Published online 28 January 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 10* ***4099–4108***
doi:10.1093/nar/gkr017

Genome-wide quantitative assessment of variation in DNA methylation patterns

Hehuang Xie[1,2,*], Min Wang[1,2], Alexandre de Andrade[1], Maria de F. Bonaldo[1,2], Vasil Galat[3], Kelly Arndt[1], Veena Rajaram[1,4], Stewart Goldman[1,5], Tadanori Tomita[1,6] and Marcelo B. Soares[1,2]

[1]Falk Brain Tumor Center, Cancer Biology and Epigenomics Program, [2]Department of Pediatrics, [3]Stem Cell Facility, Children's Memorial Research Center, [4]Department of Pathology, Division of Anatomic Pathology, [5]Division of Hematology/Oncology/Transplantation and [6]Department of Neurosurgery, Division of Pediatric Neurosurgery, Feinberg School of Medicine, Northwestern University, Chicago IL 60614-3394, USA

Received October 25, 2010; Revised December 22, 2010; Accepted January 5, 2011

ABSTRACT

Genomic DNA methylation contributes substantively to transcriptional regulations that underlie mammalian development and cellular differentiation. Much effort has been made to decipher the molecular mechanisms governing the establishment and maintenance of DNA methylation patterns. However, little is known about genome-wide variation of DNA methylation patterns. In this study, we introduced the concept of methylation entropy, a measure of the randomness of DNA methylation patterns in a cell population, and exploited it to assess the variability in DNA methylation patterns of Alu repeats and promoters. A few interesting observations were made: (i) within a cell population, methylation entropy varies among genomic loci; (ii) among cell populations, the methylation entropies of most genomic loci remain constant; (iii) compared to normal tissue controls, some tumors exhibit greater methylation entropies; (iv) Alu elements with high methylation entropy are associated with high GC content but depletion of CpG dinucleotides and (v) Alu elements in the intronic regions or far from CpG islands are associated with low methylation entropy. We further identified 12 putative allelic-specific methylated genomic loci, including four Alu elements and eight promoters. Lastly, using subcloned normal fibroblast cells, we demonstrated the highly variable methylation patterns are resulted from low fidelity of DNA methylation inheritance.

INTRODUCTION

The addition of a methyl group to the C5 position of cytosines at CpG dinucleotides is the most common covalent modification known to occur to DNA in mammalian genomes. The resulting pattern of CpG methylation is part of the epigenetic code, which is heritable albeit not encoded in the DNA sequence. DNA methylation has been recognized as a mechanism to stably silence gene transcription and inactivate transposable elements (1). During development and cellular differentiation, the establishment of tissue specific patterns of DNA methylation enables cells with same genetic composition to exhibit distinct phenotypes (2).

In recent years, a pleithora of factors were identified to be involved in the establishment and maintenance of DNA methylation patterns in mammalian genomes (3). During early development, patterns of DNA methylation are established by *de novo* DNA methyltransferases, DNMT3A and DNMT3B, with the assistance of a lymphoid-specific helicase (LSH), member of the SNF2/helicase family (4,5). During cell division, patterns of DNA methylation are faithfully copied from parental to daughter DNA strand by DNA methyltransferase 1 (DNMT1). After DNA replication, hemimethylated CpG sites are converted to fully methylated sites by DNMT1, in a complex with UHRF1 (Ubiquitin-like, containing PHD and RING finger domains; 1), and PCNA (Proliferating Cell Nuclear Antigen) (6,7).

In vitro, DNMT1 shows high processivity on hemimethylated DNA, skipping sites at a low frequency (8,9). It methylates hemimethylated DNA with fidelity of >95%, irrespective of the flanking sequence. *In vivo*, the mitotic transmission of genomic DNA methylation patterns can also be remarkably accurate (10). By the

*To whom correspondence should be addressed. Tel: +1 773 880 4000 (extn 56747); Fax: +1 773 755 6551; Email: hxie@childrensmemorial.org analysis of DNA methylation patterns derived from the genomic DNA of clonal populations of normal human mammary epithelial cells, Ushijima and colleagues reported fidelities of methylation patterns to range from 99.85–99.92% per CpG site per generation for unmethylated CpG islands (CGIs) in the promoter regions of five genes (11). Even higher fidelity in the maintenance of DNA methylation patterns was observed for two methylated CGIs. Similarly high fidelity rates (99.90–100%) were observed in cancer cells despite a 2-fold increase in the *de novo* methylation rate (12,13).

However, the fidelity of inheritance of DNA methylation patterns may vary across the genome (14), and maintenance of DNA methylation seems to be even more complex (15). In cultured mouse cells, maintenance of DNA methylation patterns of foreign methylated DNA occurred at a significantly lower fidelity, as low as 85% per site per generation (16). In addition, the fidelity of inheritance of DNA methylation varied among CpG sites. Stochastic changes in methylation have also been reported for some endogenous CpG sites (17–20). By monitoring the methylation status of a half methylated CpG site in the mouse *Igf2* upstream region, Riggs and colleagues reported highly-variable methylation in the early stage subclones, and a steady-state methylation level of 50% in all subclones after 25 generations of cell proliferation (19). Using an elegant hairpin-bisulfite PCR technique, Laird and colleagues demonstrated that the *de novo* methylation rate could be as high as 17% per site per generation (18).

Notwithstanding the progress already made in the field, much is yet to be uncovered in regard to the stability of methylation patterns in a whole genome scale. Indeed, relatively few loci were interrogated in the aforementioned research, and to date no large-scale work has been conducted to investigate diversity of methylation patterns genome-wide. In the present study, we exploited Shannon entropy to assess the variation in DNA methylation patterns of promoters/CGIs and Alu repeat-encompassing loci. Specifically, we investigated whether: (i) genomic loci differ with respect to variation in DNA methylation patterns, (ii) for a given locus, methylation patterns—be them homogeneous or heterogeneous—remain constant among cell populations, (iii) there are genomic features associated with variation in DNA methylation patterns and (iv) such epigenetic variation is associated with differential fidelity of DNA methylation inheritance.

MATERIALS AND METHODS

High-throughput bisulfite sequencing data sets for Alu elements

The high-throughput bisulfite sequencing data were derived from Alu-anchored bisulfite PCR libraries (21). Briefly, genomic DNA is first digested with AluI restriction enzyme, ligated to adaptors and then subjected to bisulfite treatment. Bisulfite treated DNA is amplified with adaptor and Alu-specific primers, the latter targeting a large pool of CpG-rich Alu elements. Thus, each PCR product contains the 5′-end of an Alu element and its (most often) unique flanking genomic sequence, which makes it possible for each sequence to be unambiguously mapped to the reference human genome. In this study, the sequence reads were generated from eight Alu amplicon libraries derived from tissues samples, including a normal cerebellum, a normal 4th ventricle lining, two primary non-aggressive, two primary aggressive and two recurrent ependymomas (21,22). Primary non-aggressive ependymomas are defined as primary tumors from patient with free of disease progression for >4 years and primary aggressive ones are defined as primary tumors from patients with recurrent disease within 3 years or deceased of disease.

Bisulfite sequencing data for promoters/CGIs on chromosome 21

Comprehensive methylation maps for the promoters and CpG islands on chromosome 21 were downloaded from http://biochem.jacobs-university.de/name21/index.html. This data set has been described in detail ref. (23). Briefly, the methylation data were generated with bisulfite conversion followed by PCR and subclone sequencing for five human cell types, including human peripheral blood, fibroblast, trisomic 21 fibroblast, human embryonic cell line HEK293 and the human hepatocellular liver carcinoma cell line HepG2. In this study, the sequence of each amplicon was scanned to identify all possible segments with six contiguous CpG dinucleotides. To determine the methylation entropy, only segments with at least sixteen sequence reads generated were included in the analysis.

Cell culture

Human lung fibroblasts MRC-5 catalog number CCL-171 (ATCC, Manassas, VA, USA) were cultured K-DMEM supplemented by 10% of fetal bovine serum (Hyclone, Logan, UT, USA) and glutamax^{TM} (Invitrogen/Gibco, Carlsbad, CA, USA). The culture was lifted using 0.05% trypsin (Cellgro, Mediatech, VA, USA). The single cells were picked up manually under the microscope with the help of finely attenuated (pulled) glass capillary pipette with a fire polished tip. Each cell was deposited in the well of 96-well plate and allowed to grow ~2 weeks before sub culturing into 24-well plate. Subcloned cells were serially transferred to a well of 6-well plate and to a T75 flask until reach the target cell number (5×10^6) for collection.

Preparation of bisulfite converted genomic DNA with hairpin linker

Genomic DNA was extracted from subcloned normal lung fibroblasts with Qiagen DNeasy tissue kit (Qiagen, Valencia, CA, USA). Two microgram of genomic DNA was digested with TaqI enzyme, and then ligated to 200-fold molar excess hairpin linker (/5′P/-CGC CGG AGC GAT GCG TTC GAG CAT CGC TCC GG) with Fast-LinkTM Kit purchased from EPICENTRE Biotechnologies Inc., Madison, WI, USA. After ligation, genomic DNA was purified with PureLink PCR Purification kit (Invitrogen) to remove excess hairpin linkers. Bisulfite modification of genomic DNA was performed with EZ DNA Methylation Gold kit (Zymo Research, Orange, CA, USA) according to the manufacturer's instructions.

PCR cloning, sequencing and multiple sequence aligments

PCR reactions were performed with Qiagen Hotstart PCR master kit (Qiagen). For each reaction, a 50 µl PCR mixture was prepared with 2 µl (100 ng) bisulfite treated DNA, 50 pmol each forward and reverse primers. The primers used in the PCR runs for genomic locus 1 (chr9:139174924-139175041) are 5′-GGT TAT TTT TTT TTT AGT TTT GGT TTA GAT ATG A-3′ and 5′-TTT CTC CAA TCT TAA CTT AAA CAT AAT TCC-3′. The primers used in the PCR runs for genomic locus 2 (chr10:134480046-134480230) are 5′-AAA TAT AAT TTA GAA GGT ATT GTA GAT GTA AAT G-3′ and 5′-CAT AAC TTA AAA AAT ATT ACA AAT ATA AAT ACC AAC-3′. The PCR products with appropriate size were gel-purified and cloned with TOPO vectors (Invitrogen). Sequencing reactions for colonies were conducted at the Sequencing Core Facility of the Children's Memorial Research Center of Northwestern University's Feinberg School of Medicine. To ensure an accurate calculation of the fidelity of inheritance of DNA methylation, the sequence reads contain unconverted cytosine at non-CpG sites, due to the incomplete bisulfite conversion, were discarded. After the removal of vector and primer sequences, the sequence reads obtained were subjected to multiple alignments together with a reference sequence for corresponding genomic locus. Multiple sequence alignments were performed with clustal W (24).

Statistical analysis of the association between methylation entropies and DNA related attributes

The statistical analyses were conducted as previously described (22). Briefly, we compiled a comprehensive list of attributes that can be linked directly to the genomic regions of interest. The data for most of these attributes were calculated based on the UCSC Genome Annotation Database (25). The attributes for DNA sequence features were directly calculated based on the DNA sequence extracted from the human genome. All the attributes are either in the numerical form or boolean form (such as present in gene or not). The non-parametric Wilcoxon ranksum test and chi-square test statistical tests were performed for each attribute in numerical form or boolean form, respectively. Significance thresholds were adjusted for multiple testing using the highly conservative Bonferroni method, and the family-wise error rate was set to be <1%.

RESULTS

The definition and statistical assessment of methylation entropy

Traditionally, DNA methylation data analysis is based on the determination of the average methylation level (the percentage of methylated CpG) of one or more contiguous CpG sites. Such conventional way is unable to dissect DNA methylation patterns, which are herein defined as the combination of methylation statuses of contiguous CpG dinucleotides in a DNA strand. In order to better decode epigenetic data, we defined 'methylation entropy' and exploited it to assess the variability of DNA methylation pattern that might be observed for a given genomic locus in a cell population. The concept of entropy was first introduced by Rudolf Clausius as a thermodynamic property and later modified as Shannon entropy in information theory to measure the degree of uncertainty associated with a stochastic event (26).

$$\mathit{Entropy}: H(X) = -\sum P(x)\log_2 P(x)$$

An important variable in entropy equation is the probability $P(x)$ for a given event x. A frequently used example to interpret the concept of Shannon entropy is tossing a coin, which has two possible outcomes. Since it is a random event, the probability for heads or tails would be 0.5. Similarly, the methylation status (methylated or unmethylated) of a CpG dinucleotide could be considered as heads or tails but may not be random. Thus, the concept of entropy could be modified to quantitatively assess the variation in DNA methylation patterns.

To calculate methylation entropy, the following parameters were introduced to the original entropy formula: (i) number of CpG sites in a given genomic locus; (ii) number of sequence reads generated for a genomic locus and (iii) frequency of each distinct DNA methylation pattern observed in a genomic locus, calculated based upon the sequence reads that were generated for the locus (Figure 1A). The probability of a given event in Shannon entropy equation was replaced with the frequency of a

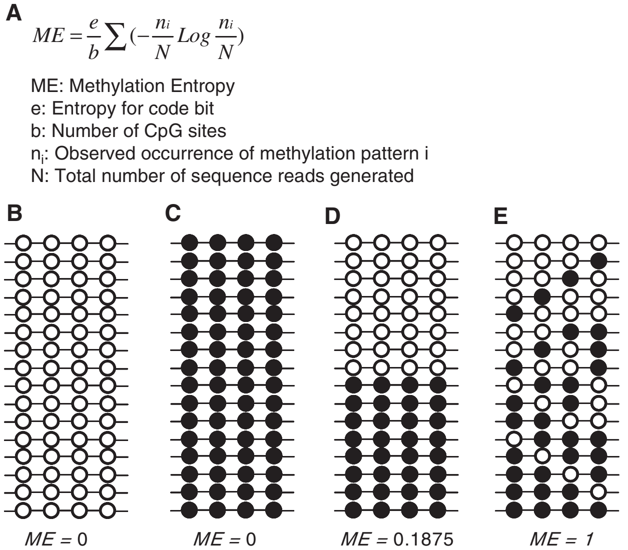
A
$ME = \frac{e}{b}\sum(-\frac{n_i}{N}Log\frac{n_i}{N})$
ME: Methylation Entropy
e: Entropy for code bit
b: Number of CpG sites
n_i: Observed occurrence of methylation pattern i
N: Total number of sequence reads generated
B
C
D
E
ME = 0
ME = 0
ME = 0.1875
ME = 1

Figure 1. The formula of methylation entropy and the examples for genomic loci with various methylation entropies in a cell population. (**A**) The formula of methylation entropy. The determination of methylation entropy requires three parameters: the number of CpG sites, the total number of sequence reads generated and the occurrence of each methylation pattern. (**B–E**) Genomic loci with various methylation entropies.

distinct methylation pattern observed for a genomic locus. The number of sequence reads generated was used to determine the frequency of a given methylation pattern. The number of CpG dinucleotides was used to normalize the increasing number of possible patterns resulting from the presence of additional CpG sites. The methylation entropy is minimal when DNA molecules in all cells share the same methylation pattern (Figure 1B and C), and is maximal when all possible DNA methylation patterns are equally represented in a population of cells (Figure 1E). Accordingly, genomic loci with the same methylation entropy might have different methylation levels on average (Figure 1B and C). In turn, genomic loci with different methylation entropies may share the same average level of methylation (Figure 1D and E). Since methylation entropy reflects the randomness in the distribution of DNA methylation patterns, it may serve as an indicator for stochastic methylation changes. Thus, methylation entropy analysis differs significantly from conventional methylation level-based analyses in that it enables assessment of methylation pattern stability and diversity.

We exploited simulations to provide statistical assessment for methylation entropy, thus enabling determination of statistical significance for a stochastic methylation variation observed in a locus. Similarly to the methylation entropy determinations made based on the actual sequence data, those utilizing simulated data take into consideration the average methylation level, the number of CpG dinucleotides, and the sequence reads generated. For example, to determine whether or not the methylation patterns shown in Figure 1D were stochastic, we randomly generated 10 000 data sets by simulation. Each data set exhibited an average methylation level of 50%, and comprised 16 random methylation patterns representing 16 sequences with 4 CpG dinucleotides per read. The distribution of methylation entropies of these 10 000 random data sets indicated that a genomic region associated with stochastic methylation change would have a methylation entropy of approximately 0.80, and a minimum methylation entropy of 0.54 (Figure 2). Based on such distribution, we may conclude that the formation of the methylation patterns depicted in Figure 1D, with a calculated methylation entropy of 0.1875, must not be stochastic (lower than the minimum methylation entropy 0.54 observed in 10 000 simulations; $P < 0.0001$).

To model allelic-specific methylation, as an example of deterministic methylation changes, we simulated another 10 000 semi-random sets of methylation patterns. To accommodate sampling errors and natural methylation variations on differentially methylated alleles, we arbitrarily assigned 6 out of 16 reads in each set to be completely methylated and another six reads to be completely unmethylated. The remaining four reads in each set were with random methylation patterns. The methylation entropy distribution of such semi-random data sets indicated that a genomic region associated with allelic-specific methylation might have a methylation entropy of approximately 0.35, and a maximum methylation entropy of 0.52. Therefore, although the average methylation levels of the genomic loci illustrated in Figure 1D and E are both 50%, their methylation entropies are different and their methylation patterns are formed through two distinct processes, deterministic and stochastic, respectively.

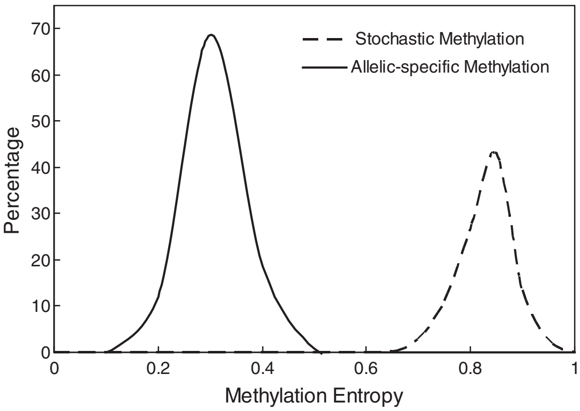

Figure 2. The distribution of methylation entropy for simulation results. For a genomic locus with four CpG dinucleotides and average methylation level as 50%, 10 000 methylation data sets were generated. Each data set comprised of 16 sequence reads with four CpG sites per read. The dashed curve represents simulation result for stochastic methylation event. For 10 000 data sets, the methylation entropy ranged from 0.54 to 0.97 with average as 0.80. The solid curve represents simulation result for allelic-specific methylation as an example of deterministic methylation event. For 10 000 data sets, the methylation entropy ranged from 0.24 to 0.52 with average as 0.35.

Comparison of Alu methylation entropies in normal and in cancer epigenomes

We applied the measure of methylation entropy to assess variation of DNA methylation patterns genome-wide. Two large data sets of bisulfite-converted genomic DNA sequences, which were previously described in detail (21,22), were explored. In these data sets, the majority of sequence reads comprise the 5′-most 80 bp of a select yet large subset of evolutionarily young, epigenetically informative, i.e. CpG-rich, AluY retrotransposons and their upstream flanking genomic sequences. The sequences were generated from eight tissues samples, including a normal cerebellum, a normal 4th ventricle lining, two primary non-aggressive, two primary aggressive and two recurrent ependymomas. Altogether, over 506 million nucleotides from 3 million sequence reads were included in this analysis. After removal of primer and adaptor sequences, 2.3 million sequence reads encompassing 348 million nucleotides were unambiguously mapped to the human genome. A total of 13 million methylation data points were generated for 289 816 distinct CpG sites that are widely distributed in the human genome. It is noteworthy that the bisulfite conversion rates attained in these data sets ranged from 99.1 to 99.7% (21,22). To ensure reliability of data analysis, for each tissue sample, only the genomic loci for which there were at least sixteen sequence reads, each containing four or more contiguous CpG dinucleotides, were included. Approximately 3000 genomic loci—it ranged from 2153 to 3730—were identified for each sample, which met the aforementioned criteria. The methylation entropies for these genomic loci were calculated based on 581 208 sequence-reads comprising over 3.5 million methylation data points (Table 1).

The distribution of methylation entropies was determined for each sample. Within a cell population, the methylation entropy varied among different genomic loci. Approximately 5–10% of genomic loci in each sample exhibited zero methylation entropy, which indicated that only one methylation pattern was found in sixteen or more sequence-reads (Figure 3A). Since Alu repeats are heavily methylated in the genome, the average methylation level for the two data sets was found to be >90% (21,22). Therefore, the majority of the genomic loci that exhibited zero methylation entropy were completely methylated. Such uniformity in the pattern of DNA methylation suggests that these genomic regions exhibit extremely high fidelity of inheritance of DNA methylation. Three out of eight tissues samples, including two normal controls and one primary non-aggressive ependymoma, had a very similar distribution of methylation entropies. Shifts to higher methylation entropies were observed in the remainder ependymomas (Figure 3A). This suggests that tumors, especially those that are most aggressive, might be characterized by an increased genome-wide disorder in DNA methylation patterns.

We further focused our analysis on the methylation entropies of genomic loci for which there were 16 or more sequence-reads in at least two samples. All possible pairwise comparisons were performed with the eight tissue samples to uncover differences in the methylation entropies of these genomic loci. Interestingly, we found that methylation entropies of most genomic loci remained constant regardless of the tissue source. To demonstrate such constancy, we determined the Pearson's correlation of methylation entropies for all pairwise comparisons. Modest corrections (ranging from 0.29 to 0.65) were identified with an average of 0.5 for a given pair (Supplementary Table S1A). In particular, the most significant correlation was observed for the methylation entropies of 935 loci from a primary aggressive and a recurrent ependymoma derived from the same individual (Figure 4). Altogether, these results indicated that for a given locus and tissue, methylation entropies can be similar among individuals. On the other hand, it also suggested that the observed variation in DNA methylation pattern is locus specific.

Comparison of methylation entropies of promoters/CGIs on chromosome 21

We extended the analysis to five comprehensive methylation maps for the promoters/CGIs on chromosome 21, generated with bisulfite conversion and subclone sequencing (23). This data set at single base resolution comprise methylation data of 190 gene promoters covered by 297 amplicons for five human cell types, including human peripheral blood, fibroblast, trisomic 21 fibroblast, human embryonic cell line HEK293 and the human hepatocellular liver carcinoma cell line HepG2. Since substantial methylation differences were found for different segments within a promoter, we scanned each sequence read to identify all possible segments with six contiguous CpG sites (same as the

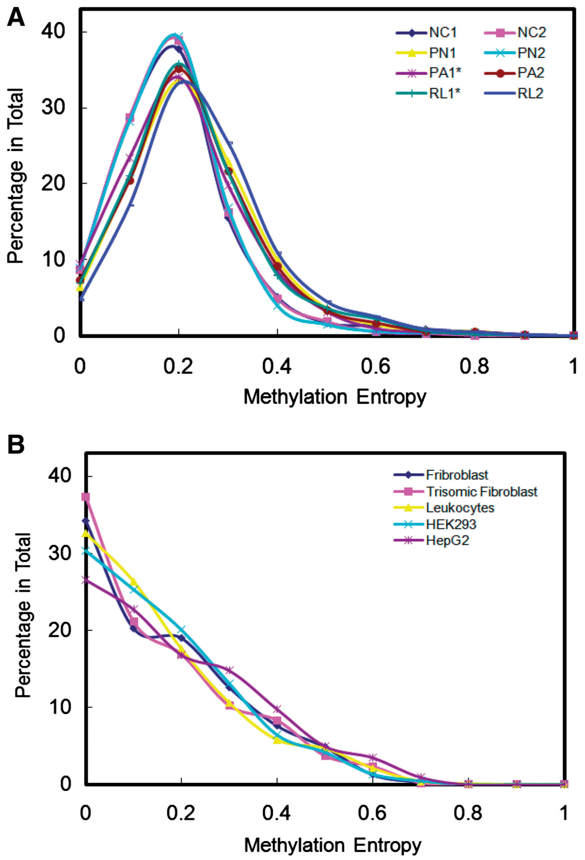

Figure 3. The distribution of methylation entropies in tissue samples. The *x*-axis represents different levels of methylation entropies. The *y*-axis represents the percentage of genomic loci examined. (**A**) The distribution of Alu methylation entropies. NC, PN, PA and RL represent normal control, primary non-aggressive, primary aggressive and relapsed ependymoma tissues, respectively. PA1* and RL1* were derived from one same individual. (**B**) The distribution of methylation entropies of all possible segments with six contiguous CpG sites in promoters and CGIs on chromosome 21.

Table 1. Statistics of high-throughput bisulfite sequencing data for Alu and flanking sequence

Sample ID	NC1	NC2	PN1	PN2	PA1	PA2	RL1	RL2
Number of Loci	2530	3678	3076	2334	3730	2958	2153	2965
Number of sequence reads	70 154	102 405	76 033	52 652	84 603	71 289	46 938	77 134
Number of CpG sites	402 506	653 650	467 705	313 967	542 630	424 410	288 981	435 017
Average reads per locus	27.7	27.8	24.7	22.6	22.7	24.1	21.8	26.0
Average CpG sites per reads	5.7	6.4	6.2	6.0	6.4	6.0	6.2	5.6

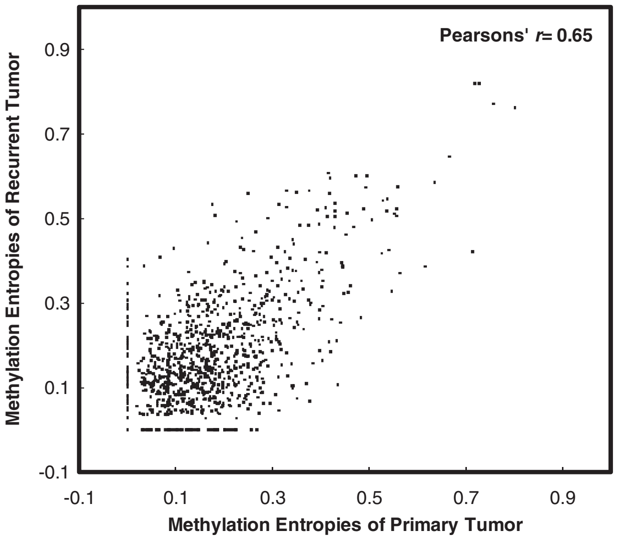

Figure 4. The correlation of methylation entropies between the primary and relapsed ependymoma tissues from one individual. Each dot represents a genomic locus with the methylation entropies calculated for primary (PA1*) and relapsed (RL1*) tumors.

average CpG dinucleotides per read for Alu elements in previous section). For each sample, approximately 3000 segments (ranged from 2663 to 3498) with methylation data at the minimum of 16-fold coverage were identified.

Within a cell population, the methylation entropy varied among different segments of promoters. Approximately 27–37% of segments in each sample exhibited zero methylation entropy (Figure 3B). Since the DNA methylation level of promoters was found to follow a bimodal distribution (23), the segments that exhibited zero methylation entropy were either completely methylated or unmethylated. Compared to normal tissues, fewer segments, 27 and 30%, respectively, were found to exhibit zero methylation entropy in the HepG2 carcinoma cells and transformed HEK293 cells. This result suggests that tumor cells tend to exhibit increased genome-wide methylation entropies—not only at Alu elements—but also at promoters and CGIs. Unexpectedly, the trisomic 21 fibroblasts showed slightly decreased methylation entropy. This suggests that the presence of an additional (partial or entire) chromosome might cause some loci to exhibit more homogenous DNA methylation patterns. Further pair-wise comparisons revealed segments in promoter regions or CGIs display modest correlations in methylation entropy between tissue samples, ranging from 0.25 to 0.56 (Supplementary Table S1B). It is not surprising to find such correlation of methylation entropies between normal and trisomic 21 fibroblasts. However, we found that the most significant correlation of methylation entropies was observed between HEK293 and HepG2 cells. HepG2 is a well differentiated hepatocellular carcinoma cell line, whereas HEK293 was derived from embryonic kidney cells transformed with partial adenovirus DNA (27). Having been derived from a mixture of embryonic cells, HEK293 cells may contain diverse types of cells including immature neurons (28). The significant correlation of methylation entropies between two such distinct cell lines suggests that tumor cells may share common loci with disordered DNA methylation patterning.

Putative identification of genomic loci exhibiting allelic-specific methylation patterns

It was shown in a previous study that the methylation statuses of some Alu elements are under parental origin effect (29). Thus, it would be of interest to examine whether any Alu element exhibits a biphasic distribution of DNA methylation. In addition, based on the reported chr21 methylation maps, a few promoters on chromosome 21 were found to be methylated in an allelic-specific manner (23,30). Since allelic-specific methylation constitutes a deterministic methylation event that is associated with a lower methylation entropy than that of a stochastic methylation event, we applied the model described in the previous section for the identification of allelic-specific methylation.

Four putative allelic-specific methylated Alu elements were identified (Supplementary Figure S1). Two of these Alu elements reside within the intronic regions of the ATP9A and the DG2L6 genes, respectively. Interestingly, according to the ENCODE transcription factor ChIP-seq data, the Alu element in the intron of the ATP9A gene hosts a binding site for FOSL2. By scanning the methylation maps of chromosome 21, we identified 54 segments within 8 promoters that show striking biphasic distribution of DNA methylation including the genomic region 176_2 at the CBR1 locus, which was identified in a previous study (23). A full list of these segments is provided in the Supplementary Table S2 and detailed methylation patterns are shown in the Supplementary Figure S2. It is noteworthy that the methylation entropy analysis alone cannot distinguish allelic-specific methylation patterns from those derived from two or more cell subpopulations. Therefore, to draw a solid conclusion in regard to the occurrence of allelic-specific methylation, further experiments including SNP analysis are needed.

The association of methylation entropy and fidelity of methylation inheritance

Previous studies indicated that stochastic changes in methylation may occur in some genomic regions (17–20). In addition, the inheritance of DNA methylation may not depend on an accurate copy of each CpG during and after DNA replication (15,31). To further understand the origin of high variations in DNA methylation pattern, we examined the fidelity of DNA methylation inheritance for the genomic loci with high methylation entropies. Since the previous Alu methylation analyses were conducted with methylation data derived from bulk tissues, the observed methylation variations could arguably be explained or in great part contributed by the presence of multiple types of cells in these tissues. To eliminate such possibility, we clonally propagated normal human fibroblast cells and derived uniform cell populations.

Using a hairpin-linker ligated to restriction-enzyme-digested genomic DNA, Laird and colleagues successfully obtained methylation information for the two complementary DNA strands simultaneously (18). In this study, genomic DNA isolated from subcloned normal fibroblast cells was first digested with the methylation-insensitive TaqI restriction endonuclease, and then ligated to hairpin linkers. After bisulfite conversion and PCR cloning, sequencing reactions were conducted to generate methylation profiles for two genomic loci that exhibited high methylation entropies in normal cerebellum and in ependymomas (Supplementary Figure S3). More detailed methylation patterns of these two genomic loci can be visualized at http://cmbteg.childrensmemorial.org/cgi-bin/gbrowse/btech (32). To ensure an accurate calculation of the fidelity of inheritance of DNA methylation, the sequence reads with incomplete bisulfite conversion (containing unconverted cytosines at non-CpG dinucleotides) were discarded. A total of 27 sequence-reads comprising 12 distinct methylation patterns and 30 sequence-reads comprising 10 distinct methylation patterns were obtained for genomic locus 1 (chr9:139174924-139175041) and genomic locus 2 (chr10:134480046-134480230), respectively. Both genomic loci exhibited high methylation variations, similarly to those observed in normal cerebellum and in ependymomas (Figure 5).

Since each sequence read encompasses the two complementary DNA strands, the resulting methylation data also enabled determination of the fidelity with which methylation is inherited for each CpG site. A symmetrical methylation status of CpG/CpG dyads (either methylated or unmethylated) indicates a successful methylation inheritance, while an asymmetrical methylation status (hemimethylated CpG/CpG dyads) indicates a failure in methylation inheritance. For genomic locus 1, each sequence read contained six CpG/CpG dyads from the two complementary strands. Sixteen asymmetrical methylation statuses were observed for a total of 162 CpG/CpG dyads. Thus, the average fidelity of methylation inheritance for genomic locus 1 was approximately 0.90. For genomic locus 2, each sequence read contained five CpG/CpG dyads. Fourteen asymmetrical methylation statuses were observed for 150 CpG/CpG dyads. The average fidelity for methylation inheritance for genomic locus 2 was approximately 0.91. These results demonstrated that these two genomic loci exhibited much lower fidelity of DNA methylation inheritance than that previously reported for DNMT1 based on *in vitro* studies—>95% (8,9), and *in vivo* analysis of CGI methylation patterns (>99.8%) derived from subcloned normal human mammary epithelial cells (11). It is noteworthy, however, that in spite of the low fidelity observed for adjacent CpG dyads, the methylation inheritances of two CpG sites (the 5′-most CpG site in genomic locus 1, and the middle CpG site in genomic locus 2) were found to be 100% accurate in 27 and 30 reads analyzed, respectively (Figure 5). This indicated that the methylation status of certain CpG sites is maintained with much higher fidelity than that of neighboring CpGs.

Genomic features and DNA related attributes associated with Alu methylation entropy

A previous methylation study on chromosome 21 promoters revealed that genomic location and sequence features have great impact on DNA methylation (23). Substantial differences in methylation were observed for different parts within a promoter. In addition, highly methylated CpG dinucleotides in promoters/CGIs were often flanked by AT rich sequences. In the previous sections of this study, we analyzed the genome-wide distribution of Alu methylation entropies for eight tissue samples. Approximately 5–10% of genomic loci in each sample exhibited only one methylation pattern in sixteen or more sequence reads, while some genomic loci showed

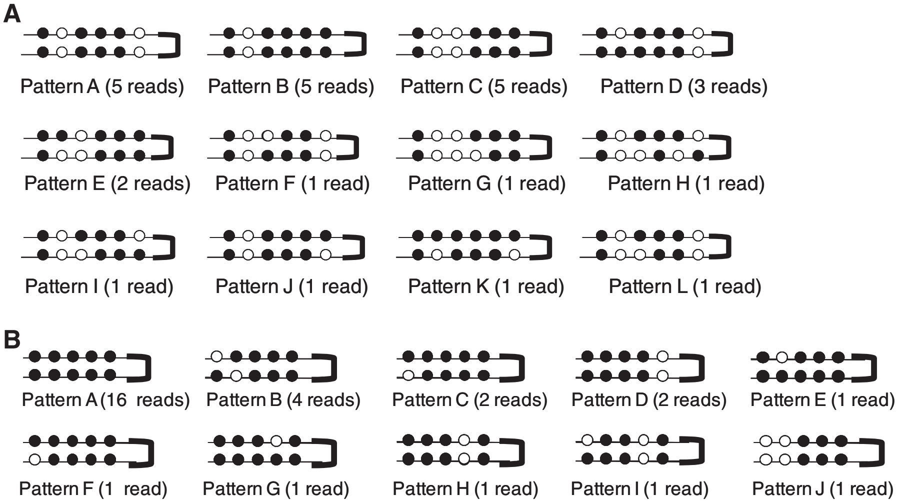

Figure 5. The methylation status of CpG/CpG dyads in subcloned human normal lung fibroblast determined by hairpin-bisulfite PCR. (**A**) Bisulfite sequencing results for genomic locus 1 (chr9:139174924-139175041). (**B**) Bisulfite sequencing results for genomic locus 2 (chr10:134480046-134480230). The methylated cytosines are indicated with filled circles while unmethylated cytosines are indicated with open circles. The bold curved lines represent hairpin linkers connected to both complementary strands. The symmetrical methylation statuses of CpG/CpG dyads indicate an accurate methylation inheritance, while asymmetrical methylation status (hemimethylated) indicates a failure in the transmission of methylation status or a *de novo* methylation event.

high methylation entropy in multiple samples. These results also suggested that in different loci, maintenance of DNA methylation might be either deterministic or stochastic. To understand such regional specificity, we examined the association between DNA-related attributes of genomic loci and their methylation entropies. Our previous study demonstrated that, in cancer epigenomes, both hypomethylation and hypermethylation do occur in Alu repeats and in their 5′-flanking sequences (22). We also showed that the differentially methylated CpG sites in ependymomas are not randomly distributed in the genome. To eliminate the influence of cancer-associated methylation changes in the present study, we limited further analysis to two normal samples in order to identify genetic factors associated with normal epigenetic variation.

Based on the level of methylation entropy observed in the two normal samples, we compiled two disparate sets of genomic loci. The first set comprising 699 genomic loci exhibited a methylation entropy lower than the cut-off for non-stochastic events ($P < 0.05$), including regions with only one methylation pattern found in 16 or more sequence reads in at least one normal sample. The second set of 678 genomic loci exhibited methylation entropy higher than the average of the methylation entropies plus one standard deviation. Based on the genomic coordinates of these loci, we ascertained each of their flanking sequences (in 1 kb windows) for 283 genomic DNA attributes, including 13 genomic features and 270 sequence characteristics (Supplementary Table S3). For all the attributes compiled, statistical comparisons were conducted for the genomic loci with high methylation entropies and the genomic loci with low methylation entropies.

After family wise error rate justification, a number of genomic features were found to be significantly associated with methylation entropy (Table 2). The distribution of CpG islands and genes in the genome contribute to the variation in DNA methylation pattern of Alu elements. The Alu and flanking sequences in the intronic regions displayed lower methylation entropy, while the ones close to CGIs exhibited higher methylation entropy. In addition, compared to the set with low methylation entropy, the high methylation entropy regions contain distinct sequence characteristics. The GC content of high methylation entropy regions is significantly higher than that of the genomic regions with low variation in DNA methylation pattern ($P < 1.3\text{E-}08$). Although there was no significant difference in the number of CpG dinucleotides per 1 kb window between the two sets of genomic loci, the CpG ratio (observed versus expected) was significantly lower in the genomic regions with high methylation entropy. Such differences in GC content were also manifested in the presentation of tetra-nucleotides. For genomic regions with low methylation entropy, ‘TA-only’ tetra-nucleotides, such as ‘AATT’ and ‘TAAT’, were enriched. In contrast, for the genomic loci showing high methylation entropy, the tetra-nucleotides ‘CCCC’ and ‘CCCT’ were enriched.

DISCUSSION

The assessment of normal intra- and inter-individual epigenetic variation is a critical step for epigenetic studies, in particular for the identification of functional epimutations associated with complex diseases (33). Recent studies on the DNA methylation patterns of specific loci revealed substantial epigenetic variation among and within individuals (34,35). In this study, we introduced the concept of methylation entropy and exploited its statistical assessment to quantitatively measure variation in the patterns of DNA methylation in a cell population. The occurrence of a uniform pattern of DNA methylation indicates a high fidelity of methylation inheritance. In contrast, a diverse pattern of DNA methylation may result either from the

Table 2. Genomic features or DNA related attributes associated with Alu methylation entropy

Attribute name	Direction of change[a]	Statistical test	Significance (not adjusted)	Significance (Bonferroni)	Significance threshold (FDR)
GC content	Increase	WilcoxonRankSum	1.24E-08	3.50E-06	3.53E-05
tetraNT_AATT	Decrease	WilcoxonRankSum	3.63E-08	1.03E-05	7.07E-05
tetraNT_TAAT	Decrease	WilcoxonRankSum	1.25E-07	3.55E-05	1.06E-04
cCount	Increase	WilcoxonRankSum	2.12E-07	5.99E-05	1.41E-04
tetraNT_CCCT	Increase	WilcoxonRankSum	3.04E-07	8.60E-05	1.77E-04
in Intron	Decrease	Chi-Square test	5.80E-06	1.64E-03	2.12E-04
tetraNT_ATTA	Decrease	WilcoxonRankSum	6.10E-06	1.73E-03	2.47E-04
tetraNT_ATTT	Decrease	WilcoxonRankSum	1.52E-05	4.29E-03	2.83E-04
tetraNT_ATAT	Decrease	WilcoxonRankSum	2.96E-05	8.37E-03	3.18E-04
distance to most adjacent CGI	Decrease	WilcoxonRankSum	5.34E-05	1.51E-02	3.53E-04
tetraNT_TAGT	Decrease	WilcoxonRankSum	5.82E-05	1.65E-02	3.89E-04
tetraNT_AAAT	Decrease	WilcoxonRankSum	6.10E-05	1.73E-02	4.24E-04
tetraNT_CCCC	Increase	WilcoxonRankSum	6.13E-05	1.73E-02	4.59E-04
tetraNT_TTTA	Decrease	WilcoxonRankSum	1.17E-04	3.30E-02	4.95E-04
tetraNT_TATA	Decrease	WilcoxonRankSum	1.40E-04	3.97E-02	5.30E-04
cgRatio	Decrease	WilcoxonRankSum	1.57E-04	4.43E-02	5.65E-04
tetraNT_AATA	Decrease	WilcoxonRankSum	1.67E-04	4.71E-02	6.01E-04

[a] The direction of change (increase or decrease) indicated the association of genomic region with high methylation entropy with the genomic features or DNA related attributes examined.

presence of multiple cell types or from a decreased fidelity of methylation inheritance. It should be emphasized, however, that the statistical assessment of methylation entropy simply enables the distinction between deterministic and stochastic patterns of DNA methylation. It is also noteworthy that this approach may be applied to any DNA methylation data set at a single base resolution.

In this study, we analyzed the methylation entropy for promoters, CGIs and Alu repeats. Despite the limitations that are inheriting to PCR bisulfite sequencing, and to the number of epigenomes that could be investigated—large-scale sequencing costs are still prohibitive—a few interesting observations could be made. Using these large sequence data sets, we were able to reconcile previous findings indicating that while certain genomic loci in CGIs showed high accuracy of DNA methylation pattern preservation (8,9,11), some repetitive elements seemed to be associated with stochastic methylation changes (17,18). We found that, in normal or cancer epigenomes, ~5–10% of Alu elements and 27–37% of promoter/CGI segments exhibiting a uniform DNA methylation pattern (Figure 3). Over 70% of promoter/CGI segments or Alu repeats were with methylation entropy <0.2 in normal tissues. Therefore, the inheritance of DNA methylation pattern is highly accurate for the majority of CGIs and Alu repeats. We also observed that ~3–5% of Alu repeats and 6–7% of promoter/CGI segments were with methylation entropy over 0.5. This indicates that, in normal tissues, the methylation variations or dynamic DNA methylation changes are limited to some genomic loci which could be the junctions that mark the boundary of hyper- and hypomethylated regions. In addition, we found that the methylation entropies of Alu repeats and CGIs remained constant regardless of the tissue source (Supplementary Table S1). This suggests different types of cells may share a general mechanism for guiding the epigenome configuration.

It has been known that morphologically homogeneous tumors could be biologically heterogeneous. The introduction of methylation entropy may provide a quantitative way to evaluate the tumor heterogeneity. Compare to normal tissues, heterogeneous tumor cells showed increased methylation entropies for both Alu repeats and CGIs. This is consistent with previous observation that tumor cells are with decreased fidelity in DNA methylation inheritance (13). The fidelity in epigenetic inheritance has been associated with many players, including protein and RNA factors (36,37). The decreased fidelity of DNA methylation inheritance in tumors suggests that tumor development and progression might be frequently accompanied by a disorder in the machinery, protein and/or RNA factors, responsible for the preservation or the establishment of DNA methylation patterns. Interestingly, substantial consistency on methylation entropy was observed among cancer cells, HEK293 and HepG2 in particular. This suggests the methylation entropy analysis might provide additional epigenetic marks lack of significant changes at the average methylation level.

In this study, we found that certain sequence characteristics within 1 kb windows of these genomic loci contribute to the maintenance of Alu methylation patterns. A high GC content but with a low CpG ratio may contribute to a variation in DNA methylation pattern. Similar to the position effect found for gene expression, we also observed that the location of genomic loci could have an impact on the preservation of their DNA methylation patterns. Although Alu elements are the primary targets for DNA methylation, which keeps them silenced and thereby prevent genomic instability, we found that Alu repeats that are closer to CGIs are more frequently associated with stochastic methylation changes. In contrast, the Alu elements in the intronic regions are more likely to be stably methylated. This is consistent with the well-known fact that the gene bodies in the mammalian genome are heavily methylated. The identification of *cis*-factors and the genomic features associated with methylation variations in this study is consistent with the scenario that recruitment of DNMTs and other factors might be region-specific (15).

Lastly, with subcloned normal fibroblast cells, we further demonstrated the association between highly variable methylation patterns and low fidelity of DNA methylation inheritance. Interestingly, in two genomic loci examined, the methylation status of certain CpG sites were found to be maintained with 100% accuracy. Such strikingly higher fidelity than that of their neighbors suggests that either some CpG sites are more accessible for DNMTs or the methylation status of these CpG sites might be essential, as previously discussed (14). We anticipate the introduction of methylation entropy and such genome-wide analysis of normal methylation variation would provide additional justification for studies to uncover epigenetic marks associated with human diseases, including cancers.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Jiping Wang (Northwestern University, Chicago) for the discussion on statistics analysis.

FUNDING

Funding for open access charge: NHLBI10279457; Everett/O'Connor Charitable Trust; Dr Ralph and Marian C. Falk Medical Research Trust; Gus Foundation; Maeve McNicholas Memorial Foundation and Medical Research Institute Council.

Conflict of interest statement. None declared.

REFERENCES

1. Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes. Dev.*, **16**, 6–21.
2. Mohn,F. and Schubeler,D. (2009) Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet.*, **25**, 129–136.

3. Ooi,S.K., O'Donnell,A.H. and Bestor,T.H. (2009) Mammalian cytosine methylation at a glance. *J. Cell Sci.*, **122**, 2787–2791.
4. Okano,M., Bell,D.W., Haber,D.A. and Li,E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.
5. Zhu,H., Geiman,T.M., Xi,S., Jiang,Q., Schmidtmann,A., Chen,T., Li,E. and Muegge,K. (2006) Lsh is involved in de novo methylation of DNA. *EMBO J.*, **25**, 335–345.
6. Arita,K., Ariyoshi,M., Tochio,H., Nakamura,Y. and Shirakawa,M. (2008) Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature*, **455**, 818–821.
7. Bostick,M., Kim,J.K., Esteve,P.O., Clark,A., Pradhan,S. and Jacobsen,S.E. (2007) UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science*, **317**, 1760–1764.
8. Goyal,R., Reinhardt,R. and Jeltsch,A. (2006) Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. *Nucleic Acids Res.*, **34**, 1182–1188.
9. Vilkaitis,G., Suetake,I., Klimasauskas,S. and Tajima,S. (2005) Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. *J. Biol. Chem.*, **280**, 64–72.
10. Ooi,S.K. and Bestor,T.H. (2008) Cytosine methylation: remaining faithful. *Curr. Biol.*, **18**, R174–176.
11. Ushijima,T., Watanabe,N., Okochi,E., Kaneda,A., Sugimura,T. and Miyamoto,K. (2003) Fidelity of the methylation pattern and its variation in the genome. *Genome Res.*, **13**, 868–874.
12. Watanabe,N., Okochi-Takada,E., Yagi,Y., Furuta,J.I. and Ushijima,T. (2006) Decreased fidelity in replicating DNA methylation patterns in cancer cells leads to dense methylation of a CpG island. *Curr. Top. Microbiol. Immunol.*, **310**, 199–210.
13. Ushijima,T., Watanabe,N., Shimizu,K., Miyamoto,K., Sugimura,T. and Kaneda,A. (2005) Decreased fidelity in replicating CpG methylation patterns in cancer cells. *Cancer Res.*, **65**, 11–17.
14. Chen,Z.X. and Riggs,A.D. (2005) Maintenance and regulation of DNA methylation patterns in mammals. *Biochem. Cell. Biol.*, **83**, 438–448.
15. Jones,P.A. and Liang,G. (2009) Rethinking how DNA methylation patterns are maintained. *Nat. Rev. Genet.*, **10**, 805–811.
16. Wigler,M., Levy,D. and Perucho,M. (1981) The somatic replication of DNA methylation. *Cell*, **24**, 33–40.
17. Riggs,A.D. and Xiong,Z. (2004) Methylation and epigenetic fidelity. *Proc. Natl Acad. Sci. USA*, **101**, 4–5.
18. Laird,C.D., Pleasant,N.D., Clark,A.D., Sneeden,J.L., Hassan,K.M., Manley,N.C., Vary,J.C. Jr, Morgan,T., Hansen,R.S. and Stoger,R. (2004) Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proc. Natl Acad. Sci. USA*, **101**, 204–209.
19. Riggs,A.D., Xiong,Z., Wang,L. and LeBon,J.M. (1998) Methylation dynamics, epigenetic fidelity and X chromosome structure. *Novartis Found. Symp.*, **214**, 214–225, discussion 225–232.
20. Pfeifer,G.P., Steigerwald,S.D., Hansen,R.S., Gartler,S.M. and Riggs,A.D. (1990) Polymerase chain reaction-aided genomic sequencing of an X chromosome-linked CpG island: methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability. *Proc. Natl Acad. Sci. USA*, **87**, 8252–8256.
21. Xie,H., Wang,M., Bonaldo,M.D., Smith,C., Rajaram,V., Goldman,S., Tomita,T. and Soares,M.B. (2009) High-throughput sequence-based epigenomic analysis of Alu repeats in human cerebellum. *Nucleic Acids Res.*, **37**, 4331–4340.
22. Xie,H., Wang,M., Bonaldo Mde,F., Rajaram,V., Stellpflug,W., Smith,C., Arndt,K., Goldman,S., Tomita,T. and Soares,M.B. (2010) Epigenomic analysis of Alu repeats in human ependymomas. *Proc. Natl Acad. Sci. USA*, **107**, 6952–6957.
23. Zhang,Y., Rohde,C., Tierling,S., Jurkowski,T.P., Bock,C., Santacruz,D., Ragozin,S., Reinhardt,R., Groth,M., Walter,J. *et al.* (2009) DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS Genet.*, **5**, e1000438.
24. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
25. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
26. Shannon,C.E. (1997) The mathematical theory of communication. 1963. *MD Comput.*, **14**, 306–317.
27. Louis,N., Evelegh,C. and Graham,F.L. (1997) Cloning and sequencing of the cellular-viral junctions from the human adenovirus type 5 transformed 293 cell line. *Virology*, **233**, 423–429.
28. Shaw,G., Morse,S., Ararat,M. and Graham,F.L. (2002) Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells. *FASEB J.*, **16**, 869–871.
29. Sandovici,I., Kassovska-Bratinova,S., Loredo-Osti,J.C., Leppert,M., Suarez,A., Stewart,R., Bautista,F.D., Schiraldi,M. and Sapienza,C. (2005) Interindividual variability and parent of origin DNA methylation differences at specific human Alu elements. *Hum. Mol. Genet.*, **14**, 2135–2143.
30. Zhang,Y., Rohde,C., Reinhardt,R., Voelcker-Rehage,C. and Jeltsch,A. (2009) Non-imprinted allele-specific DNA methylation on human autosomes. *Genome Biol.*, **10**, R138.
31. Liang,G., Chan,M.F., Tomigahara,Y., Tsai,Y.C., Gonzales,F.A., Li,E., Laird,P.W. and Jones,P.A. (2002) Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. *Mol. Cell. Biol.*, **22**, 480–491.
32. Wang,M., Xie,H., Stellpflug,W., Rajaram,V., Bonaldo Mde,F., Goldman,S., Tomita,T. and Soares,M.B. (2011) BTECH: a platform to integrate genomic, transcriptomic and epigenomic alterations in brain tumors. *Neuroinformatics*, doi:10.1007/s12021-12010-19091-12029 [Epub ahead of print, 6 January 2011].
33. Talens,R.P., Boomsma,D.I., Tobi,E.W., Kremer,D., Jukema,J.W., Willemsen,G., Putter,H., Slagboom,P.E. and Heijmans,B.T. (2010) Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. *FASEB J.*, **24**, 3135–3144.
34. Flanagan,J.M., Popendikyte,V., Pozdniakovaite,N., Sobolev,M., Assadzadeh,A., Schumacher,A., Zangeneh,M., Lau,L., Virtanen,C., Wang,S.C. *et al.* (2006) Intra- and interindividual epigenetic variation in human germ cells. *Am. J. Hum. Genet.*, **79**, 67–84.
35. Schneider,E., Pliushch,G., El Hajj,N., Galetzka,D., Puhl,A., Schorsch,M., Frauenknecht,K., Riepert,T., Tresch,A., Muller,A.M. *et al.* (2010) Spatial, temporal and interindividual epigenetic variation of functionally important DNA methylation patterns. *Nucleic Acids Res.*, **38**, 3880–3890.
36. Probst,A.V., Dunleavy,E. and Almouzni,G. (2009) Epigenetic inheritance during the cell cycle. *Nat. Rev. Mol. Cell Biol.*, **10**, 192–206.
37. Riggs,A.D. (2002) X chromosome inactivation, differentiation, and DNA methylation revisited, with a tribute to Susumu Ohno. *Cytogenet. Genome Res.*, **99**, 17–24.