UNIVERSITY OF CALIFORNIA

Los Angeles

DNA Methylation Studies in Arabidopsis: Evolution and Tissue Specificity

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Computer Science

by

Nicolas Benjamin Widman

2014

ABSTRACT OF THE DISSERTATION

DNA Methylation Studies in Arabidopsis: Evolution and Tissue Specificity

by

Nicolas Benjamin Widman

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2014

Professor Eleazar Eskin, Co-chair

Professor Matteo Pellegrini, Co-chair

High-throughput sequencing makes it possible to study epigenetic properties such as DNA methylation genome-wide with single base pair precision. Two research projects were done studying epigenetic properties in plants using high-throughput sequencing using DNA from the model organism Arabidopsis Thaliana. The first project looks at the conservation of DNA methylation on an evolutionary time scale to determine to what degree DNA methylation is conserved with respect to the underlying DNA sequence. This is done by comparing pairs of duplicated genes in Arabidopsis resulting from a genome duplication event 40-80 million years ago in addition to non-coding repeats in the genome. DNA methylation was found to be significantly conserved but methylated cytosines had a tendency to deaminate to thymine which was the most common type of mutation in the underlying DNA sequence. In the second project DNA methylation, nucleosome density and gene expression are compared between tissues in Arabidopsis. Using shoots (entire above ground section) and roots as two different tissues, significant epigenetic differences are found in certain genes, related in function, between the two tissue types. Methylation and nucleosome density are found to be periodic in nature and similar between shoot and root at a genome-wide scale. Positioning of methylation and nucleosomes across the gene body show significant change among genes that are differentially expressed between shoot and root.

The dissertation of Nicolas Benjamin Widman is approved.

Jason Ernst

Wei Wang

Eleazar Eskin, Committee Co-chair

Matteo Pellegrini, Committee Co-chair

University of California, Los Angeles

2014

Dedication

This dissertation is dedicated to everyone who provided insight and support during my PhD Studies at UCLA.

Table of Contents

Chapter 1: Introduction 1

Chapter 2: Determining the conservation of DNA methylation in Arabidopsis 4

Abstract 5

Introduction 6

Results 8

Discussion 15

Methods 16

Acknowledgements 18

Figures 19

References 25

Chapter 3: Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation 26

Abstract 26

Introduction 26

Results 28

DNA methylation in shoots and roots 28

Differential expression and DNA methylation 30

Nucleosomes in shoots and roots 31

Nucleosome positioning and methylation 31

Extensin genes and transposons 32

Discussion 33

Materials and Methods 35

Bisulfite libraries and analysis 35

RNA-seq 36

Nucleosomes 36

Acknowledgments 36

Figures 38

References 46

Chapter 4: Conclusion 48

Vita

2005	Microsoft Technical Scholarship Purdue University Dept. of Computer Science Cisco Systems Scholarship
2005-2006	Purdue University Senior Honors Research Project
2006	B.S. Computer Science B.S. Mathematics Purdue University – West Lafayette
2006-2007	Peer Advisor – HSSEAS Office of Academic and Student Affairs University of California, Los Angeles
2014	Teaching assistant, CS 32: Introduction to Computer Science II University of California, Los Angeles

Publications:

Nicolas Widman, Steve Jacobsen E. and Matteo Pellegrini Determining the conservation of DNA methylation in Arabidopsis. Epigenetics 2009 Feb 16 4(2)

Widman N, Feng S, Jacobsen SE, Pellegrini M. Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation. Epigenetics. 2013 Oct 29;9(2). [Epub ahead of print] PubMed PMID: 24169618.

Poster presentation at Keystone Symposia: Epigenetics, Development and Human Disease (January 5-10, 2009)

Chapter 1: Introduction

DNA methylation refers to the distribution of a variant of cytosine (one of four nucleotide bases in DNA) referred to as 5-methyl cytosine, which is a cytosine with an addition of a methyl group bonded to a particular location of the molecular structure. The significance of DNA methylation is that it affects gene expression and affects the structure of DNA. When ignoring methylation, DNA sequences can be represented using a 4-letter alphabet (A,C,G and T). A trivial extension of the alphabet to 5 letters, allows one to consider the methylated version of cytosine (A, C (unmethylated), c (methylated), G and T). However the occurrence of "c" (methylated) doesn't follow the same pattern as "A", "C" (unmethylated), "G" and "T". Unlike the distribution of "A", "C", "G" and "T", "C" and "c" have a distribution that correlates with related gene functions and overall gene expression levels.

Changes in DNA that do not modify the underlying sequence such as methylation are referred to as epigenetic changes. Epigenetics is the study of changes in the DNA that affect gene activity without changes in the DNA sequence. Other epigenetic changes include changes in nucleosome positioning and histone protein modification. Nucleosomes are the basic unit of large-scale DNA structure and consist of DNA wrapped around a core of four pairs of histone proteins. Epigenetics plays a major role in gene regulation and is responsible for cell type differentiation. Some epigenetic changes are temporary while others are permanent and can be inherited.

During my PhD I have been involved in two projects involving DNA methylation. I attach the resulting papers to this prospectus. The approach used to measure DNA methylation in both papers used high-throughput sequencing of DNA treated so that methylated "c" can be distinguished from unmethylated "C". However since methylation is not preserved to the same extent as the coding DNA sequence (A, C, G and T) a methylation level was calculated ranging from 0 (each instance unmethylated) to 1 (each instance methylated) at each cytosine in the sequence. This approach accounts for the fact that methylation occurs probabilistically, and so reads that map to the same location do not necessarily all have the same methylation status.

In the first paper, the conservation of DNA methylation is studied in a group of duplicated genes in Arabidopsis. This paper focuses on the extent to which overall patterns of methylation are preserved over an evolutionary time scale by comparing the DNA methylation between pairs of protein-coding genes that were duplicated. These duplications occurred between 40-80 million years ago. Additionally the transition rate between aligned gene pairs, taking into account methylated “c” and unmethylated “C” as well as “A”, “G” and “T” is studied. Similarly the methylation level transitions between aligned “C” and/or “c” positions are studied. We find that duplicated genes tend to maintain similar levels of methylation, suggesting there is a strong genetic component that determines methylation levels.

In the second paper, the relationship between DNA methylation, gene expression and nucleosomes is studied in two different tissue types in Arabidopsis. DNA methylation is first analyzed across each chromosome to find large-scale patterns in methylation. We compute the average methylation across all genes with respect to the transcription start site to find patterns of methylation across the gene body. DNA structure over the chromosome is then analyzed by measuring the density of nucleosome proteins, which the DNA wraps around. This is of interest since a higher density of nucleosome proteins along the DNA sequence results in more densely packed DNA. Gene expression is measured to estimate the rate that DNA is transcribed into mRNA. Genes are grouped using the ratio of expression levels between the two tissues and the ratio of nucleosome density between both tissues in each group of genes is compared. Next DNA methylation level is measured with respect to nucleosome positioning to reveal a pattern in the relationship between the location of DNA methylation and nucleosomes. Finally a group of genes, all with a related function, is found by selecting those with the most difference in methylation, nucleosome density and expression. The results with protein coding genes are compared with non-coding regions of the genome.

Chapter 2: Determining the conservation of DNA methylation in Arabidopsis

Nicolas Widman[1], Steven E Jacobsen[1], Matteo Pellegrini[1]*

[1]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles

***Corresponding author**

Abstract

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.

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.

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 30000 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.

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 5 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 10 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 5 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 6 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 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 10 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 (supplementary figure 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.

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 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., Cell). 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.

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 5000 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.

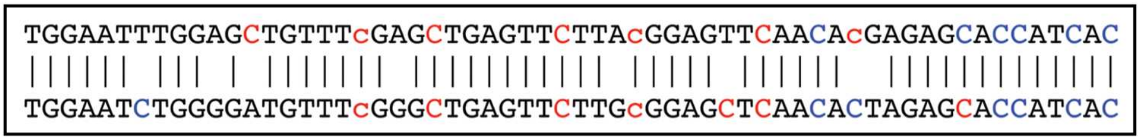
The z-score was computed by generating a distribution of correlation values based on randomly permuting one side of the methylation level pair corresponding to an aligned sequence. This was done 100 times for each methylation sequence context to form a normal distribution that can be used to calculate 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 4x4 and 5x5 matrices since the cytosine vs. methyl-cytosine matrix only counted cytosines in positions that had 5x bisulfite coverage while the 4x4 matrix counted all aligned cytosines.

Acknowledgements

NW was supported by a National Science Foundation grant. NW carried out the calculations under the supervision of MP and SEJ. SEJ is an investigator of the Howard Hughes Medical Institute. We thank Shawn Cokus for useful discussions.

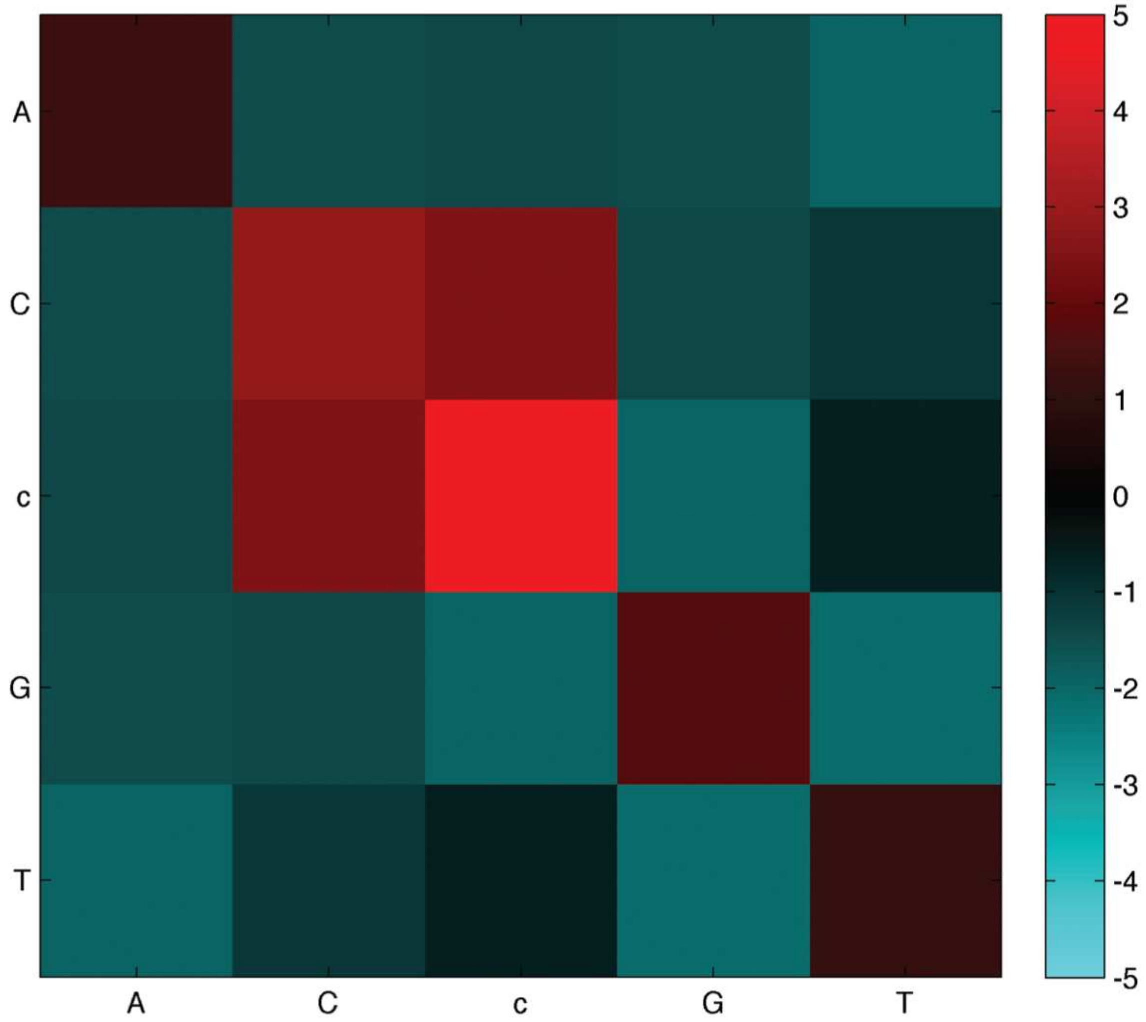
Figure 1 – Sample Alignment



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

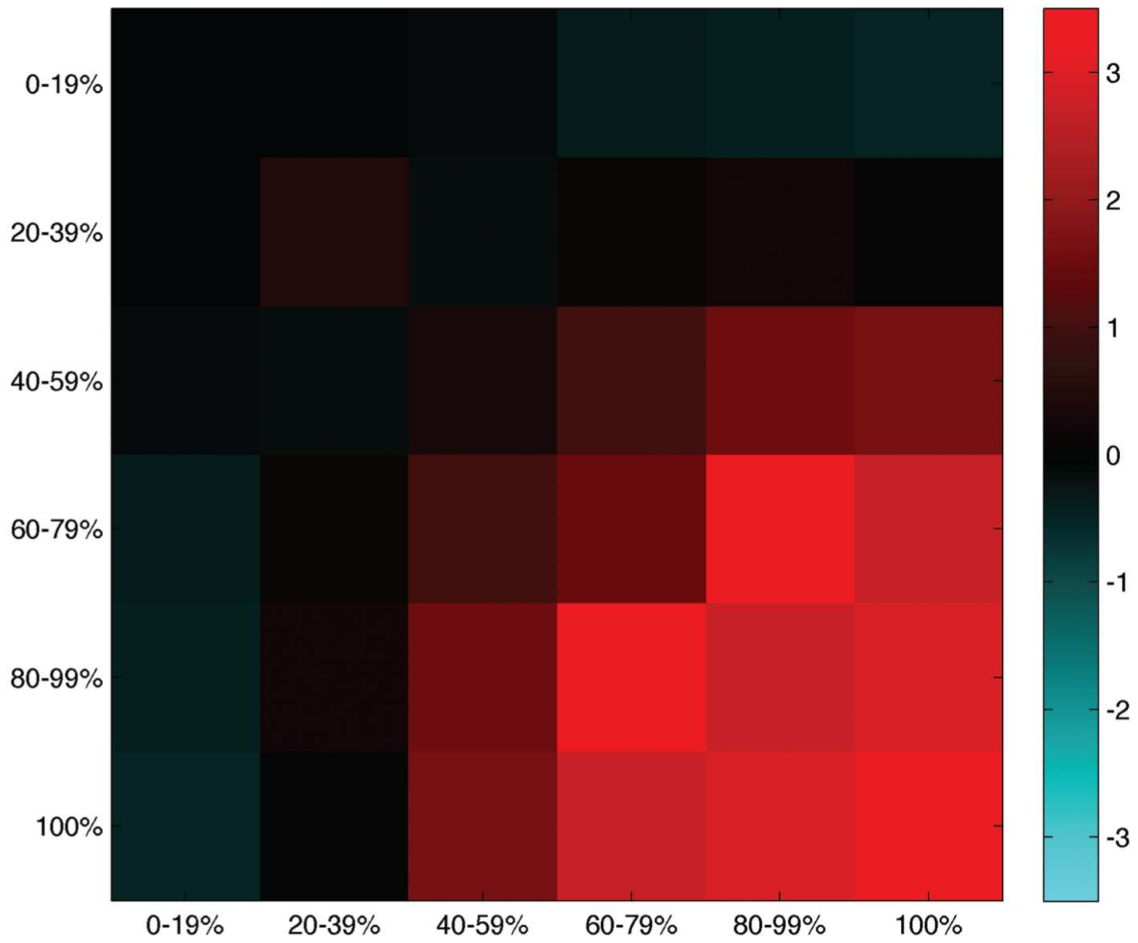
Figure 1. Sample Alignment. A fragment of a sample alignment between duplicated genes. Unmethylated cytosines are shown as "C", with methyl-cytosine as "c." Red "c" and "C" represent cytosines with ≥ 5 read coverage, blue "C" represents cytosines with <5 read coverage. (Top: AT3G47910 5555-5614 bases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

Figure 2 – Log-Odds Ratio of Nucleotide Base Transitions: Duplicated Genes



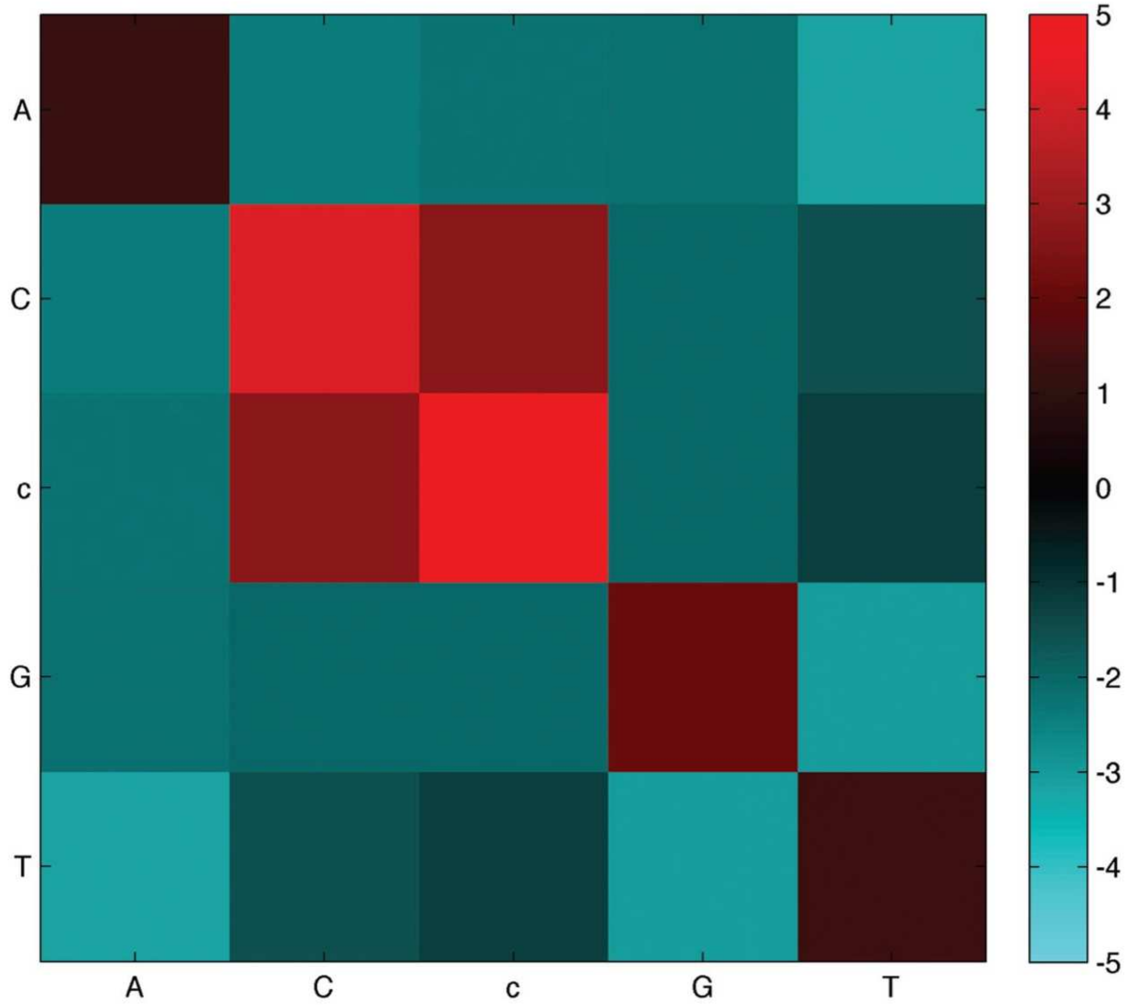
Log-odds substitution matrix of aligned duplicated gene bases (x-axis and y-axis) with methylated cytosine represented as a distinct base using lower-case c.

Figure 3 – Log-Odds Ratio of C-Methylation Percentage Transitions: Duplicated Genes



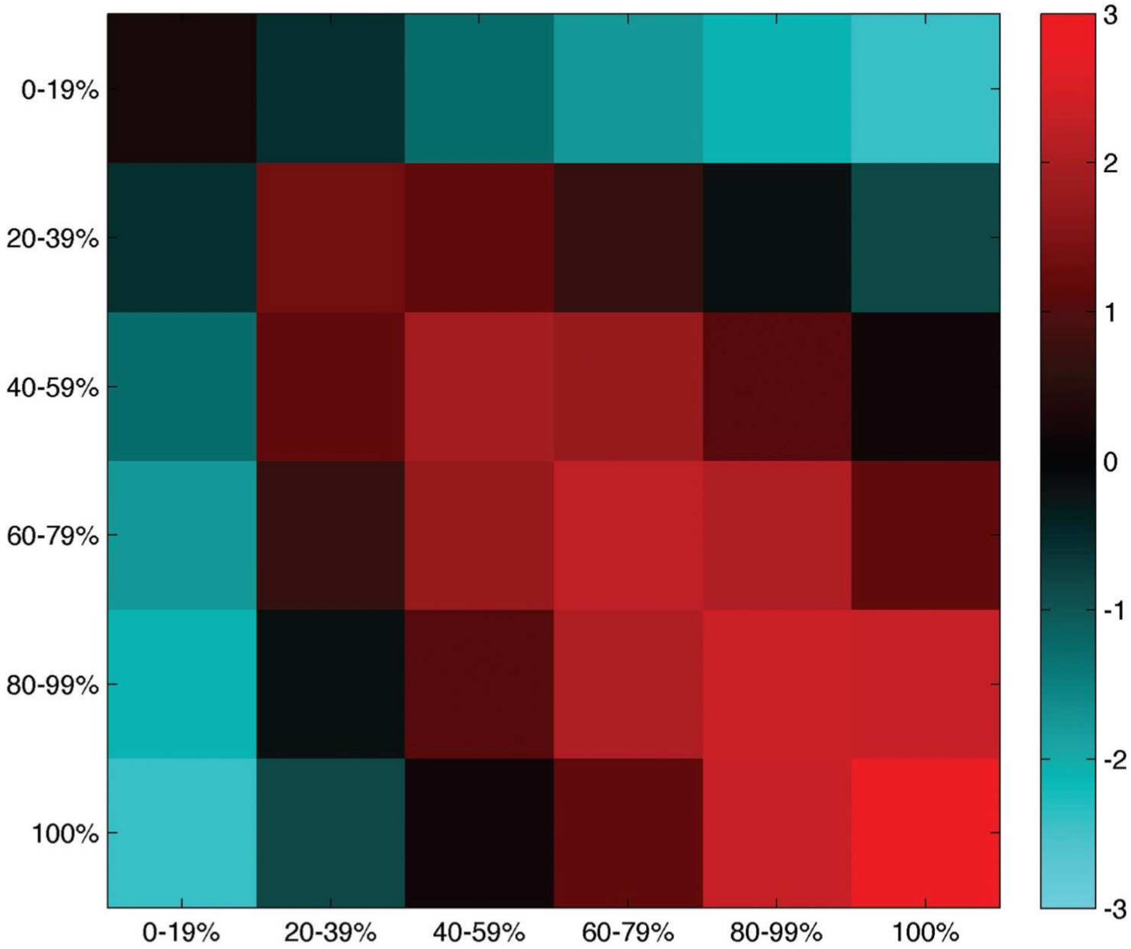
Log-odds substitution matrix of cytosine methylation levels (x-axis and y-axis) between pairs of aligned duplicated genes.

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.

Table 1: Aligned sequence methylation and Z-score

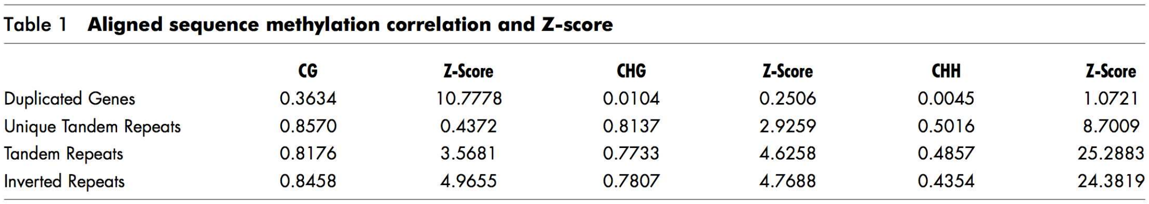


Table 1 **Aligned sequence methylation correlation and Z-score**

	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

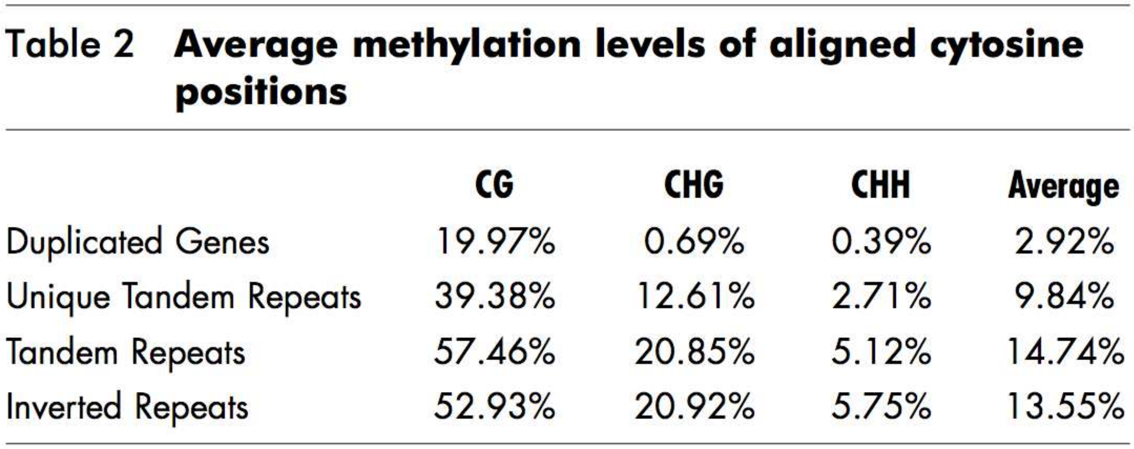


Table 2 **Average methylation levels of aligned cytosine positions**

	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%

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 U S A 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.

Chapter 3: Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation

Nicolas Widman[1], Suhua Feng[1], Steven E Jacobsen[1], and Matteo Pellegrini[1,]*

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

***Corresponding author**

Abstract

DNA methylation and nucleosome densities play a critical role in the regulation of gene expression. While much is known about the mechanisms of transcriptional control that are mediated by these, less is known about the degree to which they are tissue-specific. By comparing DNA methylation, nucleosome densities and transcriptional levels in different tissue types we can gain a clearer understanding of the extent to which these mechanisms influence gene expression in a tissue specific manner. We compared DNA methylation in Arabidopsis shoots and roots and found extensive differences across the genome. We computed DNA methylation differences between roots and shoots at single cytosines and found that one in every 173 cytosines was differentially methylated. In addition we compared DNA methylation with tissue specific gene expression and nucleosome density measurements to identify associations between these. We also identified a group of genes that are strongly correlated with these epigenetic marks and are significantly differentially methylated between roots and shoots. These root-specific genes are part of the extensin family, and are preferentially methylated and have at least 10-fold higher expression and lower nucleosome density in roots relative to shoots.

Introduction

DNA methylation is known to have a regulatory effect on gene expression and chromatin structure.[1] It is one of several epigenetic mechanisms that include chromatin structure, histone protein modifications and RNA interference.[2] In eukaryotic DNA cytosine is the only nucleotide base that is methylated and its methylation is influenced by the sequence of the two downstream bases.[3] The most highly methylated cytosines are those that are immediately followed by guanine, referred to as CG sites. CG methylation is found in both genes and repeats and can be involved in the regulation of gene expression.[4] In plant genomes, there are 2 additional sequence contexts where methylation can occur: CHG and CHH (H represents any base other than guanine). Methylation at CHG and CHH sequence contexts, in contrast to CG methylation, is absent over genes and mostly found in intergenic, repeat-rich regions of the genome and plays a critical role in silencing transposons.[5] In addition to DNA methylation, changes in chromatin structure can also affect gene expression. These changes include modifications to histone proteins and the positioning of nucleosomes.

Despite extensive research in animals, which has described changes in DNA methylation during development,[6] the degree to which these epigenetic mechanisms determine tissue differentiation in plants is still poorly understood. Prior work has shown that global methylation differences between tissue types in rice are small, with CG methylation levels being nearly identical and non-CG methylation increasing slightly with the age of the tissue.[7] The greatest methylation differences found across rice tissue types were identified in the endosperm and consist of global non-CG hypomethylation along with hypomethylation of specific genes. Similarly, Arabidopsis endosperm and pollen genomes are significantly demethylated[5]

Nucleosome positioning has also been shown to have a significant effect on transcriptional regulation. Nucleosomes appear to play a role in determining the boundaries between exons and introns and are more strongly positioned over exons than introns regardless of the gene expression level.[8] Additionally, heterochromatin in pericentromeric regions is associated with different nucleosome marks than euchromatin. However, as in the case of DNA methylation, differences in nucleosome occupancies between plant tissues have not been well characterized.

To further extend our understanding of tissue-specific epigenetic differences in plants, and their functional roles, we measured DNA methylation, nucleosome positioning and gene expression in both Arabidopsis root and shoot tissues. Our goal was to identify tissue-specific DNA methylation and nucleosome distributions and associate these with changes in gene expression between differentiated tissue types. We sought to identify genes that may be regulated by DNA methylation during tissue differentiation resulting in tissue specific expression.

Results

DNA methylation in shoots and roots

We measured DNA methylation in both root and shoot samples using whole genome bisulfite sequencing. The shoot samples were the same as those we previously published,[4] except that we employed a slightly different sequencing library generation method (see Methods).[9] The root tissues were from root culture grown in liquid medium (see Methods). This root culture method was chosen because it provides a comprehensive representation of all the different types of cells from root and produces consistent results across replicates as we can sensitively control the environment under this growth condition.

Our reads were mapped to the Arabidopsis genomes using the method described by Seeker;[10] we obtained methylation levels by computing the fraction of converted cytosines. We were able to measure methylation levels for 75–80% of all cytosines. The methylation levels for the remaining 20% of cytosines could not be accurately measured due to inadequate coverage or certain regions of the genome not being mapped uniquely using our read length. The data may be visualized at our genome browser http://genomes.mcdb.ucla.edu/ and downloaded from GEO using the accession number GSE52762.

We analyzed the root and shoot methylomes by comparing the methylation state of each individual site in the 2 samples. We used a binomial model to identify sites that were differentially methylated (see Methods). Using this single-site differential methylation analysis, we found that 1.55% of sites passed our threshold. We also found that sites that are differentially methylated are preferentially hypermethylated in shoots compared with roots with 1.85 times as many hypermethylated sites in shoots than in roots. This effect is more significant in pericentromeric regions and at sites of non-CG methylation (Fig. 1A **and** B). In contrast, single-site methylation differences in non-centromeric regions are dominated by CG methylation. This is not surprising given that CG methylation is the dominant context in euchromatic regions, while both CG and non-CG methylation is present in heterochromatic regions. We note, however, that global methylation levels do not differ significantly between roots and shoots with CG methylation being only 1 percent higher in shoots relative to roots and non-CG methylation being 5 percent higher in shoots relative to roots (Fig. 1C).

In previous studies, it was found that DNA methylation within genes is dominated by CG methylation.[4] Transcription start and termination sites are generally demethylated, as these regions are bound by regulatory factors that likely interfere with the ability of DNA methyltransferases to access and methylate DNA. CHG and CHH methylation levels generally show the opposite behavior as they are only found at very low levels within genes and at higher levels in intergenic, repeat-rich regions. We asked whether the density of single site methylation differences between shoots and roots varied across genes. To account for the changes in the average levels of methylation across genes, we normalized the densities by these averages. This normalization was performed to remove the expected bias wherein changes in highly methylated sites are more easily detected than changes to lowly methylated sites. We find that, within genes, differentially methylated sites are found primarily at CG sites and are preferentially methylated in shoot relative to root. Shoot hypermethylated sites occur about 1.85 times more frequently than sites preferentially methylated in root (Fig. 2A). Furthermore, we observe a higher fraction of these sites at the transcription initiation and termination boundaries, suggesting that these regions have more variable methylation levels than gene bodies.

Differential expression and DNA methylation

We obtained RNA-seq data from Arabidopsis shoots and roots using oligo-DT priming.[11] The reads were mapped to the genome using bowtie and the counts per gene were compared across the two samples. Overall, we identified 2424 genes that were significantly differentially expressed between the two tissues using the criteria that their fold change was at least 10.[11] Out of these differentially expressed genes, 1292 were more expressed in roots. These genes were enriched for functional groups associated with cell wall organization and biogenesis.

The root overexpressed transcripts were also strongly enriched for extensin genes, which are associated with cell wall formation (*P* = 1.41e–12). The genes that have at least 10-fold higher expression in roots are significantly more methylated in shoots than roots (with 39% more sites with greater methylation in shoots than roots than the typical gene), and this effect is more pronounced at CG sites. As shown in Figure 2B, we see that the differentially methylated CG sites of these genes (normalized by total methylation levels) are strongly enriched around the transcription start site, again suggesting these regions have more variable methylation between these tissues. In contrast, we find that non-CG methylation appears to be relatively unchanged within the genes with higher expression in roots than shoots. These results indicate that methylation changes between roots and shoots are significantly enriched over genes with differential expression, suggesting that these sites may have a direct or indirect role in the regulation of these transcriptional changes.

Nucleosomes in shoots and roots

We identified nucleosome positions across our 2 samples by digesting chromatin with micrococcal nuclease and sequencing the mononucleosome fractions. These were mapped to the genome using the method described by Bowtie.[12] Each read in this library represents a putative nucleosome start site.

We first measured nucleosome density across the genome by computing the number of reads in 1□000 kilobase windows. We find that nucleosome densities are enriched in the pericentromeric region compared with the arms (Fig. 3A). We also find that these nucleosome distributions differ slightly between root and shoot (Fig. 3A), with shoots having higher nucleosome densities in pericentromeric regions than roots.

We also measured the average nucleosome occupancy across genes (Fig. 3B). When we compared the nucleosome occupancy between shoots and roots we found that shoots had a higher nucleosome density before transcription start sites and had a second peak at the end of the gene, but were generally less nucleated over the gene body. This is a surprising result suggesting that there are global changes in genic chromatin structure between the 2 tissues with roots showing higher nucleosome densities over genes. Strikingly, genes that have at least 10-fold higher expression in root show considerably more nucleosomes around the transcriptional start and end sites in the shoot tissue samples, suggesting that increasing transcriptional rates decreases nucleosome densities over these regulatory regions. However, this effect was not seen for the shoot overexpressed genes, which have a nucleosome ratio profile that is very similar to that for all genes.

Nucleosome positioning and methylation

We measured methylation levels with respect to the start of nucleosomes and found, consistent with the results of our previous study,[8] that DNA methylation is enriched over the first nucleosome and that nucleosomes show a strong periodicity in methylation patterns that persist over several nucleosomes (Fig. 4A). This periodicity is approximately 180 base pairs long and is most pronounced in root tissue methylation. We performed a similar analysis of nucleosome periodicity by looking at the average density of nucleosome start sites centered on the start of each MNase read (Fig. 4A). We see a very similar periodicity of about 180 bases, as seen in the methylation. Finally, we asked if the 2 periodic patterns can be superimposed by shifting one with respect to the other. This allows us to determine where the peak of methylation occurs with respect to the peak of nucleosome positioning. To accomplish this we slid a 180 base window of the methylation data across the nucleosome density data from 500 bases upstream to 500 bases downstream of a nucleosome start site. We find that the 2 patterns are maximally correlated when they are shifted by 25 bases, indicating that the peak of methylation occurs 25 bases before nucleosome start sites.

Extensin genes and transposons

We performed an overlap analysis on all differentially expressed, methylated, and nucleated genes and found the overlap to be statistically significant especially for genes that have higher expression in root as well as between nucleosome density and methylation level in both shoots and roots (Fig. S1).[13] Among the root overexpressed genes, we identified a small group that have both methylation, nucleosome density and gene expression changes between roots and shoots. These genes are significantly less methylated in roots compared with shoots, have at least a 10-fold expression preference in roots over shoots and have roughly one half the nucleosome density in roots than shoots (Fig. 5). Based on Gene Ontology terms, these genes were found to be in the extensin family of genes and are involved in cell wall formation.[14,15] Nine of the 10 extensin genes are found among the 11 genes that overlap all 3 differential sets of genes (Table S2). This suggests that this group of genes might play a previously unreported role in differentiating root cells from other cell types in the plant. The differences in methylation and nucleosome densities for two of these genes are shown in Figure 5A **and** B.

The methylation levels of several extensin genes were validated using traditional bisulfite sequencing and the results are reported in the supplementary material. With respect to CG methylation, the validation shows the strongest support for 4 of the 9 genes the validation was performed on. If we only consider only those cases for which both the BS-Seq data and the validation show a decrease or increase from shoot to root, 7 of these genes have equivalent trends for CG methylation and 5 for either CHG or CHH methylation.

Transposons also show epigenetic differences between roots and shoots. Differential nucleosome density is found in numerous transposons with 11.6% of transposons showing at least a 2-fold difference and 37.4% showing at least a 50% difference. There are a total of 293 transposons with at least a 10% difference in CG methylation level with 82 having more methylation in shoots and 211 having more methylation in roots. Performing an overlap analysis on transposons found statistically significant results between differential methylation levels and nucleosome density in both shoots and roots but did not find any significance between expression and methylation or nucleosome density (Fig. S1). We also identified several transposons that showed significantly different epigenetic and transcriptional patterns between the 2 tissues. Some transposons that show differences in nucleosome density between shoot and root also have differences in expression levels. An example of a transposon with significant changes in methylation, nucleosome density and expression is shown in Figure 5C.

Discussion

We have shown that the root and shoot tissues from Arabidopsis accumulate significant epigenetic changes. We identify the largest changes in CG methylation in heterochromatic regions. Within genes, the transcriptional start (TSS) and end sites show the greatest variability in methylation when normalized by the average methylation levels. That is, although TSSs are generally hypomethylated, the levels of this hypomethylation are quite variable across these two tissues, indicating that these regions may be undergoing more chromatin changes that the bodies of genes. In contrast, within genes, our method does not have the resolution to identify significant changes in methylation at non-CG sites that are sparsely methylated. We have also identified clusters of significantly varying sites, and found that these are enriched in extensin genes, that may represent a novel class of genes involved in root development, although no additional functional characterization is available for these genes.

We also noted a large number of loci with significantly altered chromatin structure, as measured by nucleosome positioning. Overall, the density of root nucleosomes in the centromeric regions of the chromosomes appears to be lower in roots than shoots, suggesting possible changes in the density of heterochromatin between these 2 tissues. We also found that genes that were root specific tended to be denucleated around the transcriptional start sites compared with shoots. Finally, in support of previous findings, we identify that both the positioning of nucleosomes and DNA methylation show a strong periodic behavior. We conclude that nucleosomes are preferentially methylated, and that there is a peak of DNA methylation about 25 bases before the nucleosome cut site, suggesting that the DNA that is adjacent to the linker DNA is most accessible to DNA methyltransferases.

Finally we show dramatic changes in DNA, gene expression and nucleosome structure around certain transposons, suggesting that these are often reactivated in a tissue specific manner. Future studies on specific cell types will undoubtedly find much more significant variation in gene expression, methylation, and chromatin structure than we could observe in our 2 coarse sectionings of plants into roots and shoots. Nonetheless even these results suggest that significant epigenetic changes accompany the transcriptional reprogramming that results from the development of roots and shoots.

Materials and Methods

All material was collected from the Columbia strain of *Arabidopsis thaliana*. We used the genome annotation from TAIR7.

Bisulfite libraries and analysis

For the methylation analysis we used bisulfite sequencing of whole genome libraries following previously published protocols utilizing pre-methylated adapters.[9] The 2 libraries, root and shoot, were prepared from roots and all tissues of the plant growing above ground, respectively. Samples for the root library were from a root culture and are not from the same plants that were used for the shoot library. The roots were obtained from one-month-old Arabidopsis grown in Gamborg's B5 liquid medium under continuous white light. After obtaining the raw BS-seq reads, the reads were mapped using the method described by Seeker. Each base that is mapped to a cytosine can either be a cytosine or converted to a thymine by the bisulfite reagents. Methylated cytosines remain as cytosines after bisulfite treatment but unmethylated cytosines are deaminated into uracil, which is reverse transcribed into thymine in our reads. The methylation percentage at each site is determined by measuring the ratio of the cytosines mapped to a particular location over the total number of reads. We measured single-site differential methylation by comparing sites with an absolute difference of 30% methylation (e.g., 20% vs 50%, not relative difference such as 20% vs 26%). In order make sure that these differential sites were statistically significant only sites that had a binomial distribution confidence interval of 5% were included. The confidence interval was calculated by computing the binomial cumulative distribution function of the BS-seq data at each site that had an absolute methylation difference above the threshold of 30% and only incorporated sites into the single site differential data where the BS-seq data scored less than 0.025 or greater than 0.975 in the CDF.

RNA-seq

We sequenced and mapped mRNA to determine gene expression levels. The expression level was measured by normalizing the number of times mRNA mapped to a particular gene in the reference genome to a million mRNA reads. This reports the data in terms of reference sequence hits per million reads. As these libraries were oligo DT primed, they do not cover the entire transcript and we therefore did not normalize by transcript length.

Nucleosomes

Nucleosome data was obtained by sequencing MNase digested DNA. MNase cleaves DNA at both ends of nucleosomes leaving any DNA bound to the histone core of the nucleosome unaffected. This nucleosome-bound DNA was then separated from the histones and sequenced.[8] This sequenced DNA was mapped to the reference genome using the method described by Bowtie,[12] allowing for 2 mismatches and the nucleosome start sites were determined by looking at the first position of each mapped read.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

NW was supported by a National Science Foundation grant. Widman N performed the calculations under the supervision of Pellegrini M and Jacobsen SE. Feng S prepared the bisulfite-sequenced root and shoot libraries as well as performed bisulfite-sequencing validations on differentially expressed extensin genes. Feng S is a

Special Fellow of the Leukemia and Lymphoma Society. Jacobsen SE is an investigator of the Howard Hughes Medical Institute.

Supplemental Materials

Supplemental materials may be found here:

www.landesbioscience.com/journals/epigenetics/article/26869

Figure 1 – Genome-wide methylation levels

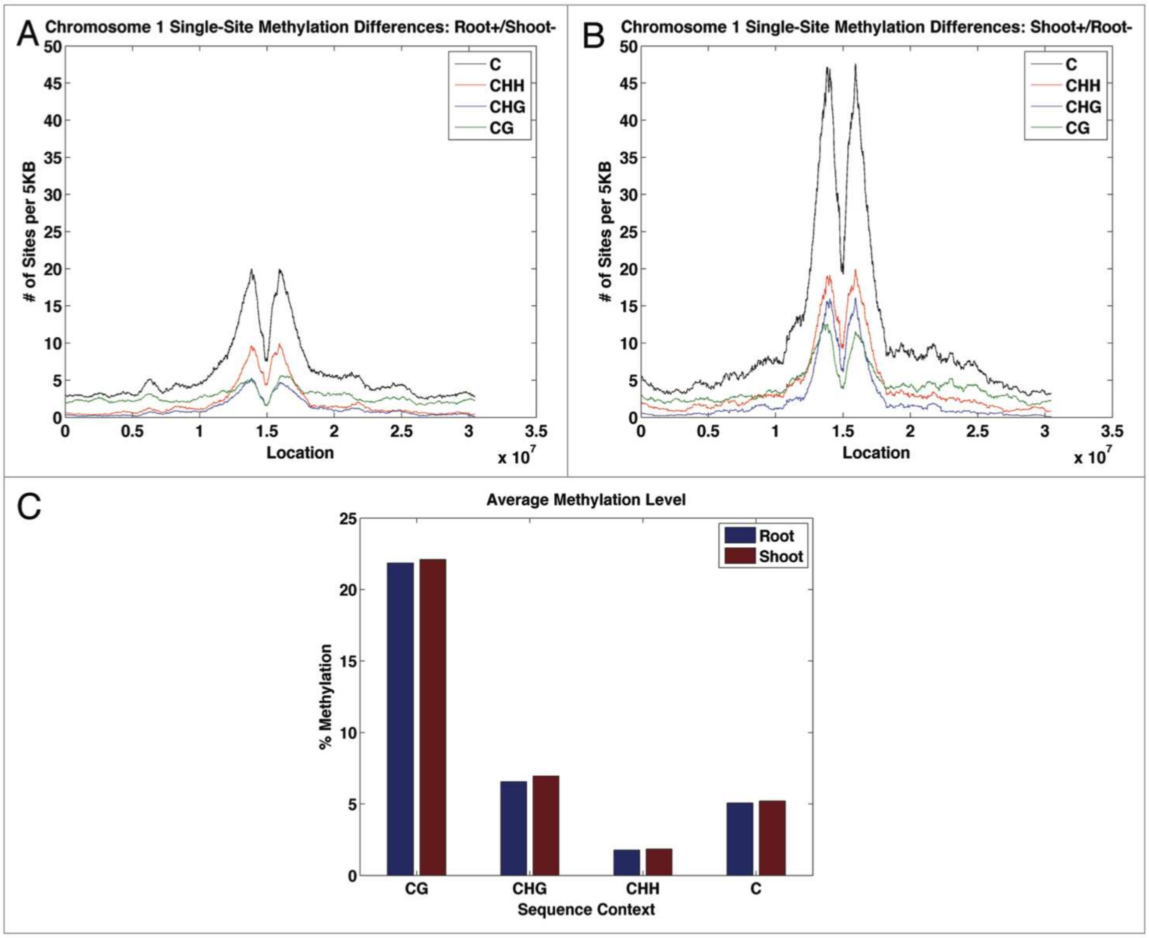


Figure 1. Genome-wide methylation levels. (**A**) Rate of single-site methylation differences by sequence context with higher methylation in root in chromosome 1. (**B**) Rate of single-site methylation differences by sequence context with higher methylation in shoot in chromosome 1. (**A and B**) X, position on chromosome; Y, number of instances of differential methylation per 5000 bases. (**C**) Average genome-wide methylation levels roots and shoots.

Figure 2 – Metagene methylation differences

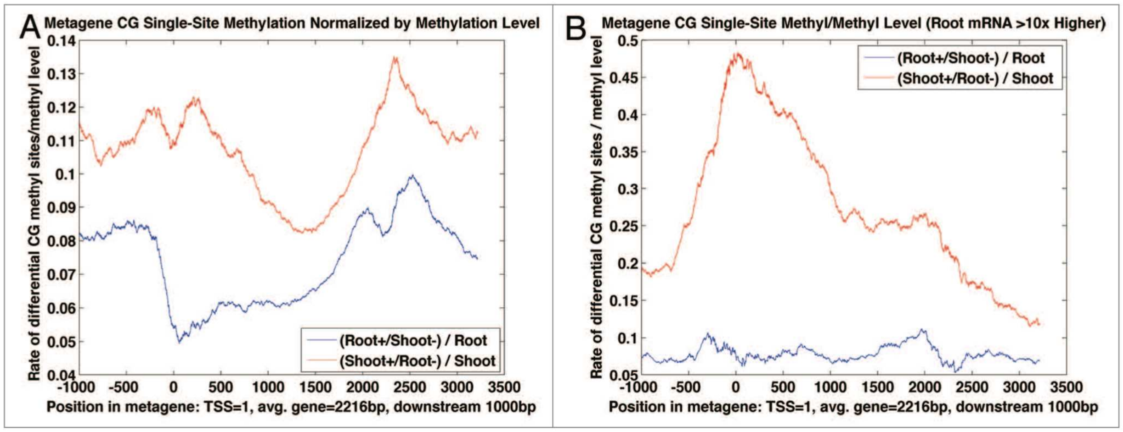


Figure 2. Metagene methylation differences. (**A**) Average rate of single-site methylation differences (CG sequence context only) across all genes. (**B**) Average rate of single-site methylation differences (CG sequence context only) across genes with 10× greater expression in roots. (**A and B**) X, metagene data normalized to average gene length of 2216 bases and 1000 bases upstream and downstream of gene, position on gene; Y, rate of differentially methylated sites divided by methylation level.

Figure 3 – Nucleosome Density and Gene Expression

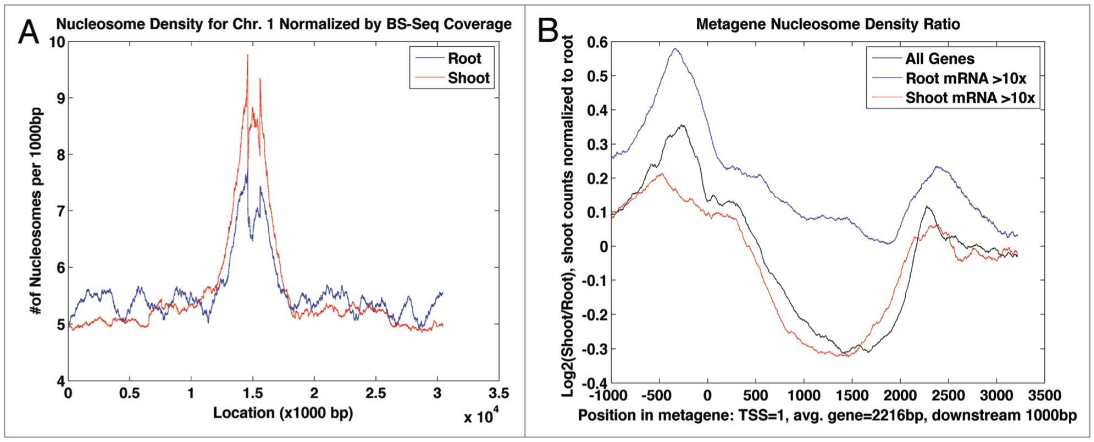


Figure 3. Nucleosome Density and Gene Expression. (**A**) Nucleosome density of roots (blue) and shoots (red). X, location on chromosome 1 (1000 bp increments); Y, nucleosome density: nucleosomes per 1000 bp with total number of nucleosomes in root and shoot normalized by BS-Seq coverage. (**B**) Average nucleosome density ratio for all genes (black), genes with 10× greater expression in shoots (red) and genes with 10× greater expression in roots (blue). X, location with respect to transcription start site, gene length normalized to 2216 bp; Y, log ratio (base 2) of shoot nucleosome density divided by root nucleosome density, shoot data normalized to root coverage level.

Figure 4 – Methylation levels relative to nucleosome start site

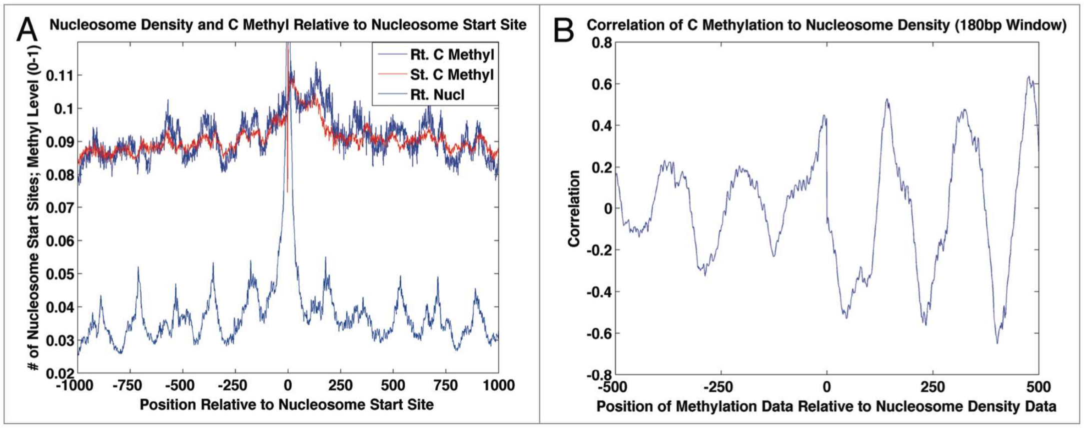


Figure 4. Methylation levels relative to nucleosome start site. (**A**) Nucleosome distribution relative to nucleosome start site compared with cytosine methylation. X, 1 kb upstream to 1 kb downstream; Y, # of nucleosomes, methylation level. (**B**) Correlation of methylation level to nucleosome density using a 180 base pair window. X, distance from start of methylation data window to start of nucleosome data window; –, upstream; +, downstream; Y, Correlation.

Figure 5 – Comparison Between Root and Shoot in UCSC Genome Browser

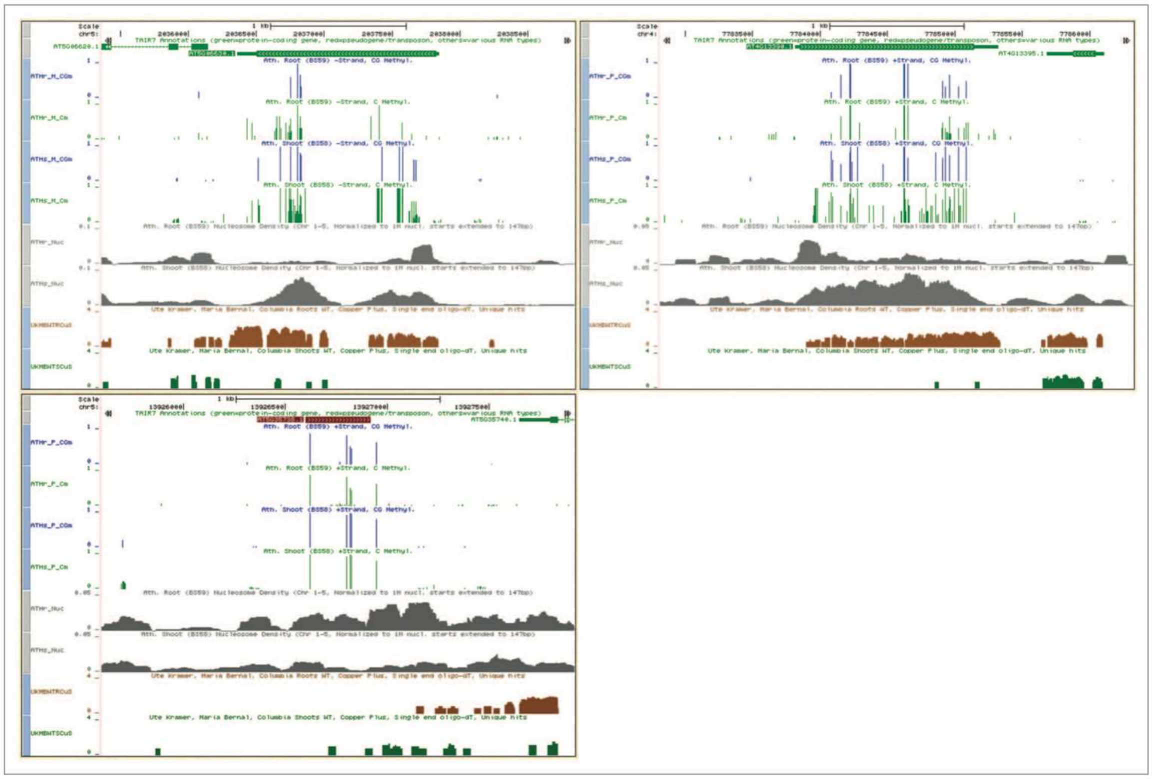
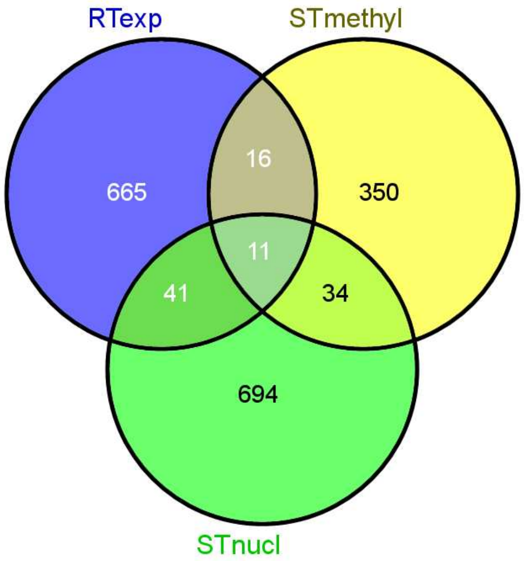
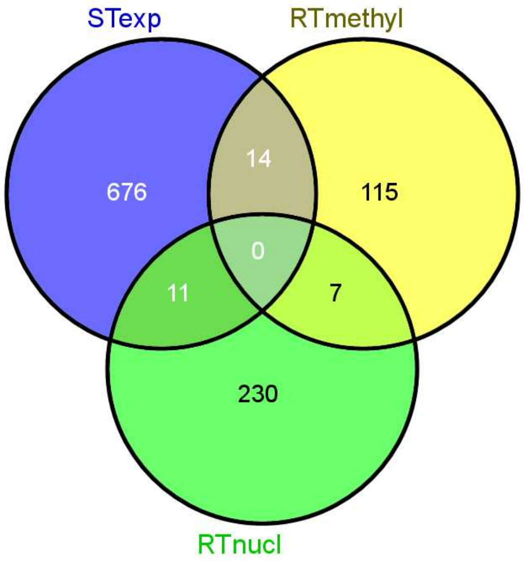

Scale
chr5:
1 kb
2036000
2036500
2037000
2037500
2038000
2038500
TAIR7 Annotations (green=protein-coding gene, red=pseudogene/transposon, others=various RNA types)
AT5G06620.1
Ath. Root (BS59) -Strand, CG Methyl.
ATHr_M_CGm
Ath. Root (BS59) -Strand, C Methyl.
ATHr_M_Cm
Ath. Shoot (BS58) -Strand, CG Methyl.
ATHs_M_CGm
Ath. Shoot (BS58) -Strand, C Methyl.
ATHs_M_Cm
Ath. Root (BS59) Nucleosome Density (Chr 1-5, Normalized to 1M nucl. starts extended to 147bp)
ATHr_Nuc
Ath. Shoot (BS58) Nucleosome Density (Chr 1-5, Normalized to 1M nucl. starts extended to 147bp)
ATHs_Nuc
Ute Kramer, Maria Bernal, Columbia Roots WT, Copper Plus, Single end oligo-dT, Unique hits
Ute Kramer, Maria Bernal, Columbia Shoots WT, Copper Plus, Single end oligo-dT, Unique hits
chr4:
Ath. Root (BS59) +Strand, CG Methyl.
ATHr_P_CGm
Ath. Root (BS59) +Strand, C Methyl.
ATHr_P_Cm
Ath. Shoot (BS58) +Strand, CG Methyl.
ATHs_P_CGm
Ath. Shoot (BS58) +Strand, C Methyl.
ATHs_P_Cm
AT4G13395.1
AT5G35740.1

Figure 5. Comparison Between Root and Shoot in UCSC Genome Browser. (**A and B**) Top: Select Extensin Genes: gene on -strand on left (**A**), gene on +strand on right (**B**). (**C**) Bottom: Example of a transposon/pseudogene. Transposons that had a significant difference in nucleosome density between root and shoot showed the most difference in expression level. Blue histograms: CG methylation on same strand as genes in roots(top)/shoots(bottom). Green histograms: C methylation on same strand as genes in roots(top)/shoots(bottom). Gray histograms: Nucleosome density in roots(top)/shoots(bottom). Brown and dark green histograms: Density of mapped mRNA in roots(brown, top)/shoots(dark green, bottom).

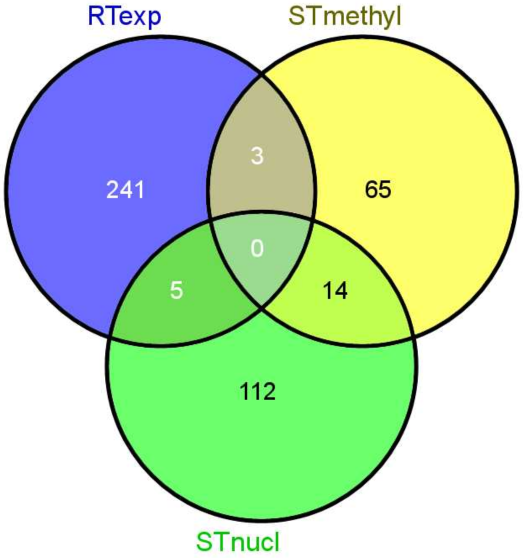
Supplementary Figure – Overlap Between Differential Genes

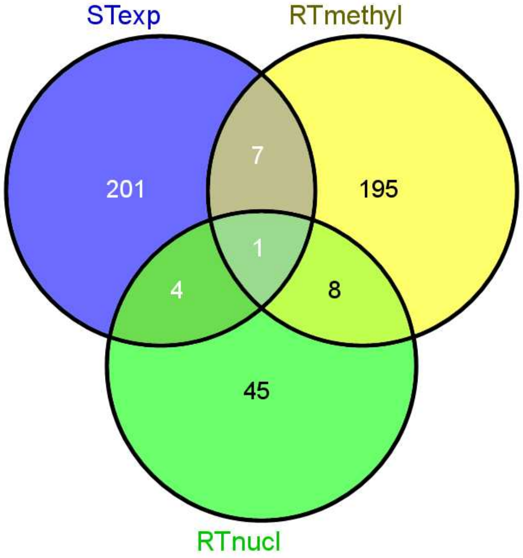




RT/STexp - Expression: >=20-fold difference
RT/STmethyl - Methylation: Average CG methylation level difference of 10% (on an absolute scale, not relative to the methylation level of the same gene between root and shoot)
RT/STnucl - Nucleosomes: >=2.5-fold difference
p values: See supplementary table 1

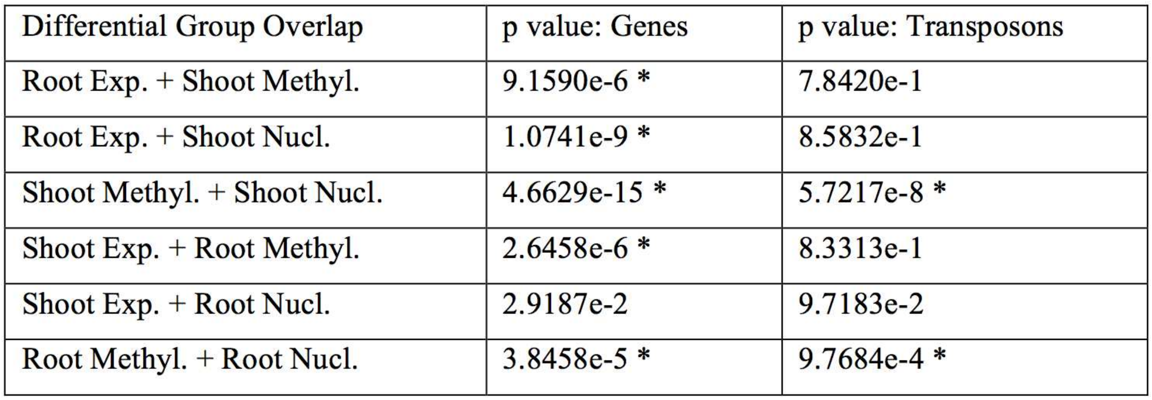
Supplementary Figure – Overlap Between Differential Transposons





RT/STexp - Expression: >=20-fold difference
RT/STmethyl - Methylation: Average CG methylation level difference of 10% (on an absolute scale, not relative to the methylation level of the same gene between root and shoot)
RT/STnucl - Nucleosomes: >=2.5-fold difference
p values: See supplementary table 1

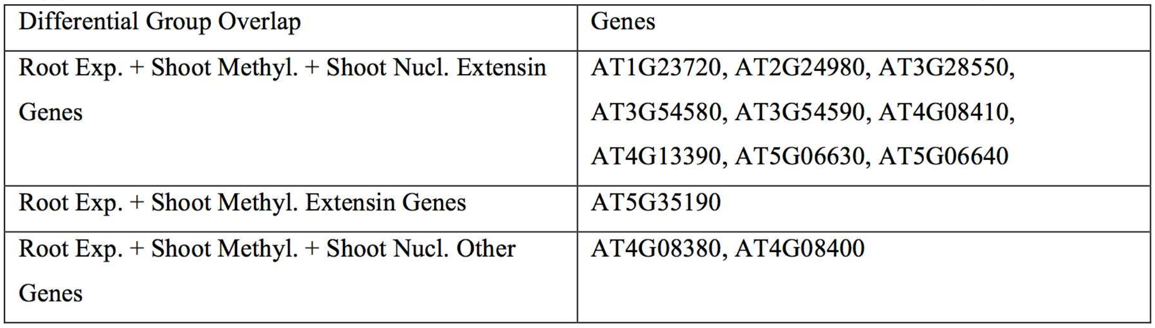
Supplementary Table 1 – p values of Overlap Between Genes Based on Hypergeometric Distribution



Differential Group Overlap	p value: Genes	p value: Transposons
Root Exp. + Shoot Methyl.	9.1590e-6 *	7.8420e-1
Root Exp. + Shoot Nucl.	1.0741e-9 *	8.5832e-1
Shoot Methyl. + Shoot Nucl.	4.6629e-15 *	5.7217e-8 *
Shoot Exp. + Root Methyl.	2.6458e-6 *	8.3313e-1
Shoot Exp. + Root Nucl.	2.9187e-2	9.7183e-2
Root Methyl. + Root Nucl.	3.8458e-5 *	9.7684e-4 *

* Significant: P < 0.025
Root(Shoot) Exp. - 20-fold higher expression in roots(shoots) Root(Shoot) Methyl. - Methylation level 10% higher in roots(shoots) Root(Shoot) Nucl. - 2.5-fold higher nucleosome density in roots(shoots)

Supplementary Table 2 – Extensin Genes and Differential Gene Overlap



Differential Group Overlap	Genes
Root Exp. + Shoot Methyl. + Shoot Nucl. Extensin Genes	AT1G23720, AT2G24980, AT3G28550, AT3G54580, AT3G54590, AT4G08410, AT4G13390, AT5G06630, AT5G06640
Root Exp. + Shoot Methyl. Extensin Genes	AT5G35190
Root Exp. + Shoot Methyl. + Shoot Nucl. Other Genes	AT4G08380, AT4G08400

Root Exp. - 20-fold higher expression in roots

Shoot Methyl. - Methylation level 10% higher in shoots Shoot Nucl. - 2.5-fold higher nucleosome density in shoots

References

1. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 2003; 33(Suppl):245-54; PMID:12610534; http://dx.doi.org/10.1038/ng1089
2. Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science 2002; 297:1833-7; PMID:12193640; http://dx.doi.org/10.1126/science.1074973
3. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 2010; 107:8689-94; PMID:20395551; http://dx.doi.org/10.1073/pnas.1002720107
4. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 2008; 452:215-9; PMID:18278030; http://dx.doi.org/10.1038/nature06745
5. Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D. Genome-wide demethylation of Arabidopsis endosperm. Science 2009; 324:1451-4; PMID:19520962; http://dx.doi.org/10.1126/science.1172417
6. Song F, Mahmood S, Ghosh S, Liang P, Smiraglia DJ, Nagase H, Held WA. Tissue specific differentially methylated regions (TDMR): Changes in DNA methylation during development. Genomics 2009; 93:130-9; PMID:18952162; http://dx.doi.org/10.1016/j.ygeno.2008.09.003
7. Zemach A, Kim MY, Silva P, Rodrigues JA, Dotson B, Brooks MD, Zilberman D. Local DNA hypomethylation activates genes in rice endosperm. Proc Natl Acad Sci U S A 2010; 107:18729-34; PMID:20937895; http://dx.doi.org/10.1073/pnas.1009695107
8. Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, et al. Relationship between nucleosome positioning and DNA methylation. Nature 2010; 466:388-92; PMID:20512117; http://dx.doi.org/10.1038/nature09147
9. Feng S, Rubbi L, Jacobsen SE, Pellegrini M. Determining DNA methylation profiles using sequencing. Methods Mol Biol 2011; 733:223-38; PMID:21431774; http://dx.doi.org/10.1007/978-1-61779-089-8_16
10. Chen PY, Cokus SJ, Pellegrini MBS. BS Seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics 2010; 11:203; PMID:20416082; http://dx.doi.org/10.1186/1471-2105-11-203
11. Bernal M, Casero D, Singh V, Wilson GT, Grande A, Yang H, Dodani SC, Pellegrini M, Huijser P, Connolly EL, et al. Transcriptome sequencing identifies SPL7-regulated copper acquisition genes FRO4/FRO5 and the copper dependence of iron homeostasis in Arabidopsis. Plant Cell 2012; 24:738-61; PMID:22374396; http://dx.doi.org/10.1105/tpc.111.090431
12. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009; 10:R25; PMID:19261174; http://dx.doi.org/10.1186/gb-2009-10-3-r25
13. Oliveros JC. An interactive tool for comparing lists with Venn Diagrams. 2007
14. Neubauer JD, Lulai EC, Thompson AL, Suttle JC, Bolton MD. Wounding coordinately induces cell wall protein, cell cycle and pectin methyl esterase genes

involved in tuber closing layer and wound periderm development. J Plant Physiol 2012; 169:586-95; PMID:22251796; http://dx.doi.org/10.1016/j.jplph.2011.12.010

15. Velasquez SM, Ricardi MM, Dorosz JG, Fernandez PV, Nadra AD, Pol-Fachin L, Egelund J, Gille S, Harholt J, Ciancia M, et al. O-glycosylated cell wall proteins are essential in root hair growth. Science 2011; 332:1401-3; PMID:21680836; http://dx.doi.org/10.1126/science.1206657

Chapter 4: Conclusion

High-throughput sequencing makes it possible to study DNA samples at a genome-wide scale with single base accuracy. The study of DNA methylation has especially benefited from high-throughput sequencing through the use of bisulfite-treated DNA. With these techniques DNA methylation can be studied in a similar manner to the DNA sequence itself. Using aligned sequences from bisulfite converted DNA, changes in methylation appear as mutations in the nucleotide sequence. Additionally DNA methylation can be studied at single-base precision so patterns in methylation levels of genes, transposons and nucleosomes become apparent. However when comparing methylation between two different samples, the differences may be subtle and require extensive analysis to find those that are statistically significant.

The first project dealt with determining the degree to which DNA methylation is conserved on an evolutionary time scale. This was done by comparing a set of genes in Arabidopsis that were duplicated between 40 and 80 million years ago. An analysis of the aligned sequences found that CG methylation is significantly conserved in genes while CHH methylation is significantly conserved in repeat regions. The highest methylation levels were found to be in CG methylation in both genes and repeats but repeats also have methylation at CHG and CHH sites which lack methylation in genes. After analyzing aligned sequences, the transition rates for individual bases in the alignment were computed. The log-odds substitution matrices that were generated indicate that cytosine methylation is significantly conserved especially within genes. The methylation level was also compared and is significantly conserved at fully methylated sites in genes and in repeats although overall conservation is lower, the level of methylation is conserved among partially methylated sites.

The focus of the second project was on the differences between the methylation patterns between tissue types and the relationship between methylation levels and nucleosome densities. The average genome-wide methylation level did not show a significant difference between shoot and root but a comparison using single-site differential methylation shows a completely different picture with nearly twice as many methylation sites in shoots that are hypermethylated relative to roots. To compare methylation levels in the gene body a metagene analysis of the level of differential methylation sites was performed and found that among genes that have tenfold higher expression in root, differential methylation is substantial especially near the transcription start site with much lower methylation in roots. Nucleosome density was also compared between shoot and root and pericentromeric regions were found to have higher nucleosome density in shoots as well as before transcription start sites in genes that are preferentially expressed in roots. An analysis with methylation relative to nucleosome start site shows that methylation levels follow the same 180-base periodicity as nucleosome density with methylation peaking around 25 bases before each nucleosome start site. Finally a group of related genes, all part of the extensin family were found to have significantly lower methylation and nucleosome density in roots while showing a much higher expression level. Methylation and nucleosome density tracks were created for the UCSC genome browser and these extensin genes were compared showing characteristics consistent with the previous findings of this project.

DNA methylation plays a major role in epigenetics and the ability to study DNA methylation on a genome-wide scale with single base precision provides valuable insight into the process by which DNA methylation interacts with nucleosome positioning, gene expression and the underlying DNA sequence. Evaluation of DNA methylation at the genome-wide level is useful in determining how methylation is conserved or changed on an evolutionary time scale or finding groups of genes that have a common function or are related when comparing samples from different tissue types. Being able to analyze DNA methylation with single base precision is useful when comparing aligned sequences or using an approach based on single-site differential methylation which is more effective at finding epigenetic changes than methods based on methylation level alone. With high-throughput sequencing, these methods allow epigenetics to be studied at the same scale as the genome sequence.