[Epigenetics 2:3, 151-154; July/August/September 2007]; ©2007 Landes Bioscience

Research Paper

Distribution of CWG and CCWGG in the Human Genome

Brian Watson[†,*]
Kristofer Munson
Jarrod Clark
Taras Shevchuk[‡]
Steven S. Smith

City of Hope National Medical Center and Beckman Research Institute; Duarte, California USA

[†]Current Address: Deptartment of Bioengineering; University of Utah; Salt Lake City, Utah USA

[‡]Current Address: Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry; Russian Academy of Sciences; Moscow Region, Russia

*Correspondence to: Brian Watson; University of Utah Warnock Engineering Building; 72 South Central Campus Dr.; Rm. 2646; Salt Lake City, Utah 84112 USA; Email: Brian.Watson@utah.edu

Original manuscript submitted: 07/17/07
Manuscript accepted: 07/17/07

Previously pulished online as an *Epigenetics* E-publication:
http://www.landesbioscience.com/journals/epigenetics/article

ABSTRACT

Expression of the bacterial CG methyltransferase M•*Hha*I in mammalian cells appears to generate significant biological effects, while biological effects of the expression of the non-CG methyltransferase M•*Eco*RII in human cells have not been detected. The association of cytosine methylation with the CG site in mammals is also associated with clustering of CG sites near 5' control regions (CG-islands) of human genes. Moreover spontaneous deamination of 5-methylcytosine at these sites is thought to lead to the well known deficiency of CG sites in genomes where endogenous CG methyltransferases are expressed. Since these associations are generally taken to imply a biological function for the CG dinucleotide that is associated with its selective methylation by endogenous DNA methylation systems, we have asked whether or not CWG or CCWGG sites are clustered in regions flanking human genes and whether or not an overall deficiency of CWG or CCWGG occurs in the human genome. Using build 36.1, of the human genome, we inspected the regions flanking the 28,501 well known gene loci in the human genome. Our analysis confirmed the expected clustering of CG sites near the 5' region of known genes and open reading frames. In contrast to the CG site, neither the CWG site nor the CCWGG site recognized by the bacterial methyltransferase M•*Eco*RII were clustered in any particular region near known genes and open reading frames. Moreover, neither the CCWGG nor the CWG site was depleted in the human genome, again in sharp contrast to the known genomic deficiency of CpG sites. Our findings suggest that in contrast to CG site recognition, human cytosine methyltransferases recognize CWG and CCWGG only at very low frequency if at all.

INTRODUCTION

Most of the methylation that occurs in the human genome does not occur at the CG dinucleotide.[1] Together, non CG methylation in the form of CC, CA and CT nearest neighbors accounts for the majority of the 5-methylcytosines in the human genome. In spite of the prevalence of this form of DNA methylation, its presence has not been directly associated with a biological function. This stands in stark contrast to methylation at CG sites which is correlated with a variety of clastic events including viral integration, genetic recombination, gene expansion, as well as transcriptional silencing.[2-9] Consistent with this contrast are the findings[10] that the expression of the bacterial *G*^{M}*CGC* methyltransferase M•*Hha*I resulted in increased numbers of soft agar foci, formed by Mouse 3T3 cells, while the expression of the bacterial *C*^{M}*CWGG* methyltransferase M•*Eco*RII had no detectable biological effect[11] on human HK293 cells.

Since the correlation of CG sites with biological function is reflected in the genomic clustering and under representation of these sites,[12] we have asked whether or not clustering or unusual levels of representation are associated with the CWG (W stands for weak bond, A or T) sites present in human DNA. Using build 36.1 of the human genome sequence, we found that the CWG site and the CCWGG site were evenly distributed in the vicinity of genes and open reading frames, in stark contrast to CG sites which are much more prevalent at the 5' ends of genes where they are often found in the high CpG class of known genes.[13]

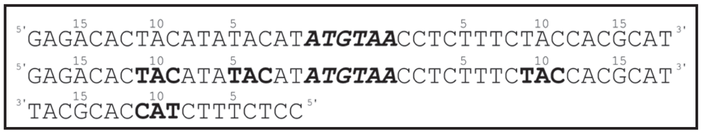
15 10 5 5 10 15
5' GAGACACTACATATACATATGTAACCTCTTTCTACCACGCAT 3'
15 10 5 5 10 15
5' GAGACACTACATATACATATGTAACCTCTTTCTACCACGCAT 3'
15 10 5
3' TACGCACCATCTTTCTCC 5'

Figure 1. Counting algorithm. In the sequence above notice in the first line the gene is given in italics with the start codon (ATG) and stop codon (TAA). In the flanking sequence above for example the program scanned for the three nucleotide sequence TAC given in bold. If the program setpoint was based off the nucleotide to the left, the two positions upstream for the TAC sequence are 11 and 5, but downstream it is 9. Looking at the sequence though, the distance of the TAC upstream at 11 and the TAC downstream at position 9 are actually the same from the gene. By reversing the order of the sequence (the downstream sequence is now 3, to 5, and reversing the sequence sought (CAT in this example), as in the third line, this method yields the correct equal correlative result.

MATERIALS AND METHODS

Materials. All computations were performed on standard desktop PCs running the Microsoft Windows operating system. All programming was performed in Python, PC version 2.4.1 (www.python.org). A sliding-window program was used to perform genomic analyses. The source code for the programs used in this report is provided in Supplemental Information. The list of known genes used[14] was taken from UCSC's Genome browser, genome.ucsc.edu. The genome was taken from NCBI build 36.1 of the human genome: ftp.ncbi.nih.gov/genomes/H_sapiens/Assembled_chromosomes/ The list of miRNAs were obtained from the Sanger miRNA database microrna.sanger.ac.uk/ (miRBase version 8.2 HSA.GFF).[15,16] A clustalW alignment was performed on the miRNA upstream set using EBI's clustalW alignment. www.ebi.ac.uk/clustalw/.

Assembly of the flanking sequence data set. A smaller subset of miRNA genes from Sanger was inspected to verify the flanking sequence assembly algorithm. T's were converted to U's to let the computer analyze its accuracy since the miRNA data set used U's. Once the algorithm was verified on the smaller data set, the chromosomal gene set was assembled.

In order to collect the regions flanking each known gene from the UCSC gene list, the program deleted the gene sequence itself (i.e., the region between the start codon and the stop codon in each gene and retained the region 20,000 nucleotides upstream to 20,000 nucleotides downstream from these sites. As per our goal, this process was designed to effectively fuse the upstream and downstream flanking sequences for each known gene. 2 x 10^4 nucleotides (nts) was taken to be the region of interest to compare upstream and downstream since CG Islands are generally between 500–1,500 nts upstream and positioning of cis-acting control elements beyond 20,000 nucleotides was considered unlikely. Moreover, the UCSC genome servers (genome.ucsc.edu/cgi-bin/hgTables) currently only allow for the capture of a maximum of 2 x 10^3 nts. This is because some genes are very close to the ends of chromosomes.

Every gene in the known list was correctly included except the mtDNAs four genes since our focus is on chromosomal methylation. The final file contained identifiers (e.g., name, location, strand) for all 39,284 genes in the UCSC gene list, followed by the sequence of flanking nucleotides. Some genes were within the window of the 20,000 basepairs from the end of the chromosome, thus the algorithm added N,s until all genes upstream and downstream flanking sequences were the same length.

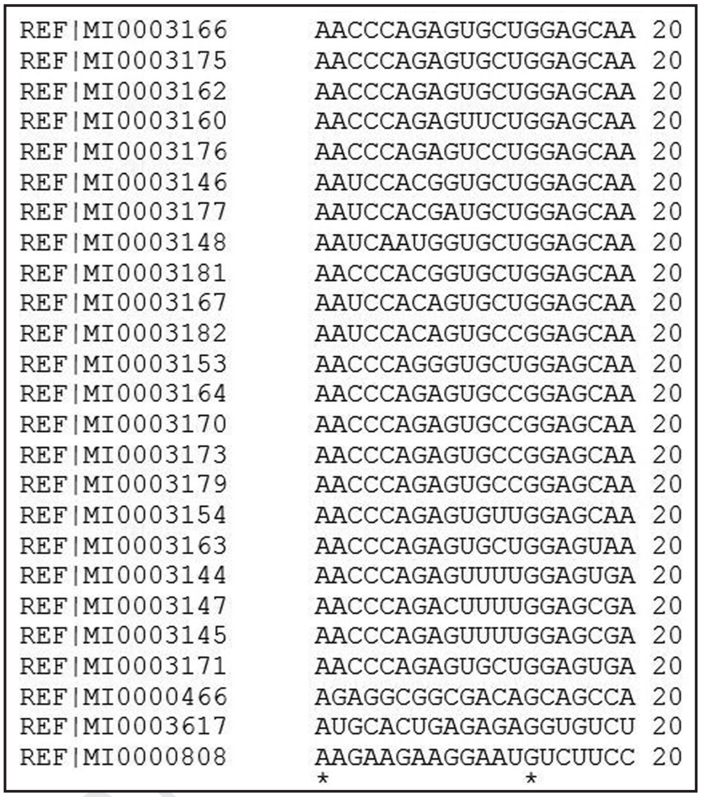
```
REF|MI0003166     AACCCAGAGUGCUGGAGCAA 20
REF|MI0003175     AACCCAGAGUGCUGGAGCAA 20
REF|MI0003162     AACCCAGAGUGCUGGAGCAA 20
REF|MI0003160     AACCCAGAGUUCUGGAGCAA 20
REF|MI0003176     AACCCAGAGUCCUGGAGCAA 20
REF|MI0003146     AAUCCACGGUGCUGGAGCAA 20
REF|MI0003177     AAUCCACGAUGCUGGAGCAA 20
REF|MI0003148     AAUCAAUGGUGCUGGAGCAA 20
REF|MI0003181     AACCCACGGUGCUGGAGCAA 20
REF|MI0003167     AAUCCACAGUGCUGGAGCAA 20
REF|MI0003182     AAUCCACAGUGCCGGAGCAA 20
REF|MI0003153     AACCCAGGGUGCUGGAGCAA 20
REF|MI0003164     AACCCAGAGUGCCGGAGCAA 20
REF|MI0003170     AACCCAGAGUGCCGGAGCAA 20
REF|MI0003173     AACCCAGAGUGCCGGAGCAA 20
REF|MI0003179     AACCCAGAGUGCCGGAGCAA 20
REF|MI0003154     AACCCAGAGUGUUGGAGCAA 20
REF|MI0003163     AACCCAGAGUGCUGGAGUAA 20
REF|MI0003144     AACCCAGAGUUUUGGAGUGA 20
REF|MI0003147     AACCCAGACUUUUGGAGCGA 20
REF|MI0003145     AACCCAGAGUUUUGGAGCGA 20
REF|MI0003171     AACCCAGAGUGCUGGAGUGA 20
REF|MI0000466     AGAGGCGGCGACAGCAGCCA 20
REF|MI0003617     AUGCACUGAGAGAGGUGUCU 20
REF|MI0000808     AAGAAGAAGGAAUGUCUUCC 20
                  *            *
```

Figure 2. Test set. A small subset of miRNAs were chosen based on the position of the Adenine and the Guanine nucleotides in front of the miRNA in the process of preparing the sliding window program. Then a clustalW was performed on the results. The accession numbers are to the left and the upstream sequence is found to the right. The "20" indicates the length of the sequence spliced. The miRNA sequence is found just to the right. Where right is the 3' and left is the 5'. The actual list of miRNAs was taken from Sanger. The Thymines were converted to Uracil in order to verify the location of the miRNA was correct, as the list of miRNAs contains U instead of T in their sequence.

Flanking sequence counts. miRNA. In making a sliding window program, the entire miRNA data set (HSA.GFF version 8.2) was first used, again due to the small size of the data set. The main purpose here was to make sure the sliding window was working effectively by looking for an upstream pattern. We thought that such a pattern may exist since a pattern previously was noted with *C. elegans* (CTCCGCCC).[17] The sliding window started from 200 nts upstream. In order to detect patterns in the upstream common sequence, (e.g., TATA box) counts of each nucleotide were recorded. Since the two consensus bases within 50 nucleotides were A at 20 nucleotides and G at seven nucleotides upstream, a list of some of the miRNAs that had that pattern upstream was prepared. A ClustalW alignment, at EBI, was run on a small cluster of the results of the miRNA 5' flanking subsequence. The pattern on chromosome 19 that emerged is depicted (Fig. 2). This test verified the effectiveness of the sliding window program.

Chromosomal gene frequencies. Counts of the subsequences GCGC, TATA, ATAT, CTG, CAG, CGCG, CG, CCTGG and CCAGG in each flanking sequence were determined using the program. Since DNA methyltransferases are thought to be sequence specific, we inspected the flanking sequences for these exact matches by using the programmed sliding window passed along each of the flanking sequences. The locations were categorized into fifty

Table 1 **Actual counts/predicted counts**

	CCWGG	GCGC	CG	CWG	TATA
Genome	**3.31372508**	**0.144782939**	**0.2356507**	**1.341659605**	**0.77622131**
10^4 bps/up	3.423061	0.296747	0.375492	1.346793	0.611728
350 bps/up	3.543773	0.618940	0.371113	1.358913	0.489993
300 bps/up	3.444138	0.671746	0.377446	1.347656	0.452849
250 bps/up	3.483417	0.729431	0.385302	1.349094	0.452595
200 bps/up	3.346418	0.818015	0.396256	1.360905	0.444326
200 bps/down	3.124633	0.137372	0.225111	1.343136	0.741477
250 bps/down	3.091102	0.132876	0.228525	1.337590	0.753943
300 bps/down	3.197923	0.135746	0.230770	1.349032	0.766282
350 bps/down	3.128944	0.118144	0.234954	1.352709	0.742494
10^4 bps/down	3.28223	0.129432	0.227323	1.348970	0.754833

When the actual count was divided by a predicted count, it is seen that the frequency of the CCWGG is about equally distributed both around genes and throughout the entire genome.

Table 2 **Observed human nucleotide frequencies**

Total nucleotides	Adenine	Cytosine	Guanine	Thymine
2,858,011,554	843,953,266	584,268,275	584,621,502	845,168,511
	%Adenine	%Cytosine	%Guanine	%Thymine
	0.295293861	0.20443174	0.204555332	0.295719067

The frequencies used to find the predicted counts were taken from the genomic counts from build 36.1 of the genome as listed here.

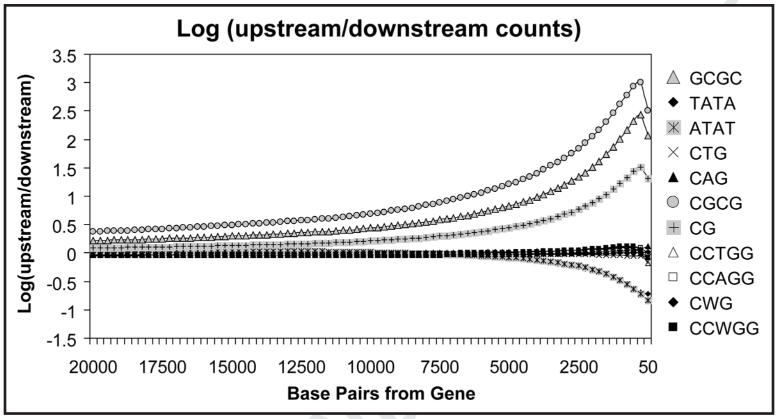

Figure 3. Upstream vs downstream frequency comparisons. The counts of each section at the same distance upstream were divided by the counts downstream. By taking the log (base 2) it makes it so if the ratio upstream is 2:1 then the value is 1, but the value is -1 if the ratio is 1:2. As noted on the graph all the sequences that contain CWG stay at a ratio of about zero meaning that the counts upstream divided by the counts downstream are about equal in distribution. The sequence CGCG is very frequent in a range up to about 2,300 nucleotides, which is consistent with evidences of CG Islands. In addition, there is a diminished frequency of the TATA sequence in upstream flanking regions which gives the dip in the graph (Table 1).

nucleotide subgroups. Thus a count was taken at 20,000–19,951 nucleotides, then from 19,950–19,901 nucleotides, and so on down to zero both upstream and downstream. A complete record of the subsequence counts recorded for each individually labeled gene was made at a1 distance of 1,000 nt to verify the results.

Since the gene location was based on the position of the first nucleotide of the start condon (ATG), the downstream sequence therefore was passed in reverse so there were correlative counts (see Fig. 1). For example, the sequence TCTTTACTAGA would be passed as is upstream, but in reverse it would be sent as AGATCATTTCT if it were downstream. The subsequence would be passed in reverse as well as the freqeuncy of each nucleotide upstream and thus CCAGG would be passed GGACC.

In the list of genes (all 39,284), generated by this method, many have the same sequences upstream and downstream, since they are splice variants with the start and stop codon in the same location. Repeated counting of these variants from the same region is expected to skew the results; thus an algorithm was designed to exclude the identical regions leaving 28,501 flanking regions out of 39,284. Many bases in this build of the human genome sequence are still ambiguous. Matches that included an unknown nucleotide (N) were discarded and assigned to a bin for later analysis.

Genome counts. In order to assess the possible existence of flanking sequence methylation frequency increases, for comparison, the counts of each nucleotide and the counts of each sequence throughout the entire genome were made on the build of the genome used (build 36.1). The frequency of each sub-sequence could be determined as well as the frequency of each nucleotide upstream and downstream. By dividing the count of each nucleotide by the total nucleotides a frequency of each nucleotide was obtained to be used for an expected subsequence frequency.

Next, by multiplying the genomic probability of each nucleotide in the subsequence by the number of possible places in the genome, an expected number can be calculated. (i.e., CpG expected frequency is the frequency of C * the frequency of G * length; (.20443174 *.204555332 * L) where L is the length of the inspected region in nt). Note here that the number of unambiguously sequenced bases in the human genome (2.858 x 10^9) is somewhat lower than the true size of the genome. Thus, by making this calculation, a comparison of the general genomic distribution of the subsequences and their distribution in the functional region can be made.

RESULTS

The miRNA 5' flanking subset alignment. We found it interesting that both a CpG and CpNpG were found in the pattern in the 7-9 nt region shown in Figure 2. Interestingly, none of the CAG sites in the consensus sequence were present as a UAG sequences in consensus in variants, while many of the CG sequences in the consensus showed UG variants suggesting that 5-methylcytosine occurs only at CG sites in this consensus sequence. This suggests the C^{m}G can deaminate to fix T in the DNA sequence of allelic variants. This consensus sequence is not found upstream from all miRNAs. While interesting, this result relies on a small subset of the miRNA genes and may not be true for all miRNA genes in general.

Upstream and downstream flanking comparisons. CWG sequences were not more frequent within the 3' or 5' regions near genes. In contrast, an increase of CG sites was observed in 5' regions as expected due to CG Islands. The graph in Figure 3 shows the log base 2 of the result of the upstream counts divided by the downstream counts. In this case all known gene flanking sequences in the human genome were inspected, and the result reflects average frequencies applicable to the entire human gene set.

Genome analysis. The total number of the occurrences of CWG, CGCG, CG and TATA in the human genome was 115,802,370, 1,652,129, 28,163,853 and 16,916,756 respectively. The ratios of the observed numbers to expected number calculated from base frequencies were 1.34166, 0.069475, 0.235651 and 0.776221 respectively. These values correspond to the limits expected as the scan moves farther away from gene control regions (Table 1). As can be seen from the table, CG and CWG are overrepresented in the region immediately upstream of known genes while the TATA sequence is underrepresented in this region. Conversely, the CG sequence is underrepresented in the downstream region while the CWG and CCWGG sequence are overrepresented downstream. The graphical representation of the data in Figure 3 shows that CG sites are clustered upstream and TATA sites are more frequent downstream, while CWG sites are uniformly overrepresented both upstream and downstream.

DISCUSSION

Both the TATA sequence and the CG sequence have a distinct depletion throughout the genome as previously noted.[12] They are both considered to be important factors in marking DNA control regions. The CWG sequence, on the other hand, showed the opposite pattern, specifically overrepresentation by about 30% genome-wide and in gene flanking sequences. This pattern is also opposite to that observed in plants where CWG methylation is more readily studied. Plant genomes demonstrate an overrepresentation of the CWG site in introns and transcribed regions[1,18] leading to the suggestion that CWG methylation in plants is more prevalent in intergenic regions.[1,18]

Further, since its frequency is not underrepresented throughout the genome or in putative control regions adjacent to genes it would appear that the methylation of the CWG sites may not have a specific biological function in human DNA that involves high levels of 5-methylcytosine. This possibility is supported by previous evidence,[11] although further experimentation will be required to answer this question. On the other hand the data does allow us to conclude that site specific systems apply methylation to CG sites in the human genome with high frequency while site specific methylation systems apply cytosine methylation to CWG and CCWGG only at very low frequency if at all. Finally, the significance of the clustalW alignment of the upstream sequence of the miRNAs[4] may merely be the result of a gene transposition. However, because the C → U transitions appear to have occurred at CpGs but not at CWGs or CCWCC's in about half of the sequences one suspects that CG methylation and not CWG or CCWGG methylation has driven the formation of variants. The miRNA's listed, except the last three, appeared in the same intergenic region. Although the alignment appears important, the same sequence isn't found throughout all miRNAs.

References

1. Hobolth A, Nielsen R, Wang Y, Wu F, Tanksley SD. CpG + CpNpG analysis of protein-coding sequences from tomato. Mol Biol Evol 2006; 23:1318-23.
2. Sneider TW. Hemimethylation of DNA: A Basis for genetic recombination? In: Usdin E, Borschardt RT, Creveling CR, eds. Transmethylation. New York: Elsevier North Holland, 1979:473-81.
3. Smith SS, Laayoun A, Lingeman RG, Baker DJ, Riley J. Hypermethylation of telomere-like foldbacks at codon 12 of the human *c-Ha-ras* gene and the trinucleotide repeat of the *FMR-1* gene of fragile X. J Mol Biol 1994; 243:143-51.
4. Oberle I, Rousseau F, Heitz D, Kretz C, Devys D, Hanauer A, Boue J, Bertheas MF, Mandel JL. Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. Science 1991; 252:1097-102.
5. Buryanov YI, Shevchuk TV. DNA methyltransferases and structural-functional specificity of eukaryotic DNA modification. Biochemistry (Mosc) 2005; 70:730-42.
6. Smith SS. Gilbert's conjecture: The search for DNA (cytosine-5) demethylases and the emergence of new functions for eukaryotic DNA (cytosine-5) methyltransferases. J Mol Biol 2000; 302:1-7.
7. Hoekstra MF, Malone RE. Expression of the *Escherichia coli* dam methylase in *Saccharomyces cerevisiae*: Effect of in vivo adenine methylation on genetic recombination and mutation. Mol Cell Biol 1985; 5:610-8.
8. Sandberg G, Schalling M. Effect of in vitro promoter methylation and CGG repeat expansion on FMR-1 expression. Nucleic Acids Res 1997; 25:2883-7.
9. Merlo A, Herman JG, Mao L, Lee DJ, Gabrielson E, Burger PC, Baylin SB, Sidransky D. 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nat Med 1995; 1:686-92.
10. Baylin SB, de Bustros AC, Ball DW, Nelkin BD. Pathobiology of the C-cells. Monogr Pathol 1993; 63-71.
11. Shevchuk T, Kretzner L, Munson K, Axume J, Clark J, Dyachenko OV, Caudill M, Buryanov Y, Smith SS. Transgene-induced CCWGG methylation does not alter CG methylation patterning in human kidney cells. Nucleic Acids Res 2005; 33:6124-36.
12. Ohno S. Grammatical analysis of DNA sequences provides a rationale for the regulatory control of an entire chromosome. Genet Res 1990; 56:115-20.
13. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA 2006; 103:1412-7.
14. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC known genes. Bioinformatics 2006; 22:1036-46.
15. Griffiths-Jones S. The microRNA registry. Nucleic Acids Res 2004; 32:D109-11.
16. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: MicroRNA sequences, targets and gene nomenclature. Nucleic Acids Res 2006; 34:D140-4.
17. Ohler U, Yekta S, Lim LP, Bartel DP, Burge CB. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA 2004; 10:1309-22.
18. Tran RK, Henikoff JG, Zilberman D, Ditt RF, Jacobsen SE, Henikoff S. DNA methylation profiling identifies CG methylation clusters in *Arabidopsis* genes. Curr Biol 2005; 15:154-9.