Body-Methylated Genes in *Arabidopsis thaliana* Are Functionally Important and Evolve Slowly

Shohei Takuno* and Brandon S. Gaut

Department of Ecology and Evolutionary Biology, University of California, Irvine

***Corresponding author:** E-mail: stakuno@uci.edu.

Associate editor: James McInerney

Abstract

DNA methylation of coding regions, known as gene body methylation, is conserved across eukaryotic lineages. The function of body methylation is not known, but it may either prevent aberrant expression from intragenic promoters or enhance the accuracy of splicing. Given these putative functions, we hypothesized that body-methylated genes would be both longer and more functionally important than unmethylated genes. To test these hypotheses, we reanalyzed single-base resolution bisulfite sequence data from *Arabidopsis thaliana* to differentiate body-methylated genes from unmethylated genes using a probabilistic approach. Contrasting genic characteristics between the two groups, we found that body-methylated genes tend to be longer and to be more functionally important, as measured by phenotypic effects of insertional mutants and by gene expression, than unmethylated genes. We also found that methylated genes evolve more slowly than unmethylated genes, despite the potential for increased mutation rates in methylated CpG dinucleotides. We propose that slower rates in body-methylated genes are a function of higher selective constraint, lower nucleosome occupancy, and a lower proportion of CpG dinucleotides.

Key words: DNA methylation, gene essentiality, substitution rate.

Introduction

Cytosine methylation is an epigenetic modification that affects both chromatin packaging and transcription. In plants, DNA methylation occurs in three sequence contexts—CG, CHG. and CHH (where H = A, C, or T)—and these contexts are affected differentially among genomic features. For example, all three contexts are methylated within repetitive elements, but only the CG context is predominantly methylated within coding regions (Cokus et al. 2008; Lister et al. 2008, 2009).

The function of DNA methylation may vary among genomic features as well. Within repetitive DNA, methylation silences transcription and functions as a host defense against transposable elements (Lisch 2009). In contrast, the function of methylation within coding regions (or "body methylation") is not yet clear. One hypothesis is that body methylation suppresses expression from cryptic promoters within coding regions, thus preventing leaky expression that could be both energetically and functionally costly (Zilberman et al. 2007; Maunakea et al. 2010). A second hypothesis is that body methylation enhances accurate splicing of primary transcripts (Lorincz et al. 2004; Luco et al. 2010). This idea is supported by the facts that body methylation, together with H3K36me, is predominantly distributed in exons, as opposed to introns, and that alternatively spliced exons tend to possess lower levels of methylation (Ball et al. 2009; Hodges et al. 2009; Kolasinska-Zwierz et al. 2009; Schwartz et al. 2009; Choi 2010; Feng et al. 2010).

A third possibility is that body methylation has no functional significance and is, perhaps, a byproduct of transcription (Roudier et al. 2009; Teixeira and Colot 2009). This viewpoint is supported by the observations that body methylation has only minor but positive effects on levels of gene expression (Zhang et al. 2006; Zilberman et al. 2007; Zemach et al. 2010) and can be highly polymorphic among individuals (Vaughn et al. 2007; Zhang et al. 2008).

These hypotheses generate differing predictions about the types of genes that should be methylated. Under the first two hypotheses, DNA methylation should be predominantly associated with essential genes because the violation of transcription—either via aberrant promotion or missplicing—would be particularly costly for genes with large phenotypic effects. Furthermore, body methylation should also be associated with gene length and exon number because long genes would have a higher probability of cryptic promotion and genes with many exons would have a potentially higher rate of splicing errors. In contrast, if body methylation has little or no functional consequence, there is no compelling reason to predict a relationship between body methylation and either gene essentiality or gene length.

A byproduct of DNA methylation is the spontaneous deamination of methyl-cytosine to thymine (Bird 1980; Pfeifer 2006); this process has been shown to accelerate evolutionary rates in both animals and plants (e.g., Bird 1980; Messeguer et al. 1991; Buckler and Holtsford 1996). As a consequence, body-methylated genes may be subjected to higher mutation rates than unmethylated genes. This possibility leads to conflicting evolutionary hypotheses. On the one hand, methylated genes may evolve quickly due to cytosine deamination. On the other hand, body-methylated genes may be essential and thus functionally and evolutionarily constrained.

Here, we examine these conflicting predictions by contrasting the structural, functional, and evolutionary characteristics of body-methylated genes against unmethylated genes. We study methylated genes in *Arabidopsis thaliana* because it is a model system for DNA methylation (Zhang et al. 2006; Zilberman et al. 2007; Cokus et al. 2008; Lister et al. 2008), gene function (Hanada, Kuromori, Myouga, Toyoda, Li, et al. 2009; Hanada, Kuromori, Myouga, Toyoda, Shinozaki, et al. 2009), gene expression (Schmid et al. 2005), and evolutionary rates (Hu et al. 2011; Yang and Gaut 2011). We begin the study by reanalyzing *A. thaliana* bisulfite sequencing (BS-Seq) data to discriminate body-methylated from unmethylated genes in accession Col-0 using a probabilistic approach. We then integrate methylation status with analyses of gene function and evolutionary rates to address four questions: First, do body-methylated genes, as a group, differ from unmethylated genes in structural characteristics like length and exon number? Second, do body-methylated genes tend to be more functionally important, as measured by gene knockouts and gene expression? Third, do body-methylated genes evolve more slowly than unmethylated genes, as expected if they are under strong constraint or do they instead evolve rapidly, perhaps as a consequence of cytosine deamination? Finally, do these analyses provide any insights into the function of body methylation?

Materials and Methods

Sequence and Methylation Data

The genomic sequences and gene annotation information for *A. thaliana* were obtained from TAIR (TAIR9 release; http://www.arabidopsis.org/). Genomic short-read sequences of *A. thaliana* Col-0 with bisulfite conversion were retrieved from the SRA (Sequence Read Archive) database (Lister et al. 2008). We followed a mapping process similar to that of Lister et al. (2008): using BRAT software (Harris et al. 2010), short reads with 32 nt were mapped to the *A. thaliana* genome without allowing any mismatches except for bisulfite conversion. Reads mapping to multiple positions were discarded. If more than one read mapped to the same start position, we assumed that it was due to clonal duplication during library preparation (a phenomenon Lister et al. (2008) called "clonal bias"). To avoid this bias, reads with the same starting location were collapsed into a single consensus in a way that each base to be retained was randomly chosen.

Following Lister et al. (2008), we estimated the total proportion of unconverted cytosine residues that mapped to the chloroplast genome, where methylation does not occur. We assumed this proportion to be the bisulfite sequencing error rate and used this error rate to test support for methylation of each nuclear cytosine residue with >1 read, after collapsing reads with clonal bias. The test was based on binomial probabilities, with a P value of 0.01.

Defining Body-Methylated Genes

The level of DNA methylation was quantified for each protein-coding region, defined as the annotated translation start to the termination codon. The levels of DNA methylation in CG, CHG, and CHH contexts were assessed independently. Taking the CG context as an example, let p_{cg} be the proportion of methylated cytosine residues at CG sites across the whole genome. Let n_{cg} and m_{cg} be the number of cytosine residues at CG sites with ≥ 2 coverage and the number of methylated cytosine residues at CG sites in a gene, respectively. Assuming a binomial probability distribution, the one-tailed P value for the departure of CG methylation level from genome average was calculated by

$$P_{CG} = \sum_{i=m_{cg}}^{n_{cg}} \binom{n_{cg}}{i} p_{cg}^{i} (1 - p_{cg})^{n_{cg}-i}, \quad (1)$$

where P_{CG} is a proxy of DNA methylation level. If the resulting P_{CG} was low, then a coding region was more densely methylated than expected at random. Using the same rationale, we calculated P_{CHG} and P_{CHH} for CHG sites and CHH sites, respectively.

Gene Characteristics

Once we distinguished methylated genes from unmethylated genes, we examined several features of *A. thaliana* genes, including gene expression. Expression analyses were based on Affymetrix expression data (Schmid et al. 2005) from 55 microarray conditions that were not based on either genetic mutants or overlapping tissues. These 55 were reanalyzed with the MAS5 method by Matsuda et al. (2010), which contain information about signal intensity and corresponding P values to test for expression.

We calculated several additional characteristics for *A. thaliana* genes. GC content was measured at both 4-fold degenerate sites (GC4) and introns (GCint). Codon usage bias was assessed using the frequency of optimal codons (F_{OP}) as an index (Ikemura 1985), using CodonW ver. 1.3 (http://www.molbiol.ox.ac.uk/cu/culong.html); optimal (or preferred) codons for each amino acid in *A. thaliana* were retrieved from Wright et al. (2004). The recombination rate of each gene was estimated by interpolating the genetic and physical distances from Singer et al. (2006). The recombination rate (cM/Mb) was estimated using the cubic splines method, implemented in MareyMap (Rezvoy et al. 2007). We also obtained information about DNA replication timing on the *A. thaliana* fourth chromosome (Lee et al. 2010), dividing regions into replication during the early-middle S phase or the late S phase.

Finally, we estimated nucleosome occupancy in each *A. thaliana* coding region. To estimate occupancy, we retrieved the genome-wide nucleosome occupancy data of Chodavarapu et al. (2010) from the SRA database. These data consist of short Illumina reads, generated from micrococcal nuclease-digested nucleosomal DNA. Using the BRAT software package (Harris et al. 2010) without the bisulfite option, the 36 nt reads were mapped to the *A. thaliana* genome with a tolerance of up to two mismatches. The level of nucleosome occupancy was assessed for each base pair, following Kaplan et al. (2009). Briefly, we first calculated short-read coverage for each position. Second, for a small fraction of sites the coverage was >10 times larger than the genome median,

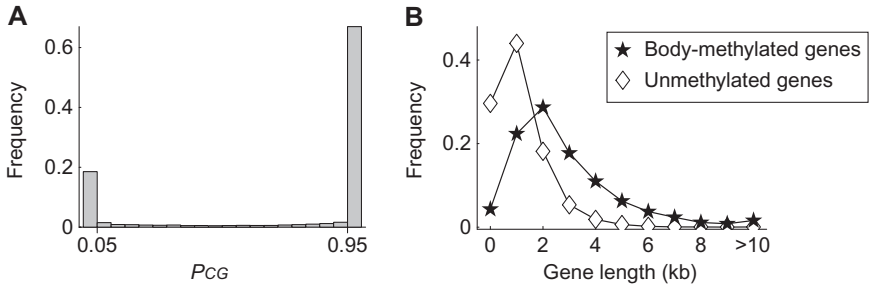

FIG. 1. (*A*) Frequency distribution of P_{CG} as a proxy of CG methylation level. Lower P_{CG} means higher methylation levels. (*B*) Frequency distribution of gene length. Black stars and white diamonds represent body- and unmethylated genes, respectively.

perhaps due to clonal bias; to reduce this bias, we reduced the coverage at these sites to be ten times the genomic median. Third, sites with nucleosome occupancy were defined as any site that had higher coverage than the genome average. Finally, we calculated the proportion of occupied sites within each gene and used that measure in analyses.

Divergence Analysis

We analyzed a set of 18,310 *Arabidopsis lyrata*/*A. thaliana* orthologs identified by Hu et al. (2011). The orthologs were aligned with CLUSTALW version 1.83 (Thompson et al. 1994). Synonymous and nonsynonymous substitution rates (K_A and K_S) between *A. thaliana* and *A. lyrata* orthologs were estimated using the Nei and Gojobori (1986) method. Distances were estimated only for the 16,447 aligned sequences having $\geq$100-bp synonymous sites. Our alignments included introns; intron divergence (K_{INT}) was estimated by the *p*-distance when the alignment of concatenated introns exceeded $\geq$100 bp.

Results

The Identification of Body-Methylated Genes

We identified body-methylated genes in the *A. thaliana* genome using previously published BS-Seq data (Lister et al. 2008). After discarding multiply mapping and clonal reads, we mapped 36,705,379 short (32 nt) reads uniquely (see Materials and Methods). These reads covered ~73% of cytosine residues (31,464,361) in the *A. thaliana* genome with read depth ≥ 2. For these cytosine residues, we applied a binomial test of support (Lister et al. 2008), assuming that the background (or error) rate was that detected in the chloroplast genome, which was that 2.42% of cytosines were falsely inferred to be methylated. Altogether, 2,262,156 cytosine residues were detected as methylated, a number similar to that of Lister et al. (2008) (2,267,447) using slightly different mapping criteria.

We calculated P_{CG}, P_{CHG}, and P_{CHH} for each gene; lower values of these metrics correspond to a smaller probability that the gene is methylated at random levels, given genome-wide levels of methylation in each sequence context. We filtered the data by considering only those genes with sufficient CG information ($n_{cg} \geq 20$) and genes for which $\geq$60% of cytosine residues were covered by at least two reads (supplementary fig. S1, Supplementary Material online), leaving 24,279 of 27,169 *A. thaliana* genes. We discarded genes with $P_{CHG} < 0.05$ and/or $P_{CHH} < 0.05$ because genes that are highly methylated in multiple contexts are atypical for coding regions (Cokus et al. 2008; Lister et al. 2008, 2009) and thus may possess transposons within coding regions, be located in highly heterochromatic regions or be misannotated. These procedures resulted in the filtering of 763 (or 3.1%) of the 24,279 analyzed coding regions.

For the remaining 23,516 genes, we calculated P_{CG} and used the distribution of P_{CG} as a proxy for the CG methylation level of a gene (fig. 1*A*). The distribution of P_{CG} was notably bimodal, indicating that CG methylation is not randomly distributed across the genome but is autocorrelated, as demonstrated previously (Cokus et al. 2008; Lister et al. 2008, 2009). We used P_{CG} to define body-methylated and unmethylated genes, using the criteria of $P_{CG} < 0.05$ and $P_{CG} > 0.95$, respectively, and discarding genes with intermediate methylation levels (i.e., with $0.05 \leq P_{CG} \leq 0.95$). By this method, we discarded another 3,402 genes from further analysis but identified 4,361 body-methylated genes and 15,753 unmethylated genes.

Body-Methylated Genes Are Longer than Unmethylated Genes

If body methylation enhances either transcription accuracy or splicing efficiency, then methylated genes should be longer and have more exons than unmethylated genes (see Introduction). Our data support these predictions. The mean length of the body-methylated genes was 3,349.5 bp, exceeding by more than 2-fold the mean length of unmethylated genes (1,595.3 bp) (fig. 1*B*). The difference in average length was significant at $P < 10^{-5}$, based on a permutation test of 100,000 trials. We obtained similar results when assessing the combined length of exons without introns (2,082.8 bp vs. 1,079.6 bp; $P < 10^{-5}$) and the number of exons (9.48 vs. 4.15; $P < 10^{-5}$).

Body-Methylated Genes Are Functionally Important

If body methylation serves a function, methylated genes should be more functionally important than unmethylated genes (see Introduction). While there is no perfect assay to test "functional importance," we tested this prediction by examining two characteristics: the phenotypic effects of gene knockouts and patterns of gene expression.

Table 1. Gene Indispensability of Body-Methylated Genes.

	Morphological Disruption	No Effect
Body-methylated gene	**476**	**378**
Unmethylated gene	**942**	**2660**
		$P < 10^{-58}$
	Expressed	**Not Expressed**
Body-methylated gene	**3686**	**79**
Unmethylated gene	**11278**	**998**
		$P < 10^{-34}$

NOTE.—*P* values by FET.

Knockout Mutants

We tallied gene "dispensability" based on the work of Hanada, Kuromori, Myouga, Toyoda, Li, et al. (2009) and Hanada, Kuromori, Myouga, Toyoda, Shinozaki, et al. (2009), who assembled data for the phenotypic effects of knockouts from >5,000 Col-0 insertional mutants (Kuromori et al. 2006). Among mutants, ~55.7% of the assessed body-methylated genes exhibited phenotypic effects (table 1). In contrast, only ~26.2% of the unmethylated genes disrupted morphology. The difference in proportion was highly significant by Fisher's exact test (FET; $P < 10^{-58}$), indicating that mutations within body-methylated genes have greater phenotypic consequences on average.

It is well known, however, that the phenotypic and functional effects of gene knockouts may be buffered by paralogs (Gu et al. 2003; Hanada, Kuromori, Myouga, Toyoda, Li, et al. 2009; Hanada, Kuromori, Myouga, Toyoda, Shinozaki, et al. 2009). We therefore sought to determine if gene duplication could be driving the apparent differences between methylated and unmethylated genes. Under this hypothesis, knockouts of unmethylated genes manifest fewer phenotypic effects because they are more often functionally buffered by paralagous gene family members. To test this hypothesis, we performed all-against-all BLASTP analysis. If a gene hit another gene with a particular *E* value, we concluded that the gene belonged to multigene families. For example, using the *E* value $\leq 10^{-50}$ as a cutoff, 67.7% and 60.2% of body- and unmethylated genes, respectively, were members of multigene families (FET; $P < 10^{-19}$). Using more stringent criteria, we obtained similar results ($P < 10^{-41}$ for $E \leq 10^{-70}$; $P < 10^{-69}$ for $E \leq 10^{-100}$). Using less stringent criteria ($E \leq 10^{-20}$ and below), the differences between methylated and unmethylated genes disappeared, but the difference was never reversed (data not shown). Thus, methylated genes are found within gene families more often than unmethylated genes, and membership within a gene family does not drive the differences in phenotypic effects between methylated and unmethylated genes.

Gene Expression

Another, albeit less direct, measure of functional importance is gene expression. We examined gene expression in Affymetrix expression data (see Materials and Methods). To test functionality on the basis of gene expression, we counted the number of unexpressed genes in all 55 assay conditions (i.e., all *P* values ≥ 0.05; table 1). Among the genes on the Affymetrix chip, the proportion of unexpressed methylated genes (2.1%) was far lower than that of unmethylated genes (8.1%) genes ($P < 10^{-34}$; FET).

We also examined the distribution of gene expression among genes by assessing both the mean and the breadth of expression across experiments. Similar to previous reports (Zhang et al. 2006; Zilberman et al. 2007), the distribution of mean signal intensity indicated that body-methylated genes are moderately expressed compared with unmethylated genes, which exhibit a broader variance in mean expression level (fig. 2*A*). This pattern emphasizes that a relatively high proportion of unmethylated genes have low expression levels, consistent with our inference that a significantly higher proportion of unmethylated genes are unexpressed. Expression breadth was measured using entropy as an index; high and low entropy values indicate broad and tissue-specific expression patterns, respectively. Figure 2*B* indicates that body-methylated genes tend to be expressed more broadly than unmethylated genes, as previously documented by Zhang et al. (2006).

Body-Methylated Genes Evolve More Slowly, on Average

DNA methylation is mutagenic (Bird 1980; Pfeifer 2006), and cytosines evolve rapidly in the context of CG dinucleotides (Bird 1980; Messeguer et al. 1991; Buckler and Holtsford 1996). It therefore seems reasonable that genes with the potential for heavy methylation by virtue of the availability of CG dinucleotides will also be exposed to a high rate of mutation. There is thus a potential paradox: on the one hand, we have shown that methylated genes are biased toward genes with phenotypic effects, suggesting that these are likely conserved genes; on the other hand, they may be subjected to higher mutation rates and hence evolve more rapidly.

To examine evolutionary rates, we estimated nonsynonymous (K_A), synonymous (K_S), and intron divergence (K_{INT}) between *A. thaliana* and *A. lyrata* orthologs (table 2). Note that we do not know methylation status in *A. lyrata* and therefore we rely on our methylation definitions in *A. thaliana*. Consistent with the high proportion of genes with phenotypic effects (table 1), estimates of K_A and the K_A/K_S ratio were significantly lower in body-methylated genes than unmethylated genes (fig. 3 and table 2). However, both K_S and K_{INT} were also significantly lower in the body-methylated genes (fig. 3 and table 2), with K_S and K_{INT} being positively correlated to each other ($r = 0.280$, $P < 10^{-5}$ by permutation test with 100,000 trails) and to K_A ($r = 0.318$, $P < 10^{-5}$ for K_S and $r = 0.113$, $P < 10^{-5}$ for K_{INT}). The low average K_S in body-methylated genes is somewhat surprising given that methylation is expected to increase mutation rates in CG dinucleotides (Bird 1980; Messeguer et al. 1991; Buckler and Holtsford 1996).

Correlates with Evolutionary Rates

Why do methylated genes have lower K_S and K_{INT}, on average? The most obvious explanation is that as a group they tend to be more essential, as suggested above, and thus are under stronger selective constraint. While this explanation

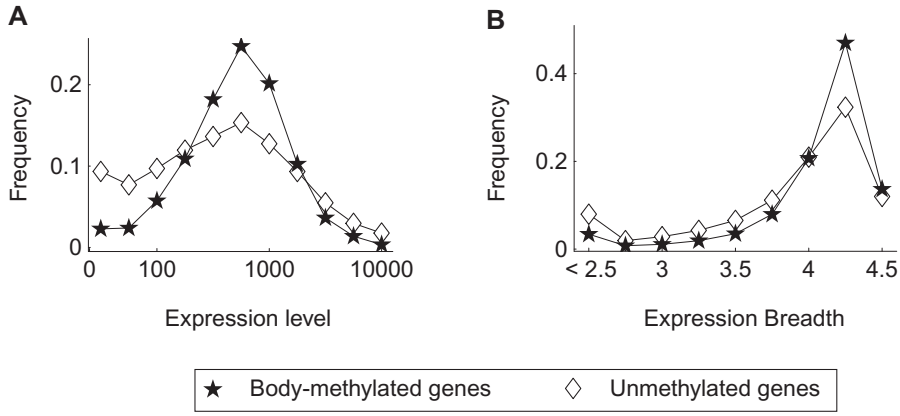

Fig. 2. (*A*) Frequency distribution of expression level (mean signal intensity across 55 tissues). (*B*) Frequency distribution of expression breadth (entropy across 55 tissues).

is satisfactory for nonsynonymous sites, it is not wholly convincing for intron and synonymous sites. We therefore examined other genic characteristics that might contribute to differences in rates between methylated and unmethylated genes.

Several genic properties either did not differ statistically between the two gene classes or differed in a way inconsistent with differences in evolutionary rate, including GC content at both exonic and intronic positions; recombination rate per base pair; translational efficiency, as measured by the frequency of optimal codons (F_{OP}); and replication timing (for details, see supplementary fig.S2 at Supplementary Material online). However, we found two additional characteristics that may help explain slower average substitution rates in methylated genes: CpG content and nucleosome occupancy.

CpG Content

We measured the proportion of CpG sites within genes and compared it with the expected proportion of CpG sites. The expected proportion was calculated from base composition (Bird 1980), and we denoted the ratio of observed to expected CpG sites as CpG[O/E]. The methylated genes had a significantly lower proportion of CpG dinucleotides and a smaller value of CpG[O/E] than unmethylated genes (2.37% vs. 3.44% for proportion of CpG; 0.585 vs. 0.785 for CpG[O/E]; fig. 4*A* and *B*). We also found that the proportion of CpG sites was significantly correlated with both K_S and K_{INT} ($r = 0.267$, $P < 10^{-5}$ for K_S; $r = 0.162$, $P < 10^{-5}$ for K_{INT}), as was CpG[O/E] ($r = 0.296$, $P < 10^{-5}$ for K_S; $r = 0.246$, $P < 10^{-5}$ for K_{INT}). In other words, methylated genes are comparatively underrepresented for CpG dinucleotides, even after correction for base composition. However, this difference was not sufficient to explain differences in rates between methylated and unmethylated genes. For example, after binning the proportion of CpG sites and CpG[O/E], both K_S and K_{INT} were still lower in body-methylated genes (fig. 4*C* and *D* for K_S, not shown for K_{INT}).

Nucleosome Occupancy

DNA sequence is wrapped around nucleosomes at well-conserved positions, with linker regions (also known as nucleosome free regions) between nucleosome units (Jiang and Pugh 2009). Presumably DNA repair machinery is more easily recruited to linker regions (Thoma 2005; Ataian and Krebs 2006), leading to lower evolutionary rates at all sites (nonsynonymous, synonymous, and intron) for genes within these regions. To investigate the relationship between nucleosome occupancy and evolutionary rate, we used the proportion of nucleosome occupancy region in *A. thaliana* as an index (see Materials and Methods) and found that K_S and K_{INT} were positively correlated to this index when all genes were considered ($r = 0.174$, $P < 10^{-5}$ for K_S; $r = 0.0372$, $P < 10^{-4}$ for K_{INT}). Moreover, body-methylated genes differed significantly in nucleosome occupancy compared with unmethylated genes (0.433 vs. 0.491; fig. 5*A*), in a direction consistent with differences in evolutionary rate. We divided nucleosome occupancy into bins and found that both K_S and K_{INT} remained lower for body-methylated genes in each bin (fig. 5*B* for K_S, not shown for K_{INT}). Therefore, nucleosome occupancy, like CpG content, may contribute to differences in rate but does not fully explain relatively low K_S and K_{INT} in body-methylated genes.

Discussion

We have reanalyzed existing BS-Seq data to classify genes as either "methylated" or "unmethylated," using a probabilistic approach. This categorical approach seems reasonable,

Table 2. The Pattern of Substitution Rate.

	K_A	K_S	K_A/K_S	K_{INT}
Body-methylated gene	**0.0235 (3456)**	**0.122 (3456)**	**0.198 (3448)**	**0.107 (2995)**
Unmethylated gene	**0.0316 (10223)**	**0.140 (10223)**	**0.230 (10142)**	**0.137 (7811)**
***P* value**	**$<10^{-5}$**	**$<10^{-5}$**	**$<10^{-5}$**	**$<10^{-5}$**

NOTE.—Numbers in parenthesis are sample size.

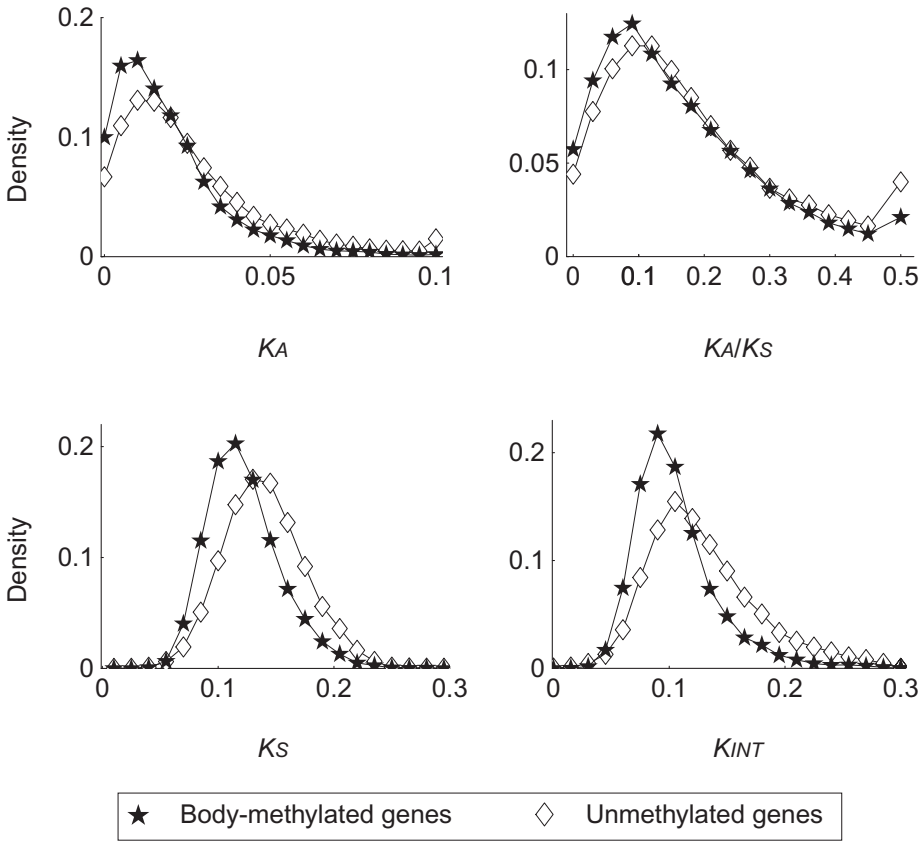

FIG. 3. Density distributions of K_A, K_A/K_S, K_S, and K_{INT}. Black stars and white diamonds represent the body- and unmethylated genes, respectively.

both because P_{CG} resulted in a strikingly bimodal distribution of genes (fig. 1A) and because it leads to results similar to those of Zhang et al. (2006), who identified methylated genes based on array data. Zhang et al. (2006) concluded that 33%, or ~8,000, genes were methylated in Col-0. We estimate that 18%, or 4,361, genes are methylated, but this number excludes the 3,402 intermediate genes with $0.05 \leq P_{CG} \leq 0.95$, which could be considered as methylated under less stringent criteria.

Our comparison of methylated and unmethylated genes has led to two primary observations. The first is that body-methylated genes in *A. thaliana* accession Col-0 tend to be longer, have more exons, and serve more important functions—as measured by phenotypic effects of insertional mutants and gene expression—than unmethylated genes. The second observation is that body-methylated genes evolve more slowly, on average, than unmethylated genes.

Potential Causes of Low Evolutionary Rates

Recent studies have revealed that evolutionary rates are a complex function of myriad gene and functional characteristics (reviewed in Pál et al. 2006; Gaut et al. 2011). Several well-known factors likely contribute to the differences in evolutionary rate between body-methylated and unmethylated genes: 1) higher selective constraint due to gene essentiality (table 1); 2) differences in gene length, which are inversely related to substitution rates (e.g., Parsch 2003; Haddrill et al. 2005; Marais et al. 2005; Halligan and Keightley 2006; Yang and Gaut 2011); and 3) differences in patterns of gene expression, which can be highly correlated with rates of nonsynonymous substitution (Drummond et al. 2006; Pál et al. 2006; Yang and Gaut 2011).

In addition, we suggest that both lower nucleosome occupancy and lower CpG[O/E] values contribute to differences in rates. Regarding the former, it has already been shown that linker regions have lower mutation rates in yeast (Washietl et al. 2008), and it is also known that distribution of methylated genes covaries with nucleosome occupancy in both animals and plants (Kolasinska-Zwierz et al. 2009; Schwartz et al. 2009; Chodavarapu et al. 2010; Choi 2010). Our observations suggest that the links among methylation, nucleosome occupancy, and evolutionary rates also pertain to *A. thaliana*.

With regard to CpG[O/E], our and previous results suggest an interesting dynamic between methylation and the prevalence of CpG sites (Bird 1980; Saxonov et al. 2006; Suzuki et al. 2007; Weber et al. 2007). One the one hand, methylated CpG sites are expected to dissipate rapidly due to high mutation rates, once methylated. In fact, CpG[O/E] values should tend to 0 for methylated regions in the absence of a countervailing force. On the other hand, the countervailing force might be selection to maintain CpG

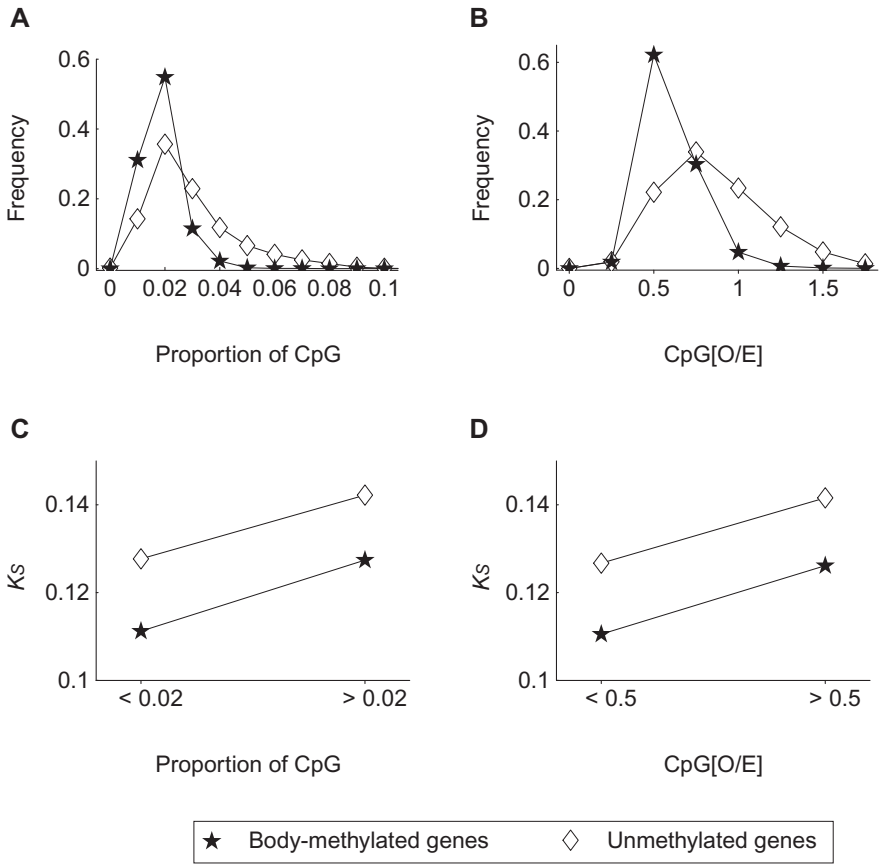

Fig. 4. (*A*) The proportion of CpG sites in the body- and unmethylated genes. (*B*) CpG[O/E] in the body- and unmethylated genes. (*C*) Relationship between the proportion of CpG sites and K_S; CpG values are binned as indicated. (*D*) Relationship between CpG[O/E] and K_S; CpG[O/E] values are binned as indicated.

sites (Buckler and Holtsford 1996), so that genes can be methylated and, as a consequence, either accurately transcribed or accurately spliced. We thus predict that there exists an equilibrium CpG maintained by the cost of high mutation rates against the benefits of body methylation. The dynamics of this equilibrium depends on accurate assessment of the magnitudes of these costs and benefits, which would be a fitting topic for future studies.

Does Body Methylation Have a Function?

Our analyses of both insertional mutants and gene expression data suggest that body-methylated genes in *A. thaliana* Col-0 are more functionally important, on average, than unmethylated genes. These results are consistent with hypotheses suggesting that body methylation has a functional role, perhaps in transcriptional accuracy or splicing efficiency. It is therefore tempting to conclude that body methylation is

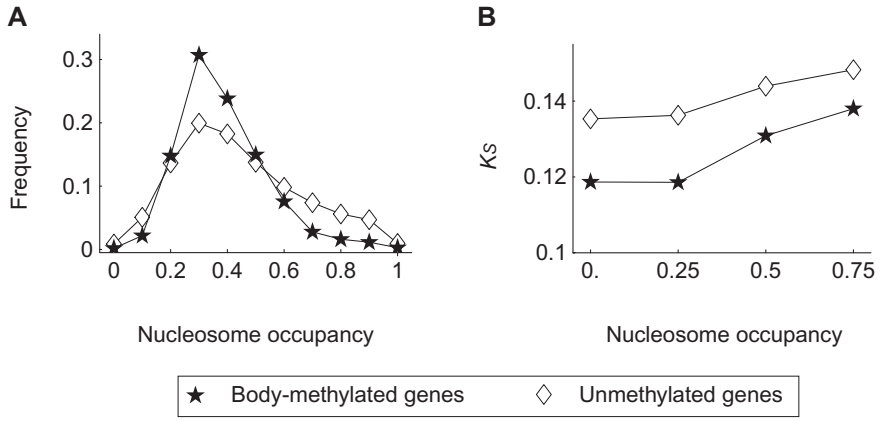

Fig. 5. (*A*) Nucleosome occupancy in body- and unmethylated genes. (*B*) The relationship between nucleosome occupancy and K_S; nucleosome occupancy is binned in from 0 to < 0.25, 0.25 to < 0.5, 0.5 to < 0.75, and 0.75 to 1.0.

indeed functional. While we favor this conclusion, it requires caution, for at least two reasons.

One reason is that we have treated methylation as a stable state, but body methylation is labile among tissues and individuals (Cedar 1988; Messeguer et al. 1991). For example, 10% of assayed CCGG sites vary between two *A. thaliana* accessions (Zhang et al. 2008), and this proportion may be higher among other accessions (e.g., Vaughn et al. 2007). Although we do not know which (perhaps all?) genes are labile, it is difficult to envision how polymorphism in body methylation would lead to the patterns observed in Col-0 without the patterns being consistent across accessions (i.e., if methylation were random across genes and individuals, we would expect none of the patterns documented here). Moreover, we anticipate that our conservative definition represents a sample of genes that is biased for constitutive (or at least consistent) body methylation across individuals. A rigorous test of this assumption will require additional body methylation data at the population level.

The second reason is that it is difficult to disentangle cause and effect. Body-methylated genes are expressed more broadly (across tissues), on average, than unmethylated genes (fig. 2*B*). Hence, if expression breadth affects body methylation, than methylation could be a functionless byproduct of transcription (Roudier et al. 2009; Teixeira and Colot 2009). Under this scenario, our results may be explained by arguing that functionally important genes evolve more slowly due to constraint and due to patterns of gene expression (Drummond et al. 2006; Pál et al. 2006; Yang and Gaut 2011), with methylation a byproduct of the latter.

However, we do not favor this interpretation for at least three reasons. First, if body methylation is a byproduct of transcription, then it is a byproduct of expression breadth (fig. 2*B*) and not mean expression level (fig. 2*A*); it is difficult to envision a mechanism to cause this distinction. Second, if methylation is a byproduct of expression breadth one might expect less overlap between the two genic classes (fig. 2*B*); the extent of overlap suggests that other factors play a role in the distinction between genic classes. Finally, the differences between body- and unmethylated genes are so consistent across functional (i.e., insertional mutants and expression) and structural (i.e., length and exon number) features that the most parsimonious explanation is, in our view, that body methylation has a functional role that has been conserved enough over time to be lead to distinct evolutionary characteristics (i.e., low CpG[O/E] and low evolutionary rates). Unfortunately, we cannot discriminate two of the hypothesized roles of body methylation—that is, suppression of intragenic transcription and splicing efficiency—because our data are consistent with both hypotheses.

Supplementary Material

Supplementary figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank two anonymous reviewers for comments; Y. Van de Peer and J.A. Fawcett for sharing the list of orthologs between *A. thaliana* and *A. lyrata*; and E. Sasaki for providing reanalyzed expression data of Affymetrix microarray. S.T. is a JSPS (the Japan Society for the Promotion of Science) Postdoctoral Fellow for Research Abroad. This work was supported by National Science Foundation grant DEB-0723860 to B.S.G.

References

Ataian Y, Krebs JE. 2006. Five repair pathways in one context: chromatin modification during DNA repair. *Biochem Cell Biol*. 84:490–504.

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol*. 27:361–368.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*. 8:1499–1504.

Buckler ES IV, Holtsford PT. 1996. *Zea* ribosomal repeat evolution and substitution patterns. *Mol Biol Evol*. 13:623–632.

Cedar H. 1988. DNA methylation and gene activity. *Cell*. 53:3–4.

Chodavarapu RK, Feng S, Bernatavichute YV, et al. (19 co-authors). 2010. Relationship between nucleosome positioning and DN A methylation. *Nature*. 466:388–392.

Choi JK. 2010. Contrasting chromatin organization of CpG islands and exons in the human genome. *Genome Biol*. 11:R70.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild C, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature*. 452:215–219.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*. 23:327–337.

Feng S, Cokus SJ, Zhang X, et al. 2010. (15 co-authors). 2010. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A. 107p. 8689–8694 .

Gaut BS, Yang L, Takuno S, Eguiarte LE. 2011. The patterns and causes of variation in plant nucleotide substitution rates. *Annu Rev Ecol Evol Syst*. doi:10.1146/annurev-ecolsys-102710-145119

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* 421:63–66.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol*. 6:R67.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res*. 16:875–884.

Hanada K, Kuromori T, Myouga F, Toyoda T, Li W, Shinozaki K. 2009. Evolutionary persistence of functional compensation by duplicate genes in *Arabidopsis*. *Genome Biol Evol*. 2009:409–414.

Hanada K, Kuromori T, Myouga F, Toyoda T, Shinozaki K. 2009. Increased expression and protein divergence in duplicate genes is associated with morphological diversification. *PLoS Genet*. 5:e1000781.

Harris EY, Ponts N, Levchuk A, Roch KL, Lonardi S. 2010. BRAT: bisulfite-treated reads analysis tool. *Bioinformatics*. 26:572–573.

Hodges E, Smith AD, Kendall J, et al. (14 co-authors). 2009. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res*. 19:1593–1605.

Hu TT, Pattyn P, Bakker EG, et al. (30 co-authors). 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet*. 43:476–481.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*. 2:13–34.
Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet*. 10:161–172.
Kaplan N, Moore IK, Fondufe-Mittendorf Y, et al. (11 co-authors). 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*. 458:362–366.
Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet*. 41:376–381.
Kuromori T, Wada T, Kamiya A, et al. (11 co-authors). 2006. A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of *Arabidopsis*. *Plant J*. 47:640–651.
Lee TJ, Pascuzzi PE, Settlage SB, et al. (16 co-authors). 2010. *Arabidopsis thaliana* chromosome 4 replicates in two phases that correlate with chromatin state. *PLoS Genet*. 6:e1000982.
Lisch D. 2009. Epigenetic regulation of transposable elements in plants. *Annu Rev Plant Biol*. 60:43–66.
Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133: 523–536.
Lister R, Pelizzola M, Dowen R, et al. (18 co-authors). 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462:315–322.
Lorincz MC, Dickerson DR, Schmitt M, Groudine M. 2004. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol*. 11:1068–1075.
Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. 2010. Regulation of alternative splicing by histone modifications. *Science*. 327:996–1000.
Marais G, Nouvellet P, Keightley PD, Charlesworth B. 2005. Intron size and exon evolution in *Drosophila*. *Genetics*. 170:481–485.
Matsuda F, Hirai MY, Sasaki E, Akiyama K, Yonekura-Sakakibara K, Provart NJ, Sakurai T, Shimada Y, Saito K. 2010. AtMetExpress development: a phytochemical atlas of *Arabidopsis* development. *Plant Physiol*. 152:566–578.
Maunakea AK, Nagarajan RP, Bilenky M, et al. (26 co-authors). 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*. 466:253–257.
Messeguer R, Ganal MW, Steffens JC, Tanksley SD. 1991. Characterization of the level, target sites and inheritance of cytosine methylation in tomato nuclear DNA. *Plant Mol Biol*. 16:753–770.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 3:418–426.
Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet*. 7:337–348.
Parsch J. 2003. Selective constraints on intron evolution in *Drosophila*. *Genetics*. 165:1843–1851.
Pfeifer GP. 2006. Mutagenesis at methylated CpG sequences. *Curr Top Microbiol Immunol*. 301:259–281.
Rezvoy C, Charif D, Guéguen L, Marais GA. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. *Bioinformatics* 23:2188–2189.
Roudier F, Teixeira FK, Colot V. 2009. Chromatin indexing in *Arabidopsis*: an epigenomic tale of tails and more. *Trends Genet*. 25:511–517.
Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*. 103:1412–1417.
Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU. 2005. A gene expression map of *Arabidopsis thaliana* development. *Nat Genet*. 37: 501–506.
Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*. 16:990–995.
Singer T, Fan Y, Chang HS, Zhu T, Hazen SP, Briggs SP. 2006. A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet*. 2:e144.
Suzuki MM, Kerr ARW, De Sousa D, Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res*. 17:625–631.
Teixeira FK, Colot V. 2009. Gene body DNA methylation in plants: a means to an end or an end to a means? *EMBO J*. 28:997–998.
Thoma F. 2005. Repair of UV lesions in nucleosomes—intrinsic properties and remodeling. *DNA Repair (Amst)*. 4:855–869.
Thompson JD, Higgins DG, Gibson TJ. 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.
Vaughn MW, Tanurdzi M, Lippman Z, et al. (13 co-authors). 2007. Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biol*. 5:e174.
Washietl S, Machné R, Goldman N. 2008. Evolutionary footprints of nucleosome positions in yeast. *Trends Genet*. 24:583–587.
Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*. 39:457–466.
Wright SI, Yau CB, Looseley M, Meyers BC. 2004. Effects of gene expression on molecular evolution in *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Mol Biol Evol*. 21:1719–1726.
Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Mol Biol Evol*. 28:2359–2369.
Zemach A, McDaniel IE, Silva P, Zilberman D. (13 co-authors). 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328:916–919.
Zhang X, Shiu S, Shiu S, Cal A, Borevitz JO. 2008. Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. *PLoS Genet*. 4:e1000032.
Zhang X, Yazaki J, Sundaresan A, et al. (11 co-authors). 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell*. 126:1189–1201.
Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet*. 39:61–69.