[Epigenetics 4:2, 119-124; 16 February 2009]; ©2009 Landes Bioscience

Research Paper

Determining the conservation of DNA methylation in Arabidopsis

Nicolas Widman, Steven E. Jacobsen and Matteo Pellegrini*

Department of Molecular; Cell and Developmental Biology; University of California, Los Angeles; Los Angeles, CA USA

Key words: arabidopsis, sequence conservation, sequence divergence, DNA methylation, duplicated genes, epigenetics, repetitive sequences

A high-resolution map of DNA methylation in Arabidopsis has recently been generated using high-throughput sequencing of bisulfite-converted DNA. This detailed profile measures the methylation state of most of the cytosines in the Arabidopsis genome, and allows us for the first time to address questions regarding the conservation of methylation across duplicated regions of the genome. To address these questions we measured the degree to which methylation is conserved in both duplicated genes and duplicated non-coding regions of the genome. Methylation is controlled by different mechanisms and methyltransferases depending on the genomic location. Methylation in genes occurs primarily at CG sites and is controlled by the maintenance methyltransferase MET1. In contrast, an RNAi mediated methylation pathway that leads to de novo methylation of asymmetric CHH sites along with CG and CHG sites by the methyltransferase DRM2, drives methylation at tandem and inverted repeats. We find that the cytosine methylation profile is strongly preserved between duplicated genes and repeat regions. The highest level of conservation can be found at CG sites in genes and CHH sites in repeat regions. By constructing substitution matrices between aligned genes we see that methylated cytosines often pair with thymines, which may be explained by the spontaneous deamination of methyl-cytosine to thymine. Despite this observation, we find that methylated cytosines are less often paired with other nucleotides than non-methylated cytosines within gene bodies indicating that they may play an important functional role.

Introduction

High-throughput DNA sequencing of bisulfite converted DNA using next-generation sequencers (BS-seq) has recently been used to determine the methylation state of nearly all the cytosines in the plant *Arabidopsis thaliana*.[1,2] Bisulfite sequencing makes it possible to measure cytosine methylation at individual sites, in contrast to immunoprecipitation-based methodologies that use microarrays to measure methylation of large fragments of several hundred nucleotides.[3,4] Furthermore, using the BS-seq approach methylation can be measured in repetitive parts of the genome that are not usually included in tiling arrays.

*Correspondence to: Matteo Pellegrini; University of California, Los Angeles; Molecular, Cell and Developmental Biology; 621 Charles E. Young Dr. South; Los Angeles, CA 90095-1606 USA; Email: matteop@mcdb.ucla.edu

Submitted: 11/21/08; Accepted: 02/16/09

Previously published online as an *Epigenetics* E-publication:
http://www.landesbioscience.com/journals/epigenetics/article/8214

In mammalian genomes only cytosines that are followed by guanines (CpGs) tend to be methylated by the enzymes DNMT1 and DNMT3. In contrast plant genomes contain three methyltransferases, MET1, CMT3 and DRM2 that are capable of methylating CG, CHG and CHH sites, respectively. MET1 is a maintenance DNA methyltransferase that methylates hemimethylated CG sites during replication. DRM2 is part of an RNAi mediated pathway that performs de novo methylation of CG, CHG and CHH sites but shows a preference for CHH sites. Finally, CMT3 methylates CHG sites that tend to be associated with a particular histone mark, dimethylated histone 3 lysine 9.

Previous studies have shown that the bodies of protein-coding genes tend to only be methylated at CG sites. In contrast, repetitive regions of the genome, as well as transposons and heterochromatic regions tend to be methylated at all three types of sites. Overall, the level of CG methylation in the Arabidopsis genome is approximately 24%, while CHG and CHH methylation occur at 7% and 2% respectively. It was also found that while CG cytosines are either fully methylated or fully unmethylated, the other sites typically show only fractional methylation levels indicating that their methylation state differs across tissues in the plant and individual cells.

The purpose of this study is to identify the degree of conservation of DNA methylation in the Arabidopsis genome. To accomplish this we measured the degree of conservation of cytosine methylation in duplicated regions of the genome. Some of these regions involve ancient duplication of genes while others reflect more recent duplication of non-coding regions of the genome. There are several mechanisms that may lead to loss of conservation of methylation in duplicated regions. First of all, methylation patterns are known to vary across different tissues in plants indicating that this epigenetic mark is inherently more variable than the genetic code itself. Moreover, the deamination of methyl-cytosines can lead to a progressive loss of methylation over time. Previous studies have found that the half-life of methyl-cytosine in double-stranded DNA at 37 degrees Celsius is approximately 30,000 years.[5] Finally, in plants a great deal of methylation is deposited by de novo methylation pathways that depend on siRNA production at specific loci, and the degree to which these pathways depend on the underlying sequence is not known.

In this study, the methylation of the entire above ground portion of the plant is measured and as a result the data represents the average methylation levels of shoots and no tissue-specific methylation was measured. We set out to measure the conservation of methylation in repeated or duplicated regions in the genome since these allow us to estimate the degree to which methylation is conserved across

similar sequences of the genome. Most of the duplicated genes in the Arabidopsis genome are due to the most recent polyploidy event that occurred between 40 and 80 million years ago.[6,7] These duplicated genes have diverged in function and expression levels due to mutations. In the Arabidopsis genome there were 984 pairs of duplicated genes found using an all-by-all sequence similarity search.[8] Approximately 500 of the most functionally divergent of these were used in determining the level of methylation conservation in Arabidopsis. Functional divergence was evaluated by comparing the level of synonymous divergence (nucleotide substitutions resulting in the same peptide sequence) against non-synonymous divergence. Sufficient functional divergence was determined by synonymous divergence being below a threshold determined by the non-synonymous divergence of a gene.

```
TGGAATTTGGAGCTGTTTcGAGCTGAGTTCTTAcGGAGTTCAACAcGAGAGCACCATCAC
|||||| ||| | ||||||| |||||||||| ||||| ||||||  ||||||||||||
TGGAATCTGGGGATGTTTcGGGCTGAGTTCTTGcGGAGCTCAACACTAGAGCACCATCAC
```

Figure 1. Sample Alignment. A fragment of a sample alignment between duplicated genes. Unmethylated cytosines are shown as "C", with methyl-cytosine as "c." Red "c" and "C" represent cytosines with ≥ 5 read coverage, blue "C" represents cytosines with <5 read coverage. (Top: AT3G47910 5555-5614 bases downstream Bottom: AT3G47890 4962-5021 bases downstream).

We find that the methylation patterns in duplicated genes as well as in repeat regions show strong conservation within specific sequence contexts. In duplicated genes, methylation is conserved at CG positions and in repeat regions methylation is most conserved at CHH sites with CG and CHG methylation conserved to a lesser degree. One of the leading causes of methylation divergence in these regions appears to be the deamination of methyl-cytosine to thymine. Additionally, cytosine-thymine substitutions have the highest log-odds of all nucleotide substitutions between different bases, most likely as a result of methyl-cytosine deamination to thymines, as well as the spontaneous deamination of unmethylated cytosines to uracils. Repeat regions have a higher degree of sequence conservation than duplicated genes due to the fact that the duplications in repeat regions are more recent as well as duplicated genes having a tendency to evolve as active genes and to lose sequence conservation as a result.

Results

In order to determine the degree of conservation of cytosine methylation, we used two sets of sequences. The first set is pairs of protein-coding genes (approximately 500) obtained from Ganko et al. 2007. These represent genes that duplicated at various points in the evolutionary history of Arabidopsis between 40 and 80 million years ago. The second set of sequences includes various repeat regions throughout the genome including tandem and inverted repeats. The tandem repeats were located using the program Tandem Repeat Finder[9] and the inverted repeats were found using Inverted Repeat Finder.[10] Since repeat regions are rarely contained within coding portions of genes, these two sets allow us to study methylation conservation in both coding and non-coding sequence contexts.

Solexa high-throughput sequencing of bisulfite converted DNA from plant shoots was used to obtain the methylation profiles.[1] For each cytosine position in the genome, the methylation estimate is obtained by counting the number of cytosines in reads that align to the genomic position and dividing by the total number of reads that align to that position. The data set has an average ten fold coverage of each cytosine and therefore when measuring the level of methylation, at each particular cytosine position, a minimum of five reads was required to consider the level of methylation.

The sequences were locally aligned using the Smith-Waterman algorithm and then for each aligned cytosine-cytosine position the methylation percentage of each pair was stored. A fragment of a small alignment is shown in Figure 1, indicating which cytosine is methylated and which is not. Cytosine positions without taking sequence context into account were considered methylated if at least 10% of reads were methylated. When considering sequence context, cytosines in a CG context were considered methylated when 80% of reads were methylated, for CHG 25% and for CHH 10%. We next constructed three arrays with these paired values, one for each of the three different cytosine types (CG, CHG and CHH), with the alignments concatenated. For the construction of these arrays we require that the two (for CG) or three bases (for CHG and CHH) of the sequence context must be perfectly conserved in the alignment. To estimate the conservation of methylation in these aligned sequences the Pearson correlations between the paired values of each gene or repeat was then computed.

To estimate the statistical significance of the observed correlation, we generated a random model by permuting one side of the data pairs for each set of sequence contexts separately. The data was permuted 100 times in order to obtain an estimate of the mean and variance of the distribution that was used to determine the z-score of the correlation in the aligned data pairs. In order to obtain a reliable z-score, only pairs of sequences that generated alignments with at least ten cytosines were used for each sequence context for both duplicated genes and inverted repeats. In the case of tandem repeats, alignments with as few as five cytosines were considered since tandem repeats tended to be much shorter in overall length.

We find that overall cytosine methylation patterns in Arabidopsis are strongly conserved with respect to our random model. As shown in Table 1, within duplicated genes, the CG methylation sites have a z-score of 10 and are therefore significantly conserved with respect to random permutations. In contrast we find that CHG and CHH sites show no significant conservation with z-scores less than 1. This result is consistent with the observation that the bodies of genes are methylated at about 30% of CG sites, while CHG and CHH sites are methylated less than 1% of the time (see Table 2).

We also asked whether the conservation of CG methylation in gene bodies is simply due to the conservation of methylation domains or whether the conservation is preserved at the level of single cytosines within these domains. To answer this question we first used the following criteria to define methylation domains: regions with at least six methylated CG sites within a window of 100 bases. We next asked whether the correlation between aligned CG sites in domains is more significant in the original sequences compared to sequences in which the CG sites have been randomly permuted. We found that the z-score for this correlation is approximately 3, indicating that the precise pattern of methylation in methylated domains of duplicated

Table 1 **Aligned sequence methylation correlation and Z-score**

	CG	Z-Score	CHG	Z-Score	CHH	Z-Score
Duplicated Genes	0.3634	10.7778	0.0104	0.2506	0.0045	1.0721
Unique Tandem Repeats	0.8570	0.4372	0.8137	2.9259	0.5016	8.7009
Tandem Repeats	0.8176	3.5681	0.7733	4.6258	0.4857	25.2883
Inverted Repeats	0.8458	4.9655	0.7807	4.7688	0.4354	24.3819

Table 2 **Average methylation levels of aligned cytosine positions**

	CG	CHG	CHH	Average
Duplicated Genes	19.97%	0.69%	0.39%	2.92%
Unique Tandem Repeats	39.38%	12.61%	2.71%	9.84%
Tandem Repeats	57.46%	20.85%	5.12%	14.74%
Inverted Repeats	52.93%	20.92%	5.75%	13.55%

genes is conserved although to a lesser degree than the overall CG methylation conservation in the entire gene. Therefore we conclude that the patterns of CG methylation we see in gene bodies are not simply due to the presence of methylated domains within the genes.

Our assumption is that CG methylation in the bodies of protein coding genes is maintained by the MET1 DNA methyltransferase during replication. Our results indicate that this process is in fact very stable and that similar methylation patterns persist in duplicated genes even after tens of millions of years of evolution. Furthermore, our results indicate that this conservation is maintained at the single cytosine level, and does not simply reflect the fact that gene bodies contained conserved methylated domains.

Repetitive regions of non-coding DNA show a distinctly different conservation pattern than that found in duplicated protein coding genes. As seen in Table 1, in tandem and inverted repeats CHH methylation sites are by far the most conserved with a z-score of 25. CG sites show the lowest level of conservation in these repetitive regions, with a low z-score in unique tandem repeats and only a modestly significant z-score in non-unique and inverted repeats. CHG sites show a level of conservation that is intermediate between CG and CHH sites in these regions.

We know that repetitive regions of the genome are targets for de novo methylation pathways via an RNAi mediated mechanism. The primary DNA methyltransferase implicated in this pathway is DRM2. It is known that while this enzyme is capable of methylating cytosines in all three sequence contexts, it has a distinct affinity for asymmetric CHH sites. Although only about 5% of CHH sites are methylated in these repeats (see Table 2), the methylation pattern of these is highly conserved in the two repeated sequences. The fact that methylation is constantly targeted at these sites, rather than copied during each replication cycle, and that methylation at these sites is driven by the repetitive nature of the underlying sequence might explain why the degree of conservation of CHH sites here is even stronger than the conservation of CG sites in duplicated genes.

We next performed a detailed study of the patterns of substitutions that occur at all sites along the duplicated sequences. Unlike traditional substitution analyses, here we keep track of the methylation states of cytosines and so are able to measure different substitution propensities as a function of methylation state.

We used log-odds matrices to determine the tendency for a given substitution to occur with respect to the null-hypothesis assumption of entirely random substitutions. A large positive value in the matrix indicates that the associated substitution is occurring more often than expected by chance, and therefore may show a strong selection pressure for this substitution in the alignments. However we note that other factors such as more efficient DNA repair mechanisms for some types of DNA damage (e.g., cytosine and/or methyl-cytosine deamination products), may limit the spontaneous rate of specific base changes, and thus explain part of the effects seen in log odds ratios as well.

The first log-odds matrix we computed was a nucleotide substitution matrix with the addition of methyl-cytosine as a fifth type of base. A cytosine was considered methylated if at least ten percent of the bisulfite-treated reads detected methylation. The matrix for nucleotide substitutions in duplicated genes is shown in Figure 2. We see from the fact that the diagonal of the matrix has positive entries that many aligned positions are conserved in the alignments, which is not surprising since we are aligning duplicated genes. The only positive off-diagonal entries are between methylated and unmethylated cytosines, indicating that there is a tendency for the methylation state of the cytosine to change over time. Nonetheless, as we have discussed above, the overall methylation profile of the gene is conserved to a statistically significant degree between these genes.

We also note that the most strongly conserved nucleotides are methylated cytosines. These are more strongly conserved than the other nucleotides in the five-letter log-odds matrix as well as in the four-letter log-odds matrix that combines methylated and unmethylated cytosines (Suppl. Fig. 1). This result indicates that despite the tendency for methylated cytosines to convert to either unmethylated cytosines or thymines through deamination, we find methylated cytosines to be paired in our alignments far more often than we expect by chance. Although the biological implication for this surprising conservation is not yet known, it may indicate that these sites are conserved because they play a particular functional role. Finally, we also note that although alignments between methylated cytosines and thymines (indicating deamination) are occurring less often than we expect by chance since the genes have not sufficiently diverged following their duplication, they are found more often than alignments between methylated cytosines and adenines or guanines. We note that the complimentary substitution (from G to A) that would occur on the opposite strand of the gene if methylated Cs convert to Ts, has a lower odds score than that of methylated Cs to Ts. The reason for this asymmetry is due to the fact that the frequencies of methylated Cs are different from those of As, since the frequencies of As are equal to the combined frequencies of methylated and

unmethylated cytosines, and this affects the log odds ratios which are a measure of the relative frequencies of paired nucleotides to the product of individual nucleotides.

In a second type of log-odds matrix we considered only cytosines with at least 5 read coverage and compared methylation levels. Six specific levels of methylation were considered in computing the log-odds matrix in 20 percent intervals from unmethylated to fully methylated. As seen in Figure 3, for duplicated genes the matrix shows that conservation is strongest among highly methylated cytosines. This is consistent with the known distribution of CG methylation in genes, which is effectively bimodal, with some sites being close to 100% methylated and others 0%.[1]

We also computed the same two types of log-odds matrices using the alignments between tandem and inverted repeats. The five-nucleotide substitution matrix is shown in Figure 4. In contrast to duplicated genes, here we see that methylated cytosines are as conserved as unmethylated cytosines. We also note that the substitution of methylated cytosines to thymines is not as significant as in duplicated genes. These two results are consistent with the notion that these duplicated regions are quite recent in comparison to gene duplication. The sequences have not diverged to the same extent, and cytosine deamination is less frequent in these alignments. In Figure 5, we show the substitution log odds ratios for the six different levels of methylated cytosines. Unlike the pattern seen in genes, here we see that even cytosines with relatively low levels of methylation are strongly conserved. This is consistent with the observation that non-CG sites are heavily methylated in theses regions and that these sites tend to be only partially methylated in our samples. Nonetheless, even cytosines with fractional levels of methylation are strongly conserved in the repeated segments. Separate matrices for the three different types of repeats are shown in the supplementary figures.

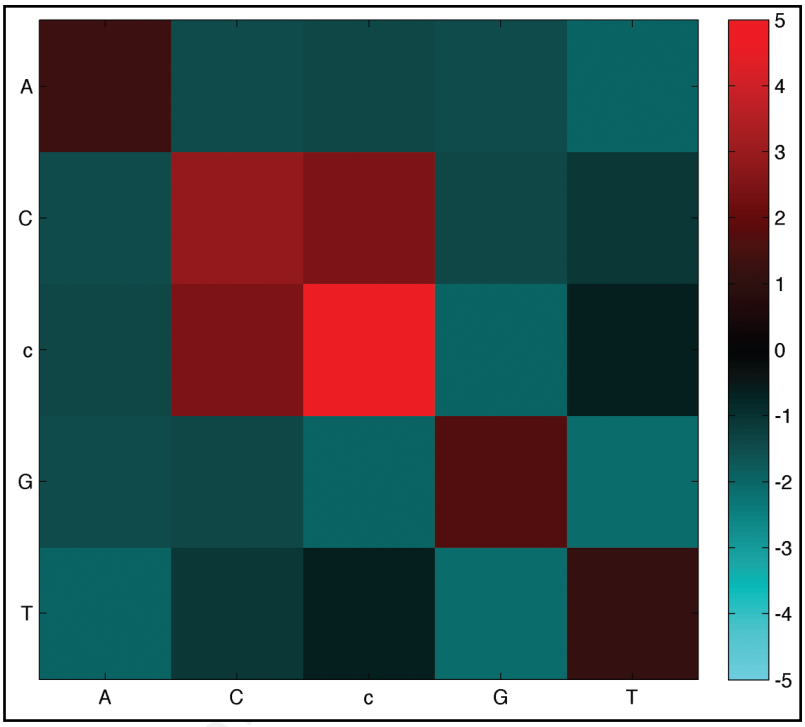

Figure 2. Log-odds ratio of nucleotide base transitions: duplicated genes. Log-odds substitution matrix of aligned duplicated gene bases (x axis and y axis) with methylated cytosine represented as a distinct base using lower-case c.

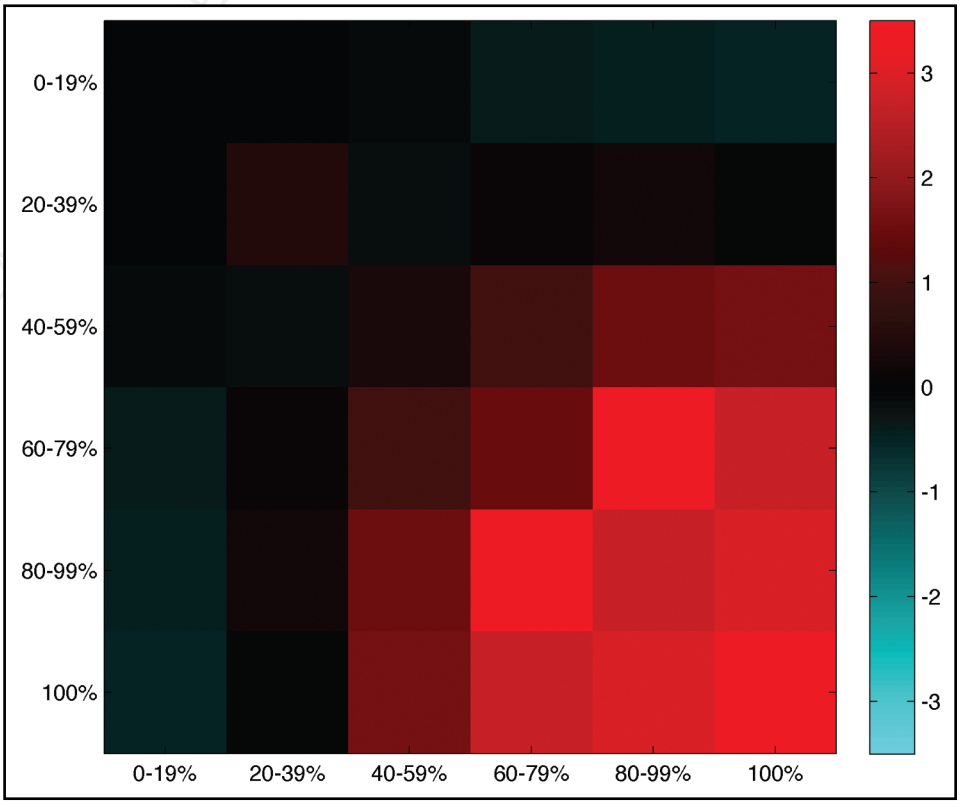

Figure 3. Log-odds ratio of c-methylation percentage transitions: duplicated genes. Log-odds substitution matrix of cytosine methylation levels (x axis and y axis) between pairs of aligned duplicated genes.

Discussion

By analyzing patterns of DNA methylation in duplicated regions of the genome we have been able to estimate the extent of conservation of methylated cytosines. In gene bodies, which are predominantly methylated at CG sites, we find that the degree of conservation of CG methylation is very significant with respect to a random model. In contrast, the low levels of CHH and CHG methylation found in gene bodies are not conserved in a statistically significant manner. This indicates that despite tens of millions of years of evolution, and the fact that the function of these genes has partially diverged, the methylation profile within the gene body is strongly preserved. This observation is in contrast to the notion that

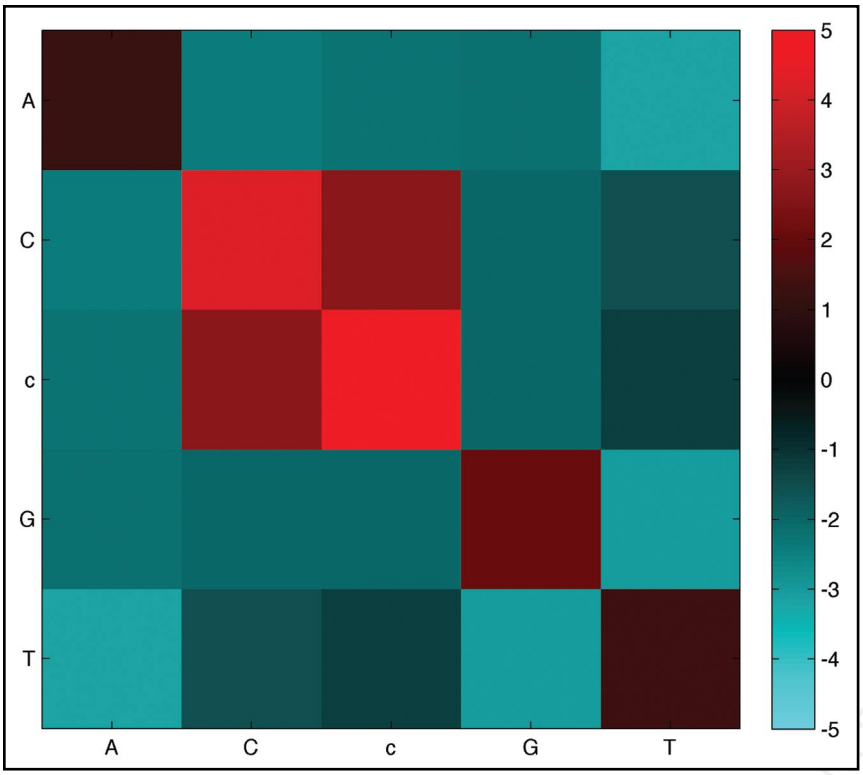

Figure 4. Log-odds ratio of nucleotide base transitions: repeat regions. Log-odds substitution matrix of repeat region bases (x axis and y axis) with methylated cytosine represented as a distinct base using lower-case c.

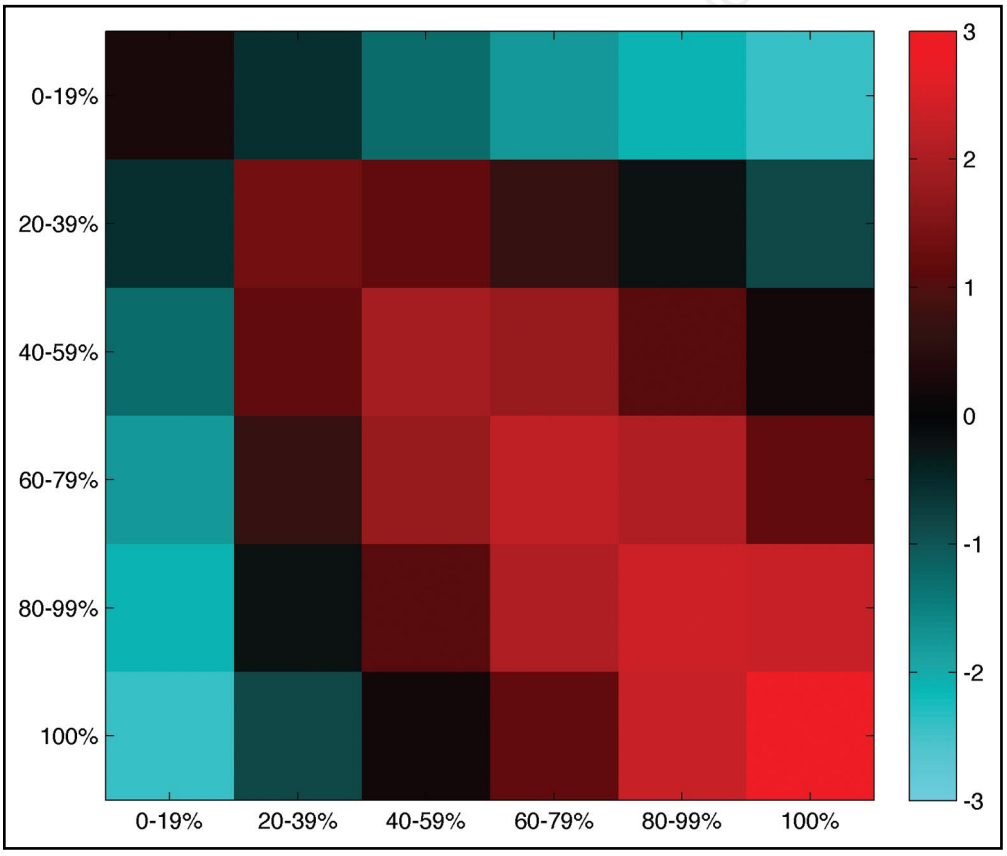

Figure 5. Log-odds ratio of c-methylation percentage transitions: repeat regions. Log-odds substitution matrix of cytosine methylation levels (x-axis and y-axis) within repeat regions.

cytosine methylation is rapidly lost during evolution due to random mutations, deamination of cytosines and tissue specific methylation mechanisms. We observe that despite the fact that all these mechanisms are at play, the methylation state of cytosines is quite robust.

By constructing log-odds matrices between aligned positions of duplicated genes that consider the methylation state of cytosines, we are also able to measure the degree of conservation of aligned nucleotides. We find that of all possible pairs of nucleotides, methylated cytosines are the most conserved. This surprising result may indicate that the methylation of cytosines plays a functional role in genes, and that their mutations are selected against more strongly than mutations to other nucleotides or unmethylated cytosines. The nature of this function is not known, and plant mutants that are defective in CG methylation have not shown strong phenotypes in the transcription of genes (Zhang et al.). Nonetheless it is possible that the methylation of these sites plays some yet undiscovered role that leads to these strong selection pressures.

In contrast to the conservation found in gene bodies, we observed a strong conservation of CHH sites in repetitive non-coding regions of the genome. Here, CHG and CG site were conserved to a lesser degree. These regions are methylated de novo by the methyltransferase DRM2 and RNAi mediated pathway. Thus the conservation we see is not due to a passive conservation of mutations over millions of years, but rather to active methylation that is most likely mediated by the production of siRNAs. Thus, in contrast to protein-coding genes, here we find that methyl-cytosines are not more strongly conserved than unmethylated cytosines in substitution matrices. Rather, these results indicate that the methylation of these regions is strongly dependent on the underlying sequence template.

The ability to measure the methylation profile of an organism on a genome-wide basis with single base accuracy has opened up the possibility of detailed studies into the evolution of DNA methylation. The accumulation of such profiles for multiple organisms will open up the possibility of extending these studies to measure the degree of methylation over much longer timescales. We anticipate that over the next few years, methylation profiles of a large number of additional organisms will become available, thus allowing us to extend and possibly further explain some of our initial observations.

Materials and Methods

Sequence alignments were performed using the Matlab implementation of Smith-Waterman. Due to memory addressing limitations and the N^2 space usage of Matlab's alignment implementation, pairs of genes whose length multiplied together is over 80 million and repeat region sequences longer than 5,000 bases were not aligned. The alignment of methylation sites was determined by recording a pair of methylation levels at each site that had at least 5x bisulfite sequence coverage and the sequence context (CG, CHG, CHH) aligned perfectly. Three separate sets of methylation level pairs were maintained, one for each sequence context.

The z-score was computed by generating a distribution of correlation values based on randomly permuting one side of the methylation level pair corresponding to an aligned sequence. This was done 100 times for each methylation sequence context to form a normal distribution that can be used to calculate a mean and standard deviation that can then be used to compute the z-score. This method effectively allows the permutation of a sequence in an aligned pair without loss of methylation site alignment or sequence context.

The nucleotide substitution and methylation transition log-odds matrices were computed by first obtaining a count of each combination of nucleotides possible for each aligned position. This count was then used to infer the amount of each type of nucleotide which was then used to calculate a probability for a particular nucleotide to occur. These probabilities were then used to calculate the null-hypotheses probability for each combination of nucleotides to occur at an alignment. The real probability was then taken by dividing the count of occurrences by the total number of aligned bases. Finally the log-odds score was determined by taking the log base-2 of the real probability divided by the null hypothesis probability. With regard to the cytosine counts for the base substitution matrices, it should be noted that the cytosine count is different between the 4 x 4 and 5 x 5 matrices since the cytosine vs. methyl-cytosine matrix only counted cytosines in positions that had 5x bisulfite coverage while the 4 x 4 matrix counted all aligned cytosines.

Acknowledgements

N.W. was supported by a National Science Foundation grant. N.W. carried out the calculations under the supervision of M.P. and S.E.J. S.E.J. is an investigator of the Howard Hughes Medical Institute. We thank Shawn Cokus for useful discussions.

Note

Supplementary materials can be found at:
www.landesbioscience.com/supplement/WidmanEPI4-2-Sup.pdf

References

1. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 2008; 452:215-9.
2. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 2008; 133:523-36.
3. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. Cell 2006; 126:1189-201.
4. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet 2007; 39:61-9.
5. Frederico LA, Kunkel TA, Shaw BR. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 1990; 29:2532-7.
6. Blanc G, Wolfe KH. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 2004; 16:1679-91.
7. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, et al. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA 2005; 102:5454-9.
8. Ganko EW, Meyers BC, Vision TJ. Divergence in expression between duplicated genes in Arabidopsis. Mol Biol Evol 2007; 24:2298-309.
9. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999; 27:573-80.
10. Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res 2004; 14:1861-9.