Published online 17 May 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 16* ***6919–6931***
doi:10.1093/nar/gkr324

Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis

Ikhlak Ahmed[1], Alexis Sarazin[1], Chris Bowler[1], Vincent Colot[1,*] and Hadi Quesneville[2]

[1]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique (CNRS) UMR8197 - Institut National de la Santé et de la Recherche Médicale (INSERM) U1024, 46 rue d'Ulm, 75230 Paris cedex 05 and [2]Unité de Recherches en Génomique-Info, Institut National de la Recherche Agronomique (INRA) UR1164, Centre de recherche de Versailles, Route de Saint Cyr, 78026 Versailles cedex, France

Received December 22, 2010; Revised March 31, 2011; Accepted April 20, 2011

ABSTRACT

Transposable elements (TEs) and their relics play major roles in genome evolution. However, mobilization of TEs is usually deleterious and strongly repressed. In plants and mammals, this repression is typically associated with DNA methylation, but the relationship between this epigenetic mark and TE sequences has not been investigated systematically. Here, we present an improved annotation of TE sequences and use it to analyze genome-wide DNA methylation maps obtained at single-nucleotide resolution in Arabidopsis. We show that although the majority of TE sequences are methylated, ~26% are not. Moreover, a significant fraction of TE sequences densely methylated at CG, CHG and CHH sites (where H = A, T or C) have no or few matching small interfering RNA (siRNAs) and are therefore unlikely to be targeted by the RNA-directed DNA methylation (RdDM) machinery. We provide evidence that these TE sequences acquire DNA methylation through spreading from adjacent siRNA-targeted regions. Further, we show that although both methylated and unmethylated TE sequences located in euchromatin tend to be more abundant closer to genes, this trend is least pronounced for methylated, siRNA-targeted TE sequences located 5′ to genes. Based on these and other findings, we propose that spreading of DNA methylation through promoter regions explains at least in part the negative impact of siRNA-targeted TE sequences on neighboring gene expression.

INTRODUCTION

Transposable elements (TEs) are ubiquitous components of genomes and their differential accumulation is responsible for most of the large variations in genome size seen among eukaryotes. However, mobilization of TEs is inherently mutagenic and is therefore a rare event. Repression of transposition involves a variety of mechanisms, including DNA methylation in plants and mammals (1,2). Moreover, TEs are among the fastest evolving sequences, leading over time to the accumulation of degenerate, non-mobile relics.

In plants, TE sequences are typically methylated at CG, CHG and CHH sites (where H = A, T or C) in a process that requires numerous factors, including small interfering RNAs (siRNAs) to guide methylation of homologous DNA sequences, and so called *de novo* and maintenance DNA methyltransferases (2,3). The model plant Arabidopsis offers several advantages for the detailed exploration of the relationship between DNA methylation and TE sequences, such as a small, almost fully sequenced genome (4) and a large collection of mutants affected in the establishment, maintenance or removal of DNA methylation (3). However, despite the fact that DNA methylation in Arabidopsis has been studied genome wide using a variety of approaches, including bisulphite treatment of genomic DNA followed by whole genome sequencing (5,6), patterns of DNA methylation associated with Arabidopsis TE sequences have not been investigated systematically so far.

We previously described the development of a highly sensitive TE annotation pipeline that doubled the fraction of the Arabidopsis genome detected as TE sequences compared to the initial annotation (7). In the present study, we have refined this pipeline further and have used the resulting set of annotated TE sequences,

*To whom correspondence should be addressed. Tel: +33 1 44 32 35 38; Fax: +33 1 44 32 39 35; Email: colot@biologie.ens.fr

which now cover 21% of the genome sequence, to re-analyze publicly available genome-wide DNA methylation and siRNA datasets. Our analysis indicates that although the majority of TE sequences are densely methylated, >25% are unmethylated at most or all sites, or show significant DNA methylation only over one or two of the three types of sites (CG, CHG and CHH). Furthermore, methylated TE sequences are less often characterized by an abundance of matching siRNAs when located in heterochromatin than in euchromatin. These methylated TE sequences with no or few matching siRNAs tend to show higher levels of DNA methylation towards their extremities and are typically flanked on both sides by methylated TE sequences that are targeted by siRNAs. These observations suggest the existence of local spreading of DNA methylation from siRNA-targeted TE sequences. Further, we show that in euchromatin, both methylated and unmethylated TE sequences are most abundant close to genes. However, this preference is less pronounced for methylated, siRNA-targeted TE sequences upstream of genes. Based on these findings, we propose that the negative impact of siRNA-targeted TE sequences on the expression of neighboring genes which has been observed in *Arabidopsis thaliana* and *Arabidopsis lyrata* (8) results at least in part from local spreading of DNA methylation into promoter regions.

MATERIALS AND METHODS

Sequences

The *A. thaliana* Release 5 genomic sequence was downloaded from TIGR web site (http://ftp.tigr.org/pub/data/a_thaliana/ath1/). Annotations Release 7 was obtained from TAIR as a dump of their database. The three TE reference sequence sets (Opt, Maxsize and OptCoding) used in addition to Repbase Update (RU) were described previously (7).

TE detection pipeline

TE sequence models were detected using the following combination of softwares: BLASTER (9,10), RepeatMasker (11), Censor (12,13). Satellite repeats were detected using RepeatMasker, Tandem Repeat Finder [TRF; (14) and mreps (15)]. The Torque resource manager was used to provide control over batch jobs and distributed compute nodes (http://www.clusterresources.com/pages/products/torque-resource-manager.php). Results were stored in a MySQL database (http://www.mysql.com/).

Each program was run independently. Parameters were chosen to make detection as sensitive as possible. The rate of false positives was minimized by running the TE detection softwares on 200-kb fragments of genomic sequence shuffled by di-nucleotides using the program shuffle [HMMer Package; (16)]. For each of the programs BLASTER, RepeatMasker and Censor, the highest score obtained for these di-nucleotide shuffled chunks was used as a threshold to filter out the results obtained on the true genome chunks. Simple repeats were removed from the TE annotation. TE models <20 bp were discarded.

RepeatMasker was run with the MaskerAid (17) search engine with sensitive parameters ('–cutoff 200 –w –s –gccalc –nolow –no_is'). We found MaskerAid to be more sensitive and much faster than Cross-match, under sensitive parameters. Censor was used at high sensitivity with parameter '–s –ns'. BLASTER now uses WU-BLAST as a search engine, and has also been set to more sensitive parameters, (inspired from MaskerAid settings). We used Blaster with parameters '–W –S 4'. The RMBLR procedure (9) has been replaced by a new procedure, called 'combinedBLR', which now combines the results obtained from BLASTER, RepeatMasker and Censor and gives them to MATCHER for chaining. To do this, we normalized alignment scores to be the hit length times the identity percentage. The MATCHER program has been developed to map match results onto query sequences by first filtering overlapping hits. When two matches overlap on the genomic (query) sequence, the one with the best alignment score is kept, the other is truncated so that only non-overlapping regions remain on the match. As a result of this procedure a match is totally removed only if it is included in a longer one with a best score.

Long insertions or deletions in the query or subject could result in two matches, instead of one with a long gap. Thus the remaining matches are chained by dynamic programming. A score is calculated by summing match scores and subtracting a gap penalty (0.05 times the gap length) as well as a mismatch penalty (0.2 times the mismatch length region), as described previously (18).

The chaining algorithm [(19), pp. 325–329] is modified to produce local alignments. A match is associated with a chain of other matches only if this results in a higher score. The best-scoring chain is kept and the search is repeated minus this chain until no more chain is found. This algorithm is run independently for matches on strand +/+, +/− and −/+. A maximum of 20 bp of overlap is allowed between matches. The chaining algorithm enables the recovery of TE sequences containing long insertions.

Although BLASTER, RepeatMasker and Censor are front ends of the same WU-BLAST program, they cover respectively 21, 18 and 19 Mb of the genome sequence, when the RU TE reference set is used (Supplementary Table S3). Note that without any score threshold, they appear to have a high false positive rate (cover 90, 18 and 23 Mb, respectively). To reduce the false positive rates we rely on a statistical procedure to set their parameters at very high-sensitive values. Supplementary Table S4 shows TE-detection overlaps between the three softwares. BLASTER appears to be the most sensitive, followed by Censor and then RepeatMasker. This is a consequence of the different BLAST parameters used by these programs. When results obtained by the three programs are combined, TE coverage (excluding satellite) is increased to 21.7 Mb.

The Arabidopsis genome contains several regions where TE sequences cluster, often as a result of nested insertions. These are particularly challenging to detect and classic annotation algorithms may fail to connect fragments of split TEs. This prompted us to implement a 'long join' procedure, which is based on age estimates of TE

fragments. Fragments to be joined must be co-linear and have the same age, as estimated using the percentage of identity with the TE reference sequence (20). In the 'nest join' version of the procedure, the inner TE sequence cover >95% of the region between the two fragments to be joined and be younger. TE sequences can also be split as a result of large non-TE sequence insertions. To account for this possibility, TE fragments that have the same age, are separated by an insert of <5 kb and align within 500 bp of each other on the TE reference sequence (Figure 1) are joined in a version of the 'long join' procedure referred to as 'simple join'.

TE fragments already connected in a previous step by MATCHER are split if inner TE fragments are younger than outer joined fragments. Although the 'long join' procedure outperforms MATCHER, which relies on dynamic programming and a scoring scheme that are ill adapted for extreme situations, it did produce only few 'simple join' and no 'nest join'. Indeed, many 'long join' were denied because fragments are too dissimilar in age (>2% difference in age) or too far apart (>100 kb). Results are summarized in Supplementary Table S5.

Assigning confidence scores to TE sequence models

The pipeline provides TE sequence models composed of four lines of evidence, one for each TE reference sequence set. The longest evidence (maximum length) for each TE sequence model is recognized as 'best evidence' and is used to determine the precise genomic coordinates. A score is assigned to the model based on the origin of the best evidence supporting the model. Indeed, because small insertions, unrelated to the TE sequence, may be present in the genomic copies that were used for building the

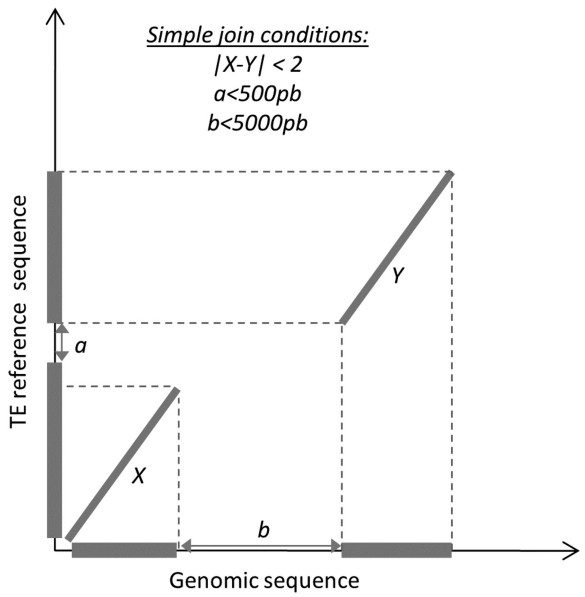

Figure 1. Schematic dot plot representation of 'simple join' conditions. Matching regions between genomic and TE reference sequence are represented by diagonals. Note that these regions might be fragments already connected by MATCHER. *X* and *Y* indicate percentage of identity to the TE reference sequence. *a* and *b* refer to the length of non-matching DNA on the TE reference and genomic sequences, respectively.

OptCoding, MaxSize and Opt TE reference sequence sets (7), evidence obtained with these different sets may not be all equally reliable. In fact, the coding constraint imposed on the OptCoding design makes this TE reference sequence set more reliable than Opt and MaxSize. A score of 3 (best) is attributed to models supported by at least RU or OptCoding, a score of 2 when support comes from Opt only and a score of 1 when support come from MaxSize alone. In cases where the longest evidence has a lower score than shorter evidence (>100 bp) and does not expand it by >50 bp, the higher score is assigned to the longest evidence.

Satellites are comprised of highly embedded tandem repeats, and the long-join procedure does not seem to work well when the TE reference sequence sets OptCoding, Opt or MaxSize are used. Indeed these TE-reference-sequence sets tend to merge many satellite units into one big unit. In contrast, RU tends to keep unit boundaries for tandem repeats. Satellites were thus annotated solely based on RU evidence and with no score attached.

RU provides best evidence for the largest set 12 922 (40%) of annotated TE sequence models, followed by Opt 12 214 (38%), MaxSize 5758 (18%) and OptCoding 981 (4%). Of the 31 245 annotated TE models, 13 752 (44%) have a score of 3; 11 773 (37.7%) a score of 2 and 5720 (18.3%) a score of 1. In addition, 3342 sequences were annotated as satellites. Note that because of the stringent statistical threshold used to detect TE sequences with high confidence, some old TE insertions are likely missed, such as those proposed to be responsible for the epigenetic regulation of the imprinted gene *FWA* (21).

DNA methylation analysis

Single-nucleotide resolution DNA methylation data were used from Cokus *et al*. (5). DNA methylation analysis was carried out separately for CG, CHG and CHH sites (Supplementary Figure S6). Sites were considered as methylated if at least 10% of the reads (CG sites) or at least one read (CHG and CHH sites) were indicative of methylation. However, because CG methylation was found to be symmetrical, as expected, methylation status was copied to the opposite strand in cases of no or insufficient coverage for that strand. Although CHG sites are symmetrical, CHG methylation was found to depart significantly from symmetry in a large number of cases, and thus methylation status of CHG sites was not copied to the opposite strand in cases of insufficient coverage for one strand. Once the methylation status of all available sites was established, DNA methylation for TE sequences or any other annotated feature is computed as a fraction of methylated Cs to the total number of covered Cs, for each of the three types of sites.

High-density tiling microarray datasets (22) were downloaded from the Gene Expression Omnibus (GSE5974). Potentially cross-hybridizing probes were identified by aligning them on the genomic sequence with nucmer from the MUMmer v3 package (23) with parameters: –maxmatch –minmatch = 10 –mincluster = 50 –nosimplify. A total of 36 993 probes (out of 382 178) were removed from the analysis because they had multiple matches with

85% identity or more. For each annotated feature, the cumulated sequence length of probes identifying a positive methylation signal was normalized by total length of probes covering the feature.

siRNA density

Small RNAs deep-sequencing data obtained from Arabidopsis Whole-aerial tissues were downloaded from GEO (accession: GSE14696) (24) and used to calculate the 24-nt siRNA density for all TE sequences with defined DNA methylation patterns. As there was a high correlation between different replicates of this library, we merged them together to achieve a ~6 million read library. Reads were mapped to the *A. thaliana* genome using MUMmer v3 and 24-nt siRNA density was calculated as follows:

$$\mathrm{ND} = \frac{\sum \frac{\mathrm{NR}_i}{\mathrm{NM}_i}}{\mathrm{TNR} \times \mathrm{Region\ length}} \times 10^8.$$

Where NR_i is the number of reads corresponding to match M_i and NM_i is the total number of matches for the sequence across the genome. TNR is the total number of reads in library and Region length is the length of TE sequence for which density is being calculated. Densities are expressed as number of reads per kilobite of the sequence per hundred thousand of the library reads.

The R package (http://cran.r-project.org) and Perl (http://www.perl.org) were used for the statistical and DNA methylation analyses, respectively.

RESULTS

Improved annotation of TE sequences

Identification of TE sequences within genomes by homology searches is challenging because many of these sequences are highly degenerate derivatives of functional TEs or occur as nested insertions. Previously, we described a method that substantially improves TE sequence detection (7). This method relies on the use of multiple sets of TE reference sequences, specifically designed to reflect diverse aspects of TE structure and evolution on the one hand, and on a TE annotation pipeline which combines several sequence-similarity search programs on the other (7). The annotation pipeline has been further refined, in particular to allow for the detection of nested insertions through a 'long join procedure'. Briefly, two or more TE sequences separated by <5 kb in the genome are joined together in the final annotation if they align in the same order and orientation within <500 bp from each other on the corresponding TE reference sequence and if they diverge from it to a similar extent (Figure 1; 'Materials and Methods' section).

Using this improved version of the TE-annotation pipeline, we now identify a total of 31 245 TE sequences, which cover 25 Mb (21%) of the 119-Mb genome sequence available. As initially reported (4), retroelements, which transpose through an RNA intermediate, represent the largest fraction of TE sequences (10 Mb), followed by helitrons (8 Mb) and DNA transposons (7 Mb), which transpose through rolling circle and cut and paste processes, respectively. A detailed description of the detected TE sequence models is provided in File 1 in Supplementary Data and the new annotation can be found at TAIR, starting with release 8. Of note, 85% and 2.5%, respectively, of sequences annotated in the TAIR release 7 as pseudogenes (3315/3897) and genes (790/31 726) show at least 75% overlap with our TE annotation (File 2 in Supplementary Data), indicating that they are in all likelihood TEs.

Defining a robust DNA methylation dataset

Two studies have combined bisulphite treatment of genomic DNA, which converts unmethylated cytosines to uracils but leaves methylated cytosines intact, with next-generation sequencing to provide single-nucleotide resolution DNA methylation maps of the Arabidopsis genome (5,6). The two studies produced essentially identical results and although no extensive analysis of TE sequences was carried out, it was concluded in both cases that repeat elements including TEs are typically methylated at CG, CHG and CHH sites. Moreover, these two studies reported that ~30% of genes are methylated, but almost exclusively at CG sites and within part of the transcribed region only.

In order to explore the genome-wide patterns of DNA methylation associated with TE sequences more systematically and in greater detail, the DNA methylation data of Cokus *et al.* (5) were first reassessed. Although the average sequencing coverage was ~20-fold (5), large variations were observed, with 66.7% and 1.3% of sequenced cytosines covered by >10 or <50 reads, respectively (Figure 2). To avoid any potential problems resulting from this uneven coverage, only those cytosines with read depths between 10 and 50 were considered for our analysis, which amounted to 32% of uniquely mapped cytosines.

Although CG sites are usually either unmethylated or methylated on over 80% of the molecules sequenced (5,6), a significant number (15% of total) have intermediate methylation levels (File 3 in Supplementary Data). Thus, among CG sites with at least one read indicative of methylation, only 53% in genes and 66% in TE sequences have methylation levels above 80% (Supplementary Figure S1). Given the possible involvement of transcription in gene body methylation (22,25), lower levels of methylation of CG sites within genes could reflect tissue-specific differences in expression. In the case of TE sequences, which are not transcribed in most cell types, intermediate levels of methylation at CG sites may rather indicate active demethylation or preferential action of the *de novo*, RNA-directed DNA methylation (RdDM) pathway over the so-called maintenance DNA methylation at these sites (3). Our analysis also confirmed that in contrast to CG methylation, CHG and CHH methylation, which are restricted almost exclusively to TE sequences and other repeat elements, rarely reach levels greater than 80% (5,6).

Based on these observations, sites were declared as methylated if at least 10% of the reads for CG sites or at least one read for CHG and CHH sites indicate methylation. Using these criteria, 82, 74 and 31% of CG, CHG

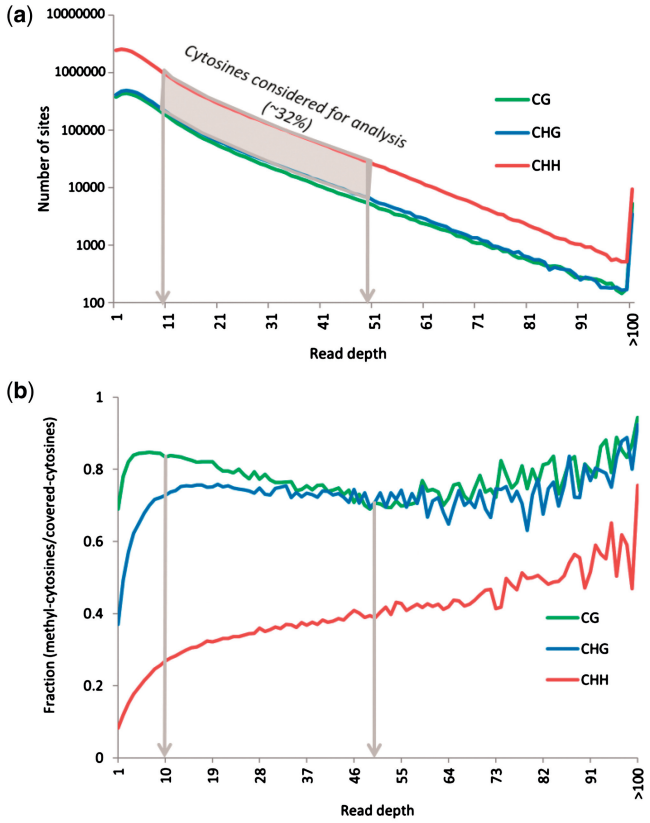

Figure 2. Read-depth coverage map of whole-genome bisulphite sequencing dataset (5). (**a**) The *x*-axis shows the number of bisulphite sequencing reads at a given cytosine and the *y*-axis represents number of sites. Most cytosines in all three sequence contexts are covered by <10 reads. For our analysis, only those cytosines were considered for which read depths were between 10 and 50. This proportion, shown in grey, represents ~32% of the original data in all three cytosine contexts. (**b**) Fraction of methyl-cytosines detected at a given sequencing coverage (Read depth). Read depths below 10 lead to an underestimation of methylated CHG and CHH sites, while read depths above 50 tend to be more often associated with methylated cytosines at all three types of sites.

and CHH sites within TE sequences are methylated, respectively, which is much higher than for genes (26, 4 and 3% for CG, CHG and CHH sites, respectively; File 3 in Supplementary Data). Furthermore, the frequency of methylated sites within the first 500 bp beyond genes decreases ~4-fold for CG sites (7%) and remains consistently low for CHG and CHH sites. In contrast, the frequency of methylated CG, CHG and CHH sites is only reduced 2-fold within the first 500 bp outside of TE sequences (42, 36 and 16%, respectively). This lower reduction in the frequency of methylated sites outside of TE sequences compared to genes could indicate that our annotation pipeline does not precisely define TE sequence boundaries or else that DNA methylation can spread from TE sequences into flanking regions (see below). We also note that among CHG sites declared as methylated, a large proportion (>30%) have statistically significant discordant methylation levels between the two strands ($P < 0.05$ in Chi-square goodness of fit test; Supplementary Figure S2). Furthermore, almost all of the latter are devoid of matching siRNAs (data not shown). These findings indicate that in at least 30% of cases, neither sequence symmetry nor siRNAs play any role in maintaining CHG methylation, which is consistent with this process relying predominantly on a reinforcing loop with dimethylation of lysine 9 of histone H3 (2).

Methylation status of TE sequences

We next determined the DNA methylation status of all individual TE sequences with sufficient information for CG, CHG and CHH sites. Given the repeated nature of TEs and the fact that only sequence reads that map to unique genomic locations with very high confidence are considered (5), cytosine coverage is reduced for TE sequences (66%) relative to the whole genome (85.6%). Nonetheless, a quarter of cytosines within TE sequences have read depths between 10 and 50, a fraction similar to that for the whole genome (27.2%) and almost identical for CG, CHG and CHH sites. Based on these observations, we only considered the 13 667 TE sequences (43.7% of total) for which information is available for >25% of each of the three distinct types of sites and the 3418 TE sequences (10.9% of total) containing only one (CHH) or two types (CHH and CG or CHG) of sites and still fulfilling the >25% coverage criterion for these sites. Two main categories of TE sequences are thus excluded from our analysis, those corresponding to recent insertions and for which reads could not be assigned unambiguously because of two or more possible matches in the genome, and those for which technical or other biases lead to <25% coverage for CG, CHG or CHH sites.

For each of the 17 085 TE sequences retained for analysis, we determined the methylation status separately for CG, CHG and CHH sites. A sequence was deemed methylated at a given type of site if at least 5% of the sites of this type had reads indicative of methylation, a value that is above the level of non-conversion of unmethylated cytosines [2–4%; (5)]. As illustrated in Figure 3, the fraction of methylated sites within individual TE sequences differs dramatically between CG, CHG and CHH sites. Thus, whereas CG sites are typically all unmethylated or all methylated within a given TE sequence, the fraction of methylated CHG sites varies almost linearly between 0% and 100% (with the exception of a peak at 98–100%) and that of CHH sites rarely exceeds 50%. To simplify the analysis, and because the reason(s) for such wide variations in the frequency of methylated CHG and CHH sites remain to be determined, the methylation status of individual TE sequences was simply summarized as either methylated (M) or un-methylated (U) for each of the three types of sites based on the 5% threshold defined above. This convention leads therefore to eight possible DNA methylation patterns (Table 1), or 10 in the case of TE sequences that are devoid of CG and/or CHG sites (Supplementary Table S1). Although all 18 patterns are observed, few predominate.

Thus, among the 13 667 TE sequences with >25% of informative CG, CHG and CHH sites, 58% have an MMM pattern (methylated in all three types of sites) and another 20% have a UUU pattern (unmethylated;

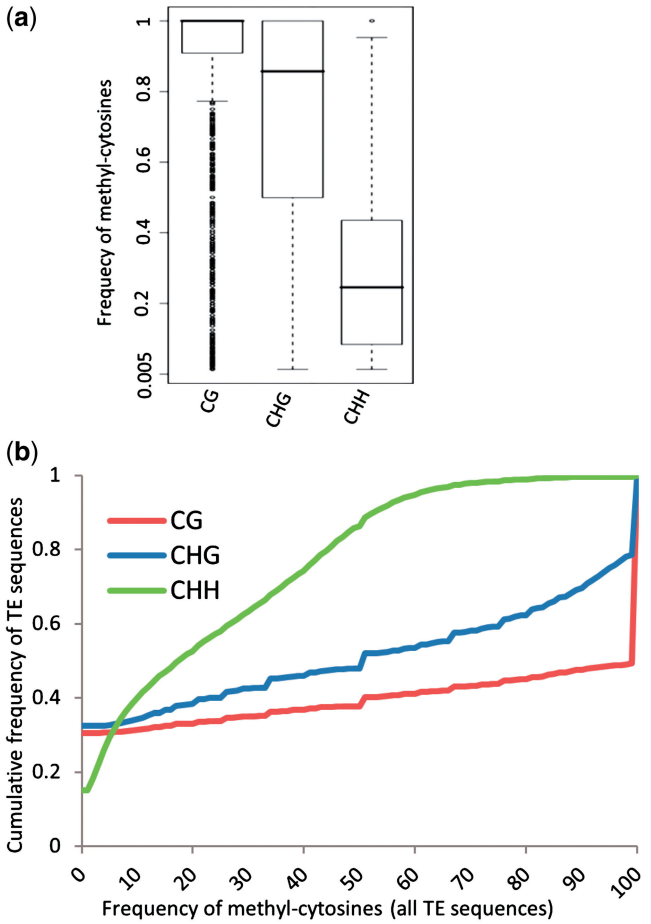

Figure 3. Frequency of methylated CG, CHG and CHH sites in TE sequences. (**a**) Boxplots showing frequency distribution of methyl-cytosines for TE sequences methylated for at least one type of site. Most of these TE sequences have almost all of their CG sites and a majority of their CHG sites methylated. (**b**) Frequency distribution of all TE sequences in relation to percentage of methylated sites, for each of the three types of sites.

Table 1). Moreover, the next most abundant pattern is UUM (6%), which is characterized by a particularly low median level of methylation for methylated CHH sites and a low-median frequency of such sites (1% and 8%, respectively, compared to 6% and 36% for the MMM pattern). Thus, whereas a majority of TE sequences for which information is available for CG, CHG and CHH sites are densely methylated, a large proportion (26%) are not significantly methylated at any of the three types of sites. Another 10% of TE sequences are methylated mainly at CG sites (MUM, MMU and MUU patterns) and have low median levels of methylation and frequency of methylated CG sites (12–40% and 38–75%, respectively, compared to 82% and 100% for the MMM pattern). These TE sequences therefore have methylation patterns resembling those of methylated genes. Finally, a small proportion of TE sequences exhibit non-CG methylation only (UMU and UMM patterns, 5% of total). Methylation data obtained by immunoprecipitation of DNA with an anti-methylcytosine antibody followed by hybridization to a high-density genome-tiling microarray [MeDIP chip; (22)] were used to validate the pertinence of the main and most contrasted patterns MMM and UUU+UUM. Out of the 382 178 probes on the array, 14 482 were extracted that covered, with little risk of cross-hybridization (see 'Materials and Methods' section), 4272 MMM and 1433 UUU + UUM TE sequences. Overall, 75 and 98% of probes corresponding to MMM and UUU + UUM patterns were declared as methylated and unmethylated, respectively, validating our classification and indicating a higher sensitivity of bisulphite sequencing over MeDIP in detecting methylated DNA, as previously reported (26).

Among the 3418 TE sequences devoid of CG and/or CHG sites and with >25% of informative sites of the other type(s), 42% have patterns (−U, -UU and U-U) clearly indicative of no, or very low, methylation (Supplementary Table S1). The -UM and U-M patterns (8% of total) also indicate very low methylation as they

Table 1. Methylation patterns for 13367 TE sequences with CG, CHG and CHH sites

	MMM	UUU	UUM	MUM	MUU	UMU	MMU	UMM
No. TEs	7983	2687	768	554	469	453	449	304
Percentage TEs	0.58	0.20	0.06	0.04	0.03	0.03	0.03	0.02
Nb TEs in Heterochromatin	5802	591	205	279	175	113	210	82
Nb TEs in Euchromatin	2181	2096	563	275	294	340	239	222
Average size (bp)	887	443	396	443	591	635	679	488
Average size in Heterochromatin	929	398	332	453	589	583	612	440
Average size in Euchromatin	777	456	419	433	593	652	738	506
Average distance*	0.24	0.27	0.27	0.26	0.26	0.27	0.26	0.27
Median Frequency of methylated sites								
CG	1.00	0.00	0.00	0.75	0.38	0.00	0.60	0.00
CHG	0.90	0.00	0.00	0.00	0.00	0.17	0.29	0.25
CHH	0.36	0.00	0.08	0.12	0.02	0.02	0.03	0.10
Median level of methylation (methylated reads/total reads)								
CG	0.82	0.00	0.00	0.40	0.12	0.00	0.37	0.00
CHG	0.28	0.00	0.00	0.00	0.00	0.01	0.03	0.02
CHH	0.06	0.00	0.01	0.01	0.00	0.00	0.00	0.01

*Jukes–Cantor distance from reference sequence.

are characterized by low median frequencies of methylated CHH sites (12–13%). In contrast, the −M pattern (20% of total) is characterized by a much higher median frequency of methylated sites (33%), close to that of the MMM pattern (36%). Similarly, the other four patterns (-MM, -MU, M-M, M-U, 30% of total) have median frequencies of methylated sites comparable to those of the MMM pattern. Thus, whereas half of the 3418 TE sequences with no CG or CHG sites are densely methylated, the other half have no or very low methylation, which is twice the fraction of TE sequences with no or very low methylation among those containing all three types of sites. This latter result indicates therefore a critical role for CG and CHG sites in dictating methylation of TE sequences.

Arabidopsis TE sequences can be classified into 13 superfamilies, four corresponding to retroelements (Copia, Gyspy, LINE and SINE), five to well-defined DNA transposons (En-Spm, Harbinger, HAT, MuDR and Pogo) and one each to Helitrons, TEs of a composite nature, Tc1/mariner and other DNA transposons. As shown in Figure 4, the Gypsy and /En-Spm superfamilies have the highest proportion (~90%) of methylated TE sequences, and RC/Helitrons and Tc1/mariner superfamilies the lowest such fraction (40–50%).

Genomic distribution of TE sequences

The 119 Mb of available Arabidopsis genome sequence can be divided into gene-rich/TE-poor and gene-poor/TE-rich regions that form the euchromatic arms of chromosomes and pericentromeric heterochromatin plus interstitial heterochromatic knobs, respectively (4,21,27). Our analysis indicates that more than two thirds of densely methylated TE sequences (MMM, -MM, M-M and −M) and a similar proportion of unmethylated or poorly methylated TE sequences are located within pericentromeric heterochromatin and euchromatin, respectively (Table 1 and Supplementary Table S1). Furthermore, the last two categories of TE sequences correspond mainly to TE relics depleted in CpGs, as indicated by their shorter size compared to their densely methylated counterparts and their higher divergence from the cognate reference TE sequence (Figure 5). Thus, the dense DNA methylation characteristic of heterochromatin results not only from the much higher density of TE sequences compared to euchromatin, but also from the larger ratio of methylated to unmethylated TE sequences within heterochromatin and the longer length of methylated TE sequences on average.

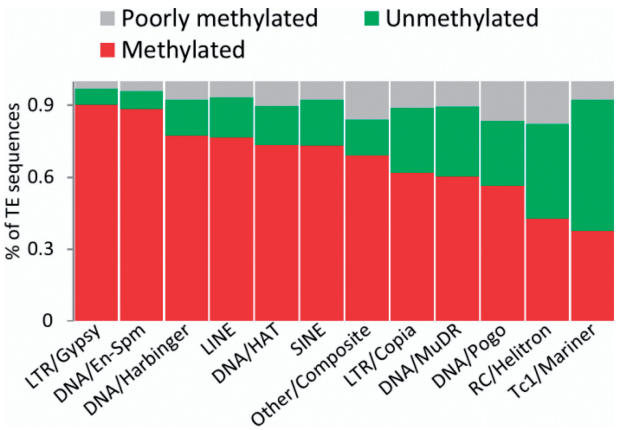

Figure 4. DNA methylation patterns within TE superfamilies. Unmethylated TE sequences are found across all classes but >90% of the sequences for LTR/Gypsy and DNA/En-Spm superfamilies are methylated. The RC/Helitron and Tc1/mariner superfamilies comprise the largest fraction (50–60%) of unmethylated TE sequences.

Although the vast majority of TE sequences are located outside of genes and cluster within heterochromatin, 17% of euchromatic TE sequences intersect with gene annotations (Table 2). Furthermore, most of these TE sequences overlay with exons, suggesting a high incidence of 'exonization' of TE sequences in Arabidopsis. Finally, whereas 53% of these exonic TE sequences are unmethylated (UUU + UUM), 24% are highly methylated (MMM), suggesting a recent origin.

TE sequences and siRNAs

Deep sequencing of small RNAs has revealed that a large fraction of methylated repeat elements present in the Arabidopsis genome are characterized by an abundance of matching 24-nt siRNAs throughout development (6). To investigate more precisely the association of different DNA methylation patterns of TE sequences with endogenous 24-nt-long siRNAs, small RNA deep sequencing data obtained from Arabidopsis whole-aerial tissues were downloaded from GEO (accession: GSE14696) (24) and used to calculate the 24-nt siRNA density for all TE sequences with defined DNA-methylation patterns (see 'Materials and Methods' section). As expected, almost all (95%) unmethylated TE sequences and many (48–58%) of those poorly methylated are devoid of matching 24-nt siRNAs, irrespective of their location in euchromatin or heterochromatin. On the other hand, 89 and 82% of densely methylated TE sequences located respectively in euchromatin and heterochromatin have matching 24-nt siRNAs (Figure 6a and b). This result confirms that most densely methylated TE sequences are associated not only with siRNAs, but also indicates a lower proportion of such sequences in heterochromatin. Indeed, differences between euchromatic and heterochromatic methylated TE sequences were even more pronounced when considering the proportion of those having an abundance of matching siRNAs (density >0.25 reads/kb/10^5 library reads), which is 61% in euchromatin but only 21% in heterochromatin.

Given the high density of TE sequences in heterochromatin, we explored the possibility that methylation of heterochromatic TE sequences with no matching 24-nt siRNAs could occur through local spreading from flanking siRNA-targeted sequences. To this end we considered the set of 749 heterochromatic MMM TE sequences longer than 200 bp and with no matching 24-nt siRNAs but flanked within 1 kb on one or both sides by sequences associated with siRNAs. Each TE sequence was split in equal halves and DNA methylation densities were calculated in non-overlapping 100-bp windows for

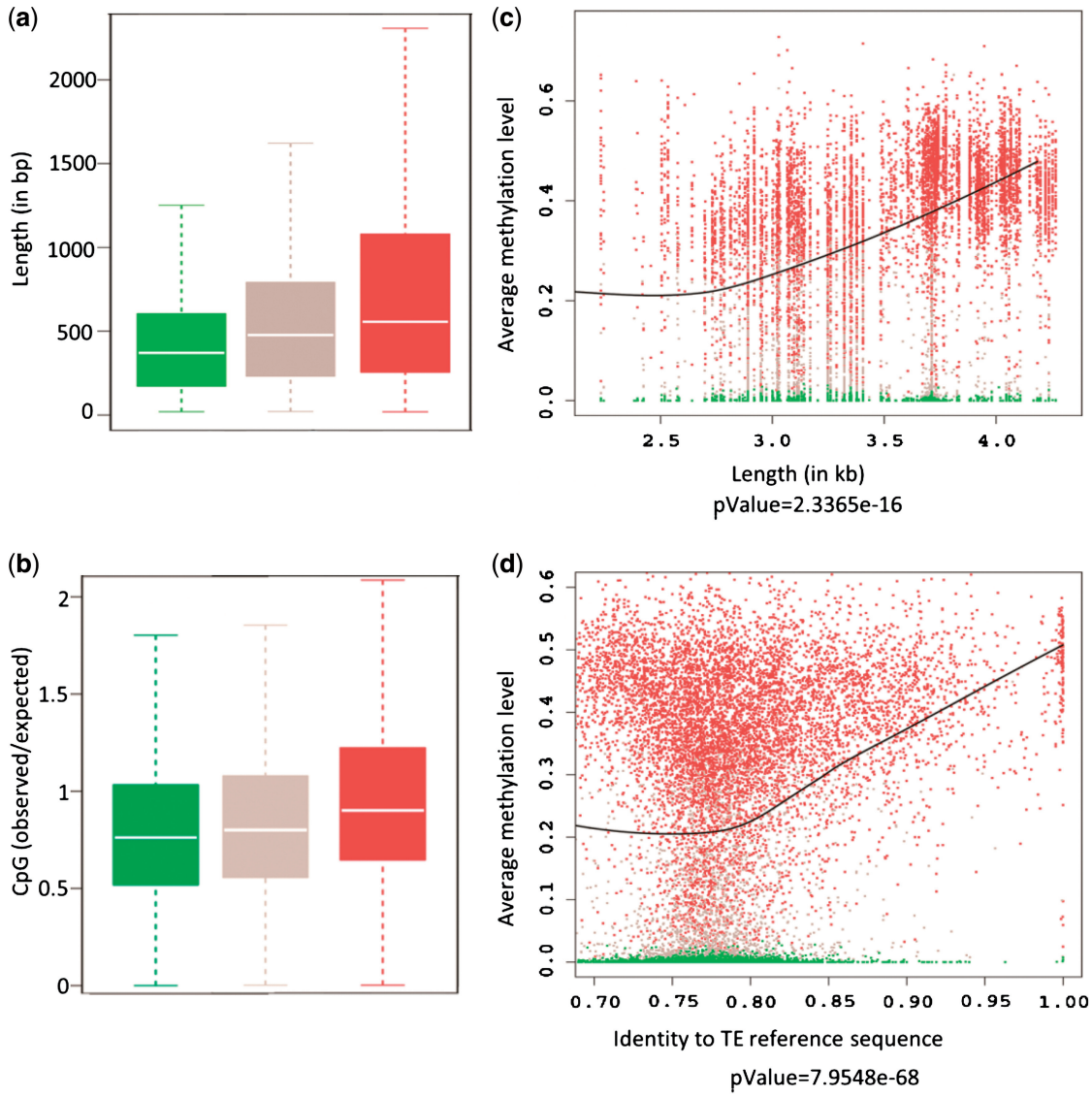

Figure 5. Relationships between DNA methylation, size, CpG content and divergence of TE sequences. Color code is as in Figure 4. (**a**) Unmethylated TE sequences tend to be smaller than their methylated counterparts. (**b**) Boxplots showing observed versus expected CpGs for the three DNA methylation patterns considered. Unmethylated TE sequences are depleted in CpGs compared to poorly methylated TEs (P-value = 0.004793, Wilcoxon rank-sum test) or methylated TEs (P-value $< 1e-10$). Poorly methylated TE sequences also have a lower CG content compared to methylated TE sequences (P-value = $5.369e-13$). (**c** and **d**) Average methylation levels of TE sequences are plotted according to length or percentage of identity with the TE reference sequence. Significant positive correlation (black curve) is observed in each case.

Table 2. TE sequences within genes

Methylation pattern	Number of TEs	Percentage pattern	Intronic	Exonic	Percentage intronic	Percentage exonic
UUU	456	0.44	116	336	0.26	0.74
MMM	249	0.24	59	189	0.24	0.76
UUM	105	0.10	22	82	0.21	0.79
UMU	64	0.06	11	53	0.17	0.83
MUU	54	0.05	12	42	0.22	0.78
MUM	48	0.05	11	37	0.23	0.77
MMU	41	0.04	10	31	0.24	0.76
UMM	30	0.03	5	25	0.17	0.83

each half and its corresponding siRNA-associated flank. As shown in Figure 6c, median DNA methylation densities are uniformly high along the 1 kb flanks, but decrease progressively within the first 500 bp of TE sequences with no matching siRNAs. Correspondingly, the 5371 siRNA-associated MMM TE sequences show higher DNA methylation than their flanks, which may or may not have matching siRNAs (Figure 6d). Taken together, these findings provide strong evidence that DNA methylation can spread over short distances (~500 bp) from siRNA-targeted regions into flanking sequences. Furthermore, analysis of additional methylomes (6) reveals that DNA methylation gradients are abolished in plants defective for the CG maintenance methyltransferase MET1 but are still detectable in plants

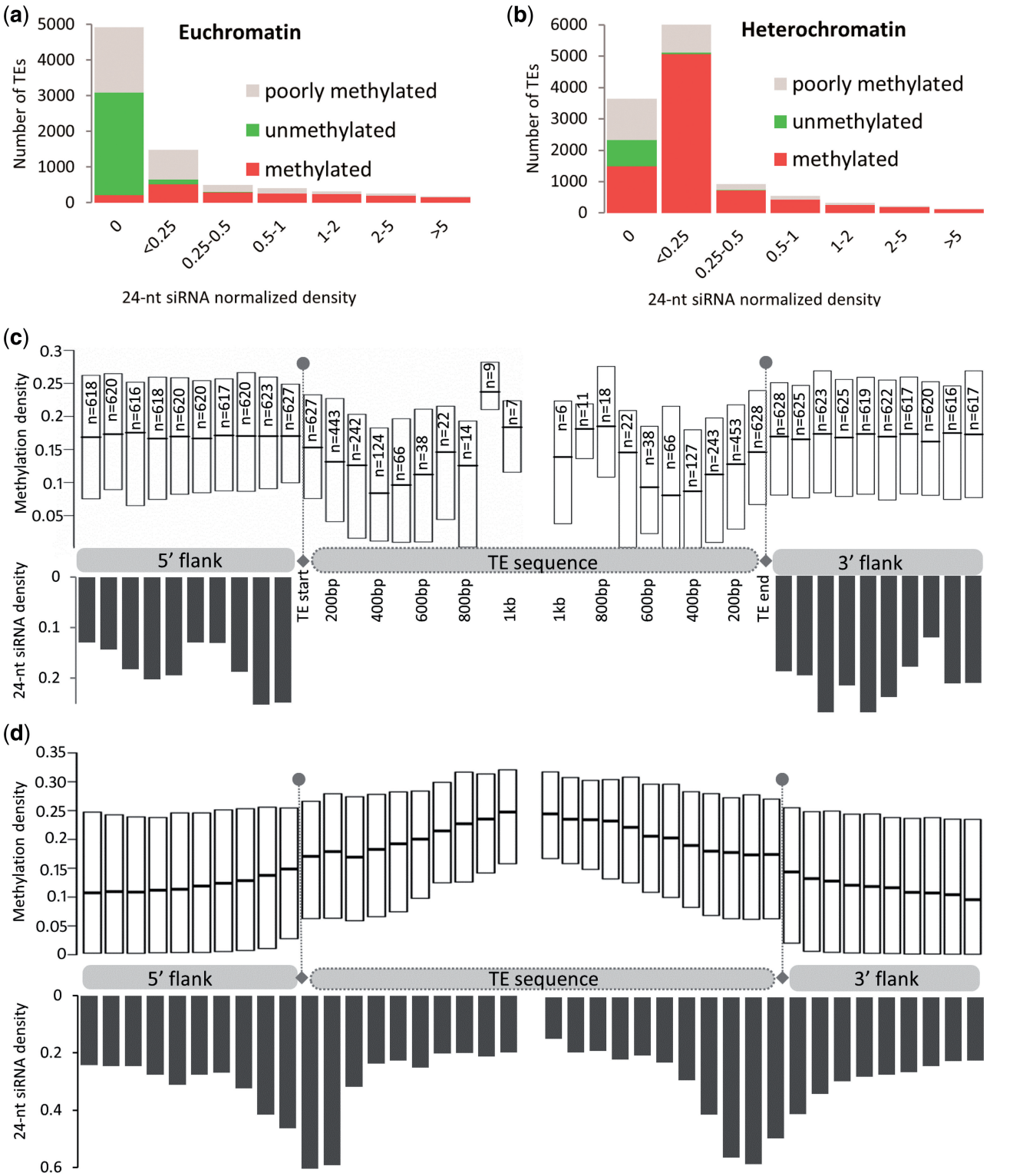

Figure 6. Relationship between methylated TE sequences and 24-nt siRNAs. (**a**) Methylated euchromatic TE sequences are almost always associated with an abundance of siRNAs. (**b**) A significant number of heterochromatic TE sequences are methylated but not associated with siRNAs (**c**) Methylated hetrochromatic TE sequences (size >200 bp) not associated with siRNAs but flanked within 1 kb on one or both sides by sequences associated with siRNAs. These TE sequences were split in half and DNA methylation densities were calculated in 100-bp windows along the two flanks and TE sequence halves by dividing the number of reads indicative of methylation at CG, CHG and CHH sites by the total number of cytosine-covering reads. Results are shown as boxplots of DNA methylation densities. Average normalized siRNA densities are also indicated for each 100-bp window. DNA methylation densities are uniform along the 1 kb flanks, but decrease progressively within the first 500 bp of TE sequences from both sides. (**d**) TE sequences associated with 24-nt siRNAs show increasing methylation from their extremities and decreasing methylation in their flanks.

that are simultaneously defective for the *de novo* DNA methyltransferases DRM1 and DRM2 and the CHG-specific DNA methyltransferase CMT3 [*drm1*, *drm2* and *cmt3* (*ddc*); Supplementary Figure S3a and b]. Nonetheless, the gradient in the *ddc* triple mutant is less steep than in wild type. Finally, plants defective for three of the four known Arabidopsis DNA demethylases [*ros1dml2dml3* or *rdd* triple mutant; (6)] display DNA methylation gradients similar to wild type (Supplementary Figure S3c). Taken together, these results rule out any significant contribution of active DNA demethylation to the gradients observed and suggest a complex set of interactions between different DNA methyltransferases in promoting or limiting DNA methylation spread.

To analyze further the local spreading of DNA methylation from siRNA-targeted TE sequences, methylation densities were plotted separately for CG, CHG and CHH sites (Figure 7). Although gradients are observed in wild type for the three types of sites, slopes are maximal for CHG, suggesting that CHG methylation spreads over shorter distances than CG and CHH methylation. Furthermore, CHG methylation is completely abolished both in the flanks and within TE sequences in the *ddc* triple mutant (Figure 7), suggesting that at least in this background the residual CG and CHH methylation gradients are contributed by DNA methyltransferases other than DRM1, DRM2 and CMT3. These results, together with the absence of any discernible methylation gradient for CHG and CHH in *met1* (Supplementary Figure S4), provide additional evidence that the extent of DNA methylation spread results from complex interactions between different DNA methyltransferases.

Whereas most MMM TE sequences with no matching siRNAs show decreasing DNA methylation towards their middle, uniform DNA methylation across the entire length is observed for some large TE sequences. This suggests either a more extensive spreading of DNA methylation in these cases or the existence of DNA methylation mechanisms not associated, directly or indirectly, with siRNAs. In agreement with the latter hypothesis, the few euchromatic MMM TE sequences

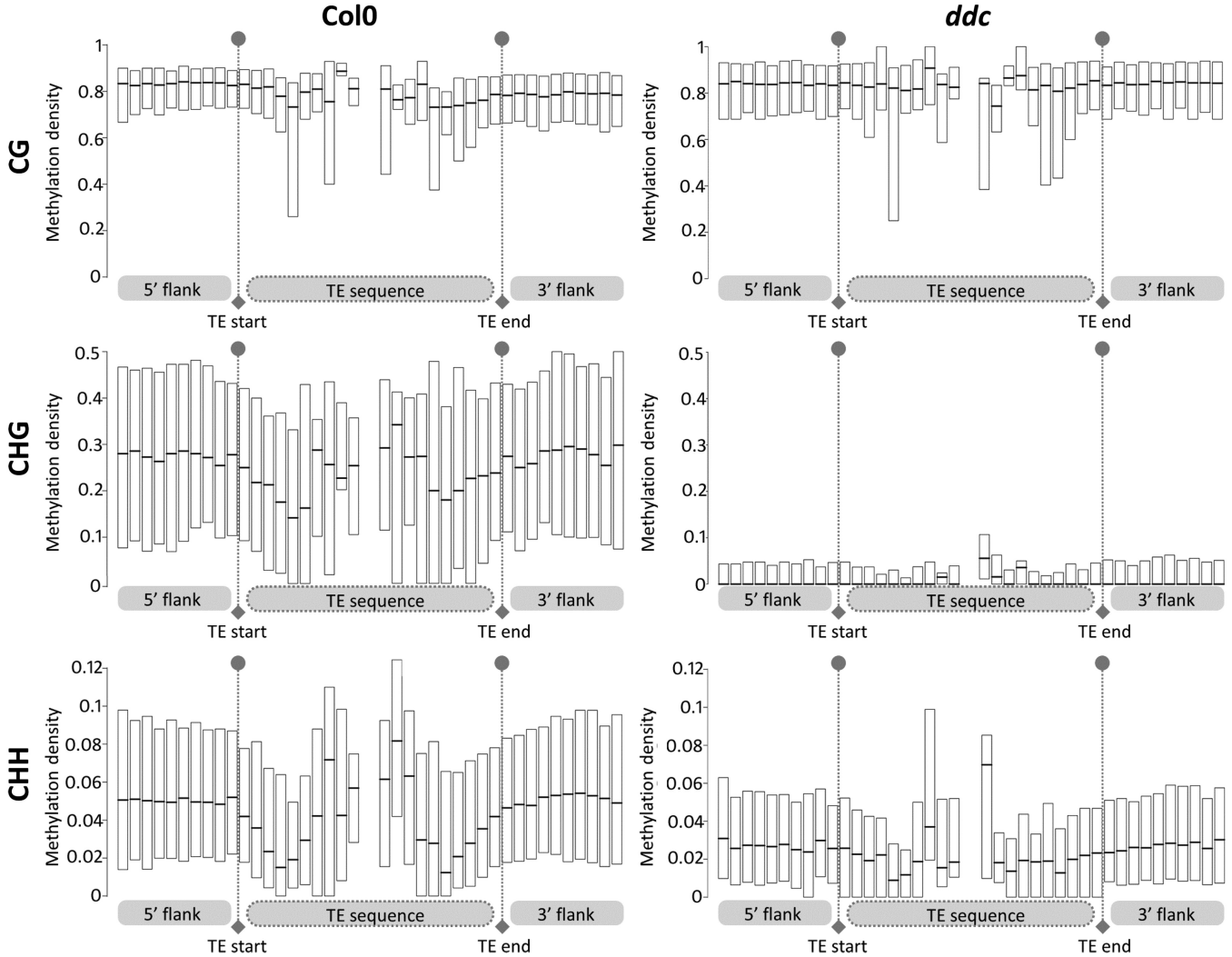

Figure 7. Analysis of methylation spreading for CG, CHG and CHH sites. The figures in the first and second columns correspond to wild type (Col0) and the *drm1*, *drm2*, *cmt3* triple mutant (*ddc*), respectively.

with no matching siRNAs tend to have uniform DNA methylation density throughout their length and are typically flanked by unmethylated sequences (data not shown).

TE sequences and flanking genes

It was previously shown that TE sequences tend to be less methylated when located close to genes, presumably because of deleterious effects of TE methylation on the expression of neighboring genes (28). The proportion of unmethylated TE sequences was reported to drop from ~55% within genes to below 20% for the first 500 bp window away from genes, with little further decrease beyond this point. However, this analysis did not distinguish euchromatic from heterochromatiic genes (28), which are characterized by dramatically distinct intergenic regions (short and TE-poor versus long and TE-rich, respectively). This prompted us to explore further the underrepresentation of methylated TE sequences near genes using our extended dataset and only considering genes within euchromatin. To this end, methylated and unmethylated TE sequences were scored in 100-bp windows for a distance of up to 1 kb upstream and downstream of genes. Our analysis reveals that in euchromatin, both methylated and unmethylated TE sequences tend in fact to over accumulate close to the 5′- and 3′-ends of genes (Figure 8a). Moreover, although the ratio of unmethylated versus methylated TE sequences drops with distance away from genes, as previously reported (28), this drop is rather limited (60% to a minimum of 40%), specific to the 5′-end of genes and less discernible when considering only methylated TE sequences with matching siRNAs, which are the least abundant overall (Figure 8b and Supplementary Table S2). These results suggest therefore that methylated TE sequences have more deleterious effects on transcription initiation than termination and that these effects are more severe when methylated TE sequences have matching siRNAs.

We next tested if spreading of DNA methylation from siRNA-targeted TE sequences could provide a plausible explanation for the deleterious effects of TE methylation on gene expression. For this, DNA methylation densities were calculated in non-overlapping 100-bp windows for all methylated euchromatic TE sequences ($n = 401$) associated with 24-nt siRNAs and flanked by sequences not associated with siRNAs. Although spreading is less pronounced than in heterochromatin, it is nonetheless clearly detectable over ~200 bp beyond siRNA-targeted sequences. Moreover, our analysis suggests that DNA methylation spreads from the center of euchromatic TE sequences towards their extremities, which are often not associated with siRNAs (Supplementary Figure S5) unlike their heterochromatic counterparts (Figure 6d).

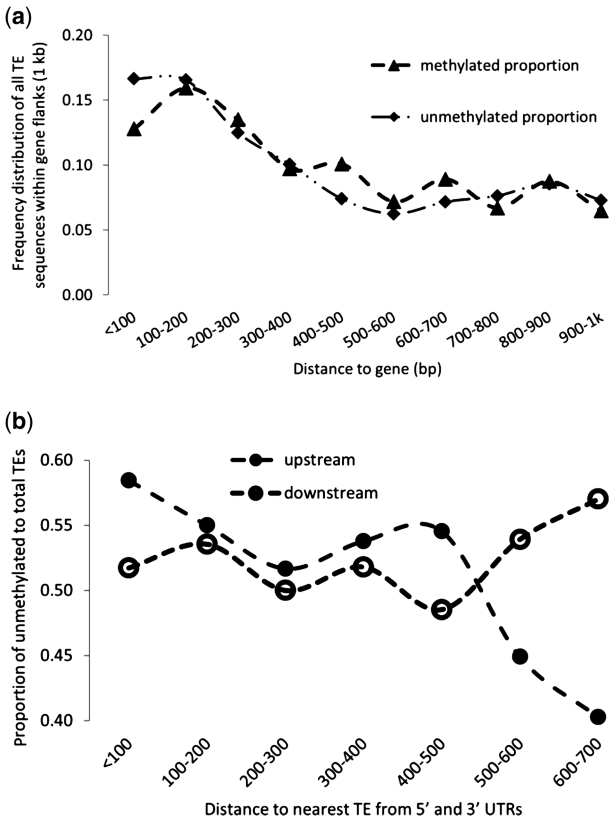

Figure 8. Distance between TE sequences and genes in euchromatin. (**a**) Both methylated and unmethylated TE sequences tend to accumulate close to genes. Note that because results do not substantially differ for the 5′- and 3′-ends of genes, they are not distinguished in the figure. (**b**) The proportion of unmethylated to total TE sequences drops slightly farther away from the 5′-end of genes. No similar drop is observed from the 3′-end of genes. Only the TE sequence closest to the start or stop codon was considered for this analysis.

DISCUSSION

Using a refined version of our previous annotation pipeline, we have obtained the most extensive dataset for TE sequences in the Arabidopsis genome to date. Given the high sensitivity and specificity of this pipeline, this dataset is not expected to evolve substantially in the future. Similarly, the use of stringent criteria for the analysis of DNA methylation of TE sequences makes our conclusions particularly robust, and should facilitate comparison of DNA methylation patterns between different conditions as well as between Arabidopsis accessions. Furthermore, the methods used to determine the DNA methylation status of individual TE sequences based on bisulphite-sequencing data are general and can be implemented to the systematic analysis of the association between DNA methylation and any annotated features of genomes for which single-nucleotide resolution methylomes are available (29).

Based on the 17 085 TE sequences (out of a total of 31 245) for which DNA methylation could be examined with high precision, we have found that 26% are unmethylated and another 15% have methylation patterns that depart significantly from the dense CG, CHG and CHH methylation typically reported. These two categories of TE sequences mainly correspond to short and highly degenerate relics located in euchromatin, many of which are missed by less sensitive detection pipelines. These relics

also tend to be depleted in CG sites, suggesting an important function for these sites in determining DNA methylation density. Thus we can propose a scenario in which TE sequences progressively lose CG sites because of their higher methylation levels compared to CHG and CHH sites and because of the higher mutability of methylcytosines compared to cytosines. This progressive loss would in turn reduce the potential for the affected sequences to perpetuate methylation at CHG and CHH sites, leading ultimately to complete loss of DNA methylation.

While our analysis confirms the association of siRNAs with DNA methylation over TE sequences, an unexpectedly high number of densely methylated TE sequences are characterized by the absence or near absence of matching siRNAs. Such TE sequences are preferentially found in heterochromatin. We have shown that in these cases, DNA methylation most likely results from local spreading (within 500 bp) from flanking, siRNA-targeted sequences (Figure 6c). We have also provided evidence that DNA methylation spread occurs in euchromatin as well, but that the extent of spreading is more limited than in heterochromatin (~200 bp versus ~500 bp; Figure 6c and Supplementary Figure S5). This could reflect either a facilitating role of heterochromatin, an inhibitory effect of euchromatin, or, as reported previously (30), a higher DNA demethylation activity in euchromatin. The spreading phenomenon we have uncovered here appears distinct from so-called secondary RdDM, which is caused by the biogenesis of secondary siRNAs from sequences adjoining those initially targeted by RdDM (31–34). Furthermore, the persistence of DNA methylation gradients for CG and CHH sites in the *ddc* triple mutant background (Figure 7) argues against an important role for RdDM in DNA methylation spreading. However, we cannot rule out that RdDM is involved, notably during the reproductive phase when it is most active (3), and that spreading of DNA methylation is maintained by MET1 and/or other DNA methyltransferases independently of RdDM during plant growth.

Finally, our study indicates that TE sequences present in euchromatin are more abundant closer to genes than away from them. This pattern is observed both upstream and downstream of genes, which could reflect a preference for TEs to insert in 'open' chromatin. Indeed, preferential insertion close to or within genes has been noted for several TE families in maize and rice (35), even though such events are unlikely to be maintained over evolutionary timescales because of their high potential to be deleterious. We have also shown that methylated TE sequences are slightly underrepresented compared to their unmethylated counterparts close to the 5′-end of genes and that methylated TE sequences with matching siRNAs are least abundant and somewhat more uniformly distributed within the 5′-end of genes than methylated sequences with no matching siRNAs (Supplementary Table S2). Given the known inhibitory effect of DNA methylation on promoter activity, it is therefore reasonable to speculate that DNA methylation spread contributes significantly to the negative impact of methylated TE sequences on neighboring gene expression. In support of this view, both in *A. thaliana* and *A. lyrata*, genes that are located <500 bp away from TE sequences tend to be expressed at lower levels than genes further away, and this reduction in gene expression is more pronounced when the TE sequences have matching siRNAs (8).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank David Swarbreck for valuable insights on TE annotation, Eva Huala for helping us to access The Arabidopsis Information Ressource, François Roudier for critical reading of the manuscript and members of the Colot group for discussions.

FUNDING

Agence Nationale de la Recherche (ANR) 'DDB1 project' (to C.B. and V.C., in part); Centre National de la Recherche Scientifique (CNRS) 'Groupement de Recherche Elements Transposables' (to V.C and H.Q., in part). PhD studentships from ANR and CNRS (I.A. and A.S., respectively). Funding for open access charges: CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Slotkin,R.K. and Martienssen,R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, **8**, 272–285.
2. Law,J.A. and Jacobsen,S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.*, **11**, 204–220.
3. Teixeira,F.K. and Colot,V. (2010) Repeat elements and the Arabidopsis DNA methylation landscape. *Heredity*, **105**, 14–23.
4. The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
5. Cokus,S.J., Feng,S., Zhang,X., Chen,Z., Merriman,B., Haudenschild,C.D., Pradhan,S., Nelson,S.F., Pellegrini,M. and Jacobsen,S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.
6. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.
7. Buisine,N., Quesneville,H. and Colot,V. (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*, **91**, 467–475.
8. Hollister,J.D., Smith,L.M., Guo,Y.L., Ott,F., Weigel,D. and Gaut,B.S. (2011) Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc. Natl Acad. Sci. USA*, **108**, 2322–2327.
9. Quesneville,H., Bergman,C.M., Andrieu,O., Autard,D., Nouaud,D., Ashburner,M. and Anxolabehere,D. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.*, **1**, 166–175.

10. Quesneville,H., Nouaud,D. and Anxolabehere,D. (2003) Detection of new transposable element families in *Drosophila melanogaster* and Anopheles gambiae genomes. *J. Mol. Evol.*, **57(Suppl. 1)**, S50–S59.
11. Smit,A.F.A., Hubley,R. and Green,P. (1996–2004) Institute for Systems Biology.
12. Jurka,J., Klonowski,P., Dagman,V. and Pelton,P. (1996) CENSOR–a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.*, **20**, 119–121.
13. Kohany,O., Gentles,A.J., Hankus,L. and Jurka,J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, **7**, 474.
14. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
15. Kolpakov,R., Bana,G. and Kucherov,G. (2003) mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.
16. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
17. Bedell,J.A., Korf,I. and Gish,W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**, 1040–1041.
18. Chao,K.M., Zhang,J., Ostell,J. and Miller,W. (1995) A local alignment tool for very long DNA sequences. *Comput. Appl. Biosci.*, **11**, 147–153.
19. Gusfield,D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York, NY.
20. Kapitonov,V. and Jurka,J. (1996) The age of Alu subfamilies. *J. Mol. Evol.*, **42**, 59–65.
21. 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.
22. Zilberman,D., Gehring,M., Tran,R.K., Ballinger,T. and Henikoff,S. (2007) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.*, **39**, 61–69.
23. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
24. Fahlgren,N., Sullivan,C.M., Kasschau,K.D., Chapman,E.J., Cumbie,J.S., Montgomery,T.A., Gilbert,S.D., Dasenko,M., Backman,T.W., Givan,S.A. *et al.* (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA*, **15**, 992–1002.
25. Miura,A., Nakamura,M., Inagaki,S., Kobayashi,A., Saze,H. and Kakutani,T. (2009) An Arabidopsis jmjC domain protein protects transcribed genes from DNA methylation at CHG sites. *EMBO J.*, **28**, 1078–1086.
26. Lister,R. and Ecker,J.R. (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res.*, **19**, 959–966.
27. Bernatavichute,Y.V., Zhang,X., Cokus,S., Pellegrini,M. and Jacobsen,S.E. (2008) Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in *Arabidopsis thaliana*. *PLoS One*, **3**, e3156.
28. Hollister,J.D. and Gaut,B.S. (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.*, **19**, 1419–1428.
29. Pelizzola,M. and Ecker,J.R. (2010) The DNA methylome. *FEBS Lett.*
30. Penterman,J., Zilberman,D., Huh,J.H., Ballinger,T., Henikoff,S. and Fischer,R.L. (2007) DNA demethylation in the Arabidopsis genome. *Proc. Natl Acad. Sci. USA*, **104**, 6752–6757.
31. Kanno,T., Bucher,E., Daxinger,L., Huettel,B., Bohmdorfer,G., Gregor,W., Kreil,D.P., Matzke,M. and Matzke,A.J. (2008) A structural-maintenance-of-chromosomes hinge domain-containing protein is required for RNA-directed DNA methylation. *Nat. Genet.*, **40**, 670–675.
32. Henderson,I.R. and Jacobsen,S.E. (2008) Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. *Genes Dev.*, **22**, 1597–1606.
33. Daxinger,L., Kanno,T., Bucher,E., van der Winden,J., Naumann,U., Matzke,A.J. and Matzke,M. (2009) A stepwise pathway for biogenesis of 24-nt secondary siRNAs and spreading of DNA methylation. *EMBO J.*, **28**, 48–57.
34. Saze,H. and Kakutani,T. (2007) Heritable epigenetic mutation of a transposon-flanked Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1. *EMBO J.*, **26**, 3641–3652.
35. Dooner,H.K. and Weil,C.F. (2007) Give-and-take: interactions between DNA transposons and their host plant genomes. *Curr. Opin. Genet. Dev.*, **17**, 486–492.