© 1992 Oxford University Press

Nucleic Acids Research, Vol. 20, No. 9 2287–2291

Isolation and characterization of the cDNA encoding human DNA methyltransferase

Ray-Whay Chiu Yen[1], Paula M.Vertino[1], Barry D.Nelkin[1], Jane J.Yu[1], Wafik El-Deiry[1], Arunthathi Cumaraswamy[1], Gregory G.Lennon[3], Barbara J.Trask[3], Paul Celano[1] and Stephen B.Baylin[1,2]

[1]Oncology Center and [2]Department of Medicine, The Johns Hopkins Medical Institutions, Baltimore, MD 21231 and [3]Human Genome Center, Biomedical Sciences Division, The Lawrence Livermore National Laboratories, Livermore, CA 94551, USA

Received January 31, 1992; Revised and Accepted March 30, 1992 EMBL accession no. X63692

ABSTRACT

We have cloned a series of overlapping cDNA clones encoding a 5194 bp transcript for human DNA methyltransferase (DNA MTase). This sequence potentially codes for a protein of 1495 amino acids with a predicted molecular weight of 169 kDa. The human DNA MTase cDNA has eighty percent homology at the nucleotide level, and the predicted protein has seventy-four percent identity at the amino acid level, to the DNA MTase cDNA cloned from mouse cells. Like the murine DNA MTase, the amino terminal two-thirds of the human protein contains a cysteine-rich region suggestive of a metal-binding domain. The carboxy terminal one-third of the protein shows considerable similarity to prokaryotic (cytosine-5)-methyltransferases. The arrangement of multiple motifs conserved in the prokaryotic genes is preserved in the human DNA MTase, including the relative position of a proline-cysteine dipeptide thought to be an essential catalytic site in all (cytosine-5)-methyltransferases. A single 5.2 kb transcript was detected in all human tissues tested, with the highest levels of expression observed in RNA from placenta, brain, heart and lung. DNA MTase cDNA clones were used to screen a chromosome 19 genomic cosmid library. The DNA MTase-positive cosmids which are estimated to span a genomic distance of 93 kb have been localized to 19p13.2-p13.3 by fluorescence *in situ* hybridization. Isolation of the cDNA for human DNA MTase will allow further study of the regulation of DNA MTase expression, and of the role of this enzyme in establishing DNA methylation patterns in both normal and neoplastic cells.

INTRODUCTION

Modification of DNA by methylation of the C-5 position of cytosine is an important process for modulating DNA compartmentalization, chromatin assembly, and gene expression (reviewed in 1–4). In prokaryotes, DNA methylation is catalyzed by a series of DNA (cytosine-5)-methyltransferases (m^5C-methylases), many of which form 5-methylcytosine within specific DNA sequences (reviewed in 5–7). In eukaryotes, DNA methylation is thought to result from the action of a single enzyme, DNA methyltransferase (DNA MTase), which recognizes and catalyzes the methylation of cytosines located 5′ to guanosines (reviewed in 1,5). This enzyme, unlike many prokaryotic m^5C-methylases, appears to act independently of the sequences surrounding the substrate CpG sites (reviewed in 1,5).

Many prokaryotic m^5C-methylase genes have been cloned and characterized (reviewed in 6,7), but only one eukaryotic DNA MTase, from mouse, has been cloned and fully sequenced (8). The carboxy terminal one-third of this enzyme appears highly similar to bacterial type II DNA cytosine methyltransferases (8).

We have recently reported cloning a partial cDNA for human DNA MTase (9) to study the role of this enzyme in the abnormal patterns of DNA methylation which occur in human neoplasia (reviewed in 4,10). We found that an increased expression of DNA MTase begins in early stages of neoplasia and continues through tumor progression (9). In the present study we report cloning of the cDNA encoding the entire 5.2 kb transcript of human DNA MTase. The nucleotide and predicted amino acid sequences of the cDNA for human DNA MTase are compared to those for the mouse. Mapping of the precise position for the human gene on chromosome 19 has been performed and the distribution of DNA MTase mRNA expression in normal human tissues is reported.

MATERIALS AND METHODS

Screening of λgt11 and λZapII cDNA libraries

Poly(A)+ RNA from the TT cell line of human medullary thyroid carcinoma, which expresses high levels of the DNA MTase transcript (9), was used to construct random hexamer primed λgt11 and λZapII cDNA libraries. A 1.3 kb cDNA clone of human DNA MTase (9) from a human small cell lung carcinoma (SCLC) cell line, NCI-H249, was initially used to screen the λgt11 cDNA library by standard protocols (11). cDNA inserts from positive phage were isolated and subcloned into the EcoR I site of pBluescript SK+ plasmid, or into λZapII phage (Stratagene, Inc., San Diego, CA). The cDNA inserts from subcloned λZapII recombinants or from positive clones of the

λZapII cDNA library were rescued as pBluescript plasmids by helper phage mediated *in vivo* excision as described by the manufacturer (Stratagene, Inc.).

Anchored PCR cloning

A 700 bp cDNA clone containing the most 3'region of human DNA MTase cDNA was obtained by a modified anchored polymerase chain reaction (PCR) procedure (12,13). Briefly, 3 μg total cellular RNA from TT cells was reverse transcribed (RT) using 5'-ATAGGAATTCC$(T)_{22}$-3' as the primer to target the poly(A) tail (13). Two μl of the 30 μl RT reaction was then PCR amplified using the above primer as the antisense primer and a human DNA MTase-specific sequence derived from clone hmt-2.5 (Figure 1), 5'-GTTTGTGAGCAACATAAC-3', as the sense primer. The products were then reamplified using the same antisense primer and a second DNA MTase-specific sense primer, 5'-CAGTTCAACACCCTCATC-3', which occurs downstream of the original sense primer. For PCR, the final reaction conditions were 0.2μM each primer, 1.5mM each dNTP, 67mM Tris pH8.8, 4.7mM $MgCl_2$, 16.6mM NH_4SO_4, 10mM β-mercaptoethanol, 6.7μM EDTA, 10% DMSO and 1 unit *Taq* DNA polymerase (Amplitaq, Cetus) with a temperature profile of 95°C for 0.5 min, 52°C for 2 min, and 70°C for 2 min. The 700 bp DNA MTase product was purified from a 4% polyacrylamide gel and blunt-end ligated into pBluescript SK+ plasmid. Four independent PCR clones were fully sequenced to eliminate any PCR introduced sequence errors.

DNA sequencing and analyses

A series of overlapping cDNA clones were obtained and sequenced using the Sequenase 2.0 kit (United States Biochemical Corporation, Cleveland, OH) according to the manufacturer's protocol. For all regions, final sequence information was derived from sequencing of both strands of at least one cDNA clone in addition to sequencing of one strand of independent overlapping cDNA clones. Comparisons for nucleotide and amino acid sequences were performed with the DNASTAR (DNASTAR Inc., Madison, WI) and the DNASIS/PROSIS (Hitachi America, Ltd., San Bruno, CA, for Pharmacia-LKB Biotechnology) programs.

Northern blot analysis

A Northern blot containing 2 μg of poly (A)+ RNA from various human tissues was purchased from Clonetech Inc. (Palo Alto, CA). Hybridizations were performed as previously described (14) using a 2.5 kb human DNA MTase cDNA clone labeled with ^{32}P by the random hexamer priming method (15). Following removal of the probe, the blot was rehybridized to a ^{32}P-labeled human β-actin cDNA probe (a gift from Dr. Donald Cleveland).

Chromosomal mapping

We have previously localized the human DNA MTase gene to chromosome 19 using analyses of a panel of hamster-human somatic cell hybrids (9). In the current study, filters carrying approximately 11,500 chromosome 19 cosmids (16) in high density arrays were hybridized with random-hexamer primed radiolabelled probes derived from cDNA clones TT6 and hmt-2.5 (Figure 1). Hybridizations were performed overnight at 65° in 0.6 M sodium chloride, 10% dextran sulphate, 2% sodium lauryl sulfate (SDS), 50 mM Tris 7.6, 10 mM EDTA, 0.1% sodium pyrophosphate and 50 μg/ml salmon sperm DNA and washed at high stringency with 0.25×SSC, 1% SDS, at 65° for one hour. Hybridization signals were quantitated by storage phosphor screen autoradiography (Molecular Dynamics, Sunnyvale, CA). The restriction enzyme fingerprints of gene-positive cosmids were compared to those of over 8000 other cosmids as a part of an ongoing effort to assemble cosmid contigs for chromosome 19 (reconstruction no. 14403) (17). Complete EcoR1 digests of cosmids in the DNA MTase-positive contig were analyzed to confirm contig assembly.

Biotinylated DNA from the identified cosmids was hybridized to metaphase chromosome spreads in the presence of unlabeled human genomic DNA as described elsewhere (18, Trask *et al.*, submitted). After hybridization, washing, and avidin-FITC (fluorescein isothiocyanate) treatment, chromosomes were incubated in 4',6-diamidino-2-phenylindole (DAPI) and actinomycin to produce a QFH-like banding pattern (Trask *et al.*, submitted). A minimum of 10 different metaphase spreads were analyzed microscopically for each mapped cosmid.

RESULTS AND DISCUSSION

Cloning of the human DNA MTase cDNA

A 1.3 kb cDNA fragment of human DNA MTase (9), previously cloned from a human SCLC cell line, NCI-H249, was used to screen the λgt11 cDNA library from poly (A)+ selected TT cell RNA. Three positive clones hmt-2.5, hmt-1.0, and hmt-2.0 (Figure 1) were obtained in the first screening. Subsequent use of clone hmt-2.5 to rescreen the λgt11 library yielded an additional series of positive clones (Figure 1). After several attempts to capture the most 5' end of the DNA MTase cDNA from the λgt11 cDNA library, a λZapII cDNA library from TT cells was also screened, and clones TT6, TT8, and TT7

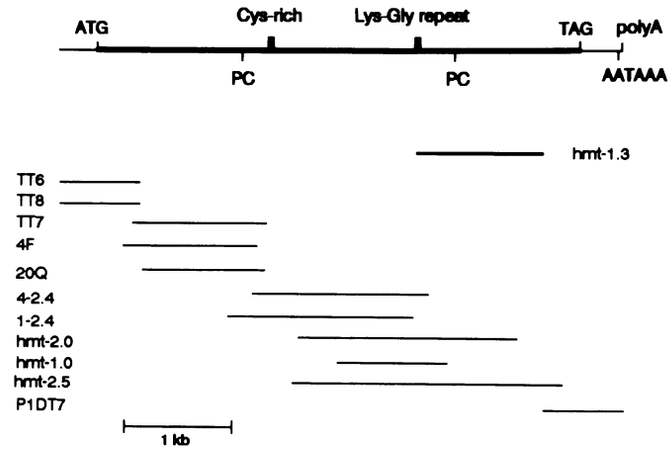

Figure 1. Diagram of the human DNA methyltransferase cDNA (*top*). The heavy line represents the region correlating with the translated portion of the mouse DNA MTase (8) including the translation start (**ATG**) and stop (**TAG**) sites. The thin line represents the untranslated region with the polyadenylation signal (**AATAAA**) and **poly(A)** tail indicated at the 3' end. Candidate functional regions for the protein, as discussed in the text, are marked as follows: **PC**, proline-cysteine dipeptide; **Cys-rich**, a cysteine-rich potential metal-binding domain; **Lys-Gly repeat**, a series of lysine-glycine repeats thought to be the junction between a carboxy region similar in structure to prokaryotic m^5C-methylases and a amino terminal domain specific to eukaryotic DNA MTase. The cDNA clones isolated for the present study (*bottom*). Clone hmt-1.3 was isolated by a RT-PCR procedure (9) and used for initial library screening. All other clones were obtained by screening the cDNA libraries, except clone P1DT7 which was isolated by an anchored PCR procedure.

(Figure 1) were obtained. The overlapping clones from the above screens yielded all but the most 3′ end of the DNA MTase cDNA. We cloned this region, including the poly(A) tail, using an anchored RT-PCR procedure (12,13) described in the Methods. A 700bp PCR product was obtained and subcloned (clone P1DT7 in Figure 1).

The overall length of the assembled cDNA sequence for human DNA MTase from the above clones is 5194 bp, including 26 nucleotides of poly(A) tail (Accession No. X63692). The nucleotide sequence contains a long open reading frame (ORF) that starts at nucleotide position 1 and extends to a stop codon (TAG) at position 4845. The first methionine residue in this ORF is present at nucleotide position 361. The sequence following this methionine codon potentially encodes a protein of 1495 amino acid residues with a predicted molecular weight of approximately 169 kDa. This

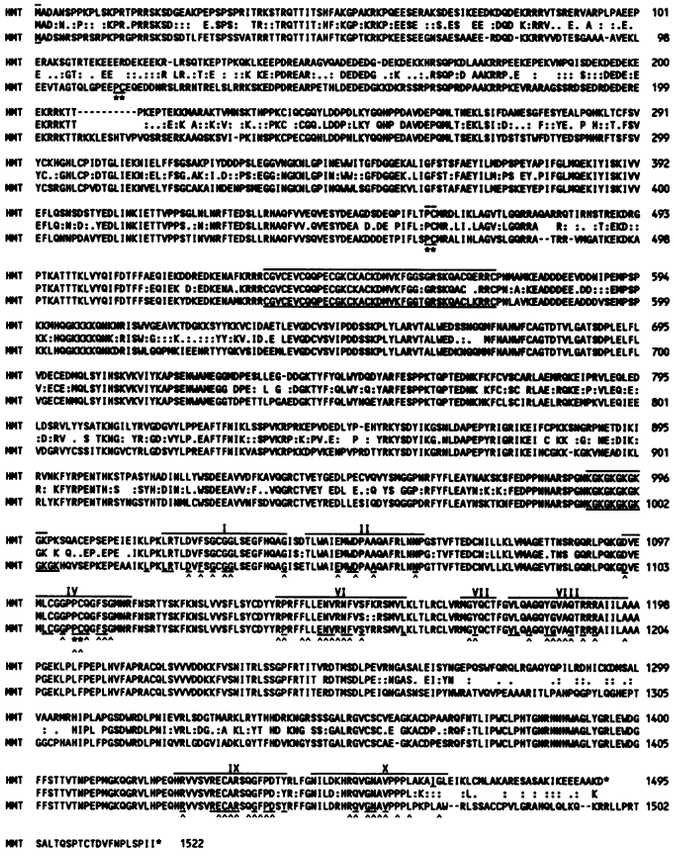

Figure 2. Comparison of the amino acid sequences of human and murine DNA MTase. HMT and MMT on the left column indicate the human and murine DNA MTases, respectively. The number on the right column represents the amino acid residue at the end of each sequence. The alignment of the two sequences was obtained by the program AALIGN in DNASTAR using the Lipman PAM250 matrix. Identical amino acids are printed between the two sequences with gaps indicated by dashes. Amino acid differences that represent conservative changes are shown with a colon (:), and non-conservative changes are shown with a blank or a period (.). Underlines regions represent the following areas discussed in the text and in Figure 1: first methionine; PC dipeptides (also marked by '**') at residues 458 and 1104 of the human protein; cysteine-rich region at residues 532−570 of the human sequence; lysine-glycine repeat area at residues 988−998 of the human peptide sequence. Underlined regions designated by Roman numerals represent sequence blocks which are highly conserved in the prokaryotic m^{5}C-methylases (6). Amino acids in these regions that are identical to the *Hha I* methylase are underlined, and those that are homologous to highly conserved amino acids (6) among all the prokaryotic enzymes are indicated by (:).

fits well with the reported size range of 150 to 190 kDa for DNA MTase polypeptides isolated from human cells (19).

Sequence comparison with the mouse

Both the human (9) and the mouse DNA MTase (8) genes encode a 5.2 kb transcript in all cell types examined to date. The nucleotide sequence of the human DNA MTase cDNA coding for this transcript was compared to the full length sequence of the mouse cDNA (8). Eighty percent homology exists between the nucleotide sequences of the two cDNA's, and 74% identity is noted at the amino acid level (Figure 2).

The mouse and the human DNA MTases show high homology for numerous candidate functional regions (Figure 2). First, the candidate translation start codon (ATG), at nucleotide position 228 in the mouse cDNA, is matched at nucleotide position 361 in the human cDNA. However, the sequence of the human cDNA is open for an additional 120 amino acids upstream from this site (nucleotides 1−360), raising the possibility of an additional or alternate translation start site for the human DNA MTase.

Second, the amino acid sequences in the carboxy terminal domain of both the human and the mouse DNA MTases share remarkable similarity to prokaryotic m^{5}C-methylases. Ten conserved sequence motifs separated by variable regions have been identified among all prokaryotic m^{5}C-methylases sequenced to date (6). Strikingly, eight of these ten conserved motifs, motifs I, II, IV, VI, VII, VIII, IX and X (Figure 2), are present in highly homologous regions of the human and the mouse DNA MTases. These motifs are arranged in the same order in the mammalian DNA MTases as in the prokaryotic m^{5}C-methylases. The invariant amino acid residues (**G-PC---S**) in motif IV, which have been conserved in all of the prokaryotic m^{5}C-methylases sequenced, are also present in the mammalian DNA MTases (residues 1102−1109 in the human sequence, Figure 2). The proline-cysteine (PC) dipeptide in this sequence

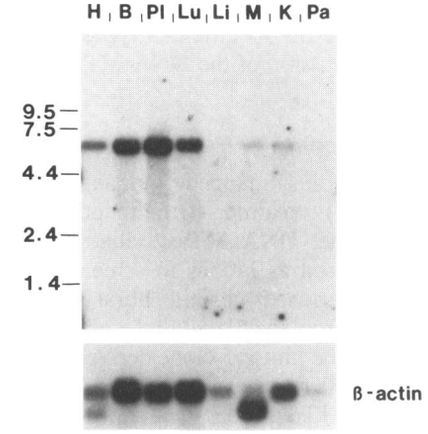

Figure 3. Expression of human DNA MTase mRNA. A northern blot (*top*) containing 2 μg poly (A)+ RNA from human heart (**H**), brain (**B**), placenta (**Pl**), lung (**Lu**), liver (**Li**), muscle (**M**), kidney (**K**) and pancreas (**Pa**) was hybridized to the ^{32}P-labeled DNA MTase clone hmt-2.5 (see Fig. 1). The migration of RNA size markers (in kb) is given along the left edge. The same blot was re-hybridized to a human β-actin cDNA probe (*bottom*). Densitometric analysis indicate the following DNA MTase to actin signal intensity ratios: heart, 0.7; brain, 0.7; placenta, 1.1; lung, 0.5; liver, 0.1; muscle, 0.1; kidney, 0.2; and pancreas, 0.2. The different sized bands for the β-actin represent known differences for transcript sizes between various tissues.

constitutes the functional catalytic domain of prokaryotic m^5C-methylases (20,21) and may play a similar role in the mammalian DNA MTases.

While the carboxy terminal one-third of the human and murine DNA MTases resembles the prokaryotic cytosine MTases, the amino terminal two-thirds of the mammalian proteins is not present in the prokaryotic enzymes. It has been proposed for the murine DNA MTase, that during evolution, this amino terminal region may have been added by gene fusion (1,8). This amino terminal domain is separated from the conserved carboxy terminal domain by a lysine-glycine repeat (Lys-Gly repeat) at residues 988–998 of the human sequence (Figure 2). This Lys-Gly repeat has been proposed to be a junction site of the putative gene fusion (8). Interestingly, the amino acid identity between mouse and human DNA MTases in the amino terminal region (70%) is less than that in the carboxy terminal one-third (83%). Nevertheless, the human and mouse proteins are highly homologous for potentially functional regions in the amino terminal domain including a cysteine-rich region (residues 532–570 in the human sequence, Figure 2), which represents a potential metal-binding domain (8).

Expression of human DNA MTase

A single 5.2 kb DNA MTase transcript was detected in eight different human tissues examined (Figure 3). While none of the tissues examined express DNA MTase in high enough levels to be detected in total RNA (data not shown), significant expression could be detected in 2 μg poly (A)+ RNA. Densitometric analysis of northern blots suggests that placenta, heart, brain and lung contain the highest levels of DNA MTase mRNA (DNA MTase to actin signal ratios of 0.5–1.1), while liver, kidney, pancreas and muscle contain significantly lower levels of DNA MTase to actin ratios of (0.1–0.2). Given that the expression of the DNA MTase is thought to be linked to active cell proliferation (22–24), it is surprising that tissues that are considered to have low proliferative potential, such as heart and brain, actually express relatively high amounts of DNA MTase mRNA. It will be of interest to determine the specific cell types responsible for the relatively high expression of the DNA MTase in these tissues.

Chromosomal mapping

Our previous studies mapped human DNA MTase to chromosome 19 (9). In order to more precisely localize the position of the gene, DNA MTase clones hmt-2.5 and TT6 (Figure. 1) were used as probes to screen a human genomic chromosome 19-enriched cosmid library. The TT6 clone (Figure 1) derived from the 5′ end of the cDNA identified seven positive cosmids. The hmt-2.5 clone derived from the 3′ end of the cDNA identified five positive cosmids, including three in common with the TT6 probe. Overlap among seven of the cosmids was detected by EcoR1 automated fingerprint analysis (25) and/or by analysis of EcoR1 digests (data not shown). The genomic distance spanned by the DNA MTase-positive cosmids indicate that the human DNA MTase gene is at least 93 kb in size. Fluorescence *in situ* hybridization was used to localize five DNA MTase-positive cosmids from this contig to cytogenetic bands on normal metaphase spreads. Four mapped to 19p13.2, and a fifth to the border between 19p13.2 and p13.3 (data not shown).

CONCLUSION

In this report, we show that the coding regions of the human and mouse DNA MTase genes have high structural conservation, particularly in a 3′ region which is structured similarly to prokaryotic DNA methyltransferases. Since the levels of DNA methylation differ dramatically among eukaryotes (1,26), becoming more abundant as a function of increasing genomic complexity, it will be of interest to compare the structure of the DNA MTase gene for these two species to those for other eukaryotes. Determining when the 5′ portion of the murine and human genes arose in evolution may aid in our understanding of the role of DNA MTase in the regulation of DNA methylation in different species.

There is increasing evidence for altered DNA methylation patterns and increased levels of DNA MTase expression in human tumors (reviewed in 4,10). Isolation of a full-length cDNA for human DNA MTase will allow us to explore the possibility that abnormalities in the structure or regulation of the gene contribute to this process.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Dr. Timothy Bestor for providing a corrected version of the murine DNA methyltransferase sequence with regard to position of the proline-cysteine dipeptide in the 3′ region. We thank Dr. Robert Casero for his assistance with the anchored PCR procedure; Ms. Kathleen Wieman, Ms. Anne Bergmann, Ms. Lorie Devlin, and Ms. Kim Lieuallen for their technical assistance; and Ms. Tammy Hess and Ms. Sandra Lund for their secretarial support. This work was funded in part by grants R01 CA43318 from the National Cancer Institute, 1987AR2 from The Council for Tobacco Research, and support from the Clayton Foundation.

REFERENCES

1. Bestor,T.H. (1990) *Philos. Trans. R. Soc. Lond. Biol. Sci.*, **326**, 179–187.
2. Riggs,A.D. (1990) *Philos. Trans. R. Soc. Lond. Biol. Sci.*, **326**, 285–297.
3. Tazi,J. and Bird,A. (1990) *Cell*, **60**, 909–920.
4. Baylin,S.B., Makos,M., Wu,J., Yen,R.-W.C., de Bustros,A., Vertino,P. and Nelkin,B.D. (1991) *Cancer Cells*, **3**, 383–390.
5. Adams,R.L.P. (1990) *Biochem. J.*, **265**, 309–320.
6. Posfai,J., Bhagwat,A.S., Posfai,G. and Roberts,R.J. (1989) *Nucleic Acids Res.*, **17**, 2421–2435.
7. Wilson,G.G. (1991) *Nucleic Acids Res.*, **19**, 2539–2566.
8. Bestor,T., Laudano,A., Mattaliano,R. and Ingram,V. (1988) *J. Mol. Biol.*, **203**, 971–983.
9. El-Deiry,W.S., Nelkin,B.D., Celano,P., Yen,R.-W.C., Falco,J.P., Hamilton,S.R. and Baylin,S.B. (1991) *Proc. Natl. Acad. Sci.*, **88**, 3470–3474.
10. Jones,P.A. and Buckley,J.D. (1990) *Adv. Cancer Res.*, **54**, 1–23.
11. Ausubel,F.M., Brent,R., Kingston,R.E., Moore,D.D., Seidman,J.G., Smith,J.A. and Struhl,K. (eds.) (1990) *Current Protocols in Molecular Biology*. Greene Publishing and Wiley Interscience, New York.
12. Frohman,M.A., Dush,M.K. and Martin,G.R. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 8998–9002.
13. Xiao,L., Celano,P., Mank,A.R., Pegg,A.E. and Casero,R.A. (1991) *Biochem. Biophys. Res. Commun.*, **179**, 407–415.
14. Mabry,M., Nakagawa,T., Baylin,S., Pentengill,O., Sorrenson,G. and Nelkin,B. (1989) *J. Clin. Invest*, **84**, 194–199.
15. Feinberg,A.P. and Vogelstein,B. (1983) *Nature*, **301**, 89–92.
16. de Jong,P.J., Yokabata,K., Chen,C., Lohman,F., Pederson,L., McNinch,J. and van Dilla,M.A. (1989) *Cytogenet. Cell Genet.*, **51**, 985.

17. Carrano,A.V., Lamerdin,J., Ashworth,L.K., Watkins,B., Branscomb,E., Slezak,T., de Jong,P.J., Keith,D., Raff,M., McBride,L., Meister,S. and Kronick,M. (1989) *Genomics*, **4**, 129–136.
18. Trask,B.J. (1990) In: Functional Organization of the Nucleus—A Laboratory Guide (*Hamkalo,B.A. and Elgin,S.C.R., eds.) Meth. Cell Biol.*, **35**, 3–35.
19. Pfeifer,G.P. and Drahovsky,D. (1986) *Biochem. Biophys. Acta.*, **868**, 238–242.
20. Wu,J.C. and Santi,D.V. (1987) *J. Biol. Chem.*, **262**, 4778–4786.
21. Wilke,K., Rauhut,E., Noyer-Weidner,M., Lauster,R., Pawlek,B., Behrens,B. and Trautner,T.A. (1988) *EMBO J.*, **7**, 2601–2609.
22. Szyf,M., Bozovic,V. and Lanigawa,G. (1987) *J. Biol. Chem.*, **266**, 10027–10030.
23. Szyf,M., Kaplan,F., Mann,V., Giloh,H., Kedar,H. and Razin,A. (1985) *J. Biol. Chem.*, **260**, 8653–8656.
24. Papadopoulos,T., Pfeifer,G.P., Hoppe,F., Drahovsky,D. and Muller-Hermelink,H.-K. (1989) *Virchows Arch. B. Cell Pathol.*, **56**, 371–375.
25. Carrano,A.V.,Branscomb,E.W., de Jong,P.J. and van Dilla,M.A. (1990) Exptl. Med., **8**, 29–35.
26. Selker,E.U. (1990) Trends Biol. Sci., **15**, 103–107.