Current Biology, Vol. 15, 154–159, January 26, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.cub.2005.01.008

DNA Methylation Profiling Identifies CG Methylation Clusters in *Arabidopsis* Genes

Robert K. Tran,[1] Jorja G. Henikoff,[1] Daniel Zilberman,[1,2] Renata F. Ditt,[1] Steven E. Jacobsen,[3,4] and Steven Henikoff[1,2,*]
[1]Fred Hutchinson Cancer Research Center
1100 Fairview Avenue North
Seattle, Washington 98109
[2]Howard Hughes Medical Institute
[3]Department of Molecular, Cell and Developmental Biology
University of California, Los Angeles
Los Angeles, California 90095
[4]Molecular Biology Institute
University of California, Los Angeles
Los Angeles, California 90095

Summary

Cytosine DNA methylation in vertebrates is widespread, but methylation in plants is found almost exclusively at transposable elements and repetitive DNA [1]. Within regions of methylation, methylcytosines are typically found in CG, CNG, and asymmetric contexts. CG sites are maintained by a plant homolog of mammalian Dnmt1 acting on hemi-methylated DNA after replication. Methylation of CNG and asymmetric sites appears to be maintained at each cell cycle by other mechanisms. We report a new type of DNA methylation in *Arabidopsis*, dense CG methylation clusters found at scattered sites throughout the genome. These clusters lack non-CG methylation and are preferentially found in genes, although they are relatively deficient toward the 5′ end. CG methylation clusters are present in lines derived from different accessions and in mutants that eliminate de novo methylation, indicating that CG methylation clusters are stably maintained at specific sites. Because 5-methylcytosine is mutagenic, the appearance of CG methylation clusters over evolutionary time predicts a genome-wide deficiency of CG dinucleotides and an excess of C(A/T)G trinucleotides within transcribed regions. This is exactly what we find, implying that CG methylation clusters have contributed profoundly to plant gene evolution. We suggest that CG methylation clusters silence cryptic promoters that arise sporadically within transcription units.

Results and Discussion

Methylation Profiling Reveals CG Targets within Genes

We previously described a strategy for profiling methylation patterns with microarrays [2]. DNA samples are treated with a methylation-sensitive restriction endonuclease and are size fractionated by sucrose gradient centrifugation. The low-molecular-weight fraction is collected and labeled with either of two fluorescent dyes, such that two samples can be compared by standard microarray analysis. When one sample is from a mutant in which methylation is reduced, affected sites will be more frequently cleaved by the restriction endonuclease than the wild-type. If the result of cleavage is that the fragment being assayed sediments faster than the 2.5 kb cut-off used in the fractionation, then there will be a stronger signal for the mutant than for the wild-type. Conversely, if the mutant causes hypermethylation of a site, then the wild-type signal will be higher than the mutant signal. In this way, we can detect changes in methylation patterns from the ratio of the two dye signals. We scored only those that are statistically significant based on repeated measurements from different biological samples [3].

In our previous methylation profiling study [2], we used the Msp I restriction endonuclease, which cleaves CCGG and is blocked by CNG methylation. We had probed a set of 240 randomly chosen single-copy loci for differences between methylation mutants and wild-types. This revealed that transposons and other repeats are preferential targets for the CMT3 DNA methyltransferase. To identify other classes of sites, we have performed identical assays on several mutants defective in DNA methylation: *kyp/suv4h*, *ago4*, *drm1/2*, and *cmt3* [4–6]. Two methyl-sensitive restriction endonucleases were used: HpyCH4 IV, which cleaves ACGT and is blocked by CG methylation, and Hpa II, which cleaves CCGG and is blocked by both CG and CNG methylation. We probed an array that included 960 loci, of which 597 were randomly selected single-copy fragments of approximately 700 bp, including those from the previous study. The remaining 363 loci were chosen as likely methylation targets for quality control but were not used in the analysis.

Using HpyCH4 IV, we detected a large number of CG methylation changes in both directions (Table 1). Nearly all of these changes were in genes: of the 28 loci detected, 26 are in annotated genes. This frequency is highly significant when one considers that 60% of the total predicted HpyCH4 IV sites in the single-copy loci used for this analysis are in genes ($\chi^2 = 8.3$, $p < 0.005$). Such a gene bias was unexpected because nearly all of the described methylation in *Arabidopsis* is confined to transposon-derived sequences and repeats [1, 7].

Changes Occur at Clusters of CG Methylation

The surprising detection of changes in CG methylation led us to verify a subset of sites detected in our microarray analysis by bisulfite sequencing. We chose three representative positives: A locus showing CG hypermethylation (At2g14255 in Table 1), one showing CG hypermethylation in the HpyCH4 IV dataset (At4g36550 in Table 1), and one showing CG hypermethylation in the Hpa II dataset (At2g15270 in Table S2). New plants were grown and processed for DNA preparation and bisulfite treatment. Primers were designed to flank sites

*Correspondence: steveh@fhcrc.org

Table 1. Loci Scored as CG Methylation Targets in Mutants from the HpyCh4 IV Dataset

Gene ID[a]	*cmt3*	*kyp*	*ago4*	*drm1 drm2*	*drm1 drm2 cmt3*
At3g42630	−	−	−		−
At4g36550	−	−	−		−
At4g14240		−	−		
At3g10390	−				
At1g09910		−			
At1g34350			−		
Intergenic			−		
At1g76510	+	+	+		+
At2g14255	+	+	+		+
At2g16860	+		+		+
At2g40630	+		+		+
At1g73840	+				+
At4g27340		+	+		
At4g34820		+		+	
At4g25290	+				
At3g66658		+			
At527000		+			
At2g36850			+		
At5g28740					+
At2g44950					+
Intergenic					+
At5g10140					+
At5g18490					+
At3g04910				+	
At1g58250				+	
At2g19010				+	
At3g10420				+	
At4g39120				+	

Probe preparation, microarray construction, hybridization, and data processing were described in our previous methylation profiling study [2]. The array size was increased from 360 to 960 loci for a total of 597 randomly chosen single-copy loci and 363 selected control loci. See Table S1 for genomic location, TIGR designation, and further experimental details. Columns on the right indicate whether methylation increased (+) or decreased (−) in the designated mutant relative to the matched control.

[a] The location of a blocked site was determined to be one that would cause depletion of the fragment spanning the locus because of an increase in size to > approximately 2.5 kb. In cases of ambiguity, a choice was made based on the greatest degree of hybridization or fragment size increase.

on either side of the 700 bp locus and therefore were likely candidates for differential methylation. After amplification and cloning, individual bacterial colonies were chosen for sequencing. For each fragment, 19–20 sequences were obtained. This provided us with a percentage of methylation occupancy for each cytosine.

Our analysis revealed unexpected clusters of dense CG methylation in all lines, both mutant and wild-type. For example, At2g14255 showed methylation of ten CG residues within a 230 bp region that spans an exon and an intron, with a median density of 90% methylation (Figure 1). Similar methylation clusters were found for At4g36550 (Figure S1 in the Supplemental Data available with this article online) and At2g15670 (Figure S2). None of these loci exhibited detectable CNG or CNN methylation in wild-type or mutant lines.

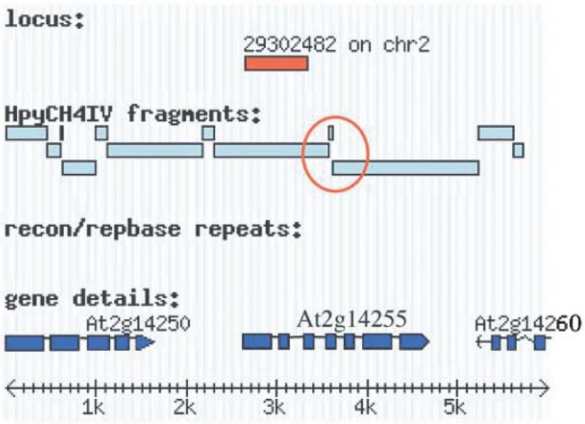
It is important to realize that our methylation profiling method does not detect methylation clusters per se because it is only sensitive to changes that occur at restriction sites when they are differentially methylated. For example, there are two closely spaced HpyCH4 IV sites in locus At2g14255 (Figure 1). One site was found to be methylated at a 5% level (1/20) in the wild-type; this level increased to 75%–90% in all mutant lines tested. This is consistent with detection of hypermethylation for *cmt3*, *kyp*, *ago4*, and *drm1/2 cmt3* by microarray analysis, but only if the nearby site, just 45 bp away, was also methylated in the same mutant lines. Indeed, we found 90%–100% methylation of this second HpyCH4 IV site in both wild-type and *cmt3*, *kyp*, *ago4*, and *drm1/2 cmt3* mutant lines. In *drm1/2*, this second HpyCH4 IV site was completely unmethylated (0%), which could account for the fact that At2g14255 was not detected as positive in the *drm1/2* line. Similar line-to-line variation that could explain detection in our assay was observed for loci At4g36550 (Figure S1) and At2g15270 (Figure S2).

These scattered methylation differences between lines are not attributable to differences between mutants and the wild-type because we found the same differences between the Ler line used as the standard for profiling and the Ler-derived *clk-st* parent of the *cmt3*, *kyp* and *ago4* mutant lines (e.g., line 2 in Figure 1). The cause of this difference between Ler and *clk-st* is unclear. Nevertheless, the consistent detection of CG hypo- or hypermethylation at sites of CG methylation clusters within genes suggests that the clusters are stably inherited even though methylation at any given site is not. This stability is evident from the fact that the same clusters are present in both Ler and WS (the parent line for *drm1/2*), two unrelated accessions that represent independent samples of the species. The fact that dense CG methylation clusters are found in the *drm1/2 cmt3* line that lacks all known de novo methylation indicates that the clusters have been stably maintained for multiple generations.

CG methylation clusters appear to be fundamentally different from targets of DNA methylation elsewhere in the *Arabidopsis* genome. The vast majority of DNA methylation consists of a mixture of methylated CG, CNG, and CNN sites at transposable elements and repeats [1, 7], whereas none of the clusters that we have identified shows any CNG or CNN methylation. Furthermore, the genic location of these CG methylation sites indicates that these clusters are unlikely to be sites of cryptic mobile elements that have evaded detection by either Repbase or RECON analyses. In fact, none of the HpyCH4 IV sites was in an element detected by Repbase analysis (Table 1), in contrast to the results of our previous study, in which all four of the randomly chosen loci detected as CMT3 targets were identified as mobile elements.

A Pronounced CG Dinucleotide Deficiency in Introns

A long-term consequence of cytosine DNA methylation is spontaneous deamination of 5-methylcytosine to thymine, which has been proposed to contribute to the well-established deficiency of CG dinucleotides relative to chance expectation [8, 9]. The localized nature of DNA methylation in plant genomes provides an opportunity



Sites / Lines	1	29	32*	44	77	129	149*	168	179	205	246
Ler	0	95	5	85	100	100	0	95	90	70	90
clk-st	0	90	75	85	95	80	80	95	95	80	95
cmt3	0	90	75	85	90	85	100	70	70	90	90
kyp	0	85	95	95	95	95	60	95	90	80	90
ago4	0	95	95	95	100	100	90	95	100	90	100
drm1/2	0	90	90	85	0	85	0	90	100	75	100
drm1/2/cmt3	0	95	90	95	95	80	55	90	100	70	90

Figure 1. Map and CG Methylation Occupancy of a Representive Gene-Rich Fragment Showing CG Hypermethylation for the HpyCH4 IV Dataset

(A) Schematic display of locus 2:6045801–6046476.

(B) Table shows bisulfite sequencing results (% methylation) for individual CG sites in the region circled in red in (A). The arrow marks the HpyCH4 IV restriction sites. Asterisks mark sites showing changes in methylation occupancy. Bisulfite treatment of DNA, cloning into a Topo TA vector (Invitrogen), and DNA sequencing were performed as described [18]. DNA was extracted by the CTAB method of Saghai-Maroof et al. [19], except that 2% CTAB, 1% βME, and 1% sodium bisulfite were used in the extraction buffer and the procedure was scaled up for 4 g plant tissue, which was ground in liquid nitrogen with a mortar and pestle. After ethanol precipitation, DNA samples were treated with DNase-free ribonuclease (Roche) and precipitated by addition of 3M sodium acetate and ethanol, then pelleted by centrifugation and air dried. Primers used for amplification were 5′-GTAGTGATTTTGAGAGAGGTGTATGTGGTGTAATGGTT-3′ and 5′-TAAATCAAACCTTTCAAAACAAACC TTACRACTATTCAATTA-3′.

to test whether there are CG or CNG deficiencies that correspond to the location of CG and/or CNG methylation. Before the present study, one might have expected that any CG deficiency resulting from deamination of 5-methylcytosine would be most pronounced in intergenic regions because genes were thought to be devoid of DNA methylation. However, if CG methylation clusters materialize from time to time at random genic locations and persist long enough for significant deamination of 5-methylcytosine to thymine, then we might find that genes also show a CG deficiency. In coding regions and 5′ and 3′ untranslated regions (UTRs), purifying selection would be expected to obscure such an effect, so that a mutation-driven CG deficiency should be most pronounced in introns. Furthermore, the absence of CNG methylation in genic clusters predicts a CNG deficiency in non-genic regions relative to genes.

To test these predictions, we tabulated all di- and trinucleotides in the *Arabidopsis* genome as classified by TAIR (http://www.arabidopsis.org) according to whether they are found in genes, coding regions, UTRs, introns, or non-genic regions. We found that relative to expectation, CG dinucleotides are indeed the most deficient in all classes, and TA is the next-most deficient (Figure 2A). Importantly, the CG deficiency is most pronounced in introns, such that a CG is only two-thirds as likely to be present in an intron as in a non-genic region. Evidence that these differences are attributable to CG DNA methylation and 5-methylcytosine deamination comes from a similar analysis of dinucleotides in *Drosophila*, an organism that lacks CG methylation but is similarly gene rich. We find that the dinucleotide log-odds profiles for *Drosophila* are much the same as those for *Arabidopsis* and include a prominent TA deficiency; however,

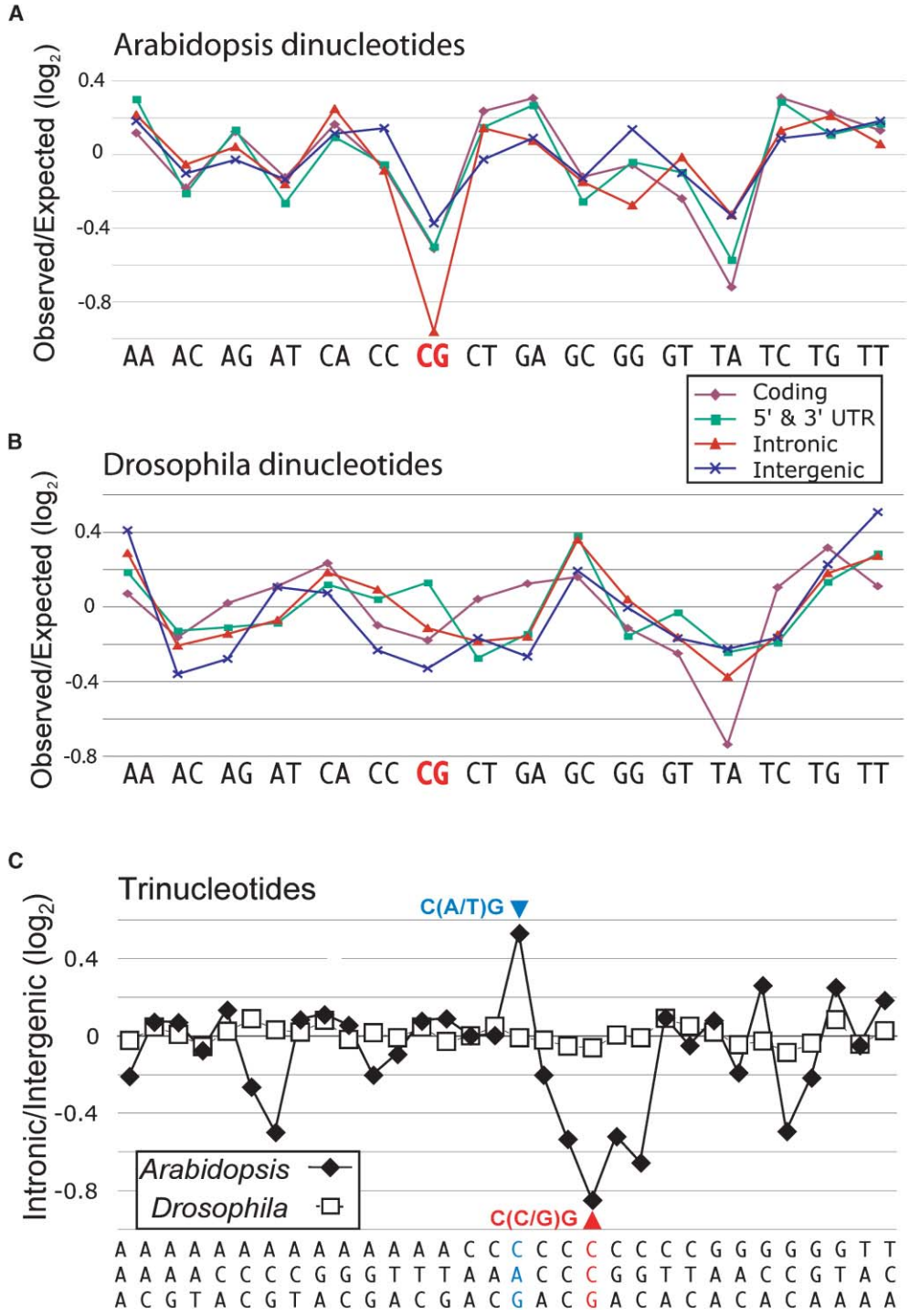


Figure 2. Di- and Tri-Nucleotide Log-Odds Profiles of Subsets of the *A. thaliana* and *D. melanogaster* Genomes

(A) Ratio of $\log_2$ observed and expected frequencies of dinucleotides for all annotated coding (purple), 5′ and 3′ intergenic (green), intronic (red), and intergenic (blue) segments in the entire *Arabidopsis* genome. Expected frequencies are calculated as the product of the frequencies of the individual mononucleotides.

(B) Same as (A) except for the *Drosophila* genome.

(C) Log_2 ratio of intronic and intergenic trinucleotide ratios (calculated as observed/expected) for the entire *Arabidopsis* and *Drosophila* genomes. Sequence files were downloaded from http://www.arabidopsis.org for *A. thaliana* and from http://www.flybase.net for *Drosophila melanogaster*. Among these were separate genome-wide FASTA-formatted sequence files for genes, annotated transcription units, coding sequences, introns, intergenic regions, and whole chromosomes. We calculated mono-, di-, and tri-nucleotide frequencies by using a C program running under Unix.

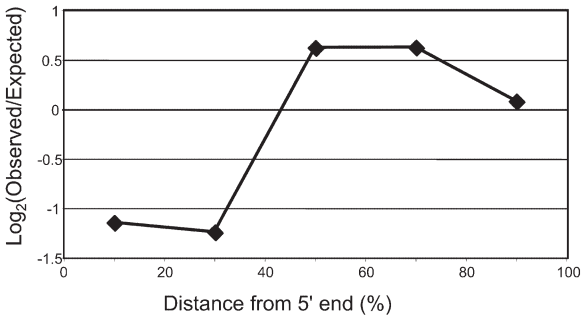


Figure 3. Distribution of CG Methylation Sites within Target Genes

Each gene target was divided into five equal bins based on length from TIGR annotations, and HpyCH4 IV and Hpa II target sites were mapped to bins. Expected values are proportional to the total number of HpyCH4 IV and Hpa II sites within each bin.

the corresponding *Drosophila* CG frequencies show only an average deviation from expectation (Figure 2B).

Analysis of *Arabidopsis* trinucleotides provides confirming evidence that the excess CG deficiency in genes results from CG-specific DNA methylation and 5-methylcytosine deamination. The CNG triplet that includes a CG dinucleotide (CCG/CGG) shows the lowest intronic/intergenic ratio of all trinucleotides, whereas the other CNG triplet (CTG/CAG) shows by far the highest intronic/intergenic ratio (Figure 2C). No such biases are seen in the *Drosophila* genome. Extreme CNG biases are expected from preferential deamination of 5-methylcytosine in genic CGs; such deamination would convert CCG/CGG triplets to CTG/CAG and cause these trinucleotides to diverge in abundance. In addition, preferential deamination of non-genic CNGs that are methylated by CMT3 would lower the denominator of the genic-to-intergenic ratio for CTG/CAG. This precise correspondence of observation to expectation provides compelling evidence that CG DNA methylation in *Arabidopsis* genes has been occurring genome-wide in transcribed regions over long evolutionary periods.

What Process Generates CG Methylation Clusters?

The genic nature of CG methylation targets and the evidence for their profound effect on the nucleotide composition of introns led us to ask whether they are located at random relative to 5′ and 3′ ends of genes. They are not (Figure 3). By mapping the inferred site of methylation blockage for each gene detected in the HpyCH4 IV and HpaII sets, we find that comparatively few fall into the 5′-most 40% of the length of the gene ($p < 0.005$). Methylation clusters near the 5′ end of *Arabidopsis* genes might be deleterious because they would interfere with normal promoter activity [10], which would account for their deficiency in our assay. However, methylation clusters within the body of genes should have no deleterious effect on transcriptional elongation because transcription through methylated regions is commonplace, for example in vertebrates [11].

How can an epigenetic feature of a gene be linked to transcriptional elongation? Nucleosomes present barriers to elongation; however, transiting RNA polymerases destabilize nucleosomes [12]. Evidence from yeast suggests that cryptic promoters that might be present within transcription units can fire in the wake of transiting polymerases, but these are normally repressed by the chromatin regulator Spt6p [13]. More generally, it is thought that prevention of unscheduled gene expression is important genome-wide [11]. For example, the nearly universal deficiency of TA dinucleotides (Figures 2A and 2B) is thought to be an adaptation for prevention of unscheduled expression [14]; TA has the lowest free energy of helix disruption of any dinucleotide [15, 16] and is especially frequent in both prokaryotic and eukaryotic regulatory elements, so that mutations in TA-poor regions would be relatively unlikely to give rise to cryptic regulatory elements. Cryptic promoters resulting from chance mutations within genes that initiate transcription toward the wild-type promoter would be most damaging to gene expression because they would cause collisions between polymerases (Figure 4A). In such cases, nascent antisense transcripts would be produced in close proximity to sense transcripts from the wild-type promoter, which in plants could lead to RNA-directed DNA methylation (Figure 4B) [10, 17]. Such homeostatic silencing of the cryptic promoter and loss of the RNA signal would result in disappearance of non-CG methylation, which depends on an active RNA signal. In contrast, CG methylation clusters can be maintained after DNA replication without an active signal (Figure 4C). This is confirmed by the lack of effect of mutations in genes responsible for de novo DNA methylation (Figures 1, S1, and S2). In this way, stable CG methylation clusters would appear sporadically within genes and persist until random 5meCG-to-TG mutations eventually disable the underlying promoter. Homeostatic silencing of cryptic promoters by DNA methylation would thus represent a general mechanism for preventing transcriptional noise in large eukaryotic genomes [11].

Acknowledgments

We thank members of our laboratories for helpful discussions, Terri Bryson for technical assistance, Samson Kwong and Tom Boyle for software, the Hutchinson Center microarray facility for slide preparation and processing, and Zhirong Bao for making the RECON *Arabidopsis* library available to us. R.K.T. was supported by a National Institutes of Health Chromosome Metabolism and Cancer Biology training grant (T32CA09657), J.G.H. by a grant from the National Science Foundation (DBI 0234960), and S.E.J. by National Institutes of Health grant GM60398.

Received: October 1, 2004
Revised: November 11, 2004
Accepted: November 12, 2004
Published: January 26, 2005

References

1. Bender, J. (2004). DNA Methylation and Epigenetics. Annu. Rev. Plant Physiol. Plant Mol. Biol. *55*, 41–68.
2. Tompa, R., McCallum, C.M., Delrow, J., Henikoff, J.G., van Steensel, B., and Henikoff, S. (2002). Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. Curr. Biol. *12*, 65–68.
3. Baldi, P., and Long, A.D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics *17*, 509–519.

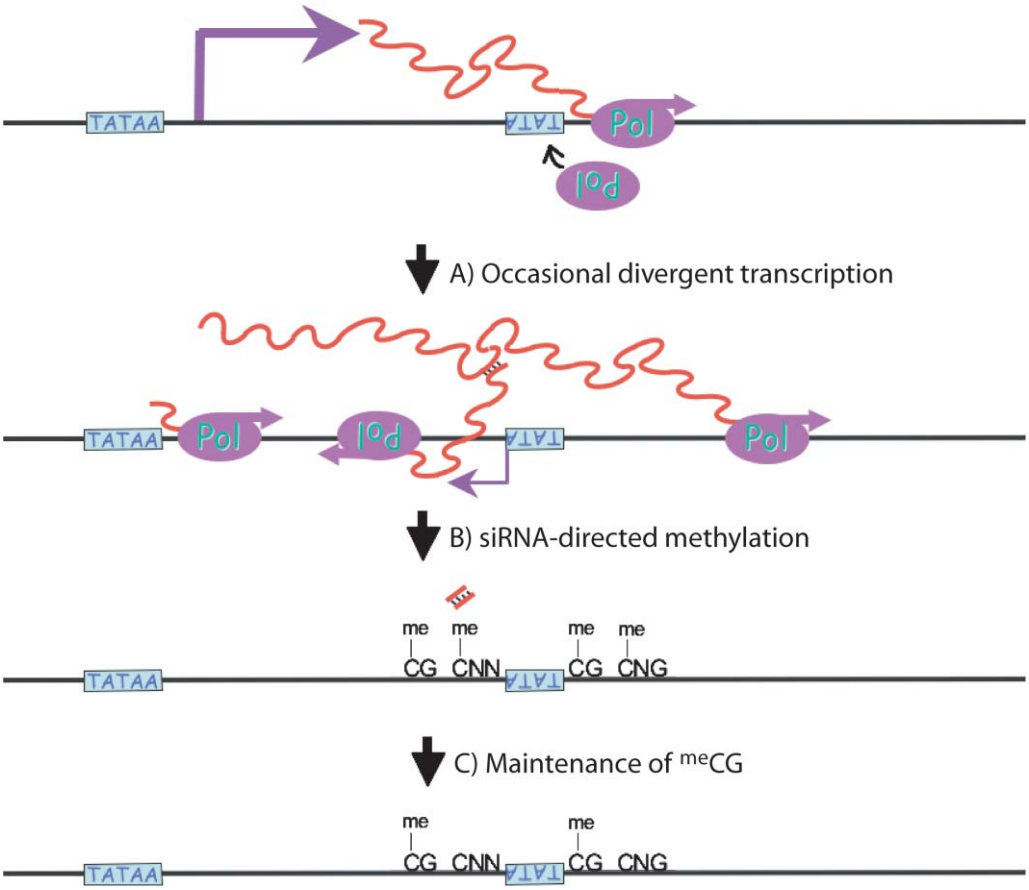


Figure 4. Model for How Occasional Firing of Cryptic Promoters Can Result in Clusters of CG Methylation by siRNA Targeting of DNA Methylation

(A) Transcription through a cryptic promoter can result in its activation [13], leading to divergent transcription.
(B) Nascent chains in proximity anneal and are processed to yield siRNAs, which target DNA methylation [5].
(C) Without continued generation of siRNAs, non-CG methylation is lost, whereas CG methylation is maintained. RNA polymerases (magenta), nascent transcripts (red), native promoter (TATAA), cryptic promoter (TATA), methyl (me).

4. Cao, X., Aufsatz, W., Zilberman, D., Mette, M.F., Huang, M.S., Matzke, M., and Jacobsen, S.E. (2003). Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. Curr. Biol. *13*, 2212–2217.
5. Zilberman, D., Cao, X., Johansen, L.K., Xie, Z., Carrington, J.C., and Jacobsen, S.E. (2004). Role of Arabidopsis ARGONAUTE4 in RNA-directed DNA methylation triggered by inverted repeats. Curr. Biol. *14*, 1214–1220.
6. Zilberman, D., Cao, X., and Jacobsen, S.E. (2003). ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. Science *299*, 716–719.
7. Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., et al. (2004). Role of transposable elements in heterochromatin and epigenetic control. Nature *430*, 471–476.
8. Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. *8*, 1499–1504.
9. Salser, W. (1977). Globin mRNA sequences: Analysis of base pairing and evolutionary implications. Cold Spring Harb. Symp. Quant. Biol. *42*, 985–1002.
10. Mette, M.F., Aufsatz, W., van der Winden, J., Matzke, M.A., and Matzke, A.J. (2000). Transcriptional silencing and promoter methylation triggered by double-stranded RNA. EMBO J. *19*, 5194–5201.
11. Bird, A.P. (1995). Gene number, noise reduction and biological complexity. Trends Genet. *11*, 94–100.
12. Henikoff, S., Furuyama, T., and Ahmad, A. (2004). Histone variants, nucleosome assembly and epigenetic inheritance. Trends Genet. *20*, 320–326.
13. Kaplan, C.D., Laprade, L., and Winston, F. (2003). Transcription elongation factors repress transcription initiation from cryptic sites. Science *301*, 1096–1099.
14. Karlin, S., and Burge, C. (1995). Dinucleotide relative abundance extremes: A genomic signature. Trends Genet. *11*, 283–290.
15. Breslauer, K.J., Frank, R., Blocker, H., and Marky, L.A. (1986). Predicting DNA duplex stability from the base sequence. Proc. Natl. Acad. Sci. USA *83*, 3746–3750.
16. Delcourt, S.G., and Blake, R.D. (1991). Stacking energies in DNA. J. Biol. Chem. *266*, 15160–15169.
17. Jones, L., Ratcliff, F., and Baulcombe, D.C. (2001). RNA-directed transcriptional gene silencing in plants can be inherited independently of the RNA trigger and requires Met1 for maintenance. Curr. Biol. *11*, 747–757.
18. Lindroth, A.M., Cao, X., Jackson, J.P., Zilberman, D., McCallum, C.M., Henikoff, S., and Jacobsen, S.E. (2001). Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. Science *292*, 2077–2080.
19. Saghai-Maroof, M.A., Soliman, K.M., Jorgensen, R.A., and Allard, R.W. (1984). Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA *81*, 8014–8018.