Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Nucleotide distance influences co-methylation between nearby CpG sites

Ornella Affinito[a,b,*], Domenico Palumbo[b], Annalisa Fierro[c], Mariella Cuomo[a,b], Giulia De Riso[a,b], Antonella Monticelli[a], Gennaro Miele[d], Lorenzo Chiariotti[a,b,e], Sergio Cocozza[b]

[a] Istituto di Endocrinologia ed Oncologia Sperimentale (IEOS) "Gaetano Salvatore", Consiglio Nazionale delle Ricerche (CNR), Naples, Italy
[b] Dipartimento di Medicina Molecolare e Biotecnologie Mediche, Università degli Studi di Napoli "Federico II", Via S. Pansini 5, 80131 Naples, Italy
[c] CNR-SPIN, c/o Complesso di Monte S. Angelo, via Cinthia, 80126 Napoli, Italy
[d] Dipartimento di Fisica "E. Pancini", Università degli Studi di Napoli "Federico II", Naples, Italy
[e] Dipartimento di Farmacia, Università degli Studi di Napoli "Federico II", Naples, Italy

ARTICLE INFO

Keywords:
DNA methylation
Nearby CpG sites
Co-methylation
Intrinsic methylation susceptibility
Nucleotide distance effect

ABSTRACT

The tendency of individual CpG sites to be methylated is distinctive, non-random and well-regulated throughout the genome. We investigated the structural and spatial factors influencing CpGs methylation by performing an ultra-deep targeted methylation analysis on human, mouse and zebrafish genes. We found that methylation is not a random process and that closer neighboring CpG sites are more likely to share the same methylation status. Moreover, if the distance between CpGs increases, the degree of co-methylation decreases. We set up a simulation model to analyze the contribution of both the intrinsic susceptibility and the distance effect on the probability of a CpG to be methylated. Our finding suggests that the establishment of a specific methylation pattern follows a universal rule that must take into account of the synergistic and dynamic interplay of these two main factors: the intrinsic methylation susceptibility of specific CpG and the nucleotide distance between two CpG sites.

1. Introduction

In vertebrates, the methylation of cytosine is the most frequent epigenetic modification of DNA, consisting of the addition of a methyl group to carbon-5 of cytosine. It is mediated by specific DNA methyltransferases (DNMTs) that are responsible for de novo methylation (i.e., DNMT3a, DNMT3b) and maintenance (i.e., DNMT1) of methylation patterns during replication [1]. DNA methylation has an important role in multiple biological processes: development, stem cell differentiation [2,3], aging [4–6], regulation of gene transcription [7], genomic imprinting [8] and diseases pathogenesis [9–13].

The tendency of individual CpG sites to be methylated is distinctive [14], non-random and well-regulated throughout the genome [15]. Several factors may affect the CpGs methylation susceptibility: sequence context [14], local chromatin configuration [16] and active demethylation [15]. It has also been suggested that the methylation status of a CpG site can be affected by the methylation of neighboring CpGs. Several studies investigated the co-occurrence of methylation between neighboring and distant CpG sites both at genome-wide [17–22] and at locus-specific level [23–28]. Most of these studies support the general notion that nearby CpG sites are commonly methylated together. It has been suggested that this phenomenon could be due to the preference of the DNMTs to methylate CpG pairs at particular distance range [26,29] or to the influence of the nearby CpG site in the recruitment of DNA methylase and/or demethylase enzymes [30,31]. However, most of these studies generally suffer from a statistical limitation due to the low depth of the sequencing. The ultra-deep bisulfite amplicon sequencing overcomes these problems allowing one to obtain a very high coverage of bisulfite sequences from selected loci. By this way it is possible to investigate DNA methylation at single molecule level and at single-base resolution with sufficient statistical power. The term "epialleles" is here used to indicate different combinations of methylated CpGs in single molecules. The population of epialleles can be then computationally treated as a population of haploid organisms, allowing the use of techniques derived from population genetics and ecology [32–34].

In this study we performed an ultra-deep methylation analysis of seven different loci (DDOH, p57, SCRN1, DDO_R7, DLX6, H19, TPH1a) from three different species (human, mouse and zebrafish). The populations of the epialleles were studied by two approaches widely used in population genetics and ecology: co-occurrence analysis and Mantel test. In ecology, the co-occurrence analysis is used to evaluate the co-

Abbreviations: DNMT, DNA methyltransferase; SES, standardized effect size; CU, checkboard unit

* Corresponding author at: Via S. Pansini 5, Naples 80131, Italy.
E-mail address: ornella.affinito@gmail.com (O. Affinito).

https://doi.org/10.1016/j.ygeno.2019.05.007
Received 6 February 2019; Received in revised form 18 April 2019; Accepted 8 May 2019

existence among species inside a community [35]. In the methylation context, CpGs may represent the species, while the epialleles may represent the different sampling sites. Here we adopted the co-occurrence analysis to statistically test the randomness of the methylation process. In population genetics, the Mantel test is often used to evaluate the relationship between geographic distance and genetic divergence [36]. Here, we used the Mantel test to evaluate the relationship between CpGs nucleotide distance and the occurrence of a possible co-methylation between neighboring CpG sites. We obtained sound statistical evidences that the co-occurrence of methylation between CpGs is non-random and that the methylation state of a CpG is influenced by the status of the nearby CpGs with a strong distance effect. In order to analyze the contribution of both the intrinsic susceptibility and the distance effect on the probability of a CpG to be methylated, we set up a simulation model. We found that this probability is dependent on the synergy of both phenomena.

2. Material and methods

2.1. Ethics approval and consent to participate

C57BL/6 J mice were purchased from “The Jackson Laboratory”. All research involving animals was performed in accordance with the European directive 86/609/EEC governing animal welfare and protection, which is acknowledged by the Italian Legislative Decree no. 116 (January 27, 1992). Animal research protocols were also reviewed and consented by a local animal care committee.

All fish were treated in accordance with the Directive of the European Parliament and of the Council on the Protection of Animals Used for Scientific Purposes (Directive 2010/63/EU) and in agreement with the Bioethical Committee of University of Napoli Federico II. All experiments involving fish were approved by the Bioethical Committee of the University of Naples Federico II (authorization protocol number 47339–2013). Human tissue samples were obtained from the MRC London Neurodegenerative Disease Brain Bank of the Institute of Psychiatry, King's College London, UK. All tissues were carried out under the regulations and licenses of the Human Tissue Authority and in accordance with the Human Tissue Act of 2004.

2.2. DNA extraction and sequencing

Three samples for each human, mouse and zebrafish tissue were analyzed. Genomic DNA was extracted from each tissue using Dneasy Blood & Tissue Kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. DNA was quality checked using NanoDrop 2000 (Thermo Scientific) and quantified using Qubit 2.0 Fluorometer (Invitrogen). DNA was converted by sodium bisulfite by EZ DNA Methylation Kit (Zymo Research) according to the manufacturer's instruction. A first PCR step was performed using bisulfite-specific primers described in Supplementary Table S1. Reactions were performed as described [32]. Second PCR step was performed using Nextera XT primers (Illumina, San Diego, CA), in conditions described in [32]. All amplicons were quantified using Qubit® 2.0 Fluorometer. Paired-end sequencing was performed in 281 cycles per read (281 × 2) using Illumina MiSeq.

2.3. Data processing

The pair-end reads were merged together using the PEAR tool with a minimum of 40 overlapping residues and finally they were converted to FASTA format using PRINSEQ. An average of about 56,000 reads/sample were obtained. FASTA converted reads where analyzed using the AmpliMethProfiler pipeline [34] in order to obtain 1 or 0 values for each methylated or unmethylated CpG for each sample. We retained only those reads which satisfied the following parameters: i) length ± 50% compared with the reference length; ii) at least 80% sequence similarity of the primer with the corresponding gene; iii) at least 98% bisulfite efficiency; iv) alignment for at least 60% of their bases with the reference sequence; and v) maximum percentage of ambiguously aligned CpG sites equal to 0.

2.4. Incidence matrices

We analyzed 12 datasets from 4 genes. Each dataset was organized in an incidence matrix (i.e., a presence-absence matrix) in which each row represents a CpG site and each column represents a read. The presence of a methylation at a specific CpG site was denoted by a 1, and its absence was denoted by a 0 [37].

2.5. Co-occurrence analysis (C-score)

To assess the CpGs co-methylation for each dataset, we calculated the C-score index [38] using the sim9 co-occurrence randomization algorithm implemented in EcoSimR v 0.1.0 R package.

The C-score measures the degree of average CpGs pairwise co-methylation. This index is based on the average number of checkboard unit (CU) between all possible CpGs pairs in a matrix. The number of checkboard units (CU_{ij}) for each pair of CpGs *i* and *j* is:

$$CU_{ij} = (r_i - S)(r_j - S)$$

where r_i and r_j are the matrix row totals for CpGs *i* and *j* (the number of reads where each CpG of the pair is methylated) and S is the number of reads where both CpGs are co-methylated. The C-score is then calculated as the average of CUs per CpGs pair, for all CpGs pairs in a dataset. If the resulting C-score is significantly larger than the C-score produced by the null distribution, then at least some pairs of methylated CpGs co-occur less often than expected by chance. On the contrary, if the C-score is significantly less than the C-score for the null distribution then more methylated CpGs co-occur than expected by chance.

2.6. Randomization algorithm (null model algorithm) and standardized effect size (SES)

The significance of the observed value (C-score) was tested using the fixed-fixed null model (sim9 algorithm) [37], based on a Monte Carlo null model simulation to randomize each matrix in the dataset. For the null model, random matrices were produced by shuffling the original matrix through repeated swapping of random submatrices [39]. Following the most conservative option for null model comparisons, the random matrices retained the row and column totals as the real original matrix, thus conserving the number of species per site and the number of sites per species [37]. As the default, using this algorithm, we generated 1000 random matrices for each original dataset. The C-score was calculated for each null matrix (simulated matrix; I_{exp}) and the mean and the standard deviation for the index values thus obtained were calculated. Using the observed (I_{obs}) and simulated C-score index, we calculated the standardized effect size (SES) for each matrix [40], according to the following formula [41]:

$$SES = \frac{I_{obs} - I_{exp}}{SD_{exp}}$$

where I_{obs} is the observed C-score value, I_{exp} is the mean of the 1000 simulated C-score values calculated from the random matrices and SD_{exp} is the standard deviation of the 1000 simulated C-score values calculated from the random matrices.

SES measures the number of standard deviations that the observed C-score is above or below the mean C-score from simulated matrices. In other words, it measures the statistical amount of deviation from random co-occurrence. High SES values indicate greater C-score than expected from the observed number of species and low SES values indicate lower C-score than expected. As a consequence, high SES of the

C-score means less co-occurrence than low SES values. In the methylation context, a low SES score (near to zero) indicate a stochastic methylation and each CpG would have the same probability to be methylated, regardless of the nucleotide distance. By comparing I_{obs} with the distribution of simulated values, it is possible to assess the probability that I_{obs} does not differ from the value expected by chance. For our purposes, we consider as significant a $|SES| > 1.96$, which corresponds to the 95% confidence interval of the two-tailed distribution. The null hypothesis is that the average effect size is zero and that 95% of the observations will lie between -2 and $+2$. In our study, using the sim9 algorithm, we estimated CpGs co-occurrence with 1000 permutations; this allows us to confirm our analysis with a p-value $<$ .001.

2.7. Mantel test

A Mantel test was used to test the correlation between CpGs nucleotide distance and the occurrence of a possible co-methylation between neighboring CpG sites. The Mantel is given by:

$$Z_m = \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} \times d_{ij}$$

where g_{ij} and d_{ij} are, respectively, the epigenetic (in our case, the pairwise CpGs co- methylation) and nucleotide distances between CpG i and CpG j, considering n CpGs. Because of Z_m is given by the sum of products of distances, its value depends on the number of CpGs sites under investigation, as well as on their distances. The linear dependence between the two matrices was calculated with the Pearson correlation coefficient. The statistical significance was assessed with 1000 permutations, followed by a Bonferroni correction and using a significance level (alpha) of 0.05. “vegan” R- package (version 2.4-1) was used to perform the tests and draw correlograms.

3. Results

3.1. Co-occurrence analysis

We performed an in-depth methylation analysis using ultra-deep bisulphite amplicon sequencing of seven different loci (DDOH, p57, SCRN1, DDO_R7, DLX6, H19, TPH1a) from three different species (human, mouse and zebrafish) and different tissues (brain, cerebellum, thyroid, gut and lymphocytes). Detailed information on samples is reported in Supplementary Table S1.

As first step we estimated, for each gene analyzed, the mean methylation values of each CpG. As expected, the different CpGs showed different methylation values. These differences were well conserved among different samples of the same gene (Supplementary Fig. S1) suggesting that this phenomenon could be partially due to a sort of intrinsic proneness to be methylated.

To assess the CpGs co-methylation, we adopted a statistical approach widely used in the ecology field, named the co-occurrence analysis. In ecology, this analysis establishes if species are randomly distributed or according to some rules in a certain environment. The degree of average species pairwise co-occurrence can be evaluated by the C-score measure. The significance of observed C-score value was then evaluated by SES measure (see Material and Methods). SES score represents how much far the observed phenomenon (co-methylation) is from the randomness. Usually in ecology, a $|SES| > 1.96$ was considered as statistically significant, corresponding to the 95% confidence interval of the two-tailed distribution.

Table 1 shows C-score and SES score values obtained for each sample analyzed. SES score was highly significant for each sample analyzed. This result provides sound statistical evidence that the methylation status is not randomly distributed among CpGs belonging to the same molecule, but that a sort of a “rule” seems to exist.

Table 1
Results of the co-occurrence analysis. The analysis was based on the C-score.

Sample	Gene	Tissue	Obs C-score	Mean Sim C-score	SES
H1	DDOH	Cerebellum	453,924	374,814	51.879
H2	DDOH	Cerebellum	624,716	589,889	43.277
H3	DDOH	Cerebellum	224,274	202,449	38.335
H1	p57	Thyroid	1,515,898	1,462,635	23.342
H2	p57	Thyroid	3,370,075	3,311,756	18.022
H3	p57	Thyroid	4,736,823	4,533,619	54.136
H1	SCRN1	Lymphocytes	18,775,895	17,140,321	112.16
H2	SCRN1	Lymphocytes	27,500,479	25,114,478	128.12
H3	SCRN1	Lymphocytes	54,456,835	50,239,266	143.14
M1	DDO_R7	Brain	206,718,758	199,931,706	169.62
M2	DDO_R7	Brain	2,508,116,080	2,473,615,574	179.87
M3	DDO_R7	Brain	528,810,035	519,745,258	125.88
M1	DLX6	Brain	13,316,719	13,080,118	23.756
M2	DLX6	Brain	6,588,269	6,562,251	34.221
M3	DLX6	Brain	7,256,850	7,111,403	22.087
M1	H19	Brain	1,242,362,655	987,023,122	2141
M2	H19	Brain	14,271,939	11,632,589	532.27
M3	H19	Brain	9,675,527	8,337,075	271.99
ZF1	TPH1a	Brain	7,492,867	6,983,630	178.46
ZF2	TPH1a	Brain	6,612,375	6,255,254	148.28
ZF3	TPH1a	Brain	3,854,568	3,662,860	108.4
ZF1	TPH1a	Gut	77,971,777	74,951,078	275.16
ZF2	TPH1a	Gut	25,181,406	24,348,413	159.69
ZF3	TPH1a	Gut	57,673,366	54,223,635	342.76

Abbreviations: Obs C-score = Observed C-score; Mean sim C-score = Mean simulated C-score from 1000 random runs; SES = Standardized effect size.

3.2. Correlation between nucleotide distance and CpG co-methylation

To find the possible elements of this “rule”, we explored the role of the neighboring CpGs in influencing the methylation status of other ones. In particular, we determined if and how the nucleotide distance between two CpGs influences their co-methylation. To reach this aim, we used the Mantel test. This test divides CpGs into distance groups and checks the amount of co-methylation in the different groups. Basically, Mantel test compare two distance matrices: the CpGs nucleotide distance on one hand and the CpG co-methylation on the other. Fig. 1 shows the Mantel correlogram for each gene under investigation. For each condition tested, we found a significant ($P < 0.05$, Bonferroni-corrected) positive correlation in the first distance class. This finding suggests that the co-methylation is statistically more frequent for CpGs located under 50 base pairs of distance. The Mantel correlation was positive for all genes and all species under investigation. This result suggests that the methylation status of a CpG influences the methylation status of the closest CpGs.

3.3. Simulation model

We used a simulation model to explore the interplay between two possible factors influencing the CpG susceptibility to be methylated: the intrinsic one (without influences of the neighboring CpGs) and the other one depending on the methylation status of neighboring CpGs. As a first step, we developed an algorithm for the generation of artificial epiallele populations starting from real data. The starting epiallele population (based on real data) was then artificially methylated in silico, according to the methylation rules that we plan to check. The final step was to compare the final artificial epiallele population obtained using the tested rules with the real, experimental ones. To perform these simulations we used an experimental dataset described by us in a previous paper [33]. These data derive from the in-depth sequencing of a 398 bp region at the 5′ end of the DDO_R4 gene from three mice (lung and gut) at different three times during their development (denoted for simplicity $T_{in} < T_{inter} < T_{fin}$). During this time, the experimental data showed an increase in the total methylation of the region. The six CpGs present in the region were methylated in different combinations in each

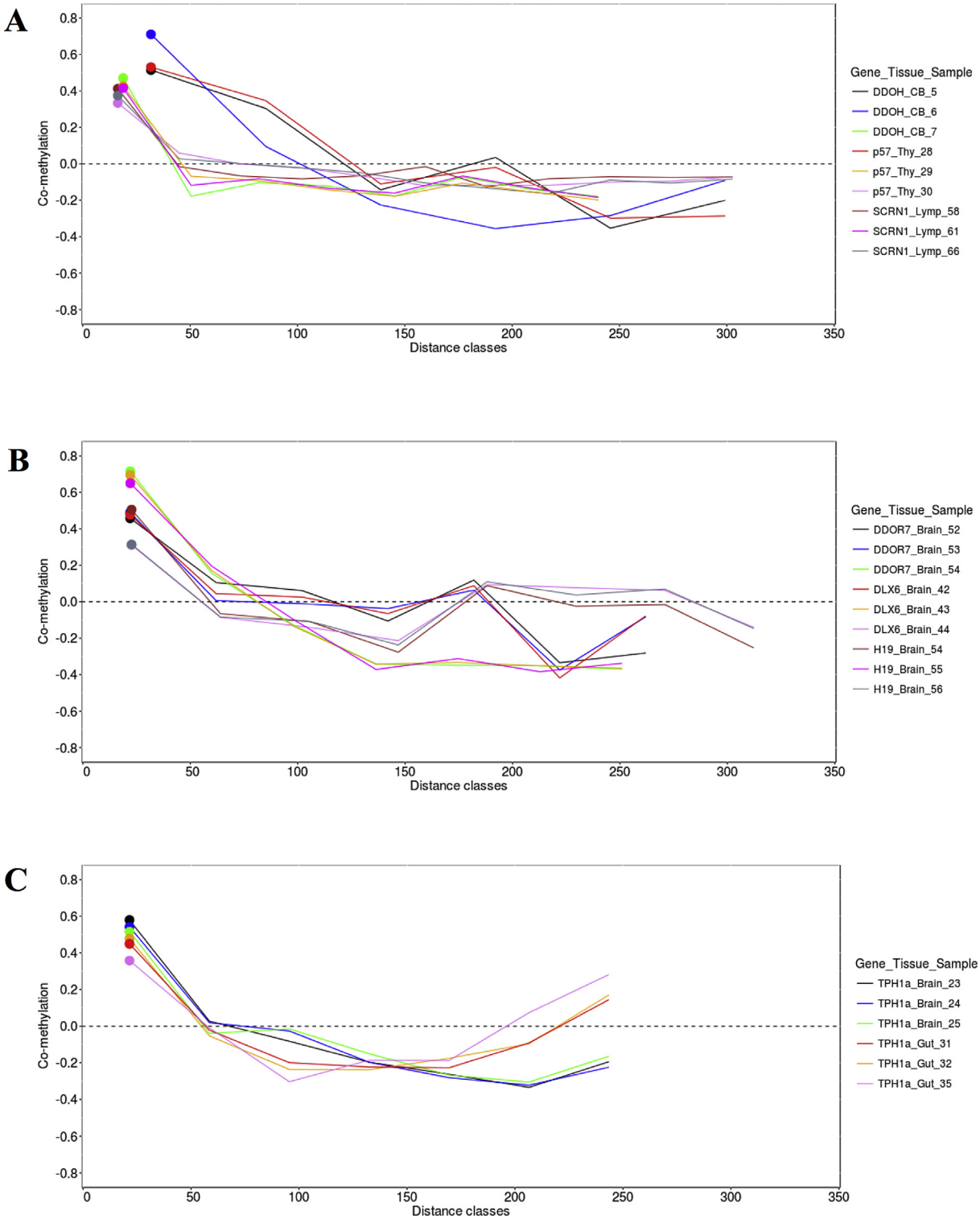

Fig. 1. Correlogram plots. The correlogram plots show the relationship between the pairwise CpGs distances classes and the co-methylation for each sample for A) human genes, B) mouse genes, and C) zebrafish genes. On the x-axis, the nucleotide distance classes are reported in bins of 50 nucleotides. On the y-axis, the co-methylation between pairwise CpGs is reported. Colored circles represent statistically significant co-methylation after Bonferroni correction (p-value < .05). Abbreviations: CB = Cerebellum; Thy = Thyroid; Lymp = Lymphocytes.

molecule, with the formation of 64 different epialleles, whose frequencies varied with time.

The simulation algorithm started from real initial conditions, namely a population of N_e = 75,648 epialleles (individuals) distributed among the 64 possible ones according to the observed frequencies measured at the starting time T_{in}. Such population was then evolved for an arbitrary number of generations N_g = 30, and the final distribution of epiallele frequencies was then compared with the real ones corresponding to T_{fin}. The changes in methylation status of a given individual CpG when passing to the next generation were driven by its methylation probability. The simulation approach allowed us to test three different biological models of methylation assuming different

criteria to define the methylation probability for each CpG.

Let us denote with $s_i(t_n) = 0,1$ the methylation status of the i-th CpG (with $i = 1,..,6$) at a certain generation time t_n, with $s_i(t_n) = 1$ for methylated status and 0 otherwise. Furthermore, we denote with P ($s_i(t_n) = 0$; $s_i(t_{n+1}) = 1$) the methylation probability for the i-th CpG once passing from generation t_n to t_{n+1}.

Model A – According to this approach we made the assumption that the $P(s_i(t_n) = 0; s_i(t_{n+1}) = 1) = p_i$, namely methylation probability is only determined by individual susceptibility of the i-th CpG, here denoted by p_i.

Model B – In this case the probability to be methylated for the i-th CpG during the simulation was only depending on the status of methylation of the surrounding CpGs, according to the following formula

$$P(s_i(t_n) = 0; s_i(t_{n+1}) = 1) = p \exp\left[-\frac{d_i^{\min}(t_n)}{D}\right]$$

where p denotes the individual susceptibility taken equally for all sites, whereas $d_i^{\min}$ is defined as follows.

$d_i^{\min}(t_n) = \min_{\forall j \neq i : s_j(t_n)=1} [d_{ij}]$ if $\sum_{j=1}^{6} s_j(t_n) \neq 0$.with d_{ij} denoting the physical distance among the i-th and j-th CpG. In case there is no methylated CpG, namely if $\Sigma_{j=1}{}^{6} s_j(t_n) = 0$, we assume $d_i^{\min}(t_n) = 300$ that is a value greater than the maximum possible distance between any couple of CpG considered.

Model C – In this case one merges Model A and B thus obtaining the methylation probability of the i-th CpG in the form

$$P(s_i(t_n) = 0; s_i(t_{n+1}) = 1) = p_i \exp\left[-\frac{d_i^{\min}(t_n)}{D}\right]$$

with the individual susceptibility p_i now depending on the particular site i, and d_i^{min} defined as before.

In all cases the free parameters D, p or p_i have to be fitted by the observations. Given a population at a generation t_n and an arbitrary epiallele contained in it, represented by a six methylation status ($s_1(t_n)$,.., $s_6(t_n)$), we evolved it through generations according to the methylation probability parameterized by the previous models. The comparison of the final distribution of epiallele frequencies so obtained with the observed ones allows determining the best fit values of free parameters (Table 2). The comparison of the epialleles frequencies achieved by the three models with the real data shows that only epiallele frequencies obtained by model C resemble the experimental data. To better evaluate the phenomenon, we calculated the differences between simulated and experimental data for the three models. Fig. 2 shows that model C shows the lowest simulated/experimental delta values, suggesting that both the individual susceptibility and the distance effect are necessary to fit simulated data with experimental ones.

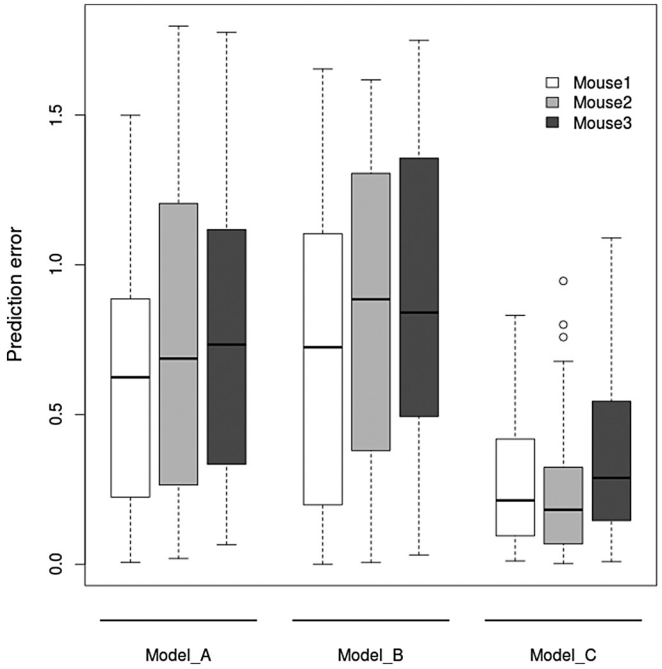

Fig. 2. Simulation models. For each sample, the box plot shows the prediction error of the three simulation models expressed as the difference over semi-sum.

4. Discussion

In this study, we investigated factors influencing CpGs' methylation by performing ultra-deep methylation analysis on seven different loci from three different species (human, mouse and zebrafish).

For any locus tested, we found that, the methylation of CpGs located on the same molecule occurs in a non-stochastic manner. We achieved this result adopting the co-occurrence analysis, a statistical approach that is widely used in ecology but never used before in epigenetics. In the ecological version of this analysis, the null hypothesis is that species are randomly distributed in the different environments. The alternative hypothesis is that some kind of rule exists determining non-random co-occurrences of species in the same environment. The degree of species co-occurrence is evaluated by the C-score measure. In our "epigenetic" version of the test, the null hypothesis is that methylated CpGs are randomly distributed on the molecules. The alternative hypothesis is that rules exist, which may determine the methylation of the CpGs located on the same DNA molecule. For each locus tested, we found very high values of C-scores. This result provides statistically sound evidence that methylated CpGs are distributed on each single molecule in an absolutely non-stochastic manner. Next, we explored the hypothesis that one of the factors influencing the methylation of a CpG can be the pre-existing methylation landscape on the molecule. Indeed, previous studies have suggested that CpGs influence each other with a distance-dependent effect, by means that nearby CpG sites tend to be methylated together [17,22,23,30,31]. Consistently with these observations, DNMT3a structure studies revealed a correlation of methylated CpG sites at distances of 8 to 10 nucleotides, indicating that DNMTs may methylate DNA in a periodic pattern [26,29]. Starting from these observations, we tested, in our experimental system, the hypothesis that the likelihood of a CpG site to be methylated could be influenced by the physical distance of the nearby methylated CpG site. Also in this case we adopted a statistical test borrowed from ecology and population genetics: the Mantel test [36]. In these scientific areas the Mantel test is one of the most popular methods used to evaluate spatial processes involved in determining population genetic structure [36]. Basically,

Table 2
Performances of the simulation models. The performances of three different simulation scenarios are here presented by reporting their reduced-χ^2 evaluated comparing the epiallele frequencies observed at T_{fin} (see *Simulation Model* section for its definition) with the predicted ones, if we start from the observed frequencies at T_{in}. In the Table we also report the range for free parameters involved in each model.

Models	Best fit Parameters	χ^2/dof
Model A	$P_1 = 0.022 \pm 0.005$, $P_2 = 0.030 \pm 0.006$, $P_3 = 0.010 \pm 0.001$, $P_4 = 0.015 \pm 0.002$, $P_5 = 0.012 \pm 0.002$, $P_6 = 0.011 \pm 0.003$	0.017 ± 0.005
Model B	$P = 0.08 \pm 0.02$, $D = 113 \pm 15$	0.013 ± 0.002
Model C	$P_1 = 0.07 \pm 0.02$, $P_2 = 0.11 \pm 0.04$, $P_3 = 0.023 \pm 0.004$, $P_4 = 0.049 \pm 0.011$, $P_5 = 0.039 \pm 0.011$, $P_6 = 0.039 \pm 0.011$, $D = 120 \pm 20$	0.0009 ± 0.0004

Mantel analysis is a statistical test of the correlation between two matrices. In population genetics, the two matrices are, in most cases, the genetic divergences and the geographical distances. In our case, the two matrices were the nucleotide distance and the methylation co-occurrence.

By using Mantel analysis, we found that closer CpG sites are more likely to be simultaneously methylated, and that, if the distance between CpG sites increases, the degree of co-methylation decreases. The co-methylation mainly appears at distances < 50 nt. This distances are in good agreement with those previously obtained for a limited number of loci by different methods [27,28]. It is intriguing to note that we found very similar results for all genes analyzed in the present study, independently of species (human, mouse and zebrafish) and tissue origin, suggesting that the “distance effect” could be a universal rule.

Previous studies have suggested that also other factors can influence likelihood of a CpG site to be methylated. Among them, it seems that the CpGs possess an intrinsic propensity to be methylated [42]. Indeed, it is well established that CpG sites are not evenly methylated, but that some sites are more prone to be methylated than others [30]. Both epigenetic and genetic factors have been invoked to explain this different susceptibility [18]. For example, it has been demonstrated that DNA methyltransferases show different preference for the CpGs' flanking sequences [43]. Furthermore it has been demonstrated that highly methylated CpG sites are more likely flanked by A/T rich sequences while unmethylated ones tend to be flanked by G/C rich sequences [23].

Therefore, in addition to the influence of the methylation status of neighboring CpGs, the intrinsic propensity for methylation of the CpG site must also be taken into account. To study the interplay of these two factors in influencing the methylation of a CpG, we set up a simulation model. The simulation approach allowed us to test three different biological scenarios by simply modifying the criteria that defined, in the simulation algorithm, the likelihood of each CpG to be methylated. In the first model, the likelihood was only dependent on the individual susceptibility of each CpG sites. In the second model, the likelihood of each CpG to be methylated depended only on the status of methylation of the surrounding CpGs. In the third model, the likelihood depended on both factors. The best fit with the experimental data was obtained only in this last case, when both parameters were introduced in the simulation algorithm. This result suggests that the combined action of intrinsic susceptibility and distance effects influence the likelihood of one CpG to be methylated.

This study has some limitations, which have to be pointed out. First, we tested our hypothesis in a limited number of loci. Some previous studies investigated the same topic by an genome-wide approach [17–22]. By this approach, the gain in universalization of the finding is balance by a loss statistical power, because of the low depth of the sequence for each genomic region. We decided to use a “targeted” strategy in our study because the in-depth single molecule DNA methylation analysis guarantees, in our opinion, a greater statistical reliability. The availibility of thousands sequences of the same locus allowed us to use statistical approaches borrowed from ecology and population genetics further increasing the statistical soundness of the results. Another limitation concerns the simulation model. We tested, in our simulations, only two parameters: intrinsic susceptibility and distance effects. It is very likely that many other factors are involved in determining in vivo the methylation of a CpG. Therefore our simulations are not oriented to give an exhaustive description of the biological system, but only to test the interaction between two variables of this system.

5. Conclusions

In this study we showed that methylation of CpGs located on the same DNA fragment occurs non-stochastic. In particular, we found that closer neighboring CpG sites are more likely to share the same methylation status, and that, if the distance between CpG sites increases, the degree of co-methylation decreases. By a simulation approach, we showed that the probability of a CpG site to be methylated is conditioned by a synergistic effect of the methylation landscape of the nearby region and the intrinsic susceptibility of that site to undergo methylation.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2019.05.007.

Declaration of competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Epigenomics Flagship Project SP-6.6 – EPIGEN, C.N.R., Italy.

Acknowledgements

OA analyzed and interpreted data and wrote the manuscript. DP helped in the analyses and in the manuscript writing. GDR contributed in the manuscript writing. AF and GM performed the simulation model. MC has prepared and sequenced samples. LC e AM contributed to the data interpretation and in writing the manuscript. SC coordinated the experimental plan and contributed to the writing of the manuscript. All authors have read and approved the manuscript.

References

[1] Z.D. Smith, A. Meissner, DNA methylation: roles in mammalian development, Nat. Rev. Genet 14 (3) (2013) 204–220, https://doi.org/10.1038/nrg3354 Mar 12.
[2] W. Reik, Epigenetic reprogramming in mammalian development, Science (80-) 293 (5532) (2001) 1089–1093, https://doi.org/10.1126/science.1063443 Aug 10.
[3] E. Li, Chromatin modification and epigenetic reprogramming in mammalian development, Nat. Rev. Genet. 3 (9) (2002) 662–673, https://doi.org/10.1038/nrg887.
[4] P. Hamet, J. Tremblay, Genes of aging, Metabolism 52 (Suppl. 2) (2003) 5–9, https://doi.org/10.1016/S0026-0495(03)00294-4 Oct.
[5] N. Ahuja, J.P. Issa, Aging, methylation and cancer, Histol. Histopathol. 15 (3) (2000) 835–842, https://doi.org/10.14670/HH-15.835.
[6] Z. Zhang, C. Deng, Q. Lu, B. Richardson, Age-dependent DNA methylation changes in the ITGAL (CD11a) promoter, Mech. Ageing Dev. 123 (9) (2002) 1257–1268, https://doi.org/10.1016/S0047-6374(02)00014-3 May.
[7] M.F. Chan, G. Liang, P.A. Jones, Relationship between transcription and DNA methylation, Curr. Top. Microbiol. Immunol. 249 (2000) 75–86.
[8] W. Reik, J. Walter, Imprinting mechanisms in mammals, Curr. Opin. Genet. Dev. 8 (2) (1998) 154–164, https://doi.org/10.1016/S0959-437X(98)80136-6 Apr.
[9] M.I. Scarano, M. Strazzullo, M.R. Matarazzo, M. D'Esposito, DNA methylation 40 years later: its role in human health and disease, J. Cell. Physiol. 204 (1) (2005) 21–35, https://doi.org/10.1002/jcp.20280.
[10] L.K. Jones, V. Saha, Chromatin modification, leukaemia and implications for therapy, Br. J. Haematol. 118 (3) (2002) 714–727, https://doi.org/10.1046/j.1365-2141.2002.03586.x Sep.
[11] P.A. Jones, S.B. Baylin, The fundamental role of epigenetic events in cancer, Nat. Rev. Genet. 3 (6) (2002) 415–428, https://doi.org/10.1038/nrg816.
[12] P.W. Laird, Principles and challenges of genome-wide DNA methylation analysis, Nat Rev Genet 11 (3) (2010) 191–203, https://doi.org/10.1038/nrg2732.
[13] M. Kulis, M. Esteller, DNA methylation and cancer, Advances in Genetics, 2010, pp. 27–56, , https://doi.org/10.1016/B978-0-12-380866-0.60002-2.
[14] S. Kim, M. Li, H. Paik, K. Nephew, H. Shi, R. Kramer, et al., Predicting DNA methylation susceptibility using CpG flanking sequences, Pac. Symp. Biocomput. (2008) 315–326.
[15] Z. Chen, A.D. Riggs, DNA methylation and demethylation in mammals, J. Biol. Chem. 286 (21) (2011) 18347–18353, https://doi.org/10.1074/jbc.R110.205286 May 27.
[16] Van Der Wijst MGP, A.Y. Van Tilburg, M.H.J. Ruiters, M.G. Rots, Experimental mitochondria-targeted DNA methylation identifies GpC methylation, not CpG methylation, as potential regulator of mitochondrial gene expression, Sci. Rep. 7 (1) (2017) 1–15, https://doi.org/10.1038/s41598-017-00263-z.
[17] D. Saito, M. Suyama, Linkage disequilibrium analysis of allelic heterogeneity in DNA methylation, Epigenetics 10 (12) (2015) 1093–1098, https://doi.org/10.1080/15592294.2015.1115176.
[18] J.T. Bell, A.A. Pai, J.K. Pickrell, D.J. Gaffney, R. Pique-Regi, J.F. Degner, et al., DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines, Genome Biol. 12 (1) (2011) R10, https://doi.org/10.1186/gb-

2011-12-1-r10.
[19] Y. Li, J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, et al., The DNA methylome of human peripheral blood mononuclear cells, PLoS Biol. 8 (11) (2010), https://doi.org/10.1371/journal.pbio.1000533.
[20] J.L. Huynh, P. Garg, T.H. Thin, S. Yoo, R. Dutta, B.D. Trapp, et al., Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains, Nat. Neurosci. 17 (1) (2014) 121–130, https://doi.org/10.1038/nn.3588.
[21] T.C. Martin, I. Yet, P.C. Tsai, J.T. Bell, coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns, BMC Bioinformatics 16 (1) (2015) 1–5, https://doi.org/10.1186/s12859-015-0568-2.
[22] F. Eckhardt, J. Lewin, R. Cortese, V.K. Rakyan, J. Attwood, M. Burger, et al., DNA methylation profiling of human chromosomes 6, 20 and 22, Nat. Genet. 38 (12) (2006) 1378–1385, https://doi.org/10.1038/ng1909.
[23] Y. Zhang, C. Rohde, S. Tierling, T.P. Jurkowski, C. Bock, D. Santacruz, et al., DNA methylation analysis of chromosome 21 gene promoters at Single Base pair and single allele resolution, PLoS Genet. 5 (3) (2009) e1000438, , https://doi.org/10.1371/journal.pgen.1000438.
[24] K. Hu, A.H. Ting, J. Li, BSPAT: a fast online tool for DNA methylation co-occurrence pattern analysis based on high-throughput bisulfite sequencing data, BMC Bioinformatics 16 (1) (2015), https://doi.org/10.1186/s12859-015-0649-2.
[25] R. Shoemaker, J. Deng, W. Wang, K. Zhang, Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome, Genome Res. 20 (7) (2010) 883–889, https://doi.org/10.1101/gr.104695.109 Jul 1.
[26] D. Jia, R.Z. Jurkowska, X. Zhang, A. Jeltsch, X. Cheng, Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation, Nature 449 (7159) (2007) 248–251, https://doi.org/10.1038/nature06146.
[27] L. Cao-Lei, S. Suwansirikul, P. Jutavijittum, S.B. Mériaux, J.D. Turner, C.P. Muller, Glucocorticoid receptor gene expression and promoter CpG modifications throughout the human brain, J. Psychiatr. Res. 47 (11) (2013) 1597–1607, https://doi.org/10.1016/j.jpsychires.2013.07.022 Nov.
[28] S.R. Witzmann, J.D. Turner, S.B. Mériaux, O.C. Meijer, C.P. Muller, Epigenetic regulation of the glucocorticoid receptor promoter 1(7) in adult rats, Epigenetics 7 (11) (2012) 1290–1301, https://doi.org/10.4161/epi.22363 Nov.
[29] R.Z. Jurkowska, N. Anspach, C. Urbanke, D. Jia, R. Reinhardt, W. Nellen, et al., Formation of nucleoprotein filaments by mammalian DNA methyltransferase Dnmt3a in complex with regulator Dnmt3L, Nucleic Acids Res. 36 (21) (2008) 6656–6663, https://doi.org/10.1093/nar/gkn747 Dec.
[30] C. Lövkvist, I.B. Dodd, K. Sneppen, J.O. Haerter, DNA methylation in human epigenomes depends on local topology of CpG sites, Nucleic Acids Res. 44 (11) (2016) 5123–5132, https://doi.org/10.1093/nar/gkw124.
[31] J.O. Haerter, C. Lövkvist, I.B. Dodd, K. Sneppen, Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states, Nucleic Acids Res. 42 (4) (2014) 2235–2244, https://doi.org/10.1093/nar/gkt1235.
[32] E. Florio, S. Keller, L. Coretti, O. Affinito, G. Scala, F. Errico, et al., Tracking the evolution of epialleles during neural differentiation and brain development: D-aspartate oxidase as a model gene, Epigenetics 12 (1) (2017) 41–54, https://doi.org/10.1080/15592294.2016.1260211.
[33] O. Affinito, G. Scala, D. Palumbo, E. Florio, A. Monticelli, G. Miele, et al., Modeling DNA methylation by analyzing the individual configurations of single molecules, Epigenetics 11 (12) (2016) 881–888, https://doi.org/10.1080/15592294.2016.1246108 Dec.
[34] G. Scala, O. Affinito, D. Palumbo, E. Florio, A. Monticelli, G. Miele, et al., ampliMethProfiler: a pipeline for the analysis of CpG methylation profiles of targeted deep bisulfite sequenced amplicons, BMC Bioinformatics 17 (1) (2016) 484, https://doi.org/10.1186/s12859-016-1380-3 Nov 25.
[35] I. Šímová, C. Violle, N.J.B. Kraft, D. Storch, J.-C. Svenning, B. Boyle, et al., Shifts in trait means and variances in north American tree assemblages: species richness patterns are loosely related to the functional space, Ecography (Cop) 38 (7) (2015) 649–658, https://doi.org/10.1111/ecog.00867 Jul.
[36] J.A.F. Diniz-Filho, T.N. Soares, J.S. Lima, R. Dobrovolski, V.L. Landeiro, C. Telles MP de, et al., Mantel test in population genetics, Genet. Mol. Biol. 36 (4) (2013) 475–485, https://doi.org/10.1590/S1415-47572013000400002.
[37] E.F. Connor, D. Simberloff, The assembly of species communities: chance or competition? Ecology 60 (6) (1979) 1132, https://doi.org/10.2307/1936961 Dec.
[38] L. Stone, A. Roberts, The checkerboard score and species distributions, Oecologia 85 (1) (1990) 74–79, https://doi.org/10.1007/BF00317345 Nov.
[39] B.F.J. Manly, A note on the analysis of species co-occurrences, Ecology 76 (4) (1995) 1109–1115, https://doi.org/10.2307/1940919 Jun.
[40] J. Gurevitch, L.L. Morrow, A. Wallace, J.S. Walsh, A meta-analysis of competition in field experiments, Am. Nat. 140 (4) (1992) 539–572, https://doi.org/10.1086/285428 Oct.
[41] N.J. Gotelli, D.J. McCabe, Species co-occurrence: a meta-analysis of J. M. Diamond's assembly rules model, Ecology 83 (8) (2002) 2091, https://doi.org/10.2307/3072040 Aug.
[42] R. Jaenisch, A. Bird, Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals, Nat. Genet. 33 (3S) (2003) 245–254, https://doi.org/10.1038/ng1089.
[43] V. Handa, A. Jeltsch, Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome, J. Mol. Biol. 348 (5) (2005) 1103–1112, https://doi.org/10.1016/j.jmb.2005.02.044 May 20.