Running head: Natural variation of DNA methylation in maize

Corresponding author:
Nathan M. Springer
Department of Plant Biology
250 Biosciences Center
1445 Gortner Ave
Saint Paul MN 55108
Phone: (612) 624-6241
Fax: (612) 625-1738
E-mail: springer@umn.edu

Research Area: Focus issue on “Chromatin and Epigenetics (June 2015)”
Examining the causes and consequences of context-specific differential DNA methylation in maize

Qing Li¹*, Jawon Song²*, Patrick T. West¹, Greg Zynda², Steven R. Eichten¹, Matthew W. Vaughn², Nathan M. Springer¹

¹Microbial and Plant Genomics Institute; Department of Plant Biology, University of Minnesota, Saint Paul, MN 55108 USA
²Texas Advanced Computing Center, University of Texas-Austin; Austin, TX 78758 USA

*These authors contributed equally to this manuscript

Summary: A small portion of the maize genome exhibits natural variation in DNA methylation patterns that can be classified based on sequence context, flanking methylation patterns and nearby genomic features.
This work was supported by a grant from the National Science Foundation to NMS and MWV. (DBI-1237931).

Corresponding author:
Nathan M. Springer
springer@umn.edu
Abstract

DNA methylation is a stable modification of chromatin that can contribute to epigenetic variation through the regulation of genes or transposons. Profiling of DNA methylation in five maize (Zea mays) inbred lines found that while DNA methylation levels for >99% of the analyzed genomic regions are similar there are still 5,000 to 20,000 context-specific differentially methylated regions (DMRs) between any two genotypes. The analysis of identical-by-state genomic regions that have limited genetic variation provided evidence that DMRs can occur without local sequence variation but they are less common than in regions with genetic variation. Characterization of the sequence-specificity of DMRs, location of DMRs relative to genes and transposons and patterns of DNA methylation in regions flanking DMRs reveals distinct subset of DMRs. RNAseq profiling of the same tissue revealed that only ~20% of genes with qualitative (on-off) differences in gene expression are associated with DMRs and there is little evidence for association of DMRs with genes that show quantitative differences in gene expression. We also identify a set of genes that may represent cryptic information that is silenced by DNA methylation in the reference B73 genome. Many of these genes exhibit natural variation in other genotypes suggesting the potential for selection to act upon existing epigenetic natural variation. This study provides insights into the origin and influences of DMRs in a crop species with a complex genome organization.
Introduction

DNA methylation occurs in both animals and plants, and has been reported to play an important role in a number of processes including the establishment of imprinting in animal and plant species (Jirtle and Skinner 2007; Gehring et al., 2009; Waters et al., 2011; Rodrigues et al., 2013; Zhang et al., 2014), the whole-chromosome inactivation in animals (Wutz 2011) and the silencing of transposons (Chan et al., 2005). DNA methylation has also been suggested to be involved in development, response to environmental stresses, human disease and other processes (Jones and Baylin 2002; Jirtle and Skinner 2007; Dowen et al., 2012; Yu et al., 2013). Although DNA methylation can play a role in epigenetic regulation that does not require alterations in DNA sequence there is also evidence that DNA methylation can be triggered by structural rearrangements (Bender and Fink 1995) or transposon insertions (Eichten et al., 2012). Differences in DNA methylation among individuals in a population are likely to include both epigenetic and genetic causes.

The establishment and maintenance and DNA methylation are complex processes involving a variety of enzymes (Law and Jacobsen 2010; Matzke and Mosher 2014). Establishment of DNA methylation is generally guided by RNA-directed DNA methylation pathway, while there are distinct maintenance pathways for cytosines at different sequence contexts. CG methylation is maintained by MET1 during DNA replication process and requires the VIM/UHRF factors to target methylation to hemi-methylated sites. CHG methylation is maintained by the plant specific chromomethylases, such as the Arabidopsis thaliana CMT3 (Law and Jacobsen 2010). The maintenance of CHG methylation involves a self-enforcing loop with H3K9me2 (Johnson et al., 2007; Du et al., 2012). CHH methylation can be maintained by either the RNA-directed DNA methylation or CMT2 pathway (Law and Jacobsen 2010; Zemach et al., 2013; Stroud et al., 2014). In plants, DNA methylation can be removed by a family of DNA glycosylase enzymes that include DEMETER and ROS1 (Zhang and Zhu 2012; Zhu 2013). The specific targeting of methylation / demethylation activities and the fidelity of these processes can influence the occurrence of DNA methylation variation. While we have focused on describing the key genes/enzymes influencing DNA methylation in Arabidopsis there are also homologs of most of these genes in maize (Li et al., 2014) and it is likely that the majority of these pathways are conserved and performing similar functions in maize.

The analysis of mutant accumulation Arabidopsis lines that were grown for 30 generations in the absence of specific selective pressures finds that the methylation levels at specific cytosines can vary at a rate substantially higher than the observed SNP rate (Becker et al., 2011; Schmitz et al., 2011). However, differentially methylated regions (DMRs) are quite rare and occur at frequencies similar to genetic
changes (Becker et al., 2011; Schmitz et al., 2011). Studies of identical-by-state (IBS) regions of the maize (\textit{Zea mays}) genome also provide evidence for stochastic changes in DNA methylation patterns in the absence of local sequence changes (Eichten et al., 2011). However, structural rearrangements and transposon insertions can also influence the DNA methylation patterns for specific alleles (Bender and Fink 1995; Durand et al., 2012; Castelletti et al., 2014). The analysis of genetically diverse populations of Arabidopsis (Schmitz et al., 2013b), rice (Chodavarapu et al., 2012), maize (Eichten et al., 2013; Regulski et al., 2013) or soybean (Schmitz et al., 2013a) finds numerous DMRs. While there are examples in which these DMRs are caused by local sequence rearrangements (Eichten et al., 2012; Schmitz et al., 2013b) it is likely that these include epigenetic changes as well.

There is evidence that some targets of epigenetic silencing acquire DNA methylation and therefore DNA methylation is often considered to be related to gene expression. However, various studies of differential methylation in plant populations find that only a subset of DMRs are associated with altered expression for nearby genes (Eichten et al., 2013; Schmitz et al., 2013b). The relationships between gene expression and DNA methylation differ depending upon the sequence context for DNA methylation and location of the DMR relative to the gene. Gene-body DMRs that occur within the central transcribed portion of the gene usually include CG methylation only (Saze and Kakutani 2011; Schmitz et al., 2013b). There is little evidence that these gene-body CG DMRs are associated with gene expression differences and this gene-body methylation is commonly found in highly expressed genes. In maize, there is evidence that many highly expressed genes contain high levels of CHH methylation within several hundred base pairs of the transcription start site (TSS) (Gent et al., 2013). In contrast, high levels of CG or CHG methylation near the TSS are often associated with reduced gene expression. However, it remains unclear what portion of differential gene expression in plant populations might be attributed to altered levels of DNA methylation.

In this study, we profiled context-specific DNA methylation patterns in five diverse maize inbred lines to identify and characterize context-specific DMRs. We were able to develop methods to robustly identify context-specific DMRs in the maize genome. DMRs are found even in IBS portions of the maize genome but are more common in regions of the genome with genetic variation. The characterization of the overlap in differences for different sequence contexts, location relative to genes and transposons and methylation patterns for flanking regions allowed us to identify discrete types of DMRs. The contribution of DMRs to differential gene expression was also evaluated by assessing gene expression patterns in the same tissue. DMRs are associated with ~20% of genes with qualitative (on-off) differences in gene expression but rarely contribute to quantitative differences in gene expression. In addition, we identified a set of genes
with potential cryptic information in B73 reference genome, some of which were already accessible to breeders in other genotypes.

Results and Discussion

Whole-genome bisulfite sequencing was performed using DNA isolated from seedling leaf tissue of five maize inbred lines: B73, Mo17, Oh43, CML322 and Tx303. B73 is the genotype used to produce the reference genome sequence of maize (Schnable et al., 2009) and represents the stiff-stalk heterotic group. Mo17 and Oh43 are classified as non-stiff stalk temperate lines while CML322 and Tx303 are tropical inbred lines (Flint-Garcia et al., 2005). These represent three of the major groups of domesticated maize germplasm (Flint-Garcia et al., 2005). These genotypes are all parents within the nested association mapping (NAM) population developed for quantitative trait mapping in maize (McMullen et al., 2009).

For each genotype 140 to 185 million reads were obtained, providing 3.5 to 6X coverage with >99% conversion rates (Supplemental Table S1). The genome-wide DNA methylation levels were quite similar for each of the five genotypes. B73 had slightly higher levels of DNA methylation but this may be due to the fact that repetitive regions will map at the highest frequency within the reference genome and may lead to slightly higher overall estimates of DNA methylation. The genome was grouped into 100bp non-overlapping sliding tiles. The coverage and context-specific DNA methylation levels were calculated for each 100bp tile in the maize genome. The use of 100bp tiles simplifies comparisons among genotypes and also likely captures the biologically meaningful units of chromatin, such as nucleosomes.

Discovery of context-specific DMRs

A primary goal was to identify a robust set of DMRs between genotypes. There are biological differences in the abundance and variation of CHH methylation compared to CG/CHG methylation and therefore different criteria were employed to discover these distinct types of DMRs. CG and CHG DMRs were discovered by comparing each of the four non-reference genotypes with B73 for 100bp tiles that had at least three sites (CG or CHG) and at least 2X coverage in the two lines being contrasted. After applying these criteria we were able to compare DNA methylation levels for 5.6 to 6.6 million 100bp tiles, representing 27% to 32% of the maize genome (Table 1). It is worth noting that the analysis of DNA methylation differences will be focused on single-copy regions of similarity between inbred lines and we cannot compare polymorphic transposon insertions using this approach. It is likely that there is DNA methylation at many of these unalignable regions (Seymour et al., 2014). A tile was classified as differentially methylated if the difference in DNA methylation levels between the two lines was greater than 60% for CG or CHG contexts. If two tiles within 100bp of each other were both classified as differentially methylated then they were merged together to create a larger DMR. In each of the four
8 of the 100bp tiles do not exhibit differential methylation (Table 1). However, there are 15,684 to 22,204 CG or CHG DMRs in each of the four contrasts (Table 1). While over 50% of the DMRs encompass a single 100bp, there are many examples of multiple adjacent 100bp tiles that exhibit differential methylation resulting in average DMR sizes of 199 to 218bp (Supplemental Fig. S1A).

The CG and CHG DMRs are distributed throughout the maize genome, with a slight enrichment on chromosome arms (Fig. 1A; Supplemental Fig. S1B).

CHH methylation tends to be present at lower levels compared to CG or CHG methylation both on a genomic scale and at specific loci (Gent et al., 2013; West et al., 2014). Often regions with “elevated” CHH methylation will only exhibit 20% to 40% methylation of CHH sites. Given the lower levels of this mark we used a different set of criteria to identify CHH DMRs. We identified 100bp tiles in which one genotype has <5% CHH methylation and the other genotype exhibits >25% CHH methylation. Even using these relaxed criteria the number of CHH DMRs was 4 to 9 fold lower than the number of CG or CHG DMRs in each contrast (Table 1). It was less common to identify two adjacent 100bp tiles that both exhibit CHH differences resulting in smaller average size of CHH DMRs (Supplemental Fig. S1A).

Validation of context-specific DMRs

Several approaches were used to validate the DMRs. The most robust validation utilized a sequence-capture bisulfite sequencing approach (Li et al., 2014) to perform targeted bisulfite sequencing for a set of ~4,000 genomic regions in independent samples of each of these genotypes. This provided coverage for a
subset of the DMRs in each of the four contrasts (Fig. 1B) and there is quite strong agreement (R=0.96 to 0.99) in the DNA methylation levels measured in the two approaches for CG and CHG DMRs (Supplemental Fig. S2). DMRs were classified as validated, supported or not supported based on the observed methylation levels in the SeqCap Epi dataset (Fig. 1B; Supplemental Fig. S2). The majority (>90%) of CG and CHG DMRs are validated or supported by analysis of DNA methylation in independent DNA samples from the same genotypes using a different method (Fig. 1B). In contrast, only 39% to 62% of CHH DMRs are supported or validated. The correlation between whole-genome bisulfite sequencing (WGBS) and SeqCap Epi was not improved when we used 50bp or 200bp tiles for DMR discovery (Supplemental Table S2). The proportion of DMRs that are validated is only slightly improved when we only assessed regions with higher (3X or 5X) coverage (Supplemental Table S3) suggesting that our use of a low-coverage cut-off is not strongly reducing our validation rate. Many of the CG and CHG DMRs can also be supported by analysis of a methylated DNA immunoprecipitation dataset (Supplemental Fig. S3).

We compared our 100bp tile approach that used simple cut-offs for coverage and methylation difference with the eDMR approach (Li et al., 2013) that utilized statistical approaches to discover DMRs. This approach does not require pre-defined windows but instead will search for context-specific DMRs across the genome. Implementing the eDMR algorithm identified a smaller number of DMRs than the tile-based method described above for CG and CHG DMRs (Supplemental Fig. S4A). A comparison of the DMRs identified using the tile-based method and eDMR reveals that most (>75%) of the CG or CHG DMRs identified by eDMR are also identified using the tile-based method (Supplemental Fig. S4A). However, only <10% of the CHH eDMRs are also found using the tile-based method (Supplemental Fig. S4A). For most of the contrasts, the DMRs found by eDMR did not have higher validation rates than the DMRs identified by the tile-based approach (Supplemental Fig. S4B). We elected to focus on the tile-based DMRs for the remainder of the study as they had validation rates that we similar to, or better than, the validation rate for the eDMR approach and the tile-based DMRs also provided a common coordinate system for cross-sample comparisons. The lower validation rates for CHH DMRs found using either the tile-based or eDMR method may reflect biological variation in CHH patterns in different plants or technical issues in the discovery. The remaining analyses focus on the CG and CHG DMRs that could be validated at high rates.

**Characterization of CG and CHG DMRs**

While the DMRs were identified for specific sequence contexts it was clear that in some cases the same genomic region exhibits changes in multiple sequence contexts (Fig. 2A). Many (~40%) of the CG
DMRs also have over 60% differences in the level of CHG methylation (Supplemental Fig. S5A). Across the four genotypes, ~25% of the CG DMRs show similar levels of CHG methylation in both genotypes. In these cases the CHG methylation levels are usually low in both lines (pie-charts in Supplemental Fig. S5A). The CG DMRs rarely exhibit differences in CHH methylation and the CHH methylation levels are quite low for both genotypes (Supplemental Fig. S5B). A similar analysis of CHG DMRs reveals that many CHG DMRs also exhibit differences in CG methylation (Supplemental Fig. S5C). However, when CHG DMRs have similar levels of CG methylation in both genotypes the level of CG methylation tends to be high in both genotypes the majority of the time (Supplemental Fig. S5C pie-charts). Similar to CG DMRs, the CHG DMRs generally don’t show difference in CHH methylation and the CHH level is very low for both genotypes (Supplemental Fig. S5D). The analyses in Figure 2A and Supplemental Figure S5 suggest that CG and CHG DMRs can be divided into three types, CG-only, CG/CHG and CHG-only. For each genotype contrast we merged the CG and CHG DMRs into a single list and classified each DMR as CG-only, CG/CHG or CHG-only.
one of the three types or unclassified. In each contrast about 50% of the DMRs that could be classified are CG/CHG and roughly 25% are CG-only or CHG-only (Supplemental Table S4). The locations of the three types of DMRs relative to genes and transposons was assessed relative to randomly selected regions (Fig. 2B-C). All three types of DMRs are enriched for being located within genes or both genes and transposable elements (TEs), and depleted for being located within TEs. The depletion of being located near TEs is less pronounced in CG/CHG DMRs. Many (~70%) of the CG/CHG DMRs are still near or in TEs although this percentage is less than the random set (Fig. 2B). Similar to findings in Arabidopsis and soybean (Schmitz et al., 2013a, 2013b), the CG-only DMRs show the most enrichment for being near filtered genes sets (FGS) genes and the CHG-only DMRs show the most enrichment for being near both genes and TEs (Fig. 2B). These patterns are similar across the four contrasts (Supplemental Fig. S5).

The enrichment of CG-only DMRs near or within genes, which generally have low CHG methylation, explains the observation that CHG levels are generally low at CG-only DMRs. Similarly, the enrichment of CHG-only DMRs near both genes and TEs, which generally have both high CG and CHG methylation, explain the observation that CHG-only DMRs usually have high CG levels. The type of nearby TEs for the CG/CHG and CHG-only DMRs are different (Fig. 2C). Though they are both less likely to be near long terminal repeat (LTR)-spreading elements compared to the random set, the CHG-only DMRs are enriched for being near TEs classified as “other” while CG/CHG DMRs are often near LTR-spreading or terminal inverted repeat (TIR) elements (Fig. 2C).

In order to understand how DMRs might arise we assessed the DNA methylation profiles in the regions that flank DMRs. We envisioned three scenarios for the creation of DMRs including (1) a local gain of methylation in one genotype in a region of low methylation (low-low or LL), (2) a local loss of DNA methylation in one genotype in a region that is highly methylated (high-high or HH) or (3) a difference in the position of the boundary between methylated and unmethylated DNA in two genotypes (high-low or HL). Examples of each of these three types of DMRs are provided in Fig. 3A-C. The DNA methylation levels for the second and third tiles on each side of the DMR were averaged to determine the flanking DNA methylation level. The closest tile was not used as we expect that the actual edge of the DMR may be located in the middle of this tile. This analysis was restricted to the DMRs that had coverage for both flanking regions in both genotypes so only a subset of the DMRs was assessed, ranging from 296 to 2324 in the four genotype contrasts. Clustering of the DNA methylation levels of the left and right flanks for DMRs (Fig. 3D-E) reveals that the HH and HL patterns are more commonly observed than the LL patterns for all contrasts (Fig. 3F; Supplemental Fig. S6). The relatively rare LL pattern may simply be due to the fact that the majority of the maize genome is methylated. Highly methylated tiles (CG and CHG) are ~10-fold more common than tiles with low CG/CHG methylation.
By combining the flanking DNA methylation patterns and the context-specificity of DNA methylation differences with the proximity to genes or TEs we found that the LL patterns are enriched for CG-only DMRs near genes (Fig. 3F; Supplemental Fig. S6). This agrees with the fact that genic regions were generally not methylated and the main methylation type in genic region is CG methylation (West et al., 2014). The HL patterns are often found for CG/CHG DMRs and in the region between genes and TEs, suggesting that there might be different stable boundaries between heterochromatin and euchromatin for the different alleles. The HH flanking patterns are enriched for CG/CHG and CHG-only DMRs and are usually near TEs (Fig. 3F; Supplemental Fig. S6). The HH DMRs are enriched for having the high methylation state in four of the five genotypes (Supplemental Fig. S7) and likely reflect a rare loss of DNA methylation in TE regions.

**Figure 3.** Characterization of methylation at DMR flanking regions. A-C, Three representative types of DNA methylation at DMR flanking regions are shown for CG methylation in B73-Mo17 using IGV (Robinson et al., 2011). In (A) the flanking regions both show low methylation (LL), while in (B) the flanking regions both show high DNA methylation (HH) and in (C) high methylation in one flanking region and low methylation for the other flanking region (HL). D-E, Hierarchical clustering (Ward’s method) of methylation levels at the left and right flanking regions of CG (D) or CHG (E) DMRs. The B73 vs Mo17 contrast was shown as an example and clustering for other genotypes provides similar patterns. F, The flanking pattern characterization of B73-Mo17 DMRs was compared with context-specificity of DNA methylation changes and genomic location for B73-Mo17 DMRs (similar plots for the other contrasts are available in Supplemental Fig. S6). The DMR sequence contexts and flanking region types are cross-tabulated to give a total of nine different DMR categories. The size of each pie chart is scaled to the number of DMRs in each category. The number above each pie chart shows the DMR number in that specific category. Gene (red) indicates that a DMR is within 500bp of a filtered gene set gene, TE (blue) indicates that a DMR is within 500bp of a TE, both (yellow) indicates that a DMR is within 500bp of both a gene and a TE and neither (black) indicates that a DMR is greater than 500bp from both annotated genes and TEs.
The CG/CHG, CG-only or CHG-only DMRs identified in each contrast were merged to form a non-redundant list of DMRs across all genotypes to investigate the frequency of the high and low DNA methylation states within the five genotypes. The number of genotypes with coverage for each DMR was determined (Supplemental Fig. S8A). The 18,362 CG/CHG DMR, 12,789 CG-only and 10,671 CHG-only DMRs with coverage in all five genotypes were used to perform hierarchical clustering of DNA methylation levels (Supplemental Fig. S8B). This reveals that there are more examples in which B73 exhibits high levels of DNA methylation relative to at least one other genotype than examples in which B73 has low methylation. A linear discriminant analysis was used to classify each genotype as high or low methylated and then the ratio of high to low methylated alleles was used to generate a histogram (Supplemental Fig. S8C). It was more common to have three or four genotypes with high methylation levels than to have only one or two genotypes that were highly methylated.

**Role of genetic variation in contributing to DMRs**

Differential methylation could occur due to purely epigenetic variation or it may be driven by local or remote sequence variation. One way to assess the role of local sequence variation is to study the rate of DMRs in IBS genomic regions relative to the genome-wide rate of DMRs. IBS regions in each of the four genotype contrasts were identified using Beagle (Browning and Browning 2011). We then identified genomic regions >1Mb in which >90% of the regions were classified as being IBS resulting in a range of 1 to 17.75 Mb of IBS regions in the four contrasts (Supplemental Fig. S9A). The presence of DMRs in these IBS regions suggests that DMRs can occur even in regions without substantial genetic variation. The DMRs located within IBS regions are not enriched for specific contexts, flanking patterns of DNA methylation or location relative to genes and transposable elements (TEs) (Supplemental Fig. S9B-D). However, the frequency of DMRs in the IBS regions is usually lower than the genome-wide proportion of DMRs (Supplemental Fig. 9B). This suggests that local sequence variation might account for a portion of the of DMRs while either epigenetic variation or trans-acting remote sequence variation would account for the remaining DMRs.

**Contribution of DMRs to differentially expressed genes**

Many studies have reported that there is a limited negative correlation between gene expression levels and DNA methylation. Previous studies have also found that a small proportion of DMRs are negatively correlated with expression levels of nearby genes (Eichten et al., 2013; Schmitz et al., 2013b). In this study we decided to assess the frequency of differentially expressed genes that contain DMRs to better understand how frequently differential gene expression might involve DNA methylation changes. RNAseq was performed on three biological replicates of RNA isolated from the same tissues used for
Figure 4. Association of DNA methylation variation with differential expression. A-C, The difference in DNA methylation levels were calculated for differentially expressed genes (methylation at high expressed gene (H) minus methylation at lower expressed gene (L)) for CG (A), CHG (B) and CHH (C) contexts. Differentially expressed genes were divided into six categories (color code indicated within the plot; nDEG, not differentially expressed genes) based on the fold change between the highest (H) and the lowest (L) genotypes. Methylation levels were calculated for each category as the average of all the genes in that category. The two vertical lines represent the -200 bp upstream of a gene and 600 bp into a gene from 5’ end. TSS, transcriptional start site. The arrow indicates the direction of transcription. D-F, The actual methylation levels for the on-off category at CG (D), CHG (E), and CHH (F) in the genotypes with the highest or the lowest gene expression level. G-H, Percentage of genes in each expression category with CG (G), CHG (H) DMRs. DEG, differentially expressed genes. I, Distribution of CG/CHG difference between off and on genotypes at -200 to 600bp regions.

WGBS for each of the five genotypes. Differentially expressed genes were identified using DEseq (Anders and Huber 2010) and then the differentially expressed genes were divided according to the fold-change between the two genotypes (Fig. 4A-C; Supplemental Table S5). Plots of the difference in CG or CHG methylation for the two genotypes reveal that there are some differences in CG or CHG methylation for genes that exhibit very strong (>10-fold) differential expression with the high expressed genotype containing lower levels of CG and CHG methylation near the TSS but higher levels of CHH methylation...
5’ of the TSS (Fig. 4A-C). Plots of the actual DNA methylation level for differentially expressed genes with qualitative change show that both alleles have similar methylation levels in flanking regions but the average CG/CHG level is low and the CHH is high for the high expressing genotype (Fig. 4D-F). One interesting finding is that the regions that show the most association between DNA methylation and gene expression are located around the TSS, corresponding to the region that show the lowest DNA methylation in the meta-methylation profiles of genic and surrounding regions. This suggests an important role of DNA methylation in this region for influencing transcription. Methylation in different contexts associates differently with gene expression. CG methylation differences for on-off genes occur only in the first ~1kb of the gene. In contrast, differences in CHG methylation persist further into the gene. While CG and CHG methylation show a negative correlation with gene expression, CHH methylation shows a positive correlation and the region that shows the highest correlation is 200 to 500bp upstream of the gene, corresponding to the CHH island described in previous studies (Gent et al., 2013; West et al., 2014). Interestingly, the same genes that exhibit elevated CG and CHG methylation in the region downstream of the TSS also exhibit reduced CHH methylation in the region upstream of the TSS, providing further evidence for a role of elevated CHH methylation in promoter in gene expression.

The number of genes with DMRs was assessed for genes with different expression patterns (Fig. 4G-H). In general, the frequency of genes with DMRs is similar for non-differentially expressed genes and genes with less than 5-fold change in expression. Genes with >10-fold change or on-off genes show significantly higher levels of DMRs. In >95% of these cases the DMR exhibits higher methylation for the genotype with low, or no, expression. While the genes with on-off change in expression are enriched for DMRs it is worth noting that the majority (~80%) of the differentially expressed genes with on-off expression differences do not have DMRs and exhibit similar methylation in both genotypes (Fig. 4I). The remaining ~20% of genes exhibit large differences in DNA methylation level between the two alleles with lower CG/CHG methylation for the higher expressed allele (Fig. 4I). These findings provide evidence that expression variation for most of the genes is not associated with DNA methylation variation.

Potential for cryptic information in the maize genome

Several studies have created epiRIL populations in Arabidopsis that segregate for DNA methylation levels in chromosomal regions without genetic variation (Johannes et al., 2009; Reinders et al., 2009). These populations have been used to provide evidence for cryptic information in a particular ecotype that can be released by removing DNA methylation. There is some evidence that many of these genes that may also exhibit natural variation for silencing (Johannes et al., 2009; Reinders et al., 2009). We were interested in characterizing a set of genes that may be subject to silencing by DNA methylation in the
reference genotype and determining whether there was natural variation for this silenced state. There are 38,726 maize genes that have coverage for the -200 to +600bp (relative to the TSS) region in B73. Many (~85%) of these genes are expressed in seedling leaf tissue of B73 or have low levels of CG/CHG methylation surrounding the TSS (Fig. 5A; group 1 genes). The tissue-specific expression for the remaining 6,451 genes was assessed using the B73 expression atlas (Sekhon et al., 2013) to identify 2,602
genes that are expressed in other tissues of B73 (Fig. 5A; group 2 genes). The final 3,849 genes that have high levels of CG/CHG methylation and are not expressed in B73 were divided into two groups based on whether they exhibit expression in seedling leaf tissue of a panel of 51 diverse maize genotypes (Eichten et al., 2013). These two groups of genes represent potentially epigenetically silenced genes that exhibit natural variation (1,578 group 3 genes expressed in some genotypes) or fixed silencing in fairly diverse germplasm (2,271 group 4 genes) (Fig. 5A). We assessed whether our data for Oh43, Mo17, CML322 and Tx303 could be used to assess whether the genes in group 3 were related to differences in DNA methylation. There are 119 of the group 3 genes that are expressed in at least one of these four genotypes (but not in B73) and had DNA methylation data for the -200 to +600bp region surrounding the TSS. We found that 28 of the 119 (23.5%) of these genes have very low (<20%) CG and CHG methylation for the genotype with expression suggesting that at least a subset of the group 3 genes are associated with altered methylation of the promoter region.

A number of properties were examined for the genes in the four different groups. The intron length and transcript length are shorter for the group 3 and group 4 genes (Fig. 5B-C). The proportion of genes located within sub-genome 1 and 2 (Schnable et al., 2011) was similar for all four groups of genes (Fig. 5D). However, the proportion of genes located in syntenic positions relative to rice or sorghum was much higher for the group 1 genes (Fig. 5E). The genes in groups 2 to 4 are enriched for being near LTR-spreading transposons, and are much less likely to have a GO annotation (Fig. 5F-G). The proportion of the genes located in the recombination-poor central region of the chromosome was assessed for each group (Fig. 5H). The group 1 genes are similar to all genes but the groups 2 to 4 genes are enriched for being located in the low-recombination central portion of maize chromosomes. This analysis provides evidence that there is potential cryptic information in the maize genome. Nearly half of the highly methylated genes exhibit natural variation for expression levels suggesting that these alleles have been accessible to plant breeders. However, given the location of many of these genes within the recombination poor-regions of the maize genome there may be challenges in isolating favorable linkages of the expressed alleles with other allelic variation. The other half of the genes that are highly methylated and not expressed in B73 do not show expression variation in this tissue of other maize genotypes.

Perturbation of DNA methylation levels may allow the expression of these genes to reveal if there is important cryptic information in the maize genome that could allow for agronomic improvement. One hurdle to this approach is the apparent lethality of strong perturbation of DNA methylation in maize (Li et al., 2014).

Conclusions
DNA methylation has the potential to provide heritable information beyond DNA sequence. In this study we profile the context-specific DNA methylation patterns for five maize genotypes to identify regions with differential methylation. The fact that genomic regions with nearly identical sequence still contain DMRs suggests that at least a portion of the variation for DNA methylation is purely epigenetic. There are distinct sub-types of DMRs in terms of the context-specificity of DNA methylation and the patterns of methylation in regions flanking the DMRs. By assessing gene expression in the same tissue used for methylation profiling we were able to carefully assess the relationship between altered expression and DNA methylation. DNA methylation differences are associated with gene expression changes only for a subset of the genes that show major differences in expression. Genes with quantitative differences in expression (<10-fold changes) rarely exhibit differential DNA methylation. This study also provided an opportunity to study the genes that might provide cryptic information in maize. A set of silent, methylated genes were identified within the reference genome. Some of these genes are expressed in other tissues or other maize genotypes while others are silenced in all samples assessed. These genes may provide a reservoir of cryptic information that could influence traits if reactivated in populations similar to the epiRILs utilized in Arabidopsis.

**Materials and Methods**

**Plant Materials**

The maize (*Zea mays*) inbred lines B73, Mo17, CML322, Oh43 and Tx303 were grown to the third leaf stage. The third leaf was harvested, frozen in liquid nitrogen, and used for DNA and RNA isolation. Genomic DNA was prepared using standard CTAB method. Total RNA was prepared using TRIzol reagent following manufacturers’ instructions.

**Whole-genome bisulfite sequencing and data analysis**

Whole genome bisulfite sequencing libraries were prepared as previously described in Eichten et al. (2013). Libraries that passed quality control (concentration > 2nM and size range between 200 to 500 bp) were sequenced on a HiSeq 2000 using 100 cycles and paired-end mode. Adapters were trimmed using Trim_Galore. Reads were mapped to the maize B73 reference genome version 2 using Bismark (0.10.1, Krueger and Andrews, 2011) allowing 1 mismatch in the alignment (-N 1). Methylated cytosines were extracted from aligned reads using the Bismark methylation extractor under standard parameters. The proportion of CG, CHG, and CHH methylation was determined as weighted methylation levels (Schultz et al., 2012) in 100bp non-overlapping windows across the genome.
Identification of DMR: tile and eDMR methods

In order to identify DMRs using the 100bp tiles the coverage for each tile was determined by dividing the total number of times cytosines of a specific sequence context were covered by a read by the total number of cytosines of a specific sequence context within the tile. CG and CHG DMRs were required to have a minimum of 3 symmetrical CG or CHG sites, at least 2X coverage, and a minimum methylation difference between two genotypes of 60%. CHH sequence context DMRs were required to have a minimum of 6 asymmetrical CHH sites, a minimum of 2X coverage, and one genotype with <5% methylation and the other genotype with >25% methylation. DMRs that were within 100 base pairs of one another had their methylation values averaged and were merged into a single DMR.

In order to identify DMRs using eDMR (Li et al., 2013) the output file from Bismark’s methylation extractor was used to identify differentially methylated cytosines using a R package methylKit using the default parameter (Akalin et al., 2012). These are then used for implementing the eDMR package (Li et al., 2013), which identifies DMRs using default distance cutoff determined automatically by bimodal normal distributions of distances between cytosines. Overlap between the tile DMR and eDMR datasets was determined using the intersectBed tool from the BEDTools genomic toolkit (Quinlan and Hall, 2010).

DMR validation

The SeqCap Epi platform was used to validate some of the DMRs identified in this study using independent biological samples (3rd leaf tissue) for each genotype. Three technical replicates for B73, 2 technical replicates for Mo17, 2 biological replicates for Oh43, 1 sample each for CML322 and Tx303 were used to generate bisulfite converted sequencing libraries that were used to perform sequence-capture. The capture pool targets ~5000 regions of the maize genome (Li et al, 2014). The samples were sequenced using a MiSeq with 100 cycles (paired-end) for B73 and Mo17 and with 150 cycles (paired-end) for Oh43, CML322 and Tx303. The methylation levels from replicates were averaged and used to compare with the methylation levels from the WGBS dataset.

Methylation profiling using methylated DNA immunoprecipitation followed by microarray hybridization had been performed on these five genotypes in a previous study (Eichten et al., 2013). The methylation values for array probes located with 300bp of a DMR identified in this study were extracted from NCBI GEO accession GSE46949. The differential DNA methylation estimates were compared to the differences in WGBS levels determined in this study for each genotype contrast and the correlation of the values was computed using Pearson correlation.

Identical-by-state regions
IBS regions between each genotype were initially identified using the BEAGLE algorithm (Browning and
Browning 2011). These IBS regions were then intersected with 250kb sliding 1Mb tiles. 1Mb tiles that
were composed of at least 90% BEAGLE identified IBS regions were considered as IBS regions for
further analyses.

DMR merging

The DMRs were discovered using a pair-wise comparison strategy by comparing B73 with each of the
four other inbred lines. To get a non-redundant list of DMRs found in all contrasts the DMRs from the
four contrasts for each cytosine context were merged using BEDTools (Quinlan and Hall, 2010). The
number of inbred lines that have data for each DMR was determined and further analyses was performed
on DMRs that have data in all five genotypes. At each DMR, the five lines were assigned to two groups
such that the intra-group variance was minimized and the inter-group variance was maximized.

The CG and CHG DMRs from each genotype contrast were merged using the outer most boundaries.
DNA methylation differences for CG and CHG contrasts were used to classify each DMR into one of the
four categories, CG-only, CHG-only, CG/CHG, and unclassified. CG-only DMRs show >60% CG and
<20% CHG difference, CHG-only DMRs show <20% CG and >60% CHG difference, CG/CHG DMRs
are those that have >60% difference for both CG and CHG, and the other DMRs that can’t satisfy any of
these criteria were classified as unclassified DMRs.

Methylation levels at DMR flanking regions

To determine the methylation level surrounding DMRs, the methylation levels for the three 100bp tiles to
the left and to the right of every DMR were extracted. These tiles were then filtered for the minimum 2X
coverage cutoff and the minimum sequence context-specific sites. The average value of the second and
third windows, if they contained data, was determined as the surrounding methylation value. These values
were used to perform a hierarchical clustering analysis using the Ward’s method. Based on the clustering,
the DMRs were grouped into 4 types: DMRs that show high (or low) methylation on both sides in both
contrasting lines for either CG or CHG methylation (HH, or LL), DMRs that show high methylation on
one side and low methylation on the other side in both contrasting lines for either CG or CHG
methylation (HL), and undefined DMRs that cannot be assigned to any of the three groups.

DMR proximity to genes and transposons

The DMR proximity to genes and TEs was determined using the closestBed tool from the BEDTools
genomic toolkit (Quinlan and Hall, 2010). TE annotations were obtained from the Maize TE Consortium
(ZmB73_5b) and gene annotations were obtained from the maize Filtered Gene Set (ZmB73_5a, FGS).

The random subsets of tiles used as a permutation control for the tile DMRs were generated using a
custom Perl script. The script obtains a specified amount of random tiles that meet the minimum coverage
and minimum sequence specific site count of CG, CHG, or CHH cutoffs.

**RNA sequencing and analysis**

We performed RNA sequencing using the same tissue that was used for methylation analysis. Three
biological replicates were done for each of the five lines. RNA libraries were prepared using Illumina
TruSeq kit following manufacturers’ instructions. We did sequencing on a HiSeq 2000 using paired-end
50 cycles. Raw sequencing reads were analyzed using FASTQC to remove bases with poor quality or
adapter sequences. Trimmed reads were then mapped to the maize reference genome version 2 allowing
for one mismatch using TopHat (Trapnell et al., 2009). Uniquely mapped reads were then filtered using
SAMtools (-q 50) (Li et al., 2009). Only reads that are properly paired, are both mapped in the right
orientation and in the same place were used. Read counts per gene model were then summarized using
BAM2COUNT that can be found on iPlant.

**Differentially expressed genes and their correlation with methylation**

The dataset with raw read counts per gene per genotype was imported into R environment. Read counts
were normalized using DESeq (Anders and Huber 2010), and were subjected to a pair-wise comparison
between B73 and each of the other four lines to identify differentially expressed genes. The differentially
expressed genes from all four contrasts were then combined, and were divided into one of the 5 groups
based on the fold change between the genotypes with the highest and the lowest expression levels: <2, 2
to 5, 5 to 10, >10, and on-off genes. On-off genes are the genes that have detectable mRNA expression in
at least one of the five lines, but have no reads (0) in at least one other line. Methylation levels at each
sequence context were calculated for each gene at a region that spans 2000 bp upstream to 3000 bp (or the
actual gene size if gene size is less than 3000 bp) into the gene, divided into 100 bins. Methylation
difference between the genotypes that show the highest and the lowest expression was calculated for each
bin for each gene.

**Identifying genes with cryptic information**

Genes that had methylation data for the region from -200 bp to 600 bp (relative to the transcription start
site) were identified. The methylation levels for CG/CHG were then averaged over this region for each
gene. Genes with at least 60% CG and CHG methylation were identified and were further filtered using
the expression data to find these that have no detectable expression (normalized read count per kb < 0.5)
in B73. Genes that meet these criteria were further filtered using the B73 expression atlas data from Sekhon et al. (2013) to find the genes that are not developmentally regulated. These genes were further classified into two groups: one group showing expression in some other inbred lines (Eichten et al., 2013) and another group that was not expressed and was considered to be genes with potential cryptic information.

Accession numbers

The WGBS data set has been deposited into NCBI under accession numbers SRR850328, SRR850332, SRX731432, SRX731433 and SRX731434. The RNAseq data set has been deposited into NCBI under accession number SRP052226.

Acknowledgements

Tyler Kent and Jeffrey Ross-Ibarra provided assistance with the discovery of IBS regions. Data analysis was performed using the tools and resources provided by the iPlant Collaborative. Computational support and data storage was provided by the Texas Advanced Computing Center at the University of Texas at Austin. The Minnesota Supercomputing Institute provided access to software and user support for data analyses.

Table

Table 1 DMRs found in this study

<table>
<thead>
<tr>
<th>Features</th>
<th>BvC</th>
<th>BvM</th>
<th>BvO</th>
<th>BvT</th>
<th>Merged</th>
</tr>
</thead>
<tbody>
<tr>
<td># CG tiles with 2X coverage (%) all tiles</td>
<td>5,711,274 (27.7%)</td>
<td>5,867,123 (28.5%)</td>
<td>5,850,207 (28.4%)</td>
<td>5,646,804 (27.4%)</td>
<td>NA</td>
</tr>
<tr>
<td># CHG tiles with 2X coverage (%) all tiles</td>
<td>5,697,086 (27.7%)</td>
<td>5,798,435 (28.2%)</td>
<td>5,835,932 (28.3%)</td>
<td>5,638,371 (27.4%)</td>
<td>NA</td>
</tr>
<tr>
<td># CHH tiles with 2X coverage (%) all tiles</td>
<td>6,462,820 (31.1%)</td>
<td>6,217,268 (30.2%)</td>
<td>6,621,837 (32.2%)</td>
<td>6,408,902 (31.1%)</td>
<td>NA</td>
</tr>
<tr>
<td># CG DMRs</td>
<td>22,204</td>
<td>16,387</td>
<td>20,163</td>
<td>20,434</td>
<td>53,816</td>
</tr>
<tr>
<td># CHG DMRs</td>
<td>20,742</td>
<td>15,684</td>
<td>17,714</td>
<td>19,036</td>
<td>50,899</td>
</tr>
<tr>
<td># CHH DMRs</td>
<td>5,272</td>
<td>6,721</td>
<td>4,410</td>
<td>4,469</td>
<td>16,736</td>
</tr>
<tr>
<td></td>
<td>CG DMRs</td>
<td>CHG DMRs</td>
<td>CHH DMRs</td>
<td>Average CG DMRs size (bp)</td>
<td>Average CHG DMRs size (bp)</td>
</tr>
<tr>
<td>------------------</td>
<td>---------</td>
<td>----------</td>
<td>----------</td>
<td>--------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>% CG DMRs</td>
<td>0.78</td>
<td>0.57</td>
<td>0.67</td>
<td>0.72</td>
<td>210</td>
</tr>
<tr>
<td>% CHG DMRs</td>
<td>0.74</td>
<td>0.50</td>
<td>0.61</td>
<td>0.67</td>
<td>211</td>
</tr>
<tr>
<td>% CHH DMRs</td>
<td>0.09</td>
<td>0.12</td>
<td>0.07</td>
<td>0.08</td>
<td>204</td>
</tr>
<tr>
<td>Average CG DMRs size (bp)</td>
<td>210</td>
<td>211</td>
<td>204</td>
<td>207</td>
<td>210</td>
</tr>
<tr>
<td>Average CHG DMRs size (bp)</td>
<td>218</td>
<td>199</td>
<td>218</td>
<td>214</td>
<td>200</td>
</tr>
<tr>
<td>Average CHH DMRs size (bp)</td>
<td>110</td>
<td>108</td>
<td>109</td>
<td>111</td>
<td>110</td>
</tr>
</tbody>
</table>

NA: not applied

**Figure Legends**

**Figure 1.** Identification and validation of context-specific DMRs among four maize genotype contrasts. A, The genome-wide distribution of DMRs, genes, TEs and coverage is shown using IGV (Robinson et al., 2011). Each track displays the number of features per 1Mb window of the maize genome. Only data from chromosome 1 is shown. B, DMR validation and support rates using an independent sequence-capture bisulfite sequencing dataset. The subset of the DMR tiles (numbers indicated at the base of each bar) that also have coverage in a sequence-capture bisulfite sequencing (SeqCap Epi) dataset are analyzed. Each DMR is classified as validated (green), supported (yellow) or not-supported (gray). Validated DMRs represent DMRs which show >60% (for CG and CHG) or >20% (for CHH) difference in the same direction in the SeqCap Epi assay. Supported DMRs are the DMRs that show 40%-60% (for CG and CHG) or 10%-20% (for CHH) difference in the same direction in the SeqCap Epi assay.

**Figure 2.** Context-specific DMRs and their proximity to genomic features. A, Context-specific methylation levels (0-100%) are plotted for several B73-Mo17 DMRs using IGV (Robinson et al., 2011). The colored region in the center shows context-specific methylation within the DMR while the black shading for the edges shows methylation levels outside of the DMR. B, Pie-charts are used to show the proportion of DMRs that are located near (<500bp) or in genes (red), TEs (blue), both genes and TEs (yellow) or DMRs located >500bp from either a gene or a TE (black). Separate pie-charts are shown for CG/CHG DMRs (>60% difference at both CG and CHG contexts), CG-only DMRs (>60% CG difference and <20% CHG difference) and CHG-only DMRs (>60% CHG difference and <20% CG difference). The pie-chart labeled “random” shows the proportions of each location type for 20,000 randomly selected 100bp tiles that had coverage in our dataset. The size of the pie-chart is scaled to the number of DMRs in the specific category except the random set. C, For the DMRs that are located within or near a TE we classified the type of TE as TIR (blue), LTR-spreading (red), LTR-nonspreading (yellow) or other (black). The spreading/non-spreading classifications are based on Eichten et al. (2012). (B) and (C) show the...
average proportions for the four genotype contrasts, the values for each contrast are available in Supplemental Fig. S5. In (B) and (C), the “*” symbol is used to indicate significant enrichment of the feature relative to 100 random sets at P<0.01, and downwards arrow indicates significant underrepresentation.

**Figure 3.** Characterization of methylation at DMR flanking regions. A-C, Three representative types of DNA methylation at DMR flanking regions are shown for CG methylation in B73-Mo17 using IGV (Robinson et al., 2011). In (A) the flanking regions both show low methylation (LL), while in (B) the flanking regions both show high DNA methylation (HH) and in (C) high methylation in one flanking region and low methylation for the other flanking region (HL). D-E, Hierarchical clustering (Ward’s method) of methylation levels at the left and right flanking regions of CG (D) or CHG (E) DMRs. The B73 vs Mo17 contrast was shown as an example and clustering for other genotypes provides similar patterns. F, The flanking pattern characterization of B73-Mo17 DMRs was compared with context-specificity of DNA methylation changes and genomic location for B73-Mo17 DMRs (similar plots for the other contrasts are available in Supplemental Fig. S6). The DMR sequence contexts and flanking region types are cross-tabulated to give a total of nine different DMR categories. The size of each pie chart is scaled to the number of DMRs in each category. The number above each pie chart shows the DMR number in that specific category. Gene (red) indicates that a DMR is within 500bp of a filtered gene set, TE (blue) indicates that a DMR is within 500bp of a TE, both (yellow) indicates that a DMR is within 500bp of both a gene and a TE and neither (black) indicates that a DMR is greater than 500bp from both annotated genes and TEs.

**Figure 4.** Association of DNA methylation variation with differential expression. A-C, The difference in DNA methylation levels were calculated for differentially expressed genes [methylation at high expressed gene (H) minus methylation at lower expressed gene (L)] for CG (A), CHG (B) and CHH (C) contexts. Differentially expressed genes were divided into six categories (color code indicated within the plot; nDEG, not differentially expressed genes) based on the fold change between the highest (H) and the lowest (L) genotypes. Methylation levels were calculated for each category as the average of all the genes in that category. The two vertical lines represent the -200 bp upstream of a gene and 600 bp into a gene from 5’ end. TSS, transcriptional start site. The arrow indicates the direction of transcription. D-F, The actual methylation levels for the on-off category at CG (D), CHG (E), and CHH (F) in the genotypes with the highest or the lowest gene expression level. G-H, Percentage of genes in each expression category with CG (G), CHG (H) DMRs. DEG, differentially expressed genes. I, Distribution of CG/CHG difference between off and on genotypes at -200 to 600bp regions.
Figure 5. Characterizing the potential for cryptic information within the B73 genome. A, The genes with coverage at the -200 to +600 region (relative to the TSS) were assessed to identify genes that are expressed or have low CG/CHG methylation (group 1), genes that are expressed in other tissues of the B73 expression atlas (group 2), genes that are expressed in seedling tissue of other maize genotypes (group 3) or genes with no detectable expression (group 4). B-C, The distribution of lengths for mRNA transcripts (B) and introns (C) for genes from groups 1-4 is shown using a boxplot. D, The proportion of genes within each group located within sub-genome 1 (Sub1) or sub-genome 2 (Sub2) is shown. E, The proportion of genes in each group that are syntenic with rice is shown. F, The closest TE for the genes in each group were classified. G, The proportion of genes with Gene Ontology annotation is shown for each group of genes. H, The proportion of genes located within the low-recombining central portion of each chromosome is assessed for each group of genes.

Supplemental Data

Supplemental Figure S1. DMR size distribution and genome-wide distribution. A, The upper set of panels shows CG DMRs, the middle panels show CHG DMRs, and the lower panels show CHH DMRs. The number of DMR (N) and the average DMR size (L) in base pairs is shown inside each plot. B, Genome-wide distribution of DMRs, gene and transposon density, number of tiles and identical-by-state (IBS) regions in the four genotype contrasts (BvC – B73 vs CML322; BvM – B73 vs Mo17; BvO – B73 vs Oh43; BvT – B73 vs Tx303).

Supplemental Figure S2. SeqCap validation of DMRs found by WGBS. The contrast was labeled above the plots. In each plot the x-axis shows the difference in methylation (B73 minus other genotype) from the WGBS dataset while the y-axis shows the difference in methylation from the SeqCap Epi dataset. The upper, middle and lower panels show comparison of CG, CHG and CHH DMRs, respectively, in WGBS and SeqCap. The Pearson correlation coefficient was shown inside each plot. Red – validated DMRs, >=60% (20% for CHH DMR) difference in SeqCap; blue – supported DMRs, 40%-60% (10%-20% for CHH DMR) difference; grey – not supported DMRs, <40% (10% for CHH DMR) difference.

Supplemental Figure S3. Comparison of DMRs with methylated DNA immunoprecipitation-microarray data. Microarray probes located within 300bp of each DMR were identified and the array hybridization data for the five genotypes used in this study were extracted from Eichten et al. (2013). The x-axis shows the difference in DNA methylation between the two genotypes (B73 minus other genotype) while the y-axis shows the difference in microarray hybridization values for the two genotypes. The upper, middle and
lower panels show comparison of CG, CHG and CHH DMRs, respectively. The Pearson correlation coefficient was shown inside each plot. The darker the color, the greater the density of data points.

**Supplemental Figure S4.** Comparison of DMRs identified using eDMR or analysis of 100bp tiles. A, Venn diagrams are used to show the overlap of DMRs found using the two different approaches for each sequence context in all four contrasts. The DMRs found using eDMR are shown in green while the DMRs identified through analysis of 100bp tiles are shown in red. B, The proportion of DMRs that are validated or supported by the SeqCap Epi dataset is shown for DMRs found by both methods (blue), DMRs identified by the analysis of 100bp tiles only (red) and DMRs identified only by eDMR (green) for each sequence context.

**Supplemental Figure S5.** Context-specific DMRs and their proximity to genomic features in the four contrasts. A, Distribution of CHG difference at CG DMRs, and the CHG levels (pie-charts) for DMRs with no CHG difference (red parts in the histogram). B, Same as A, but shows CHH difference at CG DMR and CHH levels at CG DMRs without CHH difference. C-D, CG (C) and CHH (D) difference were shown for CHG DMRs. E, Nearby (within 500bp) genomic features for DMRs that occur at different sequence contexts. CG/CHG DMRs are DMRs that have >60% difference at both CG and CHG contexts. CG-only DMRs are DMRs that have >60% CG difference and <20% CHG difference. CHG-only are DMRs that have >60% CHG difference and <20% CG difference. FGS, filtered genes. TE, transposons. Both, both TE and FGS. Neither, neither TE or FGS. F, The nearest TE type for DMRs with different sequence contexts. DMR definition is the same as that in E.

**Supplemental Figure S6.** The closest genomic features for DMRs at different sequence contexts and with different methylation levels at flanking regions. Shown here are the BvC, BvO and BvT contrasts. The DMR sequence contexts and flanking region types are cross-tabulated to give a total of nine different DMR categories. The pie size is scaled to the number of DMRs in each category. The number above each pie chart shows the DMR number in that specific category. Gene (red) indicates that a DMR is within 500bp of a filtered gene set gene, TE (blue) indicates that a DMR is within 500bp of a TE, both (yellow) indicates that a DMR is within 500bp of both a gene and a TE and neither (black) indicates that a DMR is greater than 500bp from both annotated genes and TEs.

**Supplemental Figure S7.** The frequency of the high and low methylation state for DMRs with different flanking patterns. The DMRs that have data in all five genotypes and have coverage for flanking regions were analyzed. A linear discriminant analysis was used to determine the number of genotypes with the high and low methylation state. The % of DMRs within each class of flanking patterns (HH – high-high; HL – high-low; LL – low-low) that have 1, 2, 3 or 4 genotypes with high levels of methylation is plotted.
Supplemental Figure S8. Characterization of merged DMRs across sequence contexts and across contrasts. A, For each sequence context, the DMRs from all four contrasts was merged to form a non-redundant list of DMRs. The number of non-redundant DMRs with coverage in 2, 3, 4, or 5 genotypes is shown for CG/CHG, CG-only and CHG-only DMRs. B, Hierarchical clustering (Ward’s method) was performed for DMRs that have data in all five genotypes for CG/CHG, CG-only and CHG-only DMRs. Yellow indicates high (100% methylation) while black indicates low methylation levels. C, A linear discriminant analysis was used to classify the number of genotypes that have high or low methylation states for DMRs that had coverage in all five genotypes. The number of DMRs that have the high methylation state in 1, 2, 3 or 4 genotypes is shown for CG/CHG, CG-only and CHG-only DMRs.

Supplemental Figure S9. Analysis of characteristics for DMRs that are located within Identical By State (IBS) regions. A, The size of IBS regions and the number of DMRs in each of the four genotype contrasts. B, Frequency of DMRs in IBS regions. The percent of all tiles (with coverage) that are classified as DMRs was determined for IBS regions (red) relative to the genome-wide DMR frequency (blue). C, The percent of DMRs with different patterns of flanking methylation is assessed for DMRs within (red) or outside (blue) of IBS regions is shown; HH - high DNA methylation on both sides of DMR; HL - high methylation on one side and low methylation on the other side; LL - low methylation on both side. D, The proportion of DMRs located near (within 500bp) of FGS genes, TEs, both genes and TEs or neither genes or TEs is shown for IBS and none-IBS DMRs. The height of the bar shows the average from four contrasts, and the error bars indicate standard deviation.

Supplemental Table S1. Overview of WGBS data for five genotypes.

Supplemental Table S2. Correlation (R) values between WGBS and SeqCap Epi for DMRs discovered using different tile size.

Supplemental Table S3. Effect of increased minimum coverage cut-off on DMRs validation rates.

Supplemental Table S4. Number of context-specific DMRs in each of the four contrasts.

Supplemental Table S5. DMRs present near genes.
Supplemental Table S1. Overview of WGBS data for five genotypes.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>B73</th>
<th>Mo17</th>
<th>Oh43</th>
<th>CML322</th>
<th>Tx303</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reads (initial)</td>
<td>167,762,882</td>
<td>143,155,853</td>
<td>160,563,110</td>
<td>186,299,504</td>
<td>169,478,899</td>
</tr>
<tr>
<td>Reads (trimmed)</td>
<td>165,993,413</td>
<td>141,635,754</td>
<td>154,446,717</td>
<td>178,448,613</td>
<td>162,731,207</td>
</tr>
<tr>
<td>% duplicate reads</td>
<td>37.8</td>
<td>38.0</td>
<td>38.2</td>
<td>43.7</td>
<td>42.7</td>
</tr>
<tr>
<td>Mapping efficiency (%)</td>
<td>46.5</td>
<td>31.0</td>
<td>33.9</td>
<td>31.7</td>
<td>32.8</td>
</tr>
<tr>
<td># unique best hit</td>
<td>77,248,366</td>
<td>43,838,004</td>
<td>52,354,128</td>
<td>56,571,692</td>
<td>53,347,056</td>
</tr>
<tr>
<td># no alignment</td>
<td>12,835,485</td>
<td>29,397,283</td>
<td>28,600,416</td>
<td>33,828,981</td>
<td>31,066,308</td>
</tr>
<tr>
<td># not unique</td>
<td>75,909,562</td>
<td>68,400,467</td>
<td>73,492,173</td>
<td>88,047,940</td>
<td>78,317,843</td>
</tr>
<tr>
<td>Expected coverage</td>
<td>6.18</td>
<td>3.51</td>
<td>4.19</td>
<td>4.53</td>
<td>4.27</td>
</tr>
<tr>
<td>CG%</td>
<td>85.7</td>
<td>82.4</td>
<td>79.7</td>
<td>79.9</td>
<td>79.1</td>
</tr>
<tr>
<td>CHG%</td>
<td>70.3</td>
<td>67.8</td>
<td>66.7</td>
<td>66.3</td>
<td>64.4</td>
</tr>
<tr>
<td>CHH%</td>
<td>2.2</td>
<td>2.6</td>
<td>2.2</td>
<td>2.2</td>
<td>1.9</td>
</tr>
<tr>
<td>SRA #</td>
<td>SRR850328</td>
<td>SRR850332</td>
<td>SRX731433</td>
<td>SRX731432</td>
<td>SRX731434</td>
</tr>
</tbody>
</table>
**Supplemental Table S2.** Correlation (R) values between WGBS and SeqCap Epi for DMRs discovered using different tile size.

<table>
<thead>
<tr>
<th>Tile size</th>
<th>Contexts</th>
<th>B73</th>
<th>CML322</th>
<th>Mo17</th>
<th>Ob43</th>
<th>Tx303</th>
</tr>
</thead>
<tbody>
<tr>
<td>50bp</td>
<td>CG</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>CHG</td>
<td>0.93</td>
<td>0.93</td>
<td>0.92</td>
<td>0.91</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>CHH</td>
<td>0.81</td>
<td>0.77</td>
<td>0.84</td>
<td>0.84</td>
<td>0.78</td>
</tr>
<tr>
<td>100bp</td>
<td>CG</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>CHG</td>
<td>0.94</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>CHH</td>
<td>0.86</td>
<td>0.82</td>
<td>0.86</td>
<td>0.86</td>
<td>0.75</td>
</tr>
<tr>
<td>200bp</td>
<td>CG</td>
<td>0.92</td>
<td>0.89</td>
<td>0.91</td>
<td>0.91</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>CHG</td>
<td>0.91</td>
<td>0.88</td>
<td>0.91</td>
<td>0.89</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>CHH</td>
<td>0.82</td>
<td>0.74</td>
<td>0.82</td>
<td>0.81</td>
<td>0.68</td>
</tr>
</tbody>
</table>
**Supplemental Table S3.** Effect of increased minimum coverage cut-off on DMRs validation rates.

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Contexts</th>
<th>% supported in BvC</th>
<th>% supported in BvM</th>
<th>% supported in BvO</th>
<th>% supported in BvT</th>
</tr>
</thead>
<tbody>
<tr>
<td>2X</td>
<td>CG</td>
<td>94.6</td>
<td>98.4</td>
<td>91.0</td>
<td>92.2</td>
</tr>
<tr>
<td>3X</td>
<td>CG</td>
<td>94.9</td>
<td>98.4</td>
<td>90.7</td>
<td>92.2</td>
</tr>
<tr>
<td>5X</td>
<td>CG</td>
<td>94.8</td>
<td>98.4</td>
<td>94.2</td>
<td>93.4</td>
</tr>
<tr>
<td>2X</td>
<td>CHG</td>
<td>88.6</td>
<td>95.8</td>
<td>91.1</td>
<td>93.6</td>
</tr>
<tr>
<td>3X</td>
<td>CHG</td>
<td>89.5</td>
<td>96.1</td>
<td>90.5</td>
<td>93.0</td>
</tr>
<tr>
<td>5X</td>
<td>CHG</td>
<td>92.2</td>
<td>95.6</td>
<td>93.7</td>
<td>93.7</td>
</tr>
<tr>
<td>2X</td>
<td>CHH</td>
<td>62.0</td>
<td>43.5</td>
<td>39.3</td>
<td>61.2</td>
</tr>
<tr>
<td>3X</td>
<td>CHH</td>
<td>63.9</td>
<td>54.5</td>
<td>48.7</td>
<td>70.0</td>
</tr>
<tr>
<td>5X</td>
<td>CHH</td>
<td>71.4</td>
<td>66.7</td>
<td>63.2</td>
<td>73.3</td>
</tr>
</tbody>
</table>
**Supplemental Table S4.** Number of context-specific DMRs in each of the four contrasts.

<table>
<thead>
<tr>
<th>Contrasts</th>
<th>Total</th>
<th>CG/CHG</th>
<th>CG-only</th>
<th>CHG-only</th>
<th>Unclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>BvC</td>
<td>32893</td>
<td>11557</td>
<td>6296</td>
<td>5242</td>
<td>9798</td>
</tr>
<tr>
<td>BvM</td>
<td>24261</td>
<td>8246</td>
<td>4260</td>
<td>4324</td>
<td>7431</td>
</tr>
<tr>
<td>BvO</td>
<td>28879</td>
<td>10573</td>
<td>5943</td>
<td>4137</td>
<td>8226</td>
</tr>
<tr>
<td>BvT</td>
<td>30334</td>
<td>10194</td>
<td>5853</td>
<td>4962</td>
<td>9325</td>
</tr>
</tbody>
</table>
### Supplemental Table S5. DMRs present near genes.

<table>
<thead>
<tr>
<th>Contrasts</th>
<th>Fold change</th>
<th># gene</th>
<th># gene with 2X coverage at -200 to 600 bp for CG</th>
<th># gene with 2X coverage at -200 to 600 bp for CHG</th>
<th># gene with CG DMR</th>
<th># gene with CHG DMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BvC</td>
<td>on vs off</td>
<td>577</td>
<td>358</td>
<td>92</td>
<td>371</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>&gt;10</td>
<td>1005</td>
<td>789</td>
<td>95</td>
<td>788</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>5 to 10</td>
<td>842</td>
<td>730</td>
<td>30</td>
<td>736</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>2 to 5</td>
<td>2917</td>
<td>2623</td>
<td>58</td>
<td>2633</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>&lt; 2</td>
<td>2571</td>
<td>2369</td>
<td>42</td>
<td>2400</td>
<td>22</td>
</tr>
<tr>
<td>nDEG</td>
<td></td>
<td>14629</td>
<td>13256</td>
<td>185</td>
<td>13432</td>
<td>133</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>22541</td>
<td>20125</td>
<td>502</td>
<td>20360</td>
<td>383</td>
</tr>
<tr>
<td>BvM</td>
<td>on vs off</td>
<td>549</td>
<td>333</td>
<td>50</td>
<td>344</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>&gt;10</td>
<td>1041</td>
<td>798</td>
<td>65</td>
<td>809</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>5 to 10</td>
<td>780</td>
<td>652</td>
<td>18</td>
<td>658</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>2 to 5</td>
<td>3400</td>
<td>2970</td>
<td>89</td>
<td>2991</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>&lt; 2</td>
<td>4242</td>
<td>3841</td>
<td>93</td>
<td>3860</td>
<td>31</td>
</tr>
<tr>
<td>nDEG</td>
<td></td>
<td>12563</td>
<td>11242</td>
<td>331</td>
<td>11383</td>
<td>111</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>22575</td>
<td>19836</td>
<td>646</td>
<td>20045</td>
<td>352</td>
</tr>
<tr>
<td>BvO</td>
<td>on vs off</td>
<td>512</td>
<td>301</td>
<td>58</td>
<td>313</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>&gt;10</td>
<td>804</td>
<td>621</td>
<td>73</td>
<td>622</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>5 to 10</td>
<td>620</td>
<td>504</td>
<td>23</td>
<td>511</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>2 to 5</td>
<td>2322</td>
<td>2040</td>
<td>50</td>
<td>2047</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>&lt; 2</td>
<td>1818</td>
<td>1678</td>
<td>23</td>
<td>1683</td>
<td>12</td>
</tr>
<tr>
<td>nDEG</td>
<td></td>
<td>16453</td>
<td>14799</td>
<td>215</td>
<td>15001</td>
<td>141</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>22529</td>
<td>19943</td>
<td>442</td>
<td>20177</td>
<td>319</td>
</tr>
<tr>
<td>BvT</td>
<td>on vs off</td>
<td>526</td>
<td>305</td>
<td>72</td>
<td>318</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>&gt;10</td>
<td>1107</td>
<td>825</td>
<td>88</td>
<td>841</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>5 to 10</td>
<td>808</td>
<td>663</td>
<td>32</td>
<td>663</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>2 to 5</td>
<td>3046</td>
<td>2669</td>
<td>71</td>
<td>2687</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>&lt; 2</td>
<td>3152</td>
<td>2840</td>
<td>38</td>
<td>2870</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>nDEG</td>
<td>13947</td>
<td>12471</td>
<td>192</td>
<td>12628</td>
<td>110</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td>22586</td>
<td>19773</td>
<td>493</td>
<td>20007</td>
<td>323</td>
</tr>
</tbody>
</table>


