Available online at www.sciencedirect.com
SCIENCE DIRECT
Genomics 87 (2006) 572–579
GENOMICS
www.elsevier.com/locate/ygeno

DNA motifs associated with aberrant CpG island methylation

F. Alex Feltus [a,1], Eva K. Lee [a,b], Joseph F. Costello [c], Christoph Plass [d], Paula M. Vertino [a,*]

[a] Department of Radiation Oncology and Winship Cancer Institute, Emory University School of Medicine, 1365-C Clifton Road. NE, Atlanta, GA 30322, USA
[b] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[c] The Brain Tumor Research Center and Department of Neurological Surgery, UCSF, San Francisco, CA 94143, USA
[d] Department of Medical Microbiology and Immunology, Division of Human Cancer Genetics, The Ohio State University, Columbus, OH 43210, USA

Received 29 August 2005; accepted 21 December 2005
Available online 17 February 2006

Abstract

Epigenetic silencing involving the aberrant methylation of promoter region CpG islands is widely recognized as a tumor suppressor silencing mechanism in cancer. However, the molecular pathways underlying aberrant DNA methylation remain elusive. Recently we showed that, on a genome-wide level, CpG island loci differ in their intrinsic susceptibility to aberrant methylation and that this susceptibility can be predicted based on underlying sequence context. These data suggest that there are sequence/structural features that contribute to the protection from or susceptibility to aberrant methylation. Here we use motif elicitation coupled with classification techniques to identify DNA sequence motifs that selectively define methylation-prone or methylation-resistant CpG islands. Motifs common to 28 methylation-prone or 47 methylation-resistant CpG island-containing genomic fragments were determined using the MEME and MAST algorithms (http://meme.sdsc.edu). The five most discriminatory motifs derived from methylation-prone sequences were found to be associated with CpG islands in general and were nonrandomly distributed throughout the genome. In contrast, the eight most discriminatory motifs derived from the methylation-resistant CpG islands were randomly distributed throughout the genome. Interestingly, this latter group tended to associate with *Alu* and other repetitive sequences. Used together, the frequency of occurrence of these motifs successfully discriminated methylation-prone and methylation-resistant CpG island groups with an accuracy of 87% after 10-fold cross-validation. The motifs identified here are candidate methylation-targeting or methylation-protection DNA sequences.

Keywords: Epigenetics; DNA methylation; Discriminant analysis; Classification techniques; Repetitive DNA; Alu

Introduction

CpG islands are short stretches (500–2000 bp) of genomic DNA enriched for the dinucleotide, 5′-CpG-3′, which is the substrate for methylation by DNA methyltransferases. While most CpG sites in the human genome are methylated, those in CpG islands are typically unmethylated in normal tissue. In human cancers, de novo methylation of CpG island sequences is accompanied by gene silencing and can serve as an alternative to mutation or deletion in the inactivation of tumor suppressor and other genes. Examples include the *VHL* gene in renal cell carcinomas [1], the *RB* gene in retinoblastomas [2], the cell cycle inhibitor *CDKN2A* in many epithelial cancers [3], the mismatch repair gene *MLH1* in sporadic colon cancer [4], and *CDH1* in breast, bladder, and prostate cancer [5]. In the case of *CDH1*, it has been demonstrated that monoallelic loss due to methylation-induced silencing provides the second genetic "hit" in hereditary gastric cancer [6]. Thus, aberrant methylation of CpG islands plays a critical role in the initiation and progression of cancer.

At present, it is not known why some CpG islands succumb to aberrant methylation during tumor progression while others are protected from it. One possibility is that local sequence features contribute to the susceptibility to, or protection from, de novo methylation. For example, there is evidence to suggest that methylation can spread from repetitive DNA [7,8,9]. Unusual DNA structures (e.g., non-B-DNA) are also suggested to serve as targets of de novo methylation [10,11]. Alternatively, methylation may be mistargeted in cancer cells through physical interaction of DNA methyltransferases with sequence-specific

* Corresponding author.
E-mail addresses: feltus@uga.edu (F.A. Feltus), pvertin@emory.edu (P.M. Vertino).
[1] Current address: University of Georgia, AGTEC Building, Room 225, 111 Riverbend Road, Athens, GA 30605, USA.

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.12.016

transcription factors such as RB [12] and PML-RAR [13]. Sp1-binding sites have been demonstrated to protect the murine *APRT* promoter from methylation [14,15]. Finally, there is mounting evidence that DNA-binding factors which function in the organization of chromatin domains, such as CTCF, also play a role in determining methylation status at imprinted genes and other loci [16–20]. Collectively, these studies suggest that there are local features which act in *cis* to influence the establishment and/or stability of DNA methylation patterns. The identification of such factors and their function will lead to a better understanding the molecular mechanisms underlying altered DNA methylation patterning in aging and carcinogenesis.

Using a human cell culture model in which de novo methylation of CpG islands is driven by ectopic expression of DNMT1 [21], we recently showed that CpG island loci differ in their innate susceptibility to de novo methylation [22]. By applying DNA pattern recognition and supervised learning techniques, we found that methylation-prone and methylation-resistant CpG islands could be distinguished based on their underlying sequence context [22]. These data demonstrated that there is a sequence signature associated with susceptibility to, or protection from, aberrant methylation, at least as it occurs in the DNMT1 overexpression model. We now extend these studies in an initial attempt toward identifying *cis*-acting elements that influence the propensity for aberrant DNA methylation. Using a combined approach coupling motif elicitation to define DNA patterns and classification techniques, we identified a sequence signature based on the frequency of 13 DNA motifs that can discriminate methylation-prone or methylation-resistant CpG islands. It is hypothesized that these motifs may play a role in the local susceptibility of CpG islands to aberrant DNA methylation.

Results

Methylation-prone and methylation-resistant CpG island data sets

In a previous study, we determined the methylation state of 1749 CpG islands in human fibroblasts clones overexpressing the DNA methyltransferase, DNMT1, using restriction landmark genomic scanning (RLGS) [22]. Methylation-prone (MP) loci were defined as those sequences that were consistently hypermethylated in 3 of 3 independent DNMT1 overexpressing clones but were never methylated in 3 vector-only control clones. Methylation-resistant (MR) CpG islands were defined as those that were never methylated in all 6 cell clones. The 4000-bp genomic sequence circumscribing the CpG island center (2000 bp upstream and 2000 bp downstream) was determined for 28 MP and 47 MR CpG islands for which sequence information was available. These functionally defined CpG island fragments were used in motif analysis.

Motif discovery

To identify motifs associated with methylation susceptibility, we derived sequences common to the methylation-prone or the methylation-resistant CpG islands using the program, MEME (http://meme.sdsc.edu) [23]. MEME derives sequence motifs from a training set of sequences and builds a position-specific scoring matrix wherein there is a probability associated with the occurrence of each base at each position. Twenty-eight methylation-prone or 47 methylation-resistant CpG island sequences (4 kb centered on the CpG island; see Ref. [22]) were used as input into the MEME algorithm, and the first 20 "best-fit" motifs were obtained for each data set.

These 40 motifs were then individually aligned to the entire data set of 75 sequences with the motif alignment program, MAST (http://meme.sdsc.edu) [24] to determine the number of occurrences of each motif in the methylation-prone and methylation-resistant data sets. Only those motif hits with a position-specific goodness-of-fit P value of less than 10e-6 were considered. A t test was then used to compare the frequency of each motif between methylation-prone and methylation-resistant CpG island data sets. Motifs with frequencies that were not significantly different between the methylation-prone and methylation-resistant groups ($P > 0.01$) were excluded from further analyses. This primary filter yielded 8 motifs derived from the methylation-resistant sequences (M_R) and five motifs derived from the methylation-prone sequences (M_P) (Table 1). The spatial positions of these 13 motifs within the CpG islands in both data sets are shown in Figs. 1 and 2, and the number of occurrences is listed in Table 1. The position-specific scoring matrices of these motifs are available as Supplemental Data and can be used to probe any DNA sequence set using MAST.

Motif frequencies in the human genome

We next examined the more global distribution of the motifs in CpG island and non-CpG island DNA. To accomplish this, we extracted all 2348 annotated CpG islands from human chromosome 1 (http://genome.ucsc.edu). The 4-kb CpG island genomic fragments were constructed by locating the center position of each CpG island and extracting 2000 bp of sequence flanking both sides of this position. In a similar manner, a control data set of 2348 random 4000-bp fragments was also generated. The occurrences of the Mr and Mp motifs was then determined in these data sets using MAST. The results in Table 2 show a strong association of methylation-prone motifs with CpG islands in general relative to random DNA (t test; $P = 0.035$). In contrast, the methylation-resistant motifs were similarly distributed in both CpG island and non-CpG island DNA (t test; $P = 0.243$). These data indicate that the M_P motifs are nonrandomly distributed in the genome and are selectively associated with CpG island DNA. Interestingly, although randomly distributed throughout the genome, there appeared to be a strong association between the methylation-resistant motifs and *Alu* sequences (Fig. 1B).

Methylation-prediction potential

Next we tested the potential of the 13 motifs to classify methylation-prone and methylation-resistant CpG islands using a linear-optimization-based discriminant analysis-supervised