. Weisenberger, …, P. W. Laird. 2011. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* 44:40–46.

29. Hon, G. C., R. D. Hawkins, …, B. Ren. 2012. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* 22:246–258.

30. Lister, R., M. Pelizzola, …, J. R. Ecker. 2011. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature.* 471:68–73.

31. Lieberman-Aiden, E., N. L. van Berkum, …, J. Dekker. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 326:289–293.

32. Serra, F., M. Di Stefano, …, M. A. Marti-Renom. 2015. Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.* 589 (20 Pt A):2987–2995.

33. Collings, C. K., and J. N. Anderson. 2017. Links between DNA methylation and nucleosome occupancy in the human genome. *Epigenet. Chromatin.* 10:18.

34. Gaidatzis, D., L. Burger, …, M. B. Stadler. 2014. DNA sequence explains seemingly disordered methylation levels in partially methylated domains of Mammalian genomes. *PLoS Genet.* 10:e1004143.

35. Diesinger, P. M., and D. W. Heermann. 2009. Depletion effects massively change chromatin properties and influence genome folding. *Biophys. J.* 97:2146–2153.

36. Ricci, M. A., C. Manzo, …, M. P. Cosma. 2015. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell.* 160:1145–1158.

37. Grigoryev, S. A. 2012. Nucleosome spacing and chromatin higher-order folding. *Nucleus.* 3:493–499.

38. Routh, A., S. Sandin, and D. Rhodes. 2008. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proc. Natl. Acad. Sci. USA.* 105:8872–8877.

39. Schroeder, D. I., J. D. Blair, …, J. M. LaSalle. 2013. The human placenta methylome. *Proc. Natl. Acad. Sci. USA.* 110:6037–6042.

40. Saxonov, S., P. Berg, and D. L. Brutlag. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. USA.* 103:1412–1417.

41. Costantini, M., O. Clay, …, G. Bernardi. 2006. An isochore map of human chromosomes. *Genome Res.* 16:536–541.

42. Guo, H., P. Zhu, …, J. Qiao. 2014. The DNA methylation landscape of human early embryos. *Nature.* 511:606–610.

43. Guo, F., L. Yan, …, J. Qiao. 2015. The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell.* 161:1437–1452.

Biophysical Journal, Volume *113*

Supplemental Information

DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells

Ling Zhang, Wen Jun Xie, Sirui Liu, Luming Meng, Chan Gu, and Yi Qin Gao

Supporting Material

DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells

Ling Zhang[1,2†], Wen Jun Xie[1†], Sirui Liu[1], Luming Meng[1], Chan Gu[1,2] and Yi Qin Gao[1,2*]

[1] Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China.

[2] Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, Peking University, Beijing 100871, China.

*Corresponding author. Tel: 86-10-6275-2431; E-mail: gaoyq@pku.edu.cn

†These authors contributed equally to this work

The Supporting Material includes:

1. Sources of whole-genome bisulfite sequencing data
2. Detrended Fluctuation Analysis (DFA) for different cell classes
3. The luad_t5 and stad_n4 sample used in Fig. 2B
4. Gene analysis

1. Sources of whole-genome bisulfite sequencing data

Methylomes of human somatic cells

Reference: M. D. Schultz *et al.* (1)

URL: http://neomorph.salk.edu/human_tissue_methylomes.html

Data for individual 2 were used in the text (Fig. 1 and Fig. 2). Data for individual 1 and 3 were used to test the robustness of long-range correlations among individuals (Fig. S2A).

Table S1. Sample Details for Human Somatic Cells

number	name	symbol	individual	gender	age(year)
1	bladder	BL_1	1	male	3
2	fat	FT_1			
3	gastric	GA_1			
4	lung	LG_1			
5	left ventricle	LV_1			
6	psoas	PO_1			
7	right ventricle	RV_1			
8	thymus	TH_1			
9	small bowel	SB_1			
10	sigmoid colon	SG_1			
11	spleen	SX_1			
12	adrenal	AD_2	2	female	30
13	aorta	AO_2			
14	esophagus	EG_2			
15	fat	FA_2			
16	gastric	GA_2			
17	lung	LG_2			
18	ovary	OV_2			
19	pancreas	PA_2			
20	psoas	PO_2			
21	small bowel	SB_2			
22	spleen	SX_2			
23	adrenal	AD_3	3	male	34
24	aorta	AO_3			
25	esophagus	EG_3			
26	fat	FT_3			
27	gastric	GA_3			
28	lung	LG_3			
29	left ventricle	LV_3			
30	pancreas	PA_3			
31	psoas	PO_3			
32	right atrium	RA_3			
33	right ventricle	RV_3			

34	small bowel	SB_3
35	sigmoid colon	SG_3
36	spleen	SX_3

Methylomes of human cancer cells

The results shown here are partly based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

URL: https://portal.gdc.cancer.gov/legacy-archive/search/f

Reference for colon cells: B. P. Berman *et al.* (2)

The following typical samples were used to represent different cancers in Fig. 1B and Fig. S2: brca_t5, coad_t1, gbm_t2, luad_t1, lusc_t4, read_t2, stad_t1, ucec_t3 and colon_t1.

Table S2. Sample Details for Human Cancer Cells

number	TCGA barcode	symbol	cancer type
1	TCGA-A2-A04X-01A-21D-A19F-05	brca_t1	breast invasive carcinoma
2	TCGA-A8-A07I-01A-11D-A19F-05	brca_t2	
3	TCGA-A2-A0YG-01A-21D-A19F-05	brca_t3	
4	TCGA-E2-A15H-01A-11D-A19F-05	brca_t4	
5	TCGA-A7-A0CE-01A-11D-A148-05	brca_t5	
6	TCGA-A7-A0CE-11A-21D-A148-05	brca_n5	
7	TCGA-AA-A00R-01A-01D-A22T-05	coad_t1	colon adenocarcinoma
8	TCGA-AA-3518-01A-02D-1518-05	coad_t2	
9	TCGA-AA-3518-11A-01D-1518-05	coad_n2	
10	TCGA-06-0128-01A-01D-2294-05	gbm_t1	glioblastoma multiforme
11	TCGA-14-1454-01A-01D-2294-05	gbm_t2	
12	TCGA-14-3477-01A-01D-2294-05	gbm_t3	
13	TCGA-14-1401-01A-01D-2294-05	gbm_t4	
14	TCGA-16-1460-01A-01D-2294-05	gbm_t5	
15	TCGA-19-1788-01A-01D-2294-05	gbm_t6	
16	TCGA-38-4630-01A-01D-2365-05	luad_t1	lung adenocarcinoma
17	TCGA-67-6215-01A-11D-2365-05	luad_t2	
18	TCGA-78-7156-01A-11D-2365-05	luad_t3	
19	TCGA-91-6840-01A-11D-2365-05	luad_t4	
20	TCGA-44-6148-01A-11D-2365-05	luad_t5	
21	TCGA-44-6148-11A-01D-2365-05	luad_n5	
22	TCGA-34-2600-01A-01D-1871-05	lusc_t1	lung squamous cell carcinoma
23	TCGA-60-2695-01A-01D-1871-05	lusc_t2	
24	TCGA-21-1078-01A-01D-2365-05	lusc_t3	
25	TCGA-60-2722-01A-01D-1871-05	lusc_t4	
26	TCGA-60-2722-11A-01D-1871-05	lusc_n4	
27	TCGA-AG-3593-01A-01D-2294-05	read_t1	rectum adenocarcinoma
28	TCGA-AF-2689-01A-01D-2294-05	read_t2	

29	TCGA-AF-2689-11A-01D-2294-05	read_n2	
30	TCGA-CG-5730-01A-11D-2365-05	stad_t1	stomach adenocarcinoma
31	TCGA-D7-6519-01A-11D-2365-05	stad_t2	
32	TCGA-F1-6177-01A-11D-2365-05	stad_t3	
33	TCGA-BR-6452-01A-12D-2365-05	stad_t4	
34	TCGA-BR-6452-11A-01D-2365-05	stad_n4	
35	TCGA-B5-A0K6-01A-11D-A23D-05	ucec_t1	uterine corpus endometrial carcinoma
36	TCGA-AX-A1CK-01A-11D-A23D-05	ucec_t2	
37	TCGA-AP-A05J-01A-11D-A23D-05	ucec_t3	
38	TCGA-A5-A0G2-01A-11D-A23D-05	ucec_t4	
39	TCGA-AX-A1CI-01A-11D-A17H-05	ucec_t5	
40	TCGA-AX-A1CI-11A-11D-A17H-05	ucec_n5	
41	TCGA-DK-A1AA-01A-11D-A23D-05	blca_t1	bladder urothelial carcinoma
42	TCGA-DK-A1AG-01A-11D-A23D-05	blca_t2	
43	TCGA-BL-A13J-01A-11D-A23D-05	blca_t3	
44	TCGA-BT-A2LA-01A-11D-A23D-05	blca_t4	
45	TCGA-H4-A2HQ-01A-11D-A23D-05	blca_t5	
46	TCGA-BT-A20V-01A-11D-A23D-05	blca_t6	
47	TCGA-BT-A20V-11A-11D-A23D-05	blca_n6	

colorectal cancer

number	name	symbol	cancer type
48	colon tumor	colon_t1	colorectal cancer
49	colon normal	colon_n1	

Methylomes of human and mouse brain cells

Reference: R. Lister *et al.* (3)

URL: http://neomorph.salk.edu/brain_methylomes/

Table S3. Sample Details for Human Brain Cells

number	species	symbol	brain region	cell type	gender	age
1	human	fetal	cerebral cortex	tissue	male	20 week
2	human	35do	middle frontal gyrus	tissue	male	35 day
3	human	2yr	middle frontal gyrus	tissue	male	2 year
4	human	5yr	middle frontal gyrus	tissue	male	5 year
5	human	12yr	middle frontal gyrus	tissue	male	12 year
6	human	16yr	middle frontal gyrus	tissue	male	16 year
7	human	25yr	middle frontal gyrus	tissue	male	25 year
8	human	64yr	frontal cortex	grey matter	female	64 year
9	mouse	10wk	frontal cortex	tissue	male	10 week

Methylomes of human stem cells

Reference: R. Lister *et al.* (4)

URL: http://neomorph.salk.edu/ips_methylomes/data.html

Table S4. Sample Details for Human Stem Cells

number	name	symbol
1	ADS(adipose-derived stem cells)	ads
2	adipocytes derived from the ADS cells	ads_adipose
3	ADS iPSCs	ads_ipsc
4	foreskin fibroblast(FF)	ff
5	FF iPSC 6.9	ff_ipsc_6_9
6	FF iPSC 19.7	ff_ipsc_19_7
7	FF iPSC 19.11	ff_ipsc_19_11
8	IMR90(fetal lung fibroblast)	imr90
9	IMR90-iPSC	imr90_ipsc
10	H1	h1
11	H9	h9
12	HUES6	hues6

*For hues6, the reference is R. Lister *et al.(3)*

Methylomes for human neurodegenerative diseases

Reference: J.V. Sanchez-Mut *et al.* (5)

FastQ format reads of neurodegenerative diseases methylome were kindly provided by M. Esteller and the reads were aligned to the hg19 human reference genome with the Bowtie alignment algorithm(6).

Table S5. Sample Details for Human Neurodegenerative Diseases

number	name	symbol	disease	region	age (year)	gender
1	A09	a09	Alzheimer's disease	Brodmann area 9 gray matter	81	female
2	DBL2	dbl2	Dementia with Lewy bodies	Brodmann area 9 gray matter	77	female
3	BK1207	bk1207	Parkinson's disease	Brodmann area 9 gray matter	77	female
4	31_08	31_08	Down syndrome with Alzheimer's disease	Brodmann area 9 gray matter	49	male
5	G145	g145	Control gray matter	Brodmann area 9 gray matter	64	female
6	W145	w145	Control white matter	Brodmann area 9 white matter	64	female

2. Detrended Fluctuation Analysis (DFA) for different cell

classes

DFA has been used to show the long-range correlation in DNA sequence (7). Here we use the DFA to demonstrate the long-range correlation in DNA methylome. Root mean square fluctuation F(r) of a one-dimensional sequence is an important statistical quantity. It is typically defined as

$$\tilde{F}(r)^2 = \overline{[\Delta s(r)]^2} - \overline{\Delta s(r)}^2 \tag{1}$$

where $s(r) = \sum_{i=1}^{r} u(r)$ is the sum of the methylation level for the first r<u>th</u> units, $\Delta\mathrm{s}(\mathrm{r}) = \mathrm{s}(r_0 + \mathrm{r}) - \mathrm{s}(r_0)$, the bars indicate an average over all possible r_0 in the sequence. To make comparisons simple, here we normalize $\tilde{F}(r)$ as

$$F(r)^2 = \tilde{F}(r)^2/Var \tag{2}$$

so that all detrended fluctuations start from the same point F(1)=1. Here *Var* is the variance of the methylation level of the whole sequence. This normalized *F*(*r*) is directly related to the correlation function *C*(*r*) through the equation

$$F(r)^2 \approx \prod_{j,k=1}^{r} C(j-k). \tag{3}$$

The '≈' can be replaced by a '=' as long as r is much smaller than sequence length L, which is often the case.

For purely uncorrelated random sequences, $F(r) \sim r^{1/2}$, corresponding to a ~0.5 slope in double-log plot. If the correlation of a sequence decays exponentially, indicating a finite-range correlation, the fluctuation scaling exponent will also be 0.5. Only when a long-range correlation with an infinite characteristic length is expected, will the scaling exponent deviate from 0.5, thus may be described by a power law. If the sequence holds a power law correlation when extending to infinite length, that is to say, $F(r) \sim r^{\alpha}$ and $C(r) \sim r^{-\gamma}$ when $r \to \infty$, there is a simple relation between fluctuation scaling exponent α and correlation scaling exponent γ

$$\alpha = \frac{2-\gamma}{2}\ . \tag{4}$$

However, in most cases, one can only expect a finite sequence length, thus the quantitative relation described above may not be accurate, but the qualitative property that a higher α corresponds to a lower γ still holds. For DNA methylation, the higher is the fluctuation scaling exponent, the flatter correlation double-log plot is, indicating a slower long-range decay.

We use Equation 3 to calculate the detrended fluctuation from 200-bp resolution methylation correlation. The fluctuation scaling exponent is estimated by linearly fitting the double-log plot of detrended fluctuation in 2kb~0.2Mb range. The average scaling exponents of 0.76 ± 0.01 and 0.92 ± 0.02 are observed in normal somatic cells (Fig. S4A) and cancer cells (Fig. S4B) separately. Their deviation from 0.5 and small variances indicate a uniform power law decay within certain cell states among different types of tissues. Cancer cells hold an obviously higher scaling exponent, in accordance with their flatter double-log correlation curves.

The detrended fluctuation analyses are also performed on gland cells (Fig. S4C) and brain cells (Fig. S4D). Gland cells show similar but smaller positive deviation from normal somatic cells in fluctuation as cancer cells, while the scaling exponent of brain cells demonstrates no significant difference from that of normal somatic cells, in contrast to the corresponding correlation analysis. Since DFA is based on the sum of correlations, it largely reduces the random fluctuation in correlation. However, it also loses the detailed information through summing with higher cumulative weights for shorter-range correlations and responds much slower to scaling changes than correlation. Thus the methylome landscape differences between brain and normal somatic cells could be concealed by this cumulative operation in fluctuation analyses.

To further investigate the methylome landscape in differentiation, we apply DFA on human stem cells and related samples (Fig. S4E). All the stem cells and related fibroblasts are grouped into two categories according to DFA results, one containing primary somatic cell lines like foreskin fibroblast (ff) and IMR90 as well as adult stem cell lines like adipose-derived stem cells (ads) and adipocytes derived from ads (ads_adipose), the other containing all the human embryonic stem cell (hESC) samples and induced pluripotent stem cell (iPSC) samples. The former category shows an averaged scaling exponent of 0.91±0.03, similar to that in cancer, while the latter gives an exponent of 0.69±0.02, suggesting a long-range correlation with negative deviation from somatic cells.

3. The luad_t5 and stad_n4 sample used in Fig. 2B

As can be seen from Fig. 2B, the luad_t5 sample was clustered into normal cells using the scaling exponents of all chromosomes. The somatic mutations and copy number variations of luad_t5 sample were also analyzed to identify its clinical status. The number of somatic mutations in this sample is 17, which is smaller than that of regular tumor samples. The probability distribution of CNVs in this sample is not a single-peaked distribution. These data indicate that from the perspective of somatic mutations, the sample behaves like a normal one but the CNVs proves that it is actually a tumor sample. The AJCC stage of the patient is Stage IA, so we guess luad_t5 sample has not utterly become tumor sample.

The probability distribution of CNVs in stad_n4 sample is a unimodal distribution with a high peak at 0, indicating that there are very small number of CNVs in this sample and the sample behaves as normal cells. However, the stad_n4 sample's AJCC stage is Stage IIA which is consistent with our clustering that stad_n4 might have some cancer properties.

4. Gene analysis

Table S6. Average Expression Levels of Genes in PMDs, PMD-like Regions and Non-PMDs

Cancer type	Number of genes intersecting with PMDs	PMD	PMD-like regions	Non-PMD in tumor	Non-PMD in normal
brca	473	1.69	5.38	40.69	43.17
coad	642	8.96	24.53	24.63	28.48
luad	133	2.73	3.17	39.88	37.96
ucec	1109	7.94	12.96	23.30	36.24

* Expression levels in this table are TPM (transcripts per million).

* The length of genomic regions used in this table are all greater than 0.1 M.

* The normal samples correspond to the tumor sample of the same patient.

Table S7. Fisher's Exact Test of Gene Expression Level in PMDs, PMD-like Regions and Non-PMDs

Cancer type	Tumor PMD and Tumor non-PMD	Normal PMD-like regions and Tumor non-PMD	Tumor PMD and Normal PMD-like regions
brca	$7.2077*10^{-108}$	$6.8137*10^{-86}$	$9.6417*10^{-04}$
coad	$1.9133*10^{-54}$	$5.2351*10^{-47}$	0.4660
luad	$1.0565*10^{-33}$	$1.6939*10^{-30}$	0.3846
ucec	$1.0527*10^{-93}$	$9.1349*10^{-151}$	$1.2888*10^{-09}$

* The length of genomic regions used in this table are all greater than 0.1 M.

Table S8. Classification of Genes Intersect with PMDs

Genes	Number	Classification Feature	Specific genes	Number	Ratio
genes intersecting with PMDs	473	location	genes within PMDs	305	0.645
			genes within PMD center	156	0.330
		state	specifically expressed in tumor sample	17	0.036
			specifically expressed in normal sample	55	0.116
			repressed genes	167	0.353
			activated genes	13	0.027
		GO term	disulfide bond	190	0.402
			glycoprotein	220	0.465
			membrane	231	0.488
			signal	149	0.315
			housekeeping genes	2	0.004
		Promoter type	Non-CGI promoter	273	0.577
			CGI promoter	200	0.423

* PMDs with genomic lengths greater than 0.1M are considered in this table.

* Total number of housekeeping genes is 3796.

* PMD center is defined as the central 60% regions of PMD.

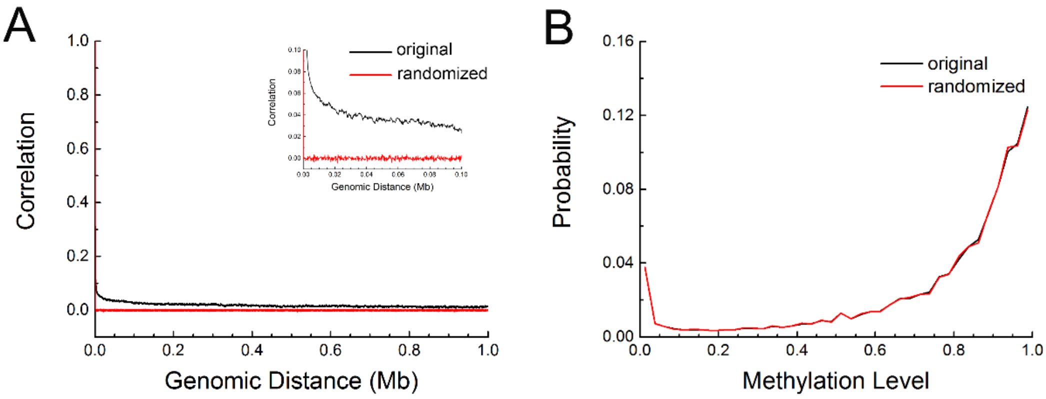

Figure S1. The non-random nature of DNA methylation. (A) The DNA methylation correlations from original experimental data and randomized data for chromosome 1 of human aorta cell (sample label: AO_2). The genomic distances below 0.10 Mb are enlarged and shown in the inset. (B) Methylation level distribution of original experimental data and randomized data. The randomized data was produced by assigning each CpG site with a random value following the overall distribution of DNA methylation level. We first generated a random number (y) following the uniform distribution between 0 and 1, and found the highest x satisfying F(x) ≤y, where F(x) is the cumulative distribution function of methylation. The value of x was assigned to each CpG site as its methylation level.

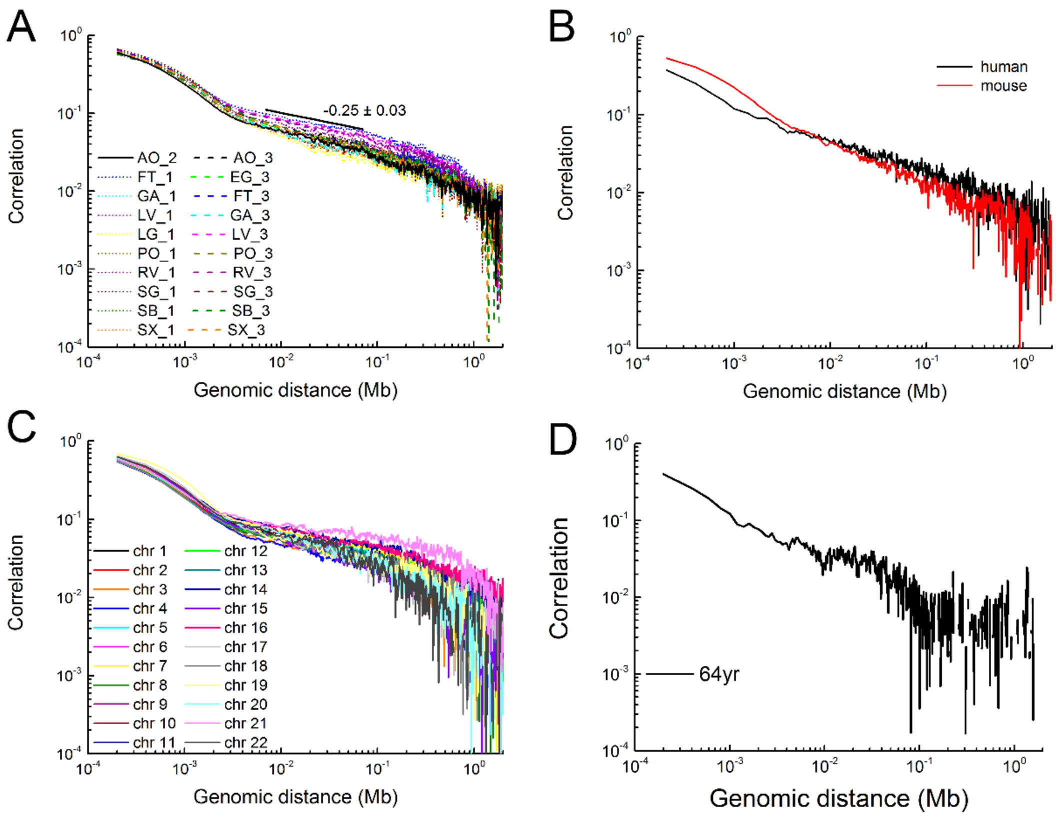

Figure S2. Power law scaling of methylation correlation in different individuals, species, chromosomes and exceptions.

(A) Robustness of the scaling exponents among different individuals. The scaling exponents for chromosome 1 of human somatic cells in three different individuals. The small standard deviations show that the scaling exponents are conserved among different individuals. (B) The power law scaling is also present in mouse brain cell. 25-year-old human brain sample and 10-week-old mouse brain sample are used as examples of human and mouse brain, respectively. All the brain data are summarized in Table S3. (C) The power law scaling behavior is observed in different chromosomes. The chromosomes in aorta from individual 2 (sample label: AO_2) are used. (D) The scaling exponents in the concerned genomic region (kilobase to megabase) are not well-defined in some chromosomes. The chromosome 22 of the 64yr human brain is taken as an example. In the kilobase to megabase region, the large fluctuation of methylation correlation makes it not feasible to calculate the scaling exponent.

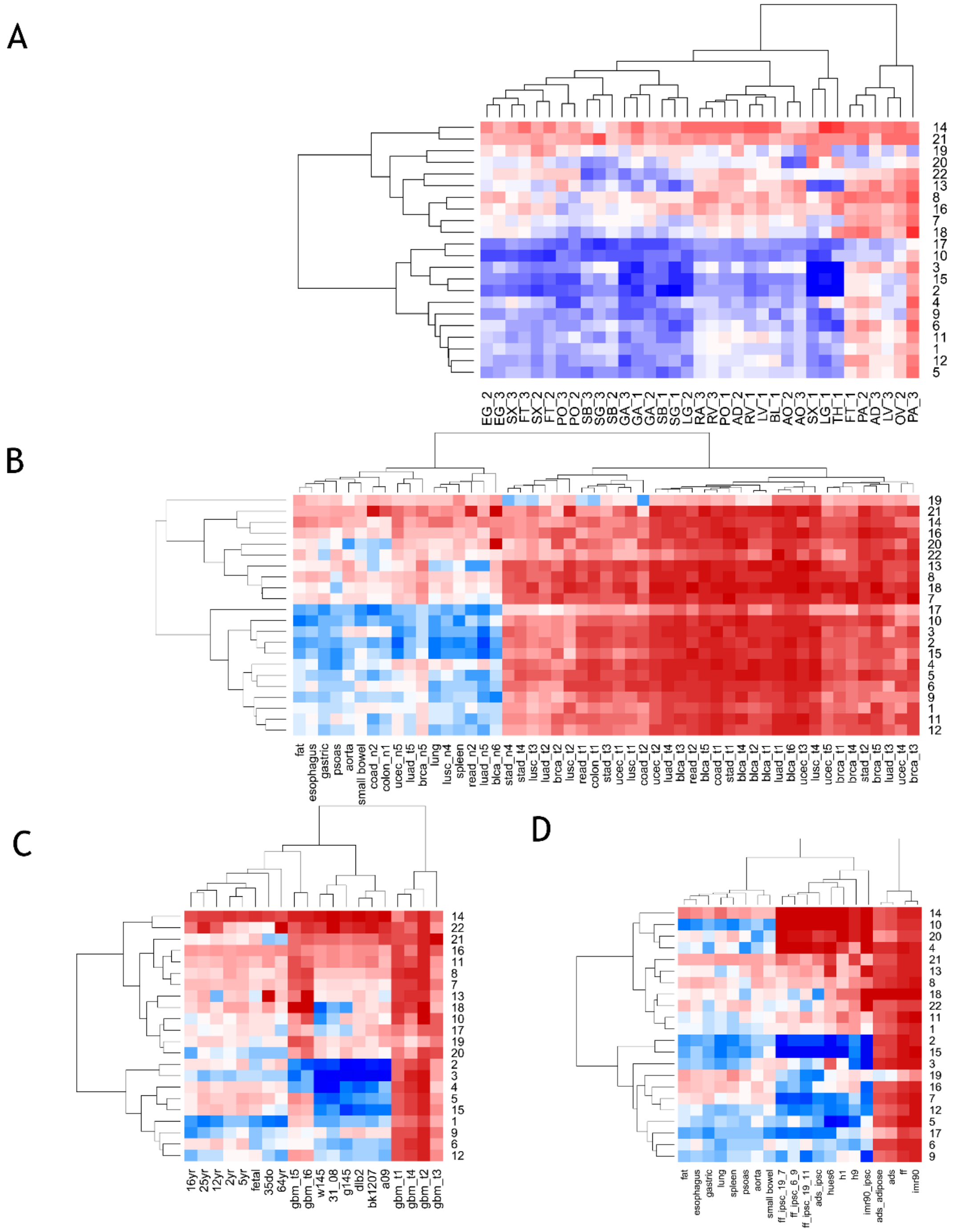

Figure S3. Heatmap clustering of scaling exponents in different chromosomes. The sample labels are the same as Fig. 2 of the main text. In this figure, we also clustered the scaling exponents on the samples. (A) The scaling exponents of normal somatic cells. (B) Normal somatic cells segregate from cancer cells. (C) Normal brain cells segregate from glioblastoma or neurodegenerative diseases. (D) ESCs and iPSCs segregate from cell lines including adult stem cell line and somatic cell lines.

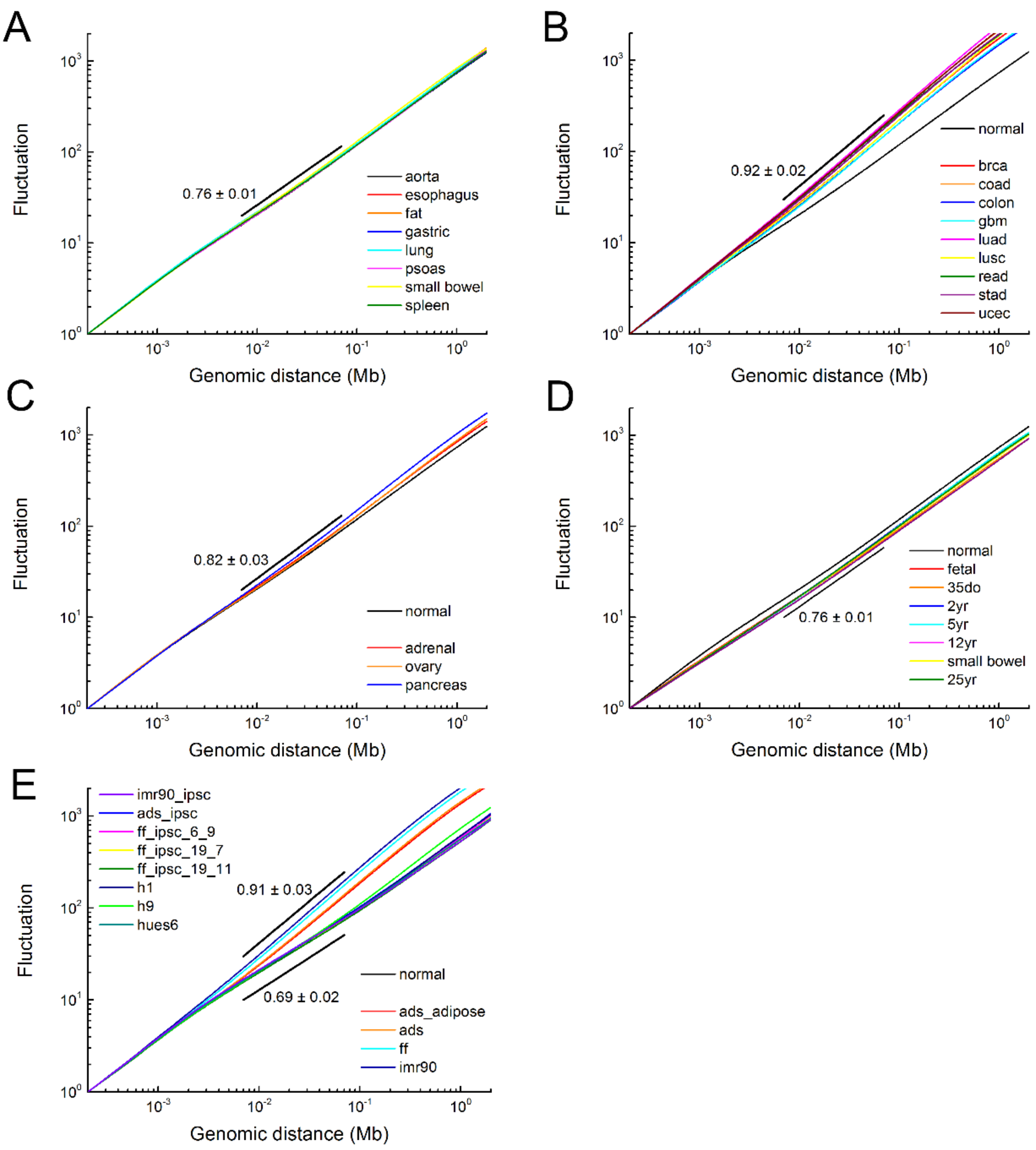

Figure S4. Detrended fluctuation analysis for chromosome 1 in different cell classes.

(A) Normal somatic cells show coherent power law scaling relationship in methylation; (B) Cancer cells. (C) Gland cells. (D) Normal brain cells. (E) Human stem cells and related cells. All these stem cells and related cells are divided into two groups, one with lower-than-normal scaling exponents including all iPSCs and hESCs, the other with cancer-like high scaling exponents including primary somatic cell lines and adult stem cell line. For all cell classes, the average scaling exponent is annotated in the figure and fluctuation for aorta is plotted as normal for comparison.

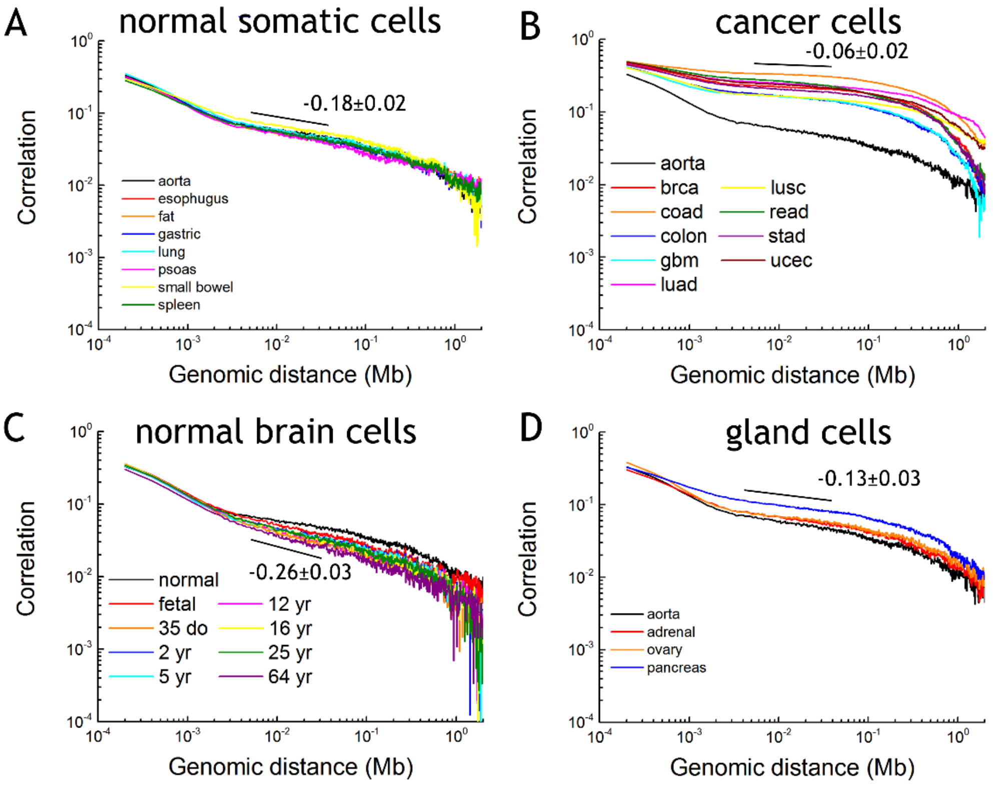

Figure S5. Long-range correlations of DNA methylation using discrete model series. The sample labels are the same as Fig. 1 of the main text. The average scaling exponents are annotated in the figure. Correlation for normal aorta cells (normal) is also plotted for comparison in (B), (C) and (D). We discretized the DNA methylation level of each sample into 1 and 0 with the methylation average as reference value. Specifically, for chromosome 1 of each cell type, we assign a value of 1 to every 200-bp unit with methylation level greater than chromosome average, and 0 to that with methylation level smaller than average.

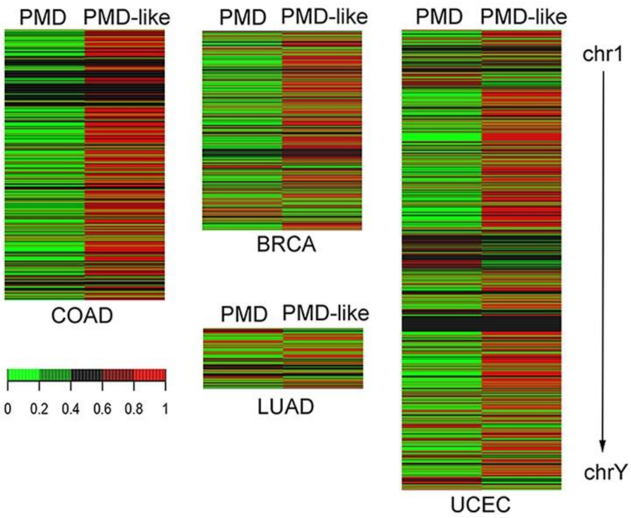

Figure S6. Heatmap of gene expression difference (TPM) in PMD and PMD-like regions in coad_t2-coad_n2, brca_t5-brca_n5, luad_t5-luad_n5 and ucec_t5-ucec_n5 sample pairs. The difference of gene expression in tumor PMDs and normal PMD-like regions for each gene ($\mathrm{d} = TPM_{PMD} - TPM_{PMD-like}$). If the difference is greater than 0 we denoted the PMD expression of this gene as 1 and the PMD-like expression as 0. If the difference is smaller than 0 we denoted the PMD expression of this gene as 0 and the PMD-like expression as 1 and if the difference equals 0, both PMD and PMD-like gene expression are denoted as 0.5. The gene expression in PMDs is lower than that in PMD-like regions.

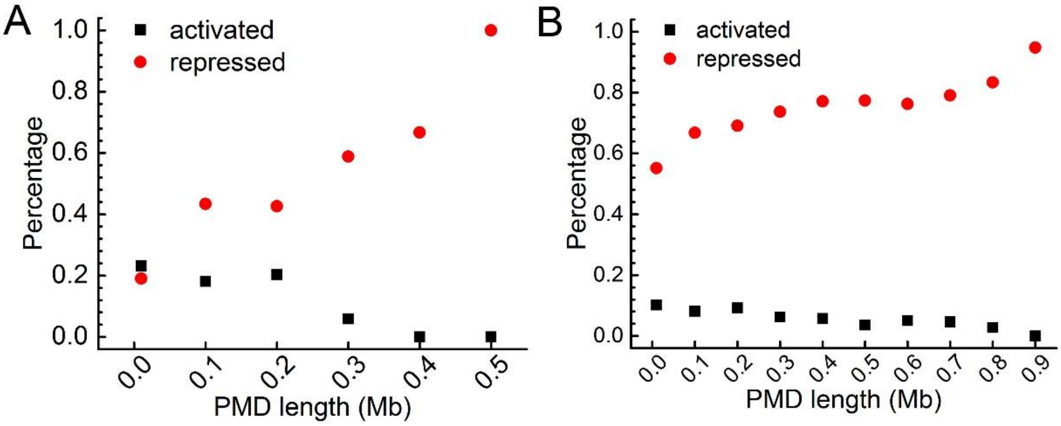

Figure S7. Percentage of genes that are transcriptional activated or repressed in oncogenesis as a function of PMD length.

(A) coad_t2-coad_n2 sample pair. (B) ucec_t5-ucec_n5 sample pair. As the most of PMDs in the luad_t5 sample is short, the luad_t5-luad_n5 sample pair is not shown.

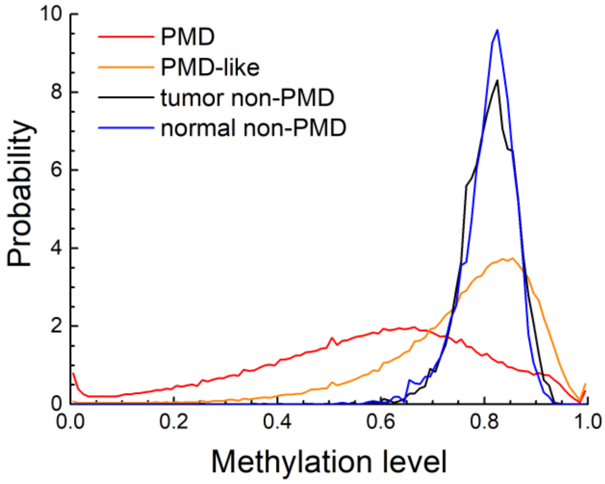

Figure S8. Distribution of methylation level of PMD, non-PMD and PMD-like genomic regions in breast cells. After oncogenesis, the methylation of PMD-like genomic region decreases and turns into the hypomethylated PMD. The methylation of non-PMD doesn’t change before and after oncogenesis.

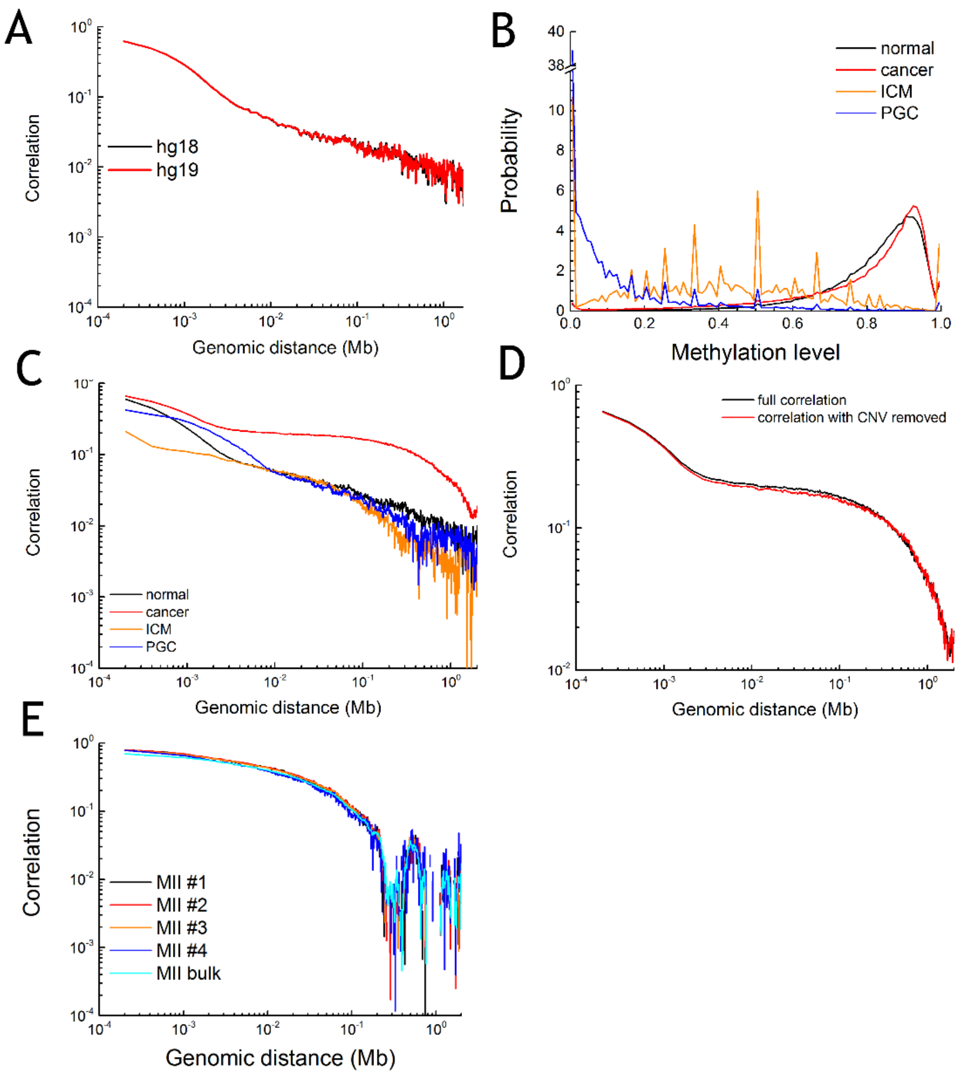

Figure S9. Influence of DNA methylation level, copy number variation and reference genome.

(A) The reference genome has no obvious effect on the methylation correlation. The DNA methylation correlation of chr1 in human brain sample (12yr) using hg18 and hg19 reference genomes respectively. The methylation data under hg19 reference genome was obtained by transforming the original hg18 data using liftover tools. (B) Methylation level of chromosome 1 in four different samples (normal, cancer, PGC and ICM). AO_2 sample is plotted as an example for normal somatic cell and brca_t5 sample for cancer cell. The 10-week female PGC sample and the ICM sequenced by WGBS were used (reference genome: hg19). The average methylation levels for normal, cancer, ICM and PGC are 0.72, 0.74, 0.38, and 0.08, respectively. (C) Methylation correlations for chromosome 1 of the 4 samples in Figure S9B. (D) Long-range DNA methylation correlations are not affected by CNVs in brca_t5 tumor sample. (E) Long-range correlation of DNA methylation is conserved among different single cells. (Data from (8))

Supporting Reference

1. Schultz, M. D., Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, S. Lin, Y. Lin, I. Jung, A. D. Schmitt, S. Selvaraj, B. Ren, T. J. Sejnowski, W. Wang, and J. R. Ecker. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 523:212-216.
2. Berman, B. P., D. J. Weisenberger, J. F. Aman, T. Hinoue, Z. Ramjan, Y. Liu, H. Noushmehr, C. P. Lange, C. M. van Dijk, R. A. Tollenaar, D. Van Den Berg, and P. W. Laird. 2012. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat. Genet. 44:40-46.
3. Lister, R., E. A. Mukamel, J. R. Nery, M. Urich, C. A. Puddifoot, N. D. Johnson, J. Lucero, Y. Huang, A. J. Dwork, M. D. Schultz, M. Yu, J. Tonti-Filippini, H. Heyn, S. Hu, J. C. Wu, A. Rao, M. Esteller, C. He, F. G. Haghighi, T. J. Sejnowski, M. M. Behrens, and J. R. Ecker. 2013. Global epigenomic reconfiguration during mammalian brain development. Science 341:1237905.
4. Lister, R., M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. A. Thomson, R. M. Evans, and J. R. Ecker. 2011. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471:68-73.
5. Sanchez-Mut, J. V., H. Heyn, E. Vidal, S. Moran, S. Sayols, R. Delgado-Morales, M. D. Schultz, B. Ansoleaga, P. Garcia-Esparcia, M. Pons-Espinal, M. M. de Lagran, J. Dopazo, A. Rabano, J. Avila, M. Dierssen, I. Lott, I. Ferrer, J. R. Ecker, and M. Esteller. 2016. Human DNA methylomes of neurodegenerative diseases show common epigenomic patterns. Transl. Psychiatry 6:e718.
6. Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.
7. Peng, C. K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. 1992. Long-range correlations in nucleotide sequences. Nature 356:168-170.
8. Smallwood, S. A., H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods 11:817-820.