Adaptive evolution of DNA methylation reshaped gene regulation in maize

GEN XU[1,2], JING LYU[1,2], QING LI[3,4], HAN LIU[5], DAFANG WANG[6], MEI ZHANG[5], NATHAN M. SPRINGER[3], JEFFREY ROSS-IBARRA[7], AND JINLIANG YANG[1,2,*]

[1] Department of Agriculture and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583, USA.
[2] Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE 68583, USA.
[3] Department of Plant Biology, Microbial and Plant Genomics Institute, University of Minnesota, Saint Paul, MN 55108, USA.
[4] National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China.
[5] Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences, Nanxincun 20, Fragrant Hill, Beijing 100093, China.
[6] Division of Math and Sciences, Delta State University, Cleveland, MS 38733-0001, USA.
[7] Department of Evolution and Ecology, Center for Population Biology and Genome Center, University of California, Davis, CA 95616, USA.
[*] Corresponding author: jinliang.yang@unl.edu

Compiled March 13, 2020

DNA methylation is a ubiquitous chromatin feature — in maize, more than 25% of cytosines in the genome are methylated. Recently, major progress has been made in describing the molecular mechanisms driving methylation, yet variation and evolution of the methylation landscape during maize domestication remain largely unknown. Here we leveraged whole-genome sequencing (WGS) and whole-genome bisulfite sequencing (WGBS) on populations of modern maize, landrace, and teosinte (*Zea mays* ssp. *parviglumis*) to investigate the adaptive and phenotypic consequences of methylation variations in maize. By using a novel estimation approach, we inferred the methylome site frequency spectrum (mSFS) to estimate forward and backward methylation mutation rates and selection coefficients. We only found weak evidence for direct selections on methylations in any context, but thousands of differentially methylated regions (DMRs) were identified in population-wide that are correlated with recent selections. Further investigation revealed that DMRs are enriched in 5′ untranslated regions, and that maize hypomethylated DMRs likely helped rewire distal gene regulation. For two trait-associated DMRs, *vgt1*-DMR and *tb1*-DMR, our HiChIP data indicated that the interactive loops between DMRs and respective downstream genes were present in B73, a modern maize line, but absent in teosinte. And functional analyses suggested that these DMRs likely served as *cis*-acting elements that modulated gene regulation after domestication. Our results enable a better understanding of the evolutionary forces acting on patterns of DNA methylation and suggest a role of methylation variation in adaptive evolution.

INTRODUCTION

Genomic DNA is tightly packed in the nucleus and is functionally modified by various chromatin marks such as DNA methylation on cytosine. DNA methylation is a heritable covalent modification prevalent in most species, from bacteria to humans [1, 2]. In mammals, DNA methylation commonly occurs in the symmetric CG context with exceptions of non-CG methylation in specific cell types, such as embryonic stem cells [3], but in plants it occurs in all contexts including CG, CHG and CHH (H stands for A, T, or C). Genome-wide levels of cytosine methylation exhibits substantial variations across angiosperms, largely due to differences in the genomic composition of transposable elements [4, 5], but broad patterns of methylation are often conserved within species [6, 7]. Across plant genomes, levels of DNA methylation vary widely from euchromatin to heterochromatin, driven by the different molecular mechanisms for the establishment and maintenance of DNA methylation in CG, CHG, and CHH contexts [8, 9].

DNA methylation is considered essential to suppress the activity of transposons [10], to regulate gene expression [11], and to maintain genome stability [8]. Failure to maintain patterns of DNA methylation in many cases can lead to developmental abnormalities and even lethality [12, 13]. Nonetheless, variation of the DNA methylation has been detected both in natural plant [14] and human populations [15]. Levels of DNA methylation can be affected by genetic variation and environmental cues [16]. Additionally, heritable *de novo* epimutation — the stochastic loss or gain of DNA methylation — can occur spontaneously and has functional consequences [17, 18]. Population methylome studies suggest that the spread of DNA methylation from transposons into flanking regions is one of the major sources of epimutation, such that 20% and 50% of the *cis*-meQTL (methylation Quantitative Trait Loci) are attributable to flanking structural variants in *Arabidopsis* [7] and maize [19].

In *Arabidopsis*, a multi-generational epimutation accumulation experiment [20] estimated forward (gain of DNA methylation) and backward (loss of methylation) epimutation rates per CG site at about 2.56×10^{-4} and 6.30×10^{-4}, respectively. Other than this *Arabidopsis* experiment, there are no systematic estimates of the epimutation rates in higher plants, making it difficult to understand the extent to which spontaneous epimutations contribute to methylome diversity in a natural population. Because the per base rates of DNA methylation variation are several orders of magnitude larger than DNA point mutation, conventional population genetic models which assume infinite sites models seemed inappropriate for epimutation modeling. As an attempt to overcome the obstacle, Charlesworth and Jain [21] developed an analytical framework to address evolution questions for a high order of mutations. Leveraging this theoretical framework, Vidalis et al. [22] constructed the methylome site frequency spectrum (mSFS) using worldwide *Arabidopsis* samples, but they failed to find evidences for selections on genic CG epimutation under benign environments. The confounding effect between DNA variation and methylation variation, as well as the high scaled epimutation rates become obstacles to further dissect the evolutionary forces in shaping the methylation patterns at different timescales under different environments.

Maize, a major cereal crop species, was domesticated from its wild ancestor Teosinte (*Z. mays* ssp. *parviglumis*) near the Balsas River Valley area in Mexico about 9,000 years ago. Genetic studies revealed that the selection of several major effect loci dramatically changed the morphology of teosinte to the modern maize [23]. About 4,000 years ago, maize was introduced to the Southwest of the United States from Mexico; after the initial introduction, however, for about 2,000 years, its territory was limited within the lowland desert. Through sequencing of ancient maize cobs from Turkey Pen Shelter in the temperate Southwest United States, evidence suggested that the ancient maize had finally adapted to the temperate environment in terms of flowering time and tillering phenotypes and eventually enabled this species to spread out to a broader area of the temperate environment [24]. Flowering time, a trait that directly affecting plant fitness, played a major role in this local adaptation process. Numerous QTLs and GWAS studies in maize suggested that flowering time in maize was predominantly controlled by a large number of additive loci [25, 26]. Occasionally, genotype-by-environment interaction was detected for flowering time traits [27]. However, the roles of methylation variation for flowering time and other adaptation-related traits in a natural population, as well as the evolutionary forces in shaping the methylome patterns in different timescales remain largely elusive.

Here, we sequenced genomes from a set of geographically widespread Mexican landraces and a natural population of teosinte collected near Palmar Chico, Mexico [28], from which we generated genomic sequences and methylomes in base-pair resolution. Additionally, we profiled the teosinte (accession no. Ames 21809) interactome using HiChIP method with two antibodies of H3K4me3 and H3K27ac. Together with the analysis from previously published genome [29], transcriptome [30], methylome [6], and interactome [31] datasets, we estimated epimutation rates and the selection pressure in different timescales, investigated the DNA methylation landscapes and variations, detected differentially methylated regions (DMRs), characterized the genomic features that are related with DMRs, and functionally validated two DMRs that are associated with adaptation-traits. Our results suggested that DNA methylation is likely under weak selection during evolution. The inferred phenotypic effects from some DMRs demonstrated that varied methylation patterns may modulate the regulation of some domestication genes and thus affect maize adaptation.

RESULTS

Genomic distribution of methylation in maize and teosinte

To investigate genome-wide methylation patterns in maize and teosinte, we performed whole-genome bisulfite sequencing from a panel of wild teosinte, domesticated maize landraces, and modern maize inbreds (**Table S1**). Using the resequenced genome of each line, we created individual pseudo-references (see **materials and methods**) that alleviated potential bias of mapping reads to a single reference genome [32] and improved overall read-mapping (**Figure S1A**). Using pseudo-references, on average about 25 million (5.6%) more methylated cytosine sites were identified than using the B73 reference (**Figure S1B**). Across populations, average genome-wide cytosine methylation levels were about 78.6%, 66.1% and 2.1% in CG, CHG, and CHH contexts, respectively, which are consistent with previous estimations in maize [12] and are much higher than observed (30.4% CG, 9.9% CHG, and 3.9% CHH) in *Arabidopsis* [5]. We observed slightly higher levels of methylation in landraces, which may due to lower sequencing depth. We found no significant differences between teosinte and maize as a group (**Figure S2**).

We found methylated cytosines in all contexts were significantly higher in the pericentromeric regions (0.54 ± 0.01 in a 10 Mb window) than in chromosome arms (0.44 ± 0.04) (Students' t test P-value $< 2.2e - 16$) (**Figure S3**). At gene level, we calculated the average methylated CGs (mCG) level across gene bodies (from transcription start site to transcription termination site, including exons and introns) in each population and observed a bimodal distribution of mCG in gene bodies (**Figure S4A**), with approximately 25% of genes ($N = 6,874$) showing evidence of gene body methylation (gbm). While the overall distribution of gbm did not differ across populations, genes with clear syntenic orthologs in *Sorghum* exhibited gbm (**Figure S4B-C**), consistent with previous reports [5, 7, 33, 34].

Genome-wide methylation is only under weak selection

As the frequency of methylation may be affected by both selections and epimutation rates, we implemented a novel MCMC approach to estimate these parameters using a population genetic model developed for highly variable loci [21]. We defined 100-bp tiles across the genome as a DNA methylation locus and categorized individual tiles as unmethylated, methylated, or heterozygous alleles

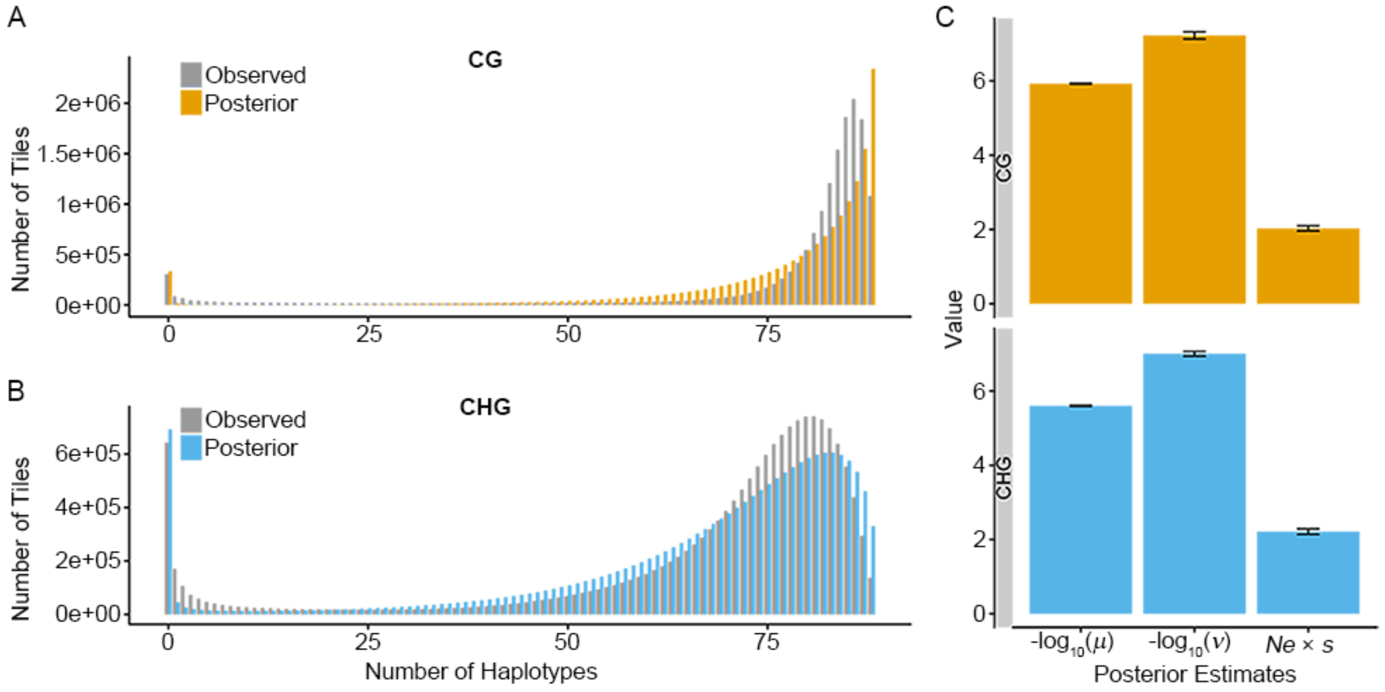

Fig. 1. Methylome site frequency spectra (mSFS) and population genetic parameters inference. (**A-B**) Observed and posterior mSFS at CG (**A**) and CHG (**B**) sites. (**C**) Posterior estimators of mean values and standard errors for μ, ν, and $Ne \times s$ for CG and CHG sites. The size of each tile was 100-bp. Values were estimated using MCMC approach with 25% burnin (see **materials and methods**).

for outcrossed populations (i.e., teosinte and landrace populations) and as unmethylated and methylated alleles for modern maize inbred lines (see **materials and methods**). To determine the thresholds for methylation calls, we employed the iterative expectation maximization algorithm to fit the data [35]. We then constructed methylome site frequency spectra (mSFS) for CG and CHG sites (**Figure 1A-B**). And sensitivity test results suggested that the mSFS were insensitive to the cutoffs used for the methylation calls (**Figure S5**). As the vast majority ($> 98\%$) of CHH sites were unmethylated (**Figure S6**), we excluded CHH sites from population genetic analysis.

Because we found little differences among populations in genome-wide patterns, we estimated parameters using the combined data; estimates from individual populations were nonetheless broadly similar (**Figure S7**). The predicted mSFS from our model was largely similar to the observed data (**Figure 1A-B**), with differences likely attributable to deviations from the simple constant-size population assumed in the model [21]. Model estimates of the epimutation rate μ for both CG (1.2×10^{-6}) and CHG (2.5×10^{-6}) sites were more than an order of magnitude higher than the back-mutation rates ($\nu = 6.0 \times 10^{-8}$ and 1.0×10^{-7}), consistent with the observed prevalence of both types of methylation. Estimates of the genome-wide selection coefficient s associated with methylation of a 100-bp tile were 1.4×10^{-5} and 1.6×10^{-5} for CG and CHG tiles, respectively. Assuming an effective population size of $\approx 150,000$ for maize [36], the population-scaled selection coefficient $Ne \times s$ for CG and CHG tiles were 2.1 and 2.3, indicating relatively weak selection for methylation in each context according to classical population genetic theory [37].

Population level DMRs are enriched in selective sweeps

Average genome-wide methylation data revealed few differences between teosinte and maize, but masks differentiation due to selection at individual loci. To investigate whether individual genomic regions exhibit differential methylation among populations, we employed metilene method [38] to identify population-based differentially methylated regions (DMRs) across the genome. This approach allowed us to precisely define the boundaries of the DMRs by merging tiles recursively. Use this approach, we identified a total of 5,278 DMRs, or about 0.08% (1.8 Mb) of the genome, including 3,900 DMRs between teosinte and modern maize, 1,019 between teosinte and landrace, and 359 DMRs between landrace and modern maize (**Table S2**).

DNA methylation can have a number of functional consequences [14, 39, 40], and thus we tested whether differences in methylation among populations were associated with selection at individual locus. To test this hypothesis, we used SNP data from each population to scan for genomic regions showing evidence of selection (see **materials and methods**). We detected a total of 1,330 selective sweeps between modern maize and teosinte (**Figure 2** and **Table S3**, see **Figure S8** for results of teosinte vs. landrace and landrace vs. modern maize). Several classical domestication genes, e.g., *tb1* [41], *ZAG2* [42], *ZmSWEET4c* [43], *RA1* [44], and *BT2* [45] were among these selective signals.

DMRs at CG and CHG sites were highly enriched in regions showing evidence of recent selection (**Figure S9**, *P*-value < 0.001), particularly in intergenic regions (**Figure S10A**). These DMRs, both hypo- and hypermethylated in maize, exhibited significantly higher allele frequency differentiations between maize and teosinte (**Figure S10B**, *P*-value < 0.001).

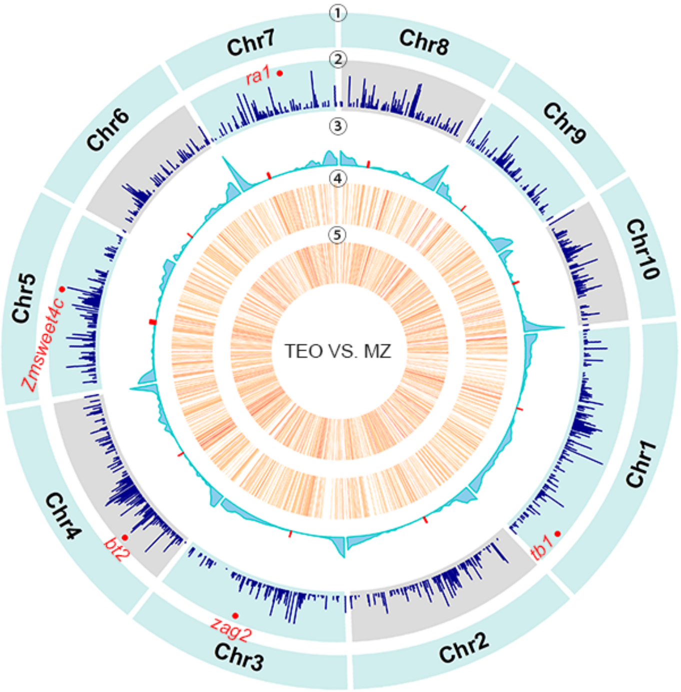

Fig. 2. Selection on differentially methylated regions. Distributions of teosinte-maize selective sweeps, DMRs and other genomic features across ten maize chromosomes. From outer to inner circles were ① Chromosome names, ② selective sweeps detected between modern maize and teosinte, ③ the recombination rate, and the density of DMRs (number per 1-Mb) between modern maize and teosinte in ④ CG and ⑤ CHG contexts. Red dots in circle ③ denote the centromeres.

Hypomethylated regions in maize are involved in distal gene regulation

Further investigation indicated that teosinte-maize CG DMRs were significantly enriched in mappable genic and intergenic (i.e., nongenic excluding 5-kb upstream and downstream of genes and transposons) regions for both hyper- and hypomethylated regions in maize, but depleted from transposon regions (**Figure 3**A). We detected maize hyper- and hypomethylated DMRs in 0.01% and 0.02% of mappable regions across the genome. In particular, 0.07% and 0.05% of maize hyper-DMR (DMR hypermethylated in maize) and hypo-DMR (DMR hypomethylated in maize) were located within mappable exonic regions, which were 14-fold and 5-fold higher than expected by chance (permutation P-values < 0.001, **Figure S11**A). These CG DMRs could be mapped to $N = 229$ unique genes (**Table S4**). After examining the mapping locations based on the collapsed gene model, we found DMRs peaked at 5′ UTR (**Figure 3**B), consistent with a pattern that was previously observed [47]. Using these DMR genes for a gene ontology (GO) analysis, we detected 15 molecular function terms were significantly enriched (**Figure S11**B). Interestingly, 14/15 of these significant terms were associated with "binding" activities, including protein, nucleoside, and ribonucleoside binding. Furthermore, we found these exonic DMRs were enriched at transcription factor binding sites which were identified using DAP-seq [48] (Permutation P-value < 0.001).

Li et al., recently profiled genomic regions colocalized with H3K4me3 and H3K27ac, two well-known chromatin marks for promoters and enhancers [49, 50], to define the interactome in maize [31]. We leveraged these interactome data to study the relationship between DMRs and physical interactions. First, we found the interactive anchor sequences were significantly enriched in DMRs that are hypomethylated in maize, especially in the regulatory regions, including upstream 5-kb, downstream 5-kb, and intergenic regions (**Figure 3**A). We also found DMRs located in transposon elements that were hypomethylated in maize more likely overlap with interactive anchors than expected by chance (Permutation P-value < 0.001).

We hypothesized that the DMRs, especially the DMRs located within the intergenic regions, will alter the up- or downstream gene expression through physical interactions. To test this hypothesis, we mapped the interactive anchors harboring maize hypomethylated DMRs to their 1st, 2nd, and 3rd levels of contacts (**Figure S12**A). Interestingly, genes ($N = 60$) directly contacted (or the 1st level contacts) with these maize hypomethylated intergenic DMRs (**Table S5**) showed significantly (Student's paired t-test, P-value < 0.05) increased expression levels in maize compared to teosinte using published data [30]. The results were insignificant for 2nd and 3rd levels contacts (**Figure S12**A). We found 5/60 genes (Enrichment test P-value < 0.01) were domestication candidate genes as reported previously [51–54]. Two of them were *Zm00001d018036* gene associated with cob length (P-value $= 6 \times 10^{-25}$) and *Zm00001d041948* gene associated with shank length (P-value $= 5.6 \times 10^{-10}$) [51]. Further investigation of these two candidates using recently published chromatin data [46] detected the STARR (Self-Transcribing Active Regulatory Region, a sequencing technology for identifying and quantifying enhancer activity [55]) and H3K27ac peaks at the DMR loci (**Figure 3**C and **Figure S13A**). Consistently with the enhancer signals, the expression levels of these two genes had been significantly increased in maize relative to teosinte (**Figure S12**B and **Figure S13B**).

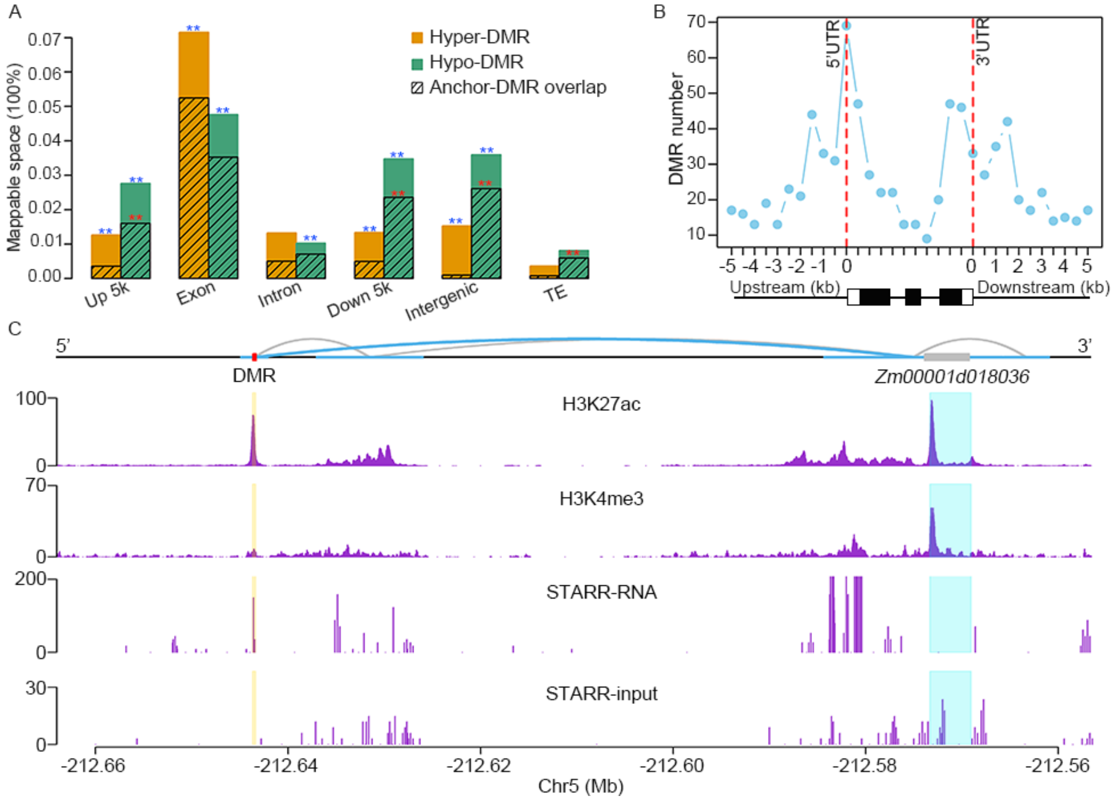

Fig. 3. Teosinte-maize CG DMRs and their associated functional features. (**A**) Breakdown of hyper-DMRs (DMR hypermethylated in maize) and hypo-DMRs (DMR hypomethylated in maize) into genomic features and their overlaps with interactive anchors using data obtained from Li et al., [31]. Blue and red stars indicated DMRs that were significantly enriched at genomic features and interaction anchors (permutation P-value < 0.001). (**B**) The distribution of the number of DMRs along the collapsed gene model. Below the figure shows a schematic gene model with three exons (black boxes). (**C**) Physical interactions (upper panel), colocalization with H3K27ac and H3K4me3 (middle panels), and STARR profiles (lower panels) around *Zm00001d018036* gene in B73. STARR-seq data obtained from [46] showed the transcriptional output (STARR-RNA) and DNA input (STARR-input) around this region. Blue curly lines indicate the interactive contacts between DMR and the candidate gene and grey curly lines indicate other interactive contacts around the region. Horizontal thick blue lines denote the interactive anchors. Red and grey boxes indicate the DMR and gene model, respectively.

DMRs explained disproportionally larger phenotypic variance and altered differential flowering time genes regulation

Next, we employed a variance component analysis approach by using SNP sets residing in DMRs (DMR-SNPs) to evaluate the relative importance of DMRs (see **materials and methods**). By using NAM population with 41 publicly available phenotypes [56], we estimated the variance explained by DMR-SNPs and explained by SNPs mapped to the rest of the genome. Results suggested that teosinte-maize CG DMR-SNPs, although only accounting for 0.01% of the genome, could explain more than 1% of the phenotypic variances for 18 traits, including ear row number (12.1%), northern leaf blight (5.5%) and, stalk strength (8.9%) (**Figure 4A**). For landrace-maize CG DMR-SNPs, we detected disproportionally larger phenotypic variances be explained for several yield and disease-related traits, including tassel branch number, ear row number, Southern and Northern leaf blight (**Figure 4B**).

Interestingly, four flowering time related traits, i.e., photoperiod growing-degree days to silk, photoperiod growing-degree days to anthesis, days to anthesis, and days to silk, showed consistently strong signals that can be explained by over 3.7% heritabilities (**Figure 4A**). We hypothesized that the disproportionally larger heritability explained by CG DMR-SNPs might be caused by some large effect flowering time related genes. To test this hypothesis, we examined several known genes in the flowering time pathway [57]. Indeed, we detected six DMRs located on four flowering time related genes (**Figure S14**) (Enrichment test P-value $= 0.02$).

Additionally, one DMR located 40-kb upstream of the *ZmRAP2.7* gene and 20-kb downstream of the *vgt1* locus that was hypomethylated in modern maize and landrace but was hypermethylated in teosinte (**Figure 4C**). A MITE transposon insertion in the *vgt1* locus was considered as the causal variation for the down regulation of the *ZmRAP2.7*, which was a transcription factor in the flowering time pathway [58]. Reanalysis of the published ChIP data revealed that the DMR colocalized with H3K27ac, the chromatin activation mark, and there existed a physical interaction between the DMR and *vgt1* locus in maize [31] (**Figure 4D**). To examine the interaction status in teosinte, we then generated HiChIP data for a teosinte sample using the same tissue and antibodies (see **materials and methods**). As expected, we failed to detect the physical interaction between *vgt1*-DMR and *vgt1* itself in teosinte (**Figure 4D**), suggesting that methylation might play a functional role in affecting physical interaction at *vgt1*-DMR locus.

To further validate the potential enhancer function of the 209-bp *vgt1*-DMR, we incorporated the *vgt1*-DMR sequence amplified from B73 into a vector constructed as shown in (**Figure 4E**) and performed the dual-luciferase transient expression assay in maize protoplasts (see **materials and methods**). The results of the transient expression assay revealed that the maize cells harboring the

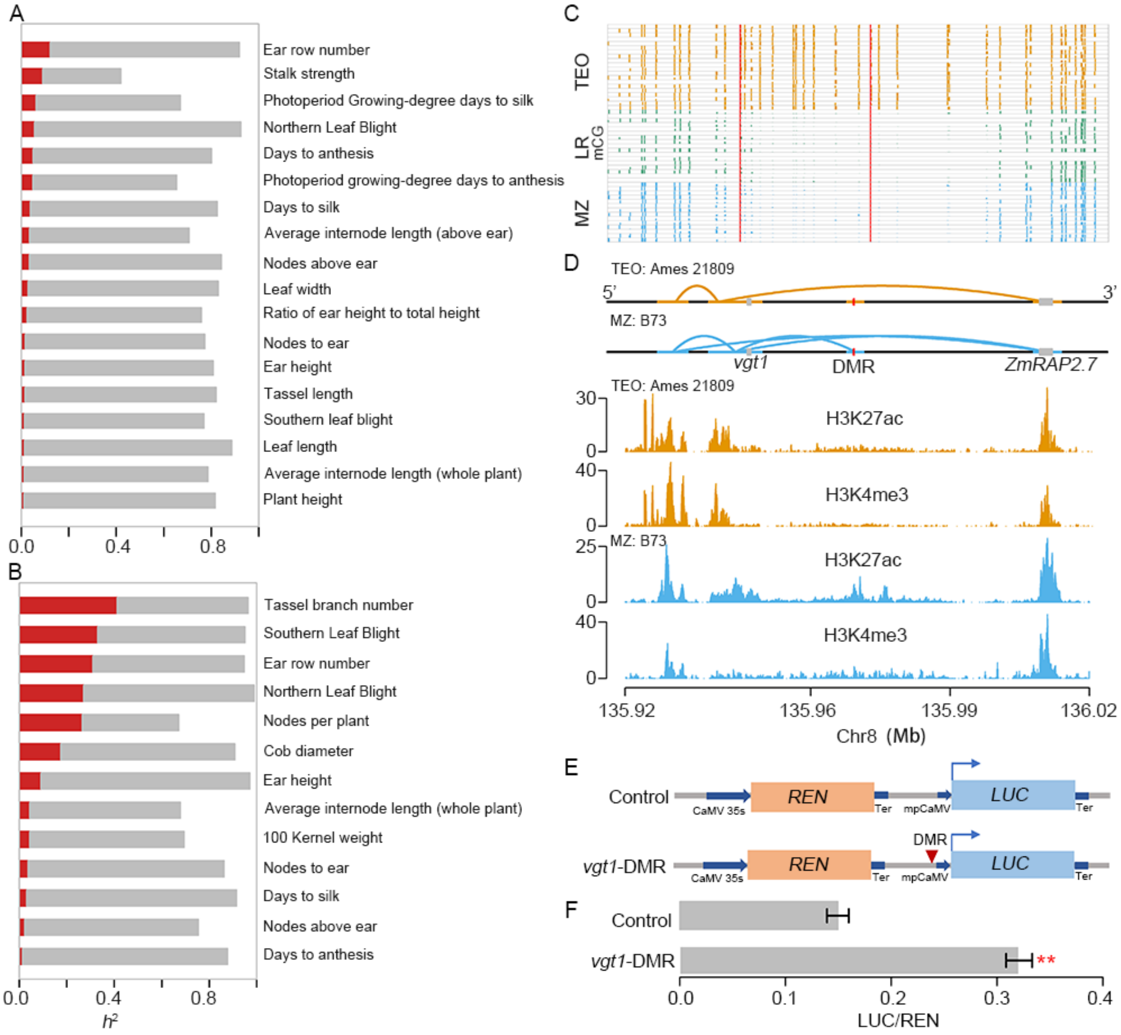

Fig. 4. Phenotypic effects of DMR-SNPs and the functional validation results. (**A-B**) Phenotypic variance explained (h^2) by SNPs residing in DMRs (red) and non-DMRs (grey) using teosinte-maize (**A**) and landrace-maize (**B**) CG DMRs. (**C**) Levels of CG methylation around *vgt1*-DMR in maize (MZ), landrace (LR), and teosinte (TEO) populations. (**D**) The interactive contacts (upper panel) and colocalization with H3K27ac and H3K4me3 (lower panel) around *vgt1*-DMR in a maize (B73) and a teosinte (Ames 21809) samples. (**E**) The vectors constructed for functional validation of the *vgt1*-DMR using the dual-luciferase transient expression assay in maize protoplasts. (**F**) The expression ratios of LUC/REN using five biological replicates. Two asterisks indicate a significant difference (Student's t test P-value < 0.001).

DMR exhibited a significantly higher LUC and REN ratio than control (fold change= 2.2, P-value= $2.4e^{-8}$, **Figure 4F**), revealing that the DMR might act as an enhancer to activate LUC expression.

A segregating *tb1*-DMR performed like a *cis*-acting element

One of the most significant teosinte-maize CG DMRs was located 30-kb upstream of the *tb1* gene, which is a transcription factor acting as a repressor of axillary branching (aka tillering) phenotype [41]. This 534-bp *tb1*-DMR was hypomethylated in modern maize, hypermethylated in teosinte, and segregating in landraces (**Figure 5A**). Phenotypic analysis indicated that the DMR was associated with the tillering phenotype using the phenotypic data observed for the 17 landraces (Fisher's exact test P-value < 0.05). And the phenotypic effect was consistent with previous observations that the hypermethylated (teosinte-like) genotypes were likely to grow tillers.

The causal variation for this locus was previously mapped to a Hopscotch TE insertion 60-kb upstream (**Figure 5B**) of the *tb1* gene. The TE was considered as an enhancer, as shown by a transient *in vivo* assay [41]. The interactome data support this claim that there was a physical contact between Hopscotch and *tb1* gene (**Figure 5B**) as have also been shown [31]. Interestingly, we also detected a direct physical contact between the *tb1*-DMR and *tb1* gene itself in maize line B73 but missing in teosinte using our newly generated HiChIP data (**Figure 5B**). This observation was consistent with our previous result for the *vgt1*-DMR locus and suggested the DMR might result in differential interactive loops. The colocalization of *tb1*-DMR with chromatin activation marks in the region also suggested the *tb1*-DMR might perform like a *cis*-acting regulatory element (**Figure 5B**).

To understand the correlation among these genomic components, i.e., the *tb1*-DMR and *tb1* gene, we conducted linkage disequilib-

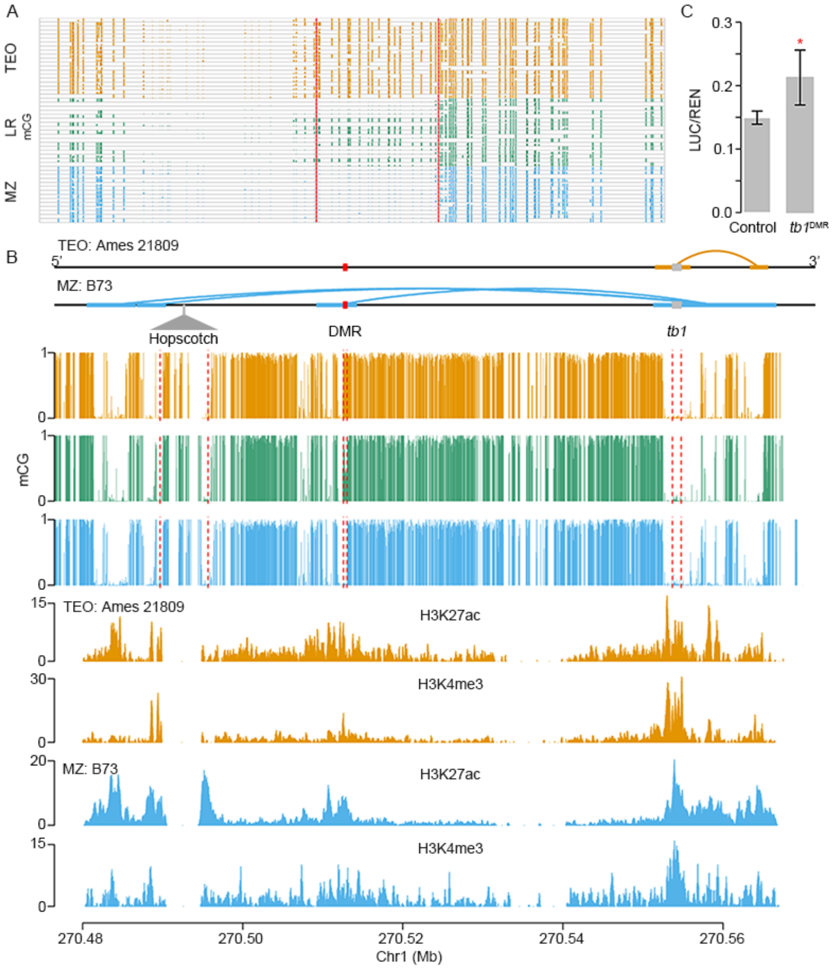

Fig. 5. A hypomethylated DMR that is upstream of *tb1* gene. (**A**) Levels of mCG for the 534-bp *tb1*-DMR in each individual methylome of the modern maize (MZ), landrace (LR), and teosinte (TEO) populations. Vertical red dashed lines indicate the boundaries of the *tb1*-DMR. (**B**) Interactive contacts (upper panel), average CG methylation levels (middle panel), and colocalization of the *tb1*-DMR with H3k27ac and H3K4me3 (lower panel). Horizontal thick lines denote the interactive anchors and solid curly lines on top of the annotations denote the interactive contacts in teosinte and maize. (**C**) Functional validation result of *tb1*-DMR using dual-luciferase transient expression assay in maize protoplasts. Asterisk indicates a significant difference (Student's t test *P*-value < 0.05).

rium (LD) analysis using landrace genomic and methylation data, which were segregating at this *tb1*-DMR locus (see **materials and methods**). As a result, we failed to detect strong LD ($R^2 = 0.1$) in this region (**Figure S15**), indicating the *tb1*-DMR might be originated independently. Further, we found the highly methylated landraces were geographically closer to the Balsas River Valley in Mexico, where maize was originally domesticated from (**Figure S16A**). As the landraces spread out from the domestication center, their CG methylation levels were gradually reduced (**Figure S16B**).

Additionally, we conducted a dual-luciferase transient assay by constructing a vector similar to (**Figure 4E**). The results indicated that the *tb1*-DMR significantly increased the LUC/REN ratio as compared to control (**Figure 5C**), suggesting that the *tb1*-DMR was potentially act as a *cis*-acting element to enhance downstream gene expression.

DISCUSSION

In this study, we employed population genetics and statistical genomics approaches to infer the rates of epimutation, selection pressure on DNA methylation, and the extent to which SNPs located within DMRs contributed to phenotypic variations. Our results revealed that the forward epimutation rate ($\sim 10^{-6}$) was about 10 times larger than the backward epimutation rate ($\sim 10^{-7}$), which is several magnitudes larger than that of the DNA mutation rate ($\sim 10^{-8}$) in maize [59]. Our estimated epimutation rates were different from the rates estimated in *Arabidopsis* using the epimutation accumulation experiments [60]. Partially because in this study, we used 100-bp tile as an inheritance unit, while the *Arabidopsis* calculated the per base epimutation rates. Additionally, the genome-wide methylation levels were dramatically different between maize and *Arabidopsis*, which may result in different epimutation rates for these two species.

Although population methylome modeling suggested that genome-wide DNA methylation was not under strong selection, we detected a large number of DMRs by conducting population-wide comparisons. These DMRs are likely to explain adaptation and domestication related phenotypes, demonstrated by both the global quantitative genetics analyses using DMR-SNPs and the functional validations of two well characterized loci *vgt1* and *tb1*. In both functional validation cases, evidences showed that methylation levels tend to affect the physical interactions. In particular, the domesticated alleles exhibiting low methylation levels in modern maize associated with newly formed interactive loops in B73 as compared to the wild ancestor teosinte. Transient expression assays demonstrated that two of these non-methylated alleles can increase the expression levels of the reporter genes. Accordingly, we identified a set of new genes (about 60) that differentially expressed between maize and teosinte, in which their exonic regions directly connect with hypo-DMRs. Taken together, these results suggested that methylations might modulate physical interactions and hence likely affect gene expression. This speculation of methylation variation affected distal regulation fitted well with the previous results from GWAS that 80% of explained variation could be attributable to trait-associated variants located in regulatory regions [61]. Our variance component analysis results further suggested that DMR-SNPs (largely mapped to the genic and intergenic regions) explained disproportionally larger phenotypic variations. Interestingly, teosinte-maize DMR-SNPs explained more phenotypic variances for domestication related phenotypes, while landrace-maize DMR-SNPs explained more for improvement related phenotypes.

Unlike the low rate of DNA mutation, the high rate of the DNA methylation might provide an alternative mechanism for plant adaptation. We showed that the hypomethylated regions tend to be involved in the distal regulations to activate or inactivate genes expressions. They likely to have greater effects on phenotypic traits and, therefore, could be serve as potential targets of recent selection. However, in this study, we could not rule out the possibility that DMRs might be the hitchhiking effects of the positive selection on genomic variations.

Collectively, the fact that a large number of DMRs overlapped with the selective sweeps during domestication and improvement processes, and the evidences that DMRs may function as *cis*-acting elements provide new insights into plant adaptive evolution. These naturally occurring adaptive DMRs could possibly be leveraged for understanding gene regulation. Since some of the DMRs exhibited favorable phenotypic consequences, they might also be the potential targets of artificial selection for further crop improvement.

MATERIALS AND METHODS

Plant materials and DNA sequencing

We obtained a set of geographically widespread open pollinated landraces across Mexico ($N = 17$) from Germplasm Resources Information Network (GRIN) (**Table S1**). The teosinte (*Zea mays* ssp. *parviglumis*; $N = 20$) were collected near Palmar Chico, Mexico [28]. We harvested the third leaf of the teosintes and Mexican landraces for DNA extraction using a modified CTAB procedure [62]. The extracted DNA was then sent out for whole genome sequencing (WGS) and whole genome bisulfite sequencing (WGBS) using Illumina HiSeq platform. Additionally, we obtained WGBS data for 14 modern maize inbred lines [6] and WGS data for the same 14 lines from the maize HapMap3 project [29].

Sequencing data analysis

The average coverage for the WGS of the 20 teosintes and 17 landraces lines was about 20 ×. For these WGS data, we first mapped the cleaned reads to the B73 reference genome (AGPv4) [63] using BWA-mem [64] with default parameters, and kept only uniquely mapped reads. Then we removed the duplicated reads using Picard tools [65]. We conducted SNP calling using Genome Analysis Toolkit's (GATK, version 4) HaplotypeCaller [66], in which the following parameters were applied: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0.

In order to improve the WGBS mapping rate and decrease the mapping bias, we replaced the B73 reference genome with filtered SNP variants using an in-house developed software — pseudoRef (https://github.com/yangjl/pseudoRef). Subsequently, we mapped reads to each corrected pseudo-reference genome using Bowtie2 [67] and kept only unique mapped reads. After filtering the duplicate reads, we extracted methylated cytosines using the Bismark methylation extractor and only retained sites with more than three mapped reads. The weighted methylation level was determined following the previously reported method [68].

Population epigenetics modeling

Spontaneous epimutation changes (i.e. gain or loss of cytosine methylation) exhibit higher rate than genomic mutation [20, 60]. The standard population genetic methods designed for SNPs are thus inappropriate for population epigenetic studies. Here, we applied the analytical framework for hypermutable polymorphisms developed by Charlesworth and Jain [21]. Under this framework, the probability density of the methylated alleles was modeled as:

$$\phi(q) = Ce^{\gamma q}(1-q)^{\alpha-1}q^{\beta-1}$$

where $\alpha = 4N_e\mu$, $\beta = 4N_e\nu$, $\gamma = 2N_es$. N_e, effective population size; q, frequency of the hypermethylation alleles; μ, forward epimutation rate (methylation gain); ν, backward epimutation rate (methylation loss); s, selection coefficient. The constant C is required so that $\int_0^1 \phi(q)dq = 1$.

We defined 100-bp tiles as a DNA methylation locus. To define the methylation status, we assumed that the methylation levels in a heterozygote individual falling into three mixture distributions (unmethylated, methylated, and heterozygote distributions). We employed an R add-on package "mixtools" and fitted the "normalmixEM" procedure to estimate model parameters [35]. Based on the converged results of the iterative expectation maximization algorithm (using the "normalmixEM" function), we decided to use 0.7 and 0.3 thresholds for heterozygote individuals (i.e., average methylation value> 0.7 for a 100-bp tile was determined as a methylated call and coded as 2; < 0.3 was determined as an unmethylated call and coded as 0; otherwise coded as 1). We also tested different cutoffs and found that the final methylation site frequency spectrum (mSFS) was insensitive to the cutoffs used. Similarly, we assumed two mixture distributions for inbred lines and used cutoff $= 0.5$ to determine methylated (coded as 1) and unmethylated (coded as 0) calls. With these cutoffs, we then constructed mSFS on genome-wide methylation loci. We also constructed interspecific (i.e., across maize, landrace, and teosinte populations) and intraspecific (i.e., within maize, landrace, and teosinte populations) mSFS.

To estimate three critical population epigenetic parameters (μ, ν, and s) from observed mSFS, we implemented a Markov Chain Monte Carlo (MCMC) method (https://rpubs.com/rossibarra/mcmcbc). In the analyses, we assumed $N_e = 150,000$ [69, 70] and used normal prior distributions for μ, ν, and s with mean 10^{-8}, 10^{-8}, 10^{-5} and standard error 0.05, 0.05, and 0.05, respectively (**Figure S17**). We ran the model using a chain length of $N = 100,000$ iterations with the first 25% as burnin.

Genome scanning to detect selective signals

We called SNPs using our WGS data and performed genome scanning for selective signals using XP-CLR method [71]. In the XP-CLR analysis, we used a 50-kb sliding window and a 5-kb step size. To ensure comparability of the composite likelihood score in each window, we fixed the number assayed in each window to 200 SNPs. We evaluated evidence for selections across the genome in three contrasts: teosinte vs landrace, landrace vs modern maize, and teosinte vs modern maize. We merged nearby windows falling into the 10% tails into the same window. After window merging, we considered the 0.5% outliers as the targets of selection.

We calculated F_{ST} using WGS data using VCFtools [72]. In the analysis, we used a 50-kb sliding window and a 5-kb step size.

DMR detection and GO term analysis

We used a software package 'metilene' for DMR detection between two populations [38]. To call a DMR, we required it contained at least eight cytosine sites with < 300-bp in distance between two adjacent cytosine sites, and the average of methylation differences between two populations should be > 0.4 for CG and CHG sites. Finally, we required a corrected P-value < 0.01 as the cutoff.

We conducted gene ontology (GO) term analysis on selected gene lists using AgriGO2.0 with default parameters [73]. We used the significance cutoff at P-value < 0.01.

HiChIP sequencing library construction

We constructed the teosinte HiChIP library according to the protocol developed by Mumbach et al. [74] with some modifications. The samples we used were two weeks aerial tissues collected from a teosinte accession (Ames 21809) that were planted in the growth chamber under the long-day condition (15h day time and 9h night time) at the temperature (25°C at day time and 20°C at night time). After tissue collection, we immediately cross-linked it in a 1.5 mM EGS solution (Thermo, 21565) for 20 min in a vacuum, followed by 10 min vacuum infiltration using 1% formaldehyde (Merck, F8775-500ML). To quench the EGS and formaldehyde, we added a final concentration of 0.15 mM glycine (Merck, V900144) and infiltrated by vacuum for 5 min. Then, cross-linked samples were washed five times in double-distilled water and flash-frozen in liquid nitrogen.

To isolate the nuclear from cross-linked tissues, we used the methods as described previously [31]. After obtaining the purified nuclear, we resuspended it in 0.5% SDS and used 10% Triton X-100 to quench it, and then performed digestion, incorporation, and proximity ligation reactions as previously described [74]. We used two antibodies H3K4me3 (Abcam, ab8580) and H3K27ac (Abcam, ab4729) to pull down the DNA. And then, we purified DNA with the MinElute PCR Purification Kit (QIAGEN) and measured the DNA concentration using Qubit. To fragment and capture interactive loops, we used the Tn5 transposase kit (Vazyme, TD502) to construct the library with 5 ng DNA. We then sent the qualified DNA libraries for sequencing using the Illumina platform.

ChIP-seq and HiChIP data analysis

We obtained ChIP-seq data from the B73 shoot tissue [31] and then aligned the raw reads to B73 reference genome (AGPv4) using Bowtie2 [75]. After alignment, we removed the duplicated reads and kept only the uniquely mapped reads. By using the uniquely mapped reads, we calculated read coverages using deepTools [76].

For the teosinte HiChIP sequencing data, we first aligned the raw reads to the B73 reference genome (AGPv4) using HiC-Pro [77], and then processed the valid read pairs to call interactive loops using hichipper pipeline [78] with a 5-kb bin size. After the analysis, we filtered out the non-valid loops with genomic distance less than 5 kb or larger than 2 Mb. By using the mango pipeline [79], we determined the remaining loops with three read pairs supports and the FDR < 0.01 as the significant interactive loops.

Kinship matrices and variance components analysis

We estimated the variance components explained by SNP sets residing in DMRs using the maize Nested Association Mapping (NAM) population [25, 80]. We downloaded the phenotypic data (/iplant/home/glaubitz/RareAlleles/genomeAnnos/VCAP/phenotypes/NAM/familyCorrected), consisting of Best Linear Unbiased Predictors (BLUPs) for 41 different traits ([56]), and imputed genotypic data (/iplant/home/glaubitz/RareAlleles/genomeAnnos/VCAP/genotypes/NAM/namrils_projected_hmp31_MAF02mnCnt2500.hmp.txt.gz) [29] from CyVerse database as described in Panzea (www.panzea.org).

In the analysis, we mapped SNPs to the DMR and non-DMR regions. We partitioned SNPs into two sets — SNPs located within DMR and SNPs that were outside of DMRs (i.e., the rest of the genome). For each SNP set, we calculated an additive kinship matrix using the variance component annotation pipeline implemented in TASSEL5 [81]. We then fed these kinship matrices along with the NAM phenotypic data to estimate the variance components explained by SNP sets using a Residual Maximum Likelihood (REML) method implemented in LDAK [82].

Dual-luciferase transient expression assay in maize protoplasts

To investigate the effect of DMRs on gene expression, we performed a dual-luciferase transient expression assay in maize protoplasts. We used the pGreen II 0800-LUC vector [83] for the transient expression assay with minor modification, where a minimal promoter from cauliflower mosaic virus (mpCaMV) was inserted into the upstream of luciferase (LUC) to drive LUC gene transcription. In the construct, we employed the *Renillia luciferase* (*REN*) gene under the control of 35S promoter from cauliflower mosaic virus (CaMV) as an internal control to evaluate the efficiency of maize protoplasts transformation. We amplified the selected DMR sequences from B73 and then inserted them into the control vector at the restriction sites *Kpn*I/*Xho*I upstream of the mpCaMV, generating the reporter constructs.

We isolated protoplasts from the 14-day-old leaves of B73 albino seedlings following the protocol [84]. Subsequently, we transformed 15 ug plasmids into the 100 ul isolated protoplasts using polyethylene glycol (PEG) mediated transformation method [84]. After 16 hours infiltration, we measured the LUC and REN activities using dual-luciferase reporter assay reagents (Promega, USA) and a GloMax 20/20 luminometer (Promega, USA). Finally, we calculated the ratios of LUC to REN. For each experiment, we included five biological replications.

Data availability

All datasets and analyzing scripts are available through GitHub (https://github.com/jyanglab/msfs_teo) and methylome data has been submitted to the NCBI SRA.

ACKNOWLEDGEMENTS

This project was partly supported by the Agriculture and Food Research Initiative Grant number 2019-67013-29167 from the USDA National Institute of Food and Agriculture, the National Science Foundation under award number OIA-1557417 for Center for Root and Rhizobiome Innovation (CRRI), Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under Grant number P20GM103476, and the University of Nebraska-Lincoln Start-up fund and the Layman seed award. This work was conducted using the Holland Computing Center of the University of Nebraska-Lincoln Start-up, which receives supports from the Nebraska Research Initiative. We would like to thank the helpful discussion in J.R.-I.'s REHAB.

AUTHOR CONTRIBUTIONS

J.Y. and J.R.-I. designed this work. J.L., Q.L., H.L., D.W., M.Z., generated the data. G.X., J.R.-I., and J.Y. analyzed the data. N.M.S. provided conceptual advice. J.Y., G.X. and J.R.-I. wrote the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

REFERENCES

1. M. A. Sánchez-Romero, I. Cota, and J. Casadesús, “Dna methylation in bacteria: from the methyl group to the methylome,” Current opinion in microbiology **25**, 9–16 (2015).
2. K. D. Robertson, “Dna methylation and human disease,” Nature Reviews Genetics **6**, 597 (2005).
3. J. Arand, D. Spieler, T. Karius, M. R. Branco, D. Meilinger, A. Meissner, T. Jenuwein, G. Xu, H. Leonhardt, V. Wolf *et al.*, “In vivo control of cpg and non-cpg dna methylation by dna methyltransferases,” PLoS Genet **8**, e1002750 (2012).
4. C. Alonso, R. Pérez, P. Bazaga, and C. M. Herrera, “Global dna cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms,” Frontiers in Genetics **6**, 4 (2015).
5. C. E. Niederhuth, A. J. Bewick, L. Ji, M. S. Alabady, K. Do Kim, Q. Li, N. A. Rohr, A. Rambani, J. M. Burke, J. A. Udall *et al.*, “Widespread natural variation of dna methylation within angiosperms,” Genome biology **17**, 194 (2016).
6. Q. Li, J. Song, P. T. West, G. Zynda, S. R. Eichten, M. W. Vaughn, and N. M. Springer, “Examining the causes and consequences of context-specific differential dna methylation in maize,” Plant physiology **168**, 1262–1274 (2015).
7. R. J. Schmitz, M. D. Schultz, M. A. Urich, J. R. Nery, M. Pelizzola, O. Libiger, A. Alix, R. B. McCosh, H. Chen, N. J. Schork *et al.*, “Patterns of population epigenomic diversity,” Nature **495**, 193 (2013).
8. H. Zhang, Z. Lang, and J.-K. Zhu, “Dynamics and function of dna methylation in plants,” Nature reviews Molecular cell biology **19**, 489 (2018).
9. N. M. Springer and R. J. Schmitz, “Exploiting induced and natural epigenetic variation for crop improvement,” Nature reviews genetics **18**, 563 (2017).
10. O. Deniz, J. M. Frost, and M. R. Branco, “Regulation of transposable elements by dna modifications,” Nature Reviews Genetics p. 1 (2019).
11. D. K. Seymour and C. Becker, “The causes and consequences of dna methylome variation in plants,” Current opinion in plant biology **36**, 56–63 (2017).
12. Q. Li, S. R. Eichten, P. J. Hermanson, V. M. Zaunbrecher, J. Song, J. Wendt, H. Rosenbaum, T. F. Madzima, A. E. Sloan, J. Huang *et al.*, “Genetic perturbation of the maize methylome,” The Plant Cell **26**, 4602–4616 (2014).
13. F.-F. Fu, R. K. Dawe, and J. I. Gent, “Loss of rna-directed dna methylation in maize chromomethylase and ddm1-type nucleosome remodeler mutants,” The Plant Cell **30**, 1617–1627 (2018).
14. Y. Shen, J. Zhang, Y. Liu, S. Liu, Z. Liu, Z. Duan, Z. Wang, B. Zhu, Y.-L. Guo, and Z. Tian, “Dna methylation footprints during soybean domestication and improvement,” Genome biology **19**, 1–14 (2018).
15. I. Hernando-Herraez, R. Garcia-Perez, A. J. Sharp, and T. Marques-Bonet, “Dna methylation: insights into human evolution,” PLoS genetics **11** (2015).
16. F. Kader and M. Ghai, “Dna methylation-based variation between human populations,” Molecular genetics and genomics **292**, 5–35 (2017).
17. K. Manning, M. Tör, M. Poole, Y. Hong, A. J. Thompson, G. J. King, J. J. Giovannoni, and G. B. Seymour, “A naturally occurring epigenetic mutation in a gene encoding an sbp-box transcription factor inhibits tomato fruit ripening,” Nature genetics **38**, 948 (2006).
18. S. Cortijo, R. Wardenaar, M. Colomé-Tatché, A. Gilly, M. Etcheverry, K. Labadie, E. Caillieux, J.-M. Aury, P. Wincker, F. Roudier *et al.*, “Mapping the epigenetic basis of complex traits,” Science **343**, 1145–1148 (2014).
19. S. R. Eichten, R. Briskine, J. Song, Q. Li, R. Swanson-Wagner, P. J. Hermanson, A. J. Waters, E. Starr, P. T. West, P. Tiffin *et al.*, “Epigenetic and genetic influences on dna methylation variation in maize populations,” The Plant Cell **25**, 2783–2797 (2013).
20. A. Van Der Graaf, R. Wardenaar, D. A. Neumann, A. Taudt, R. G. Shaw, R. C. Jansen, R. J. Schmitz, M. Colomé-Tatché, and F. Johannes, “Rate, spectrum, and evolutionary dynamics of spontaneous epimutations,” Proceedings of the National Academy of Sciences **112**, 6676–6681 (2015).
21. B. Charlesworth and K. Jain, “Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates,” Genetics **198**, 1587–1602 (2014).
22. A. Vidalis, D. Živković, R. Wardenaar, D. Roquis, A. Tellier, and F. Johannes, “Methylome evolution in plants,” Genome biology **17**, 264 (2016).
23. M. C. Stitzer and J. Ross-Ibarra, “Maize domestication and gene interaction,” New phytologist **220**, 395–408 (2018).
24. K. Swarts, R. M. Gutaker, B. Benz, M. Blake, R. Bukowski, J. Holland, M. Kruse-Peeples, N. Lepak, L. Prim, M. C. Romay *et al.*, “Genomic estimation of complex traits reveals ancient maize adaptation to temperate north america,” Science **357**, 512–515 (2017).
25. E. S. Buckler, J. B. Holland, P. J. Bradbury, C. B. Acharya, P. J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J. C. Glaubitz *et al.*, “The genetic architecture of maize flowering time,” Science **325**, 714–718 (2009).
26. J. A. R. Navarro, M. Willcox, J. Burgueño, C. Romay, K. Swarts, S. Trachsel, E. Preciado, A. Terron, H. V. Delgado, V. Vidal *et al.*, “A study of allelic diversity underlying flowering-time adaptation in maize landraces,” Nature genetics **49**, 476 (2017).
27. X. Li, T. Guo, Q. Mu, X. Li, and J. Yu, “Genomic and environmental determinants and their interplay underlying phenotypic plasticity,” Proceedings of the National Academy of Sciences **115**, 6679–6684 (2018).
28. C. J. Yang, L. F. Samayoa, P. J. Bradbury, B. A. Olukolu, W. Xue, A. M. York, M. R. Tuholski, W. Wang, L. L. Daskalska, M. A. Neumeyer *et al.*, “The genetic architecture of teosinte catalyzed and constrained maize domestication,” Proceedings of the National Academy of Sciences **116**, 5643–5652 (2019).
29. R. Bukowski, X. Guo, Y. Lu, C. Zou, B. He, Z. Rong, B. Wang, D. Xu, B. Yang, C. Xie *et al.*, “Construction of the third-generation zea mays haplotype map,” Gigascience **7**, gix134 (2017).
30. Z. H. Lemmon, R. Bukowski, Q. Sun, and J. F. Doebley, “The role of cis regulatory evolution in maize domestication,” PLoS genetics **10**, e1004745 (2014).
31. E. Li, H. Liu, L. Huang, X. Zhang, X. Dong, W. Song, H. Zhao, and J. Lai, “Long-range interactions between proximal and distal regulatory regions in maize,” Nature communications **10**, 2633 (2019).
32. P. Wulfridge, B. Langmead, A. P. Feinberg, and K. Hansen, “Choice of reference genome can introduce massive bias in bisulfite sequencing data,” bioRxiv **xxxx**, 076844 (2016).
33. P. T. West, Q. Li, L. Ji, S. R. Eichten, J. Song, M. W. Vaughn, R. J. Schmitz, and N. M. Springer, “Genomic distribution of h3k9me2 and dna methylation in a maize genome,” PLoS One **9**, e105267 (2014).
34. Y. Zhang, D. W. Ngu, D. Carvalho, Z. Liang, Y. Qiu, R. L. Roston, and J. C. Schnable, “Differentially regulated orthologs in sorghum and the subgenomes of maize,” The Plant Cell **29**, 1938–1951 (2017).
35. T. Benaglia, D. Chauveau, D. Hunter, and D. Young, “mixtools: An r package for analyzing finite mixture models,” (2009).
36. J. Ross-Ibarra, M. Tenaillon, and B. S. Gaut, “Historical divergence and gene flow in the genus zea,” Genetics **181**, 1399–1413 (2009).
37. M. W. Hahn, *Molecular population genetics* (Sinauer Associates New York, 2019).
38. F. Jühling, H. Kretzmer, S. H. Bernhart, C. Otto, P. F. Stadler, and S. Hoffmann, “metilene: Fast and sensitive calling of differentially methylated regions from bisulfite sequencing data,” Genome research **26**, 256–262 (2016).
39. L.-J. Gardiner, M. Quinton-Tulloch, L. Olohan, J. Price, N. Hall, and A. Hall, “A genome-wide survey of dna methylation in hexaploid wheat,” Genome biology **16**, 273 (2015).
40. Q. Song, T. Zhang, D. M. Stelly, and Z. J. Chen, “Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons,” Genome biology **18**, 99 (2017).
41. A. Studer, Q. Zhao, J. Ross-Ibarra, and J. Doebley, “Identification of a functional transposon insertion in the maize domestication gene tb1,” Nature

genetics **43**, 1160 (2011).

42. D. Zhao, Z. Huang, N. Umino, A. Hasegawa, and H. Kanamori, “Structural heterogeneity in the megathrust zone and mechanism of the 2011 tohoku-oki earthquake (mw 9.0),” Geophysical Research Letters **38** (2011).
43. D. Sosso, D. Luo, Q.-B. Li, J. Sasse, J. Yang, G. Gendrot, M. Suzuki, K. E. Koch, D. R. McCarty, P. S. Chourey *et al.*, “Seed filling in domesticated maize and rice depends on sweet-mediated hexose transport,” Nature genetics **47**, 1489 (2015).
44. B. Sigmon and E. Vollbrecht, “Evidence of selection at the ramosa1 locus during maize domestication,” Molecular Ecology **19**, 1296–1311 (2010).
45. S. R. Whitt, L. M. Wilson, M. I. Tenaillon, B. S. Gaut, and E. S. Buckler, “Genetic diversity and selection in the maize starch pathway,” Proceedings of the National Academy of Sciences **99**, 12959–12962 (2002).
46. W. A. Ricci, Z. Lu, L. Ji, A. P. Marand, C. L. Ethridge, N. G. Murphy, J. M. Noshay, M. Galli, M. K. Mejía-Guerra, M. Colomé-Tatché *et al.*, “Widespread long-range cis-regulatory elements in the maize genome,” Nature plants pp. 1–13 (2019).
47. J. Candaele, K. Demuynck, D. Mosoti, G. T. Beemster, D. Inzé, and H. Nelissen, “Differential methylation during maize leaf growth targets developmentally regulated genes,” Plant physiology **164**, 1350–1364 (2014).
48. M. Galli, A. Khakhar, Z. Lu, Z. Chen, S. Sen, T. Joshi, J. L. Nemhauser, R. J. Schmitz, and A. Gallavotti, “The dna binding landscape of the maize auxin response factor family,” Nature communications **9**, 4526 (2018).
49. N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching *et al.*, “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome,” Nature genetics **39**, 311 (2007).
50. M. P. Creyghton, A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp *et al.*, “Histone h3k27ac separates active from poised enhancers and predicts developmental state,” Proceedings of the National Academy of Sciences **107**, 21931–21936 (2010).
51. S. Xue, P. J. Bradbury, T. Casstevens, and J. B. Holland, “Genetic architecture of domestication-related traits in maize,” Genetics **204**, 99–113 (2016).
52. Y.-x. Li, C. Li, P. J. Bradbury, X. Liu, F. Lu, C. M. Romay, J. C. Glaubitz, X. Wu, B. Peng, Y. Shi *et al.*, “Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population,” The Plant Journal **86**, 391–402 (2016).
53. C. Xu, H. Zhang, J. Sun, Z. Guo, C. Zou, W.-X. Li, C. Xie, C. Huang, R. Xu, H. Liao *et al.*, “Genome-wide association study dissects yield components associated with low-phosphorus stress tolerance in maize,” Theoretical and applied genetics **131**, 1699–1714 (2018).
54. C. Li, B. Sun, Y. Li, C. Liu, X. Wu, D. Zhang, Y. Shi, Y. Song, E. S. Buckler, Z. Zhang *et al.*, “Numerous genetic loci identified for drought tolerance in the maize nested association mapping populations,” BMC genomics **17**, 894 (2016).
55. C. D. Arnold, D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, and A. Stark, “Genome-wide quantitative enhancer activity maps identified by starr-seq,” Science **339**, 1074–1077 (2013).
56. J. G. Wallace, P. J. Bradbury, N. Zhang, Y. Gibon, M. Stitt, and E. S. Buckler, “Association mapping across numerous traits reveals patterns of functional variation in maize,” PLoS genetics **10**, e1004845 (2014).
57. Z. Dong, O. Danilevskaya, T. Abadie, C. Messina, N. Coles, and M. Cooper, “A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling,” PLoS One **7**, e43450 (2012).
58. S. Salvi, G. Sponza, M. Morgante, D. Tomes, X. Niu, K. A. Fengler, R. Meeley, E. V. Ananiev, S. Svitashev, E. Bruggemann *et al.*, “Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize,” Proceedings of the National Academy of Sciences **104**, 11376–11381 (2007).
59. Y. Jiao, H. Zhao, L. Ren, W. Song, B. Zeng, J. Guo, B. Wang, Z. Liu, J. Chen, W. Li *et al.*, “Genome-wide genetic changes during modern breeding of maize,” Nature genetics **44**, 812 (2012).
60. C. Becker, J. Hagmann, J. Müller, D. Koenig, O. Stegle, K. Borgwardt, and D. Weigel, “Spontaneous epigenetic variation in the arabidopsis thaliana methylome,” Nature **480**, 245 (2011).
61. X. Li, C. Zhu, C.-T. Yeh, W. Wu, E. M. Takacs, K. A. Petsch, F. Tian, G. Bai, E. S. Buckler, G. J. Muehlbauer *et al.*, “Genic and nongenic contributions to natural variation of quantitative traits in maize,” Genome research **22**, 2436–2444 (2012).
62. M. Murray and W. F. Thompson, “Rapid isolation of high molecular weight plant dna,” Nucleic acids research **8**, 4321–4326 (1980).
63. P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves *et al.*, “The b73 maize genome: complexity, diversity, and dynamics,” science **326**, 1112–1115 (2009).
64. H. Li, “Aligning sequence reads, clone sequences and assembly contigs with bwa-mem,” arXiv preprint arXiv:1303.3997 (2013).
65. “Picard toolkit,” http://broadinstitute.github.io/picard/ (2019).
66. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly *et al.*, “The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data,” Genome research **20**, 1297–1303 (2010).
67. B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short dna sequences to the human genome,” Genome biology **10**, R25 (2009).
68. M. D. Schultz, R. J. Schmitz, and J. R. Ecker, “‘leveling’the playing field for analyses of single-base resolution dna methylomes,” Trends in Genetics **28**, 583–585 (2012).
69. H. Wang, T. Nussbaum-Wagler, B. Li, Q. Zhao, Y. Vigouroux, M. Faller, K. Bomblies, L. Lukens, and J. F. Doebley, “The origin of the naked grains of maize,” Nature **436**, 714 (2005).
70. F. Tian, N. M. Stevens, and E. S. Buckler, “Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10,” Proceedings of the National Academy of Sciences **106**, 9979–9986 (2009).
71. H. Chen, N. Patterson, and D. Reich, “Population differentiation as a test for selective sweeps,” Genome research **20**, 393–402 (2010).
72. P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry *et al.*, “The variant call format and vcftools,” Bioinformatics **27**, 2156–2158 (2011).
73. T. Tian, Y. Liu, H. Yan, Q. You, X. Yi, Z. Du, W. Xu, and Z. Su, “agrigo v2. 0: a go analysis toolkit for the agricultural community, 2017 update,” Nucleic acids research **45**, W122–W129 (2017).
74. M. R. Mumbach, A. J. Rubin, R. A. Flynn, C. Dai, P. A. Khavari, W. J. Greenleaf, and H. Y. Chang, “Hichip: efficient and sensitive analysis of protein-directed genome architecture,” Nature methods **13**, 919–922 (2016).
75. B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with bowtie 2,” Nature methods **9**, 357 (2012).
76. F. Ramírez, F. Dündar, S. Diehl, B. A. Grüning, and T. Manke, “deeptools: a flexible platform for exploring deep-sequencing data,” Nucleic acids research **42**, W187–W191 (2014).
77. N. Servant, N. Varoquaux, B. R. Lajoie, E. Viara, C.-J. Chen, J.-P. Vert, E. Heard, J. Dekker, and E. Barillot, “Hic-pro: an optimized and flexible pipeline for hi-c data processing,” Genome biology **16**, 259 (2015).

78. C. A. Lareau and M. J. Aryee, “hichipper: a preprocessing pipeline for calling dna loops from hichip data,” Nature methods **15**, 155 (2018).
79. D. H. Phanstiel, A. P. Boyle, N. Heidari, and M. P. Snyder, “Mango: a bias-correcting chia-pet analysis pipeline,” Bioinformatics **31**, 3092–3098 (2015).
80. J. Yu, J. B. Holland, M. D. McMullen, and E. S. Buckler, “Genetic design and statistical power of nested association mapping in maize,” Genetics **178**, 539–551 (2008).
81. P. J. Bradbury, Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler, “Tassel: software for association mapping of complex traits in diverse samples,” Bioinformatics **23**, 2633–2635 (2007).
82. D. Speed, G. Hemani, M. R. Johnson, and D. J. Balding, “Improved heritability estimation from genome-wide snps,” The American Journal of Human Genetics **91**, 1011–1021 (2012).
83. R. P. Hellens, A. C. Allan, E. N. Friel, K. Bolitho, K. Grafton, M. D. Templeton, S. Karunairetnam, A. P. Gleave, and W. A. Laing, “Transient expression vectors for functional genomics, quantification of promoter activity and rna silencing in plants,” Plant methods **1**, 13 (2005).
84. S.-D. Yoo, Y.-H. Cho, and J. Sheen, “Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis,” Nature protocols **2**, 1565 (2007).

SUPPORTING INFORMATION

SUPPORTING TABLES

Table S1. Teosinte, landrace, modern maize samples used in this study. (https://github.com/jyanglab/msfs_teo/blob/master/table/Table_S1_samples_for_sequencing.xlsx)

Table S2. Population-wide DMRs in the CG and CHG contexts. (https://github.com/jyanglab/msfs_teo/blob/master/table/Table_S2_DMR.xlsx)

Table S3. Selective sweeps detected between populations. (https://github.com/jyanglab/msfs_teo/blob/master/table/Table_S3_Selective_sweep.xlsx)

Table S4. The list of genes with CG teosinte-maize DMRs located at their exonic regions. (https://github.com/jyanglab/msfs_teo/blob/master/table/Table_S4_229_Exonic_DMR_Genes.xlsx)

Table S5. The list of genes exhibiting interactive loops between genes and hypomethylated DMRs in maize located at the intergenic regions. (https://github.com/jyanglab/msfs_teo/blob/master/table/Table_S5_60_Genes_interactive_with_intergenic_hypo_DMR.xlsx)

SUPPORTING FIGURES

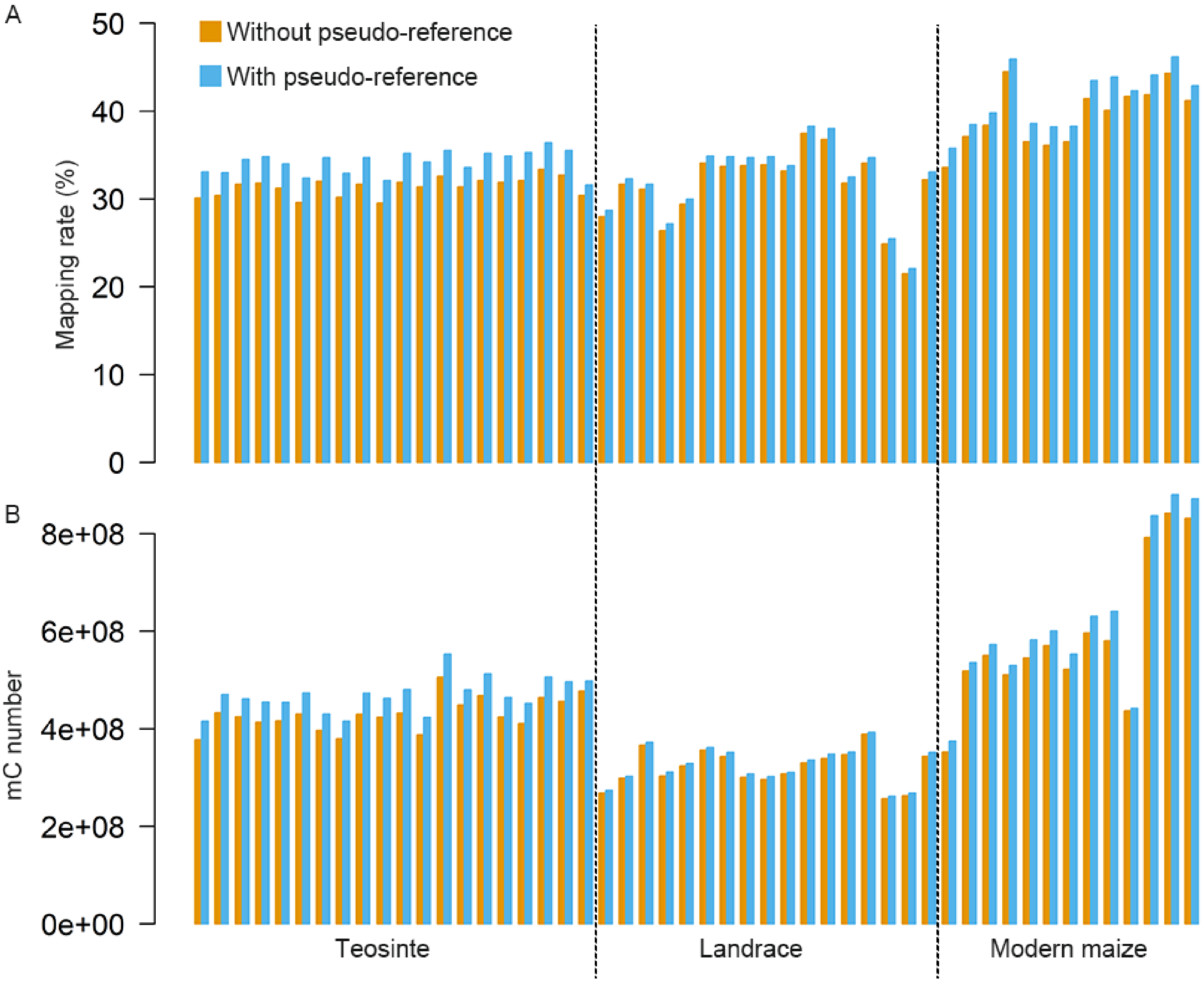

Fig. S1. Comparison of mapping rates (A) and number of methylated cytosine (mC) sites (B) with and without using pseudo-reference genome in different populations. B73 reference genome (AGPv4) was used in the analyses.

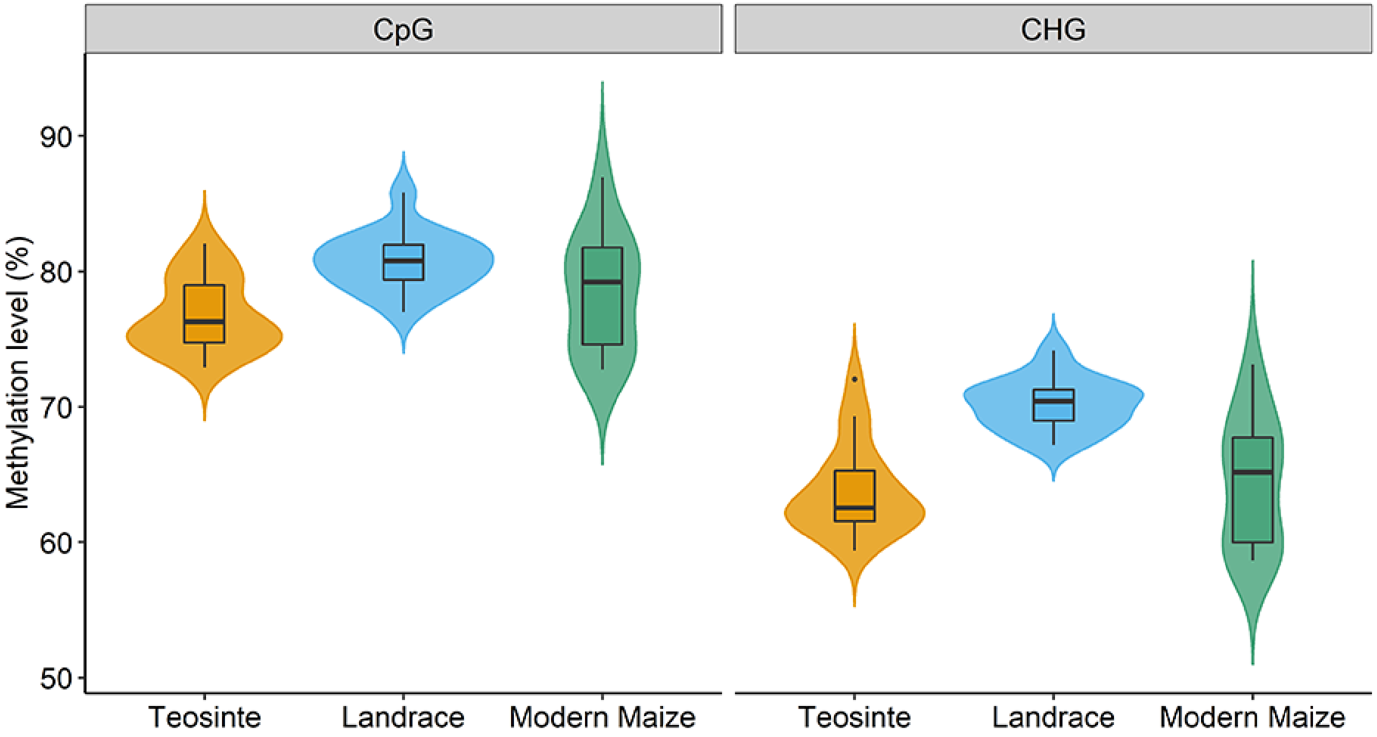

Fig. S2. Distributions of levels of DNA methylation in teosinte, landrace, and modern maize populations. Left panel denotes results from CG sites and right panel denotes results from CHG sites.

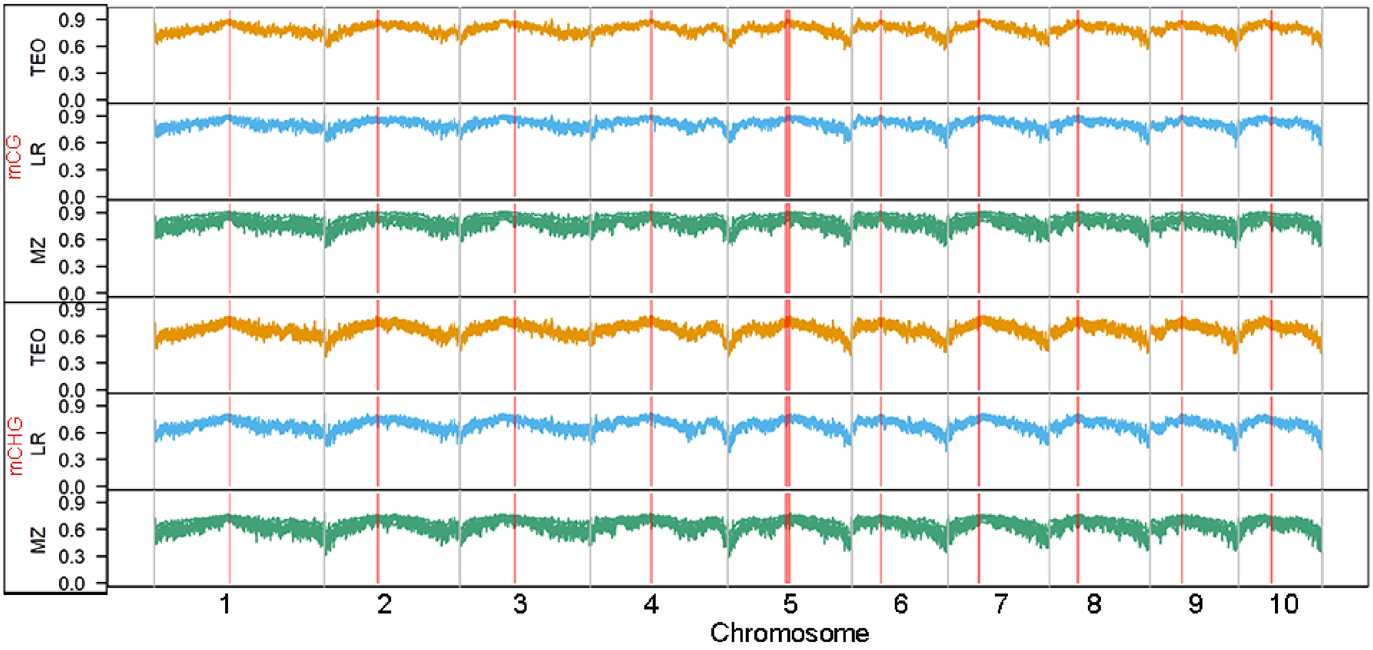

Fig. S3. Genome-wide distributions of DNA methylation across 10 maize chromosomes. TEO, LR, and MZ represent teosinte, landrace and modern maize populations, respectively. Red vertical lines indicated the pericentromeric regions.

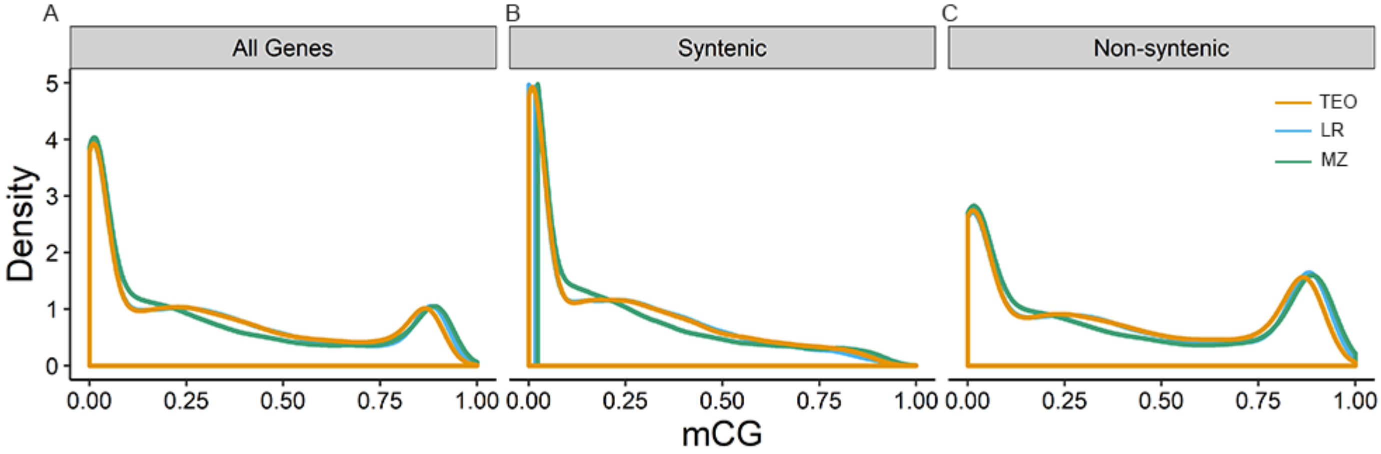

Fig. S4. Density plots of CG methylation in gene body for all the annotated maize genes (A) and for syntenic (B) and non-syntenic genes (C). TEO, LR, and MZ represent teosinte, landrace and modern maize populations, respectively. The syntenic and nonsyntenic orthologs was determined by comparing maize with Sorghum.

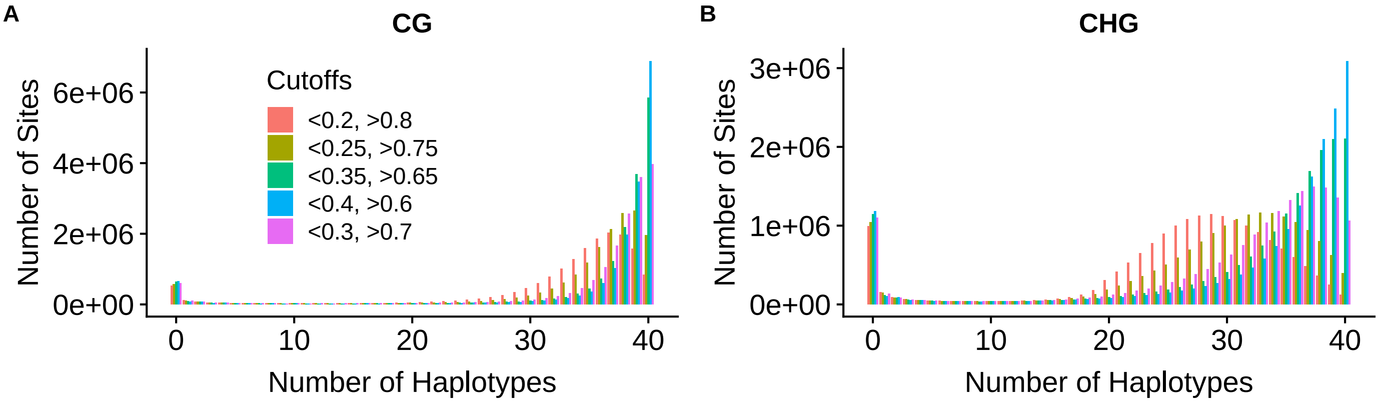

Fig. S5. Sensitivity tests using different cutoffs. Distributions of mSFS using different thresholds to determine the methylated, unmethylated, and heterozygote for the 100-bp tiles under CG (**A**) and CHG (**B**) contexts.

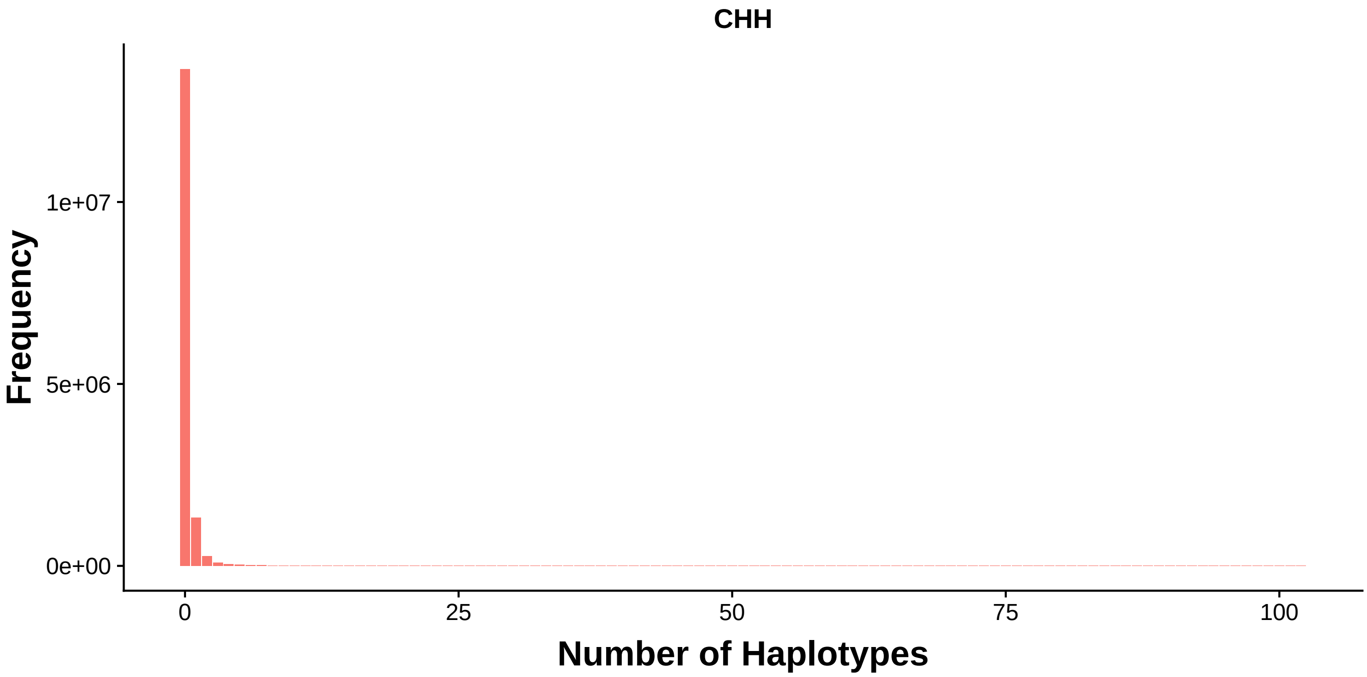

Fig. S6. Methylome site frequency spectrum under the CHH context. The distribution is highly skewed towards the unmethylated status for CHH sites.

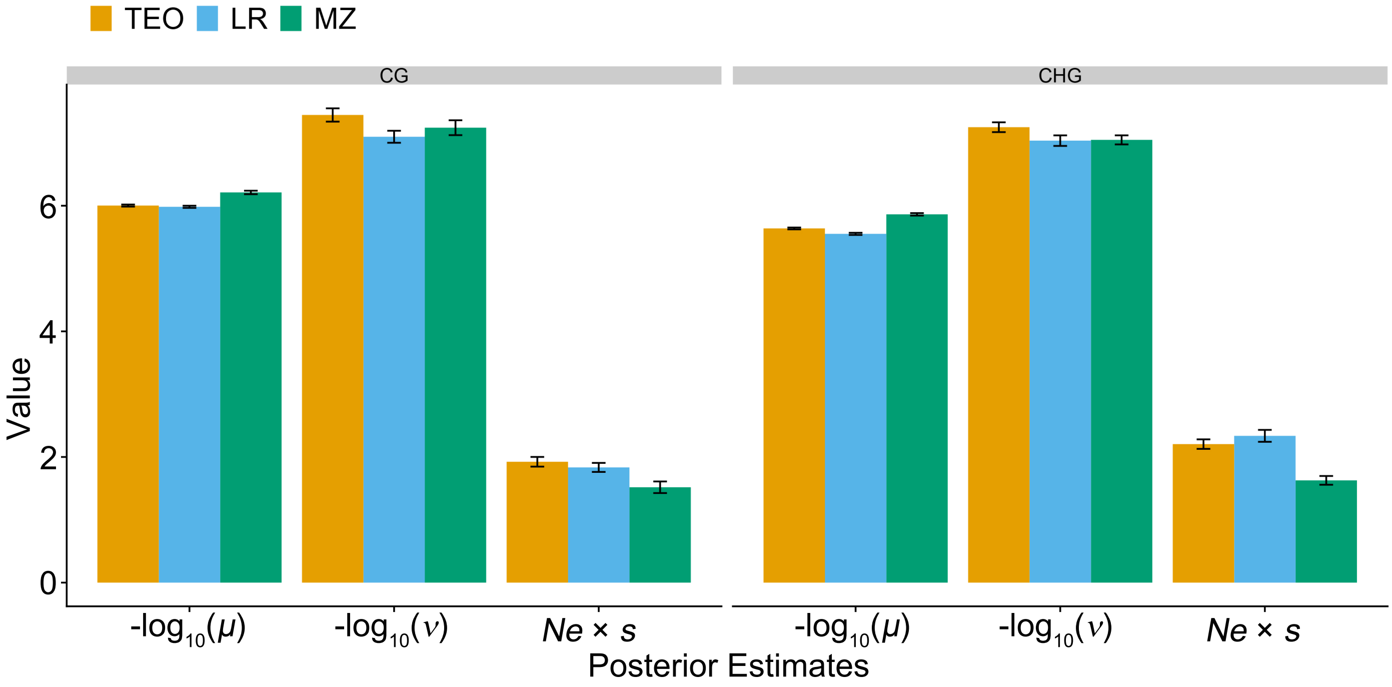

Fig. S7. Population genetic parameter inference using each individual population. Posterior estimators of mean values and standard errors for μ, ν, and Ne $\times s$ for CG and CHG sites. Values were estimated using MCMC approach with 25% burnin.

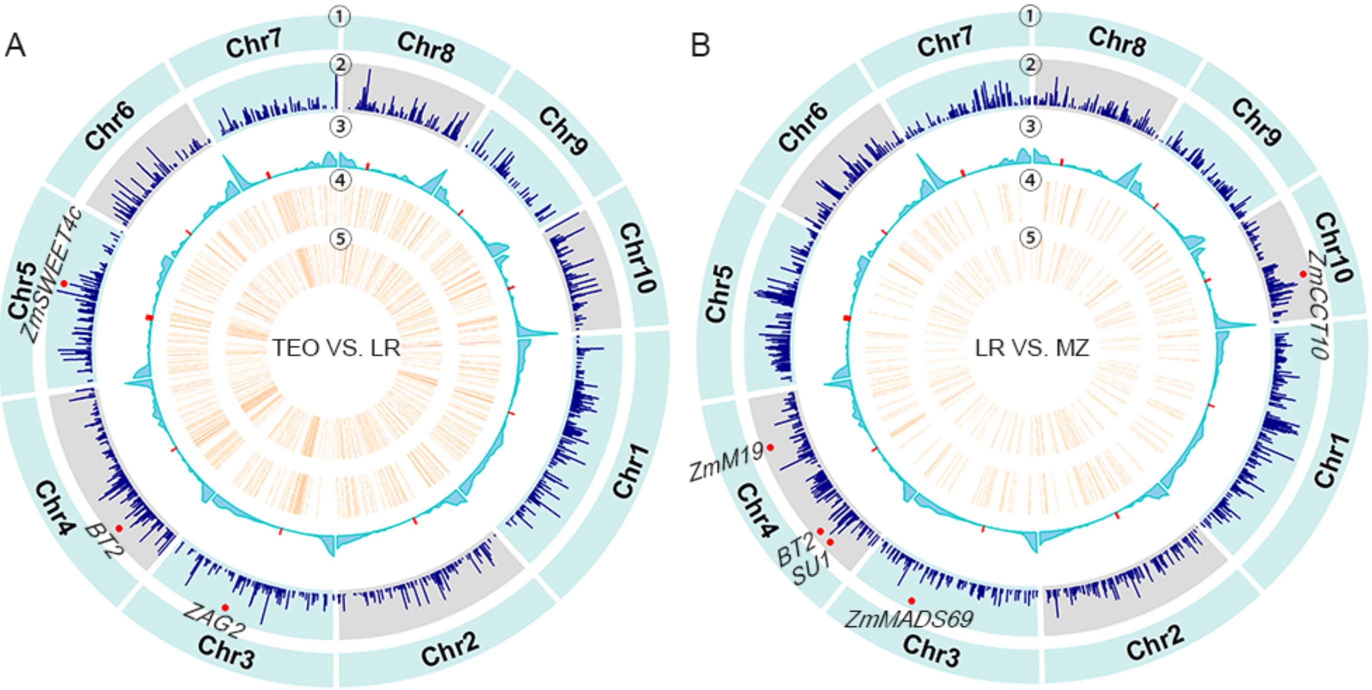

Fig. S8. Landscape of selection signals and DNA methylation variation across maize genome. Genome-wide distributions of selective sweeps, DMRs across ten maize chromosomes detected by teosinte vs. landrace (**A**) and landrace vs. modern maize (**B**). From outer to inner circles: ① chromosome names, ② selective sweeps, ③ recombination rate, and the density of DMRs (number per 1-Mb) in ④ CG and ⑤ CHG contexts. Red dots at the second track indicated the physical positions of the known genes located within the selective sweeps. Red dots at the third track indicated the centromeric regions.

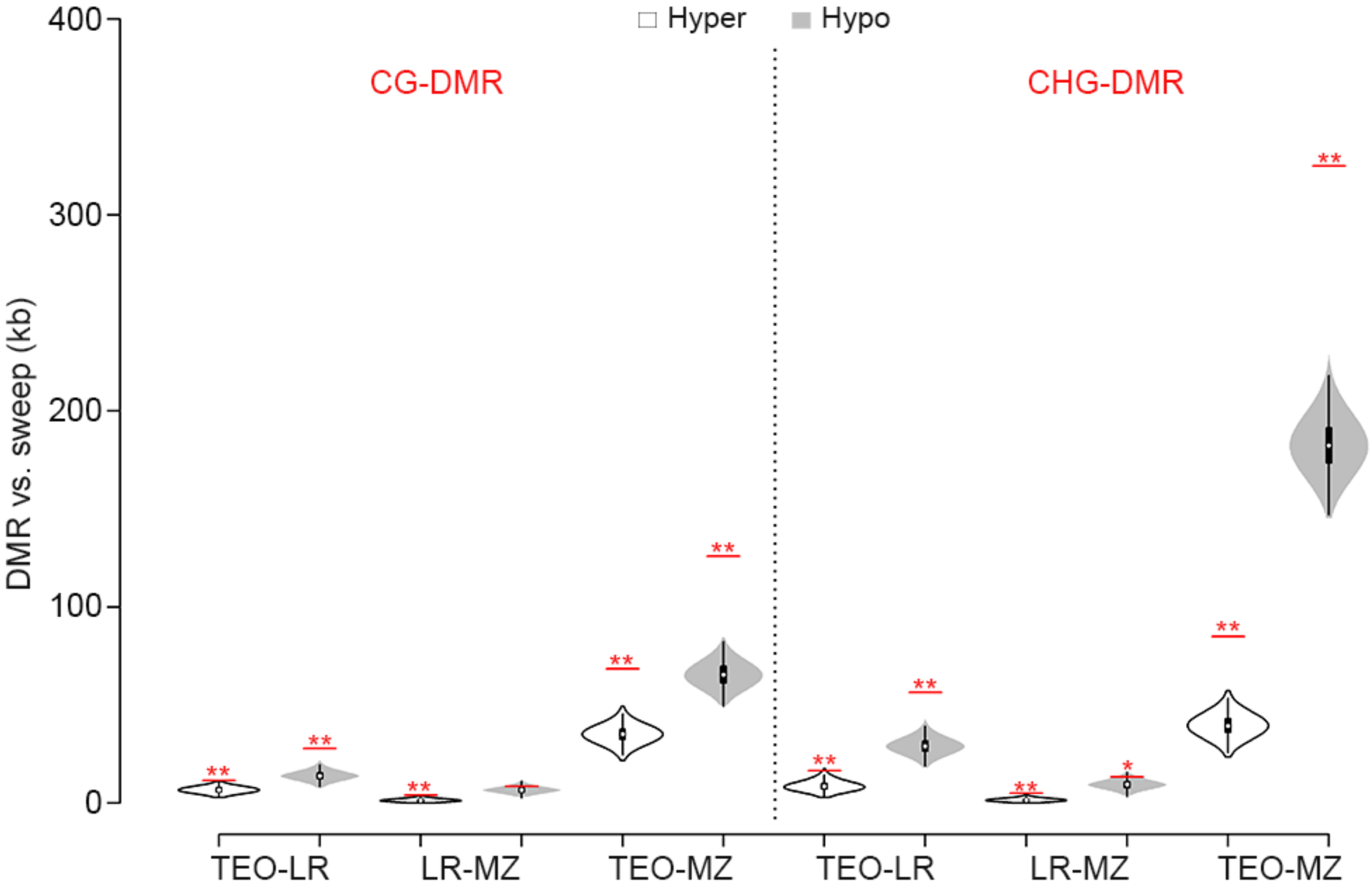

Fig. S9. The overlapped basepairs between DMRs and selective sweeps. Red horizontal bars indicated the observed values and violin plots showed the 1,000 permutation results using randomly selected regions sharing the similar genomic features as the DMRs. Red asterisks indicated the statistical significances with one asterisk denoting P-value < 0.05 and two asterisks denoting P-value < 0.01. Hyper- and hypomethylation were defined based on maize.

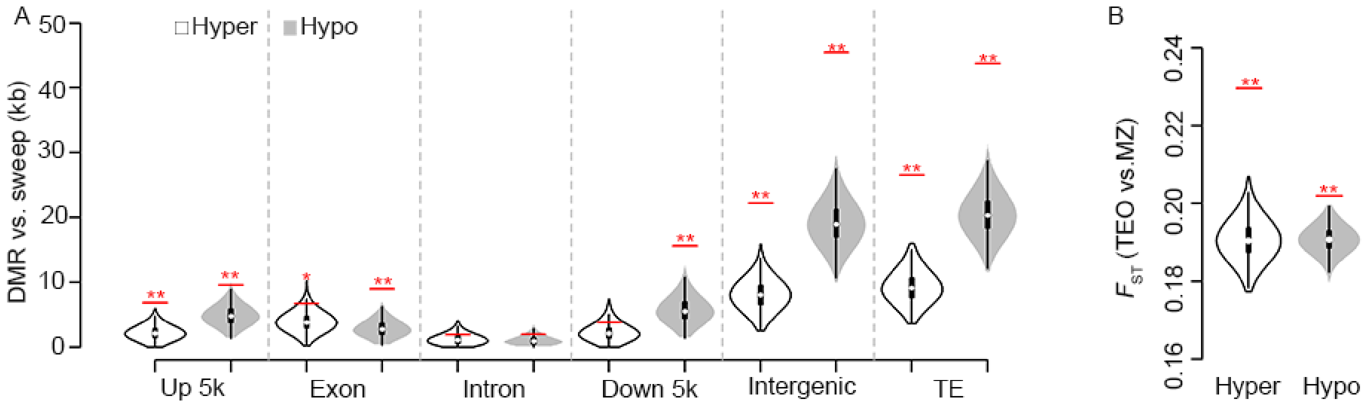

Fig. S10. Selection on differentially methylated regions. (**A**) Overlaps between teosinte-maize DMRs and selective sweeps breaking down into different genomic features. (**B**) Mean F_{ST} values of teosinte-maize DMRs that were hyper- and hypomethylated in maize. Red horizontal bars indicated the observed values and violin plots showed the 1,000 permutation results using randomly selected regions sharing the similar genomic features as the DMRs. Red asterisks indicated the statistical significances with one asterisk denoting P-value < 0.05 and two asterisks denoting P-value < 0.01.

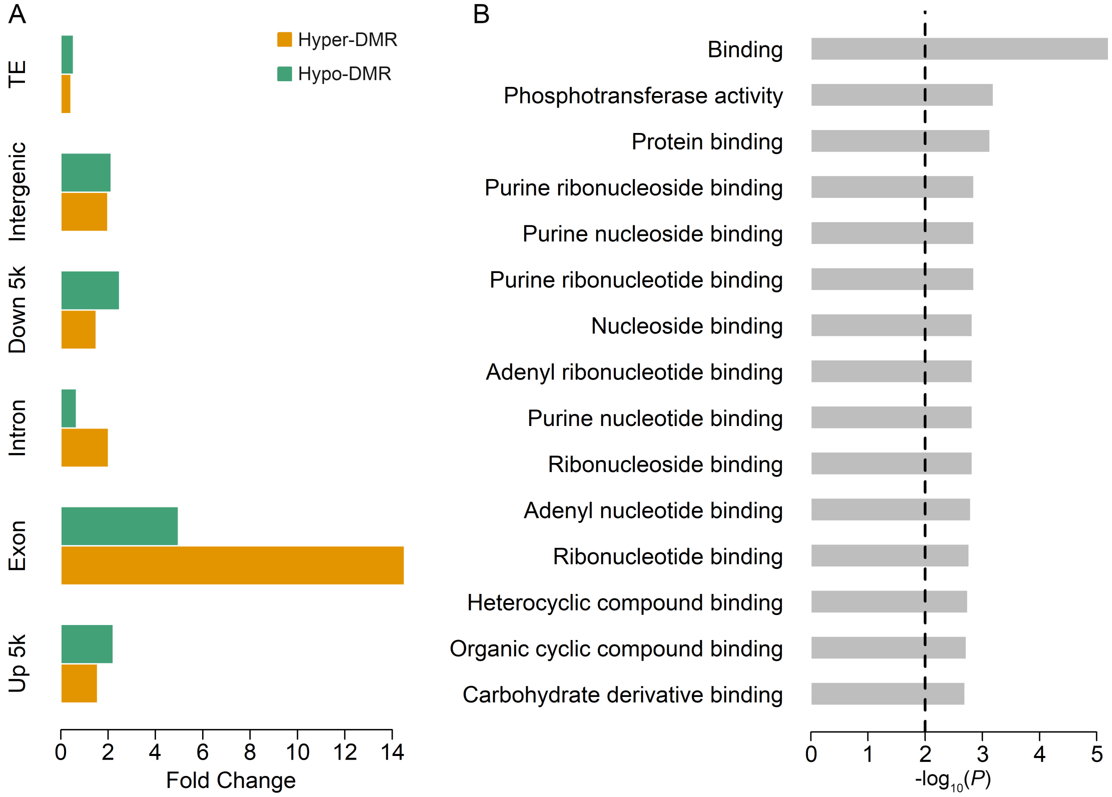

Fig. S11. Teosinte-maize CG DMRs and their associated functional features. (**A**) Fold changes of mappable DMR length relative to the mean values from 1,000 permutations. (**B**) Result of GO term enrichment test using genes exhibiting an exonic DMR. Vertical dashed line indicated the significance cutoff (*P*-value=0.01).

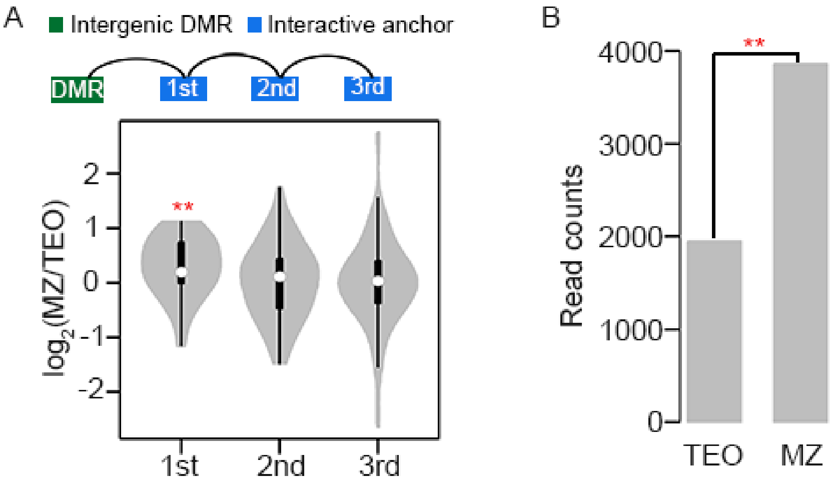

Fig. S12. Intergenic CG DMR altered downstream gene expression. (**A**) Contrast of the gene expression levels in maize relative to teosinte. In the upper panel, the schematic diagram shows the genes that involved in the 1st, 2nd, and 3rd level interactions with maize hypomethylated DMRs located in intergenic regions. Red asterisks indicated the statistically significant differences (P-value < 0.01). (**B**) Gene expression level of *Zm00001d018036* in teosinte and modern maize (P-value $= 4.6e^{-141}$ according to [30]).

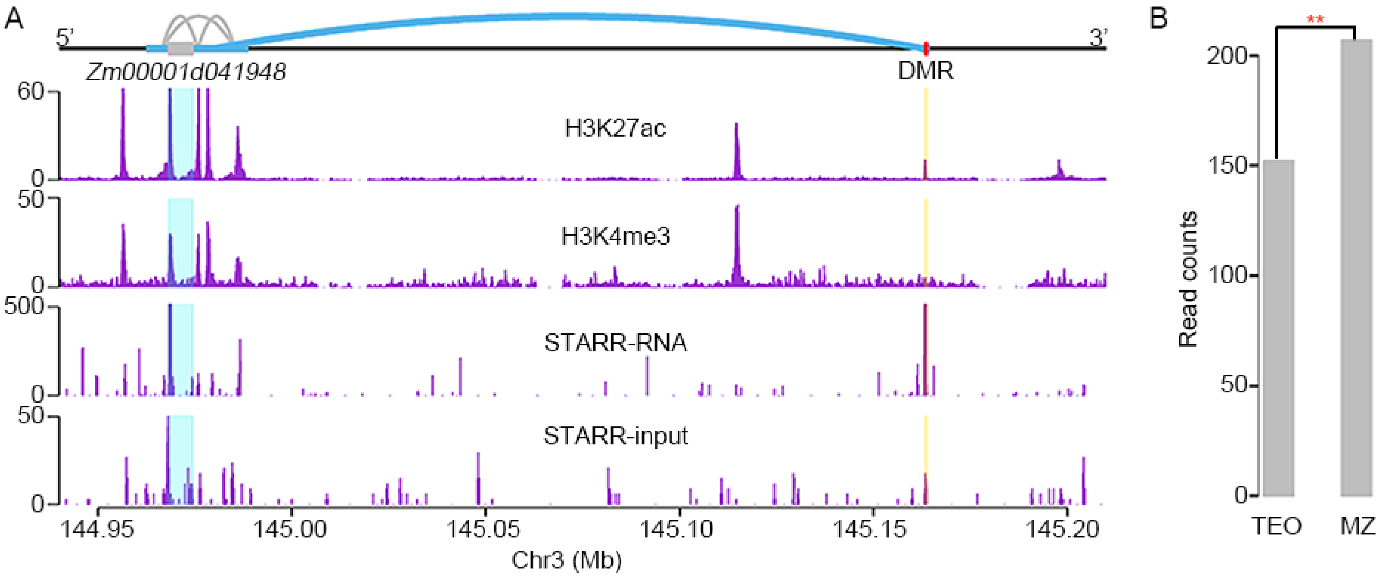

Fig. S13. Interactive loops between a DMR and a gene model *Zm0001d041948*. (**A**) Chromatin interactions (the upper panel) and ChIP-Seq profiles (the lower panels) at gene *Zm00001d041948*. Gray and red boxes indicated the physical position of the gene model and the DMR. Gray and blue lines denoted the interactive loops. (**B**) RNA-seq read counts of gene model *Zm00001d041948* in teosinte and modern maize (*P*-value = $4.3e^{-3}$ according to [30]).

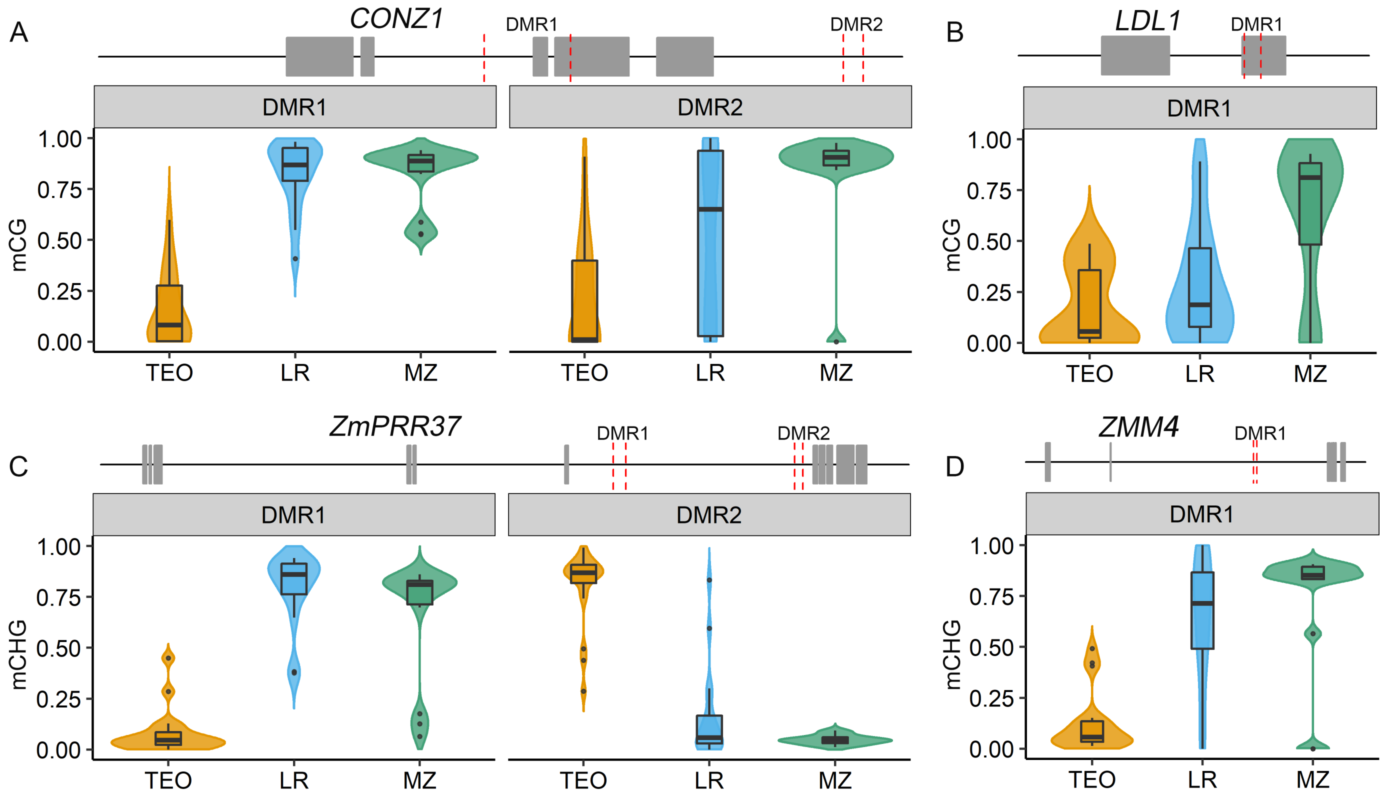

Fig. S14. Teosinte-maize DMRs located at flowering time genes. Distribution of methylation level within DMRs that located at *CONZ1* (**A**), *LDL1* (**B**), *ZmPRR37* (**C**) and *ZMM4* (**D**) in teosinte (TEO), landrace (LR), and modern maize (MZ). Two nearby vertical dashed red lines on the gene model indicated a teosinte-maize DMR.

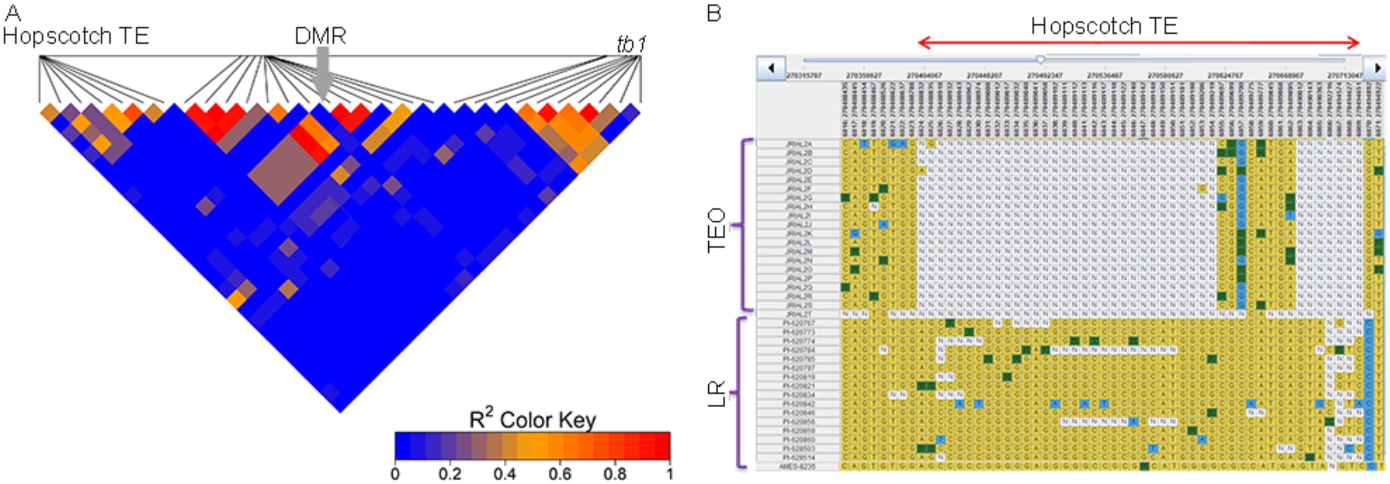

Fig. S15. The linkage disequilibrium (LD) analysis among the Hopscotch transposon, *tb1*-DMR, and *tb1* gene. (**A**) The LD heatmap using 17 landraces segregating at the *tb1*-DMR locus. The arrow indicated the position of the *tb1*-DMR. LD analysis was performed using SNPs (coded with 0, 1, and 2) called from the WGS data and the mCG level of the *tb1*-DMR. (**B**) The SNP genotypes of teosinte and landrace around the Hopscotch transposon insertion region.

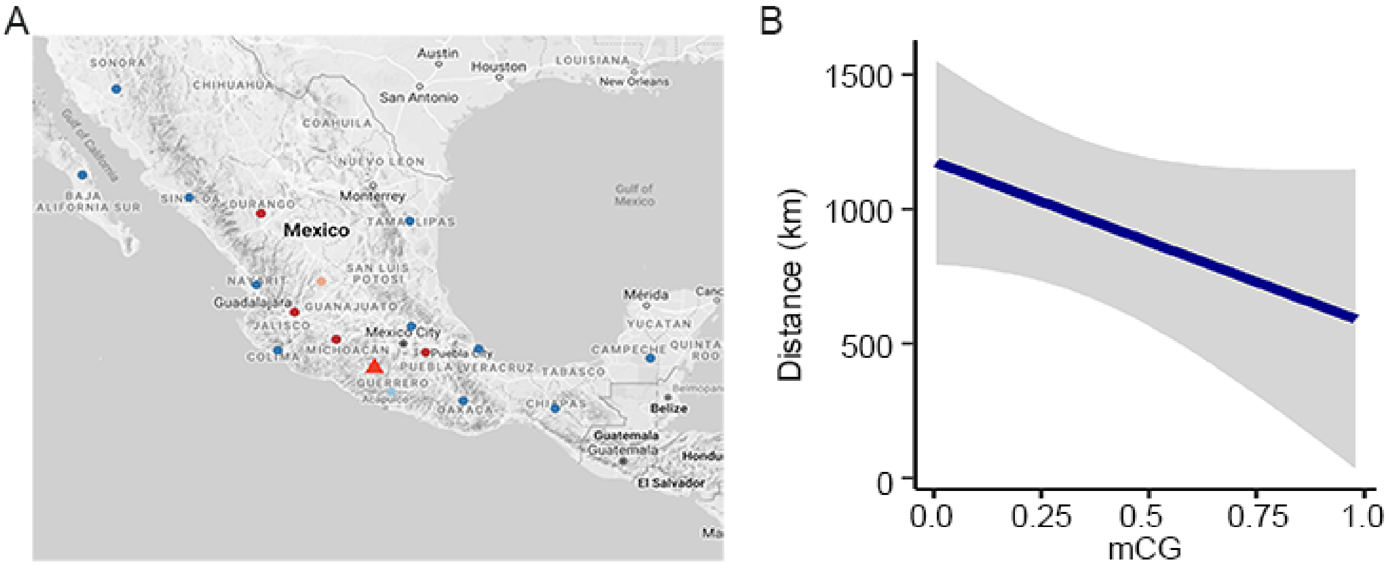

Fig. S16. Correlation analysis between geographical distributions of landrace samples and their CG methylation level at *tb1*-DMR. (**A**) Geographical distributions of the teosinte (triangle) and landrace (points) samples. Red denotes highly methylated and blue denotes lowly methylated samples in the *tb1*-DMR. (**B**) Level of mCG correlated with distance to the origin (Balsas River Valley) of maize.

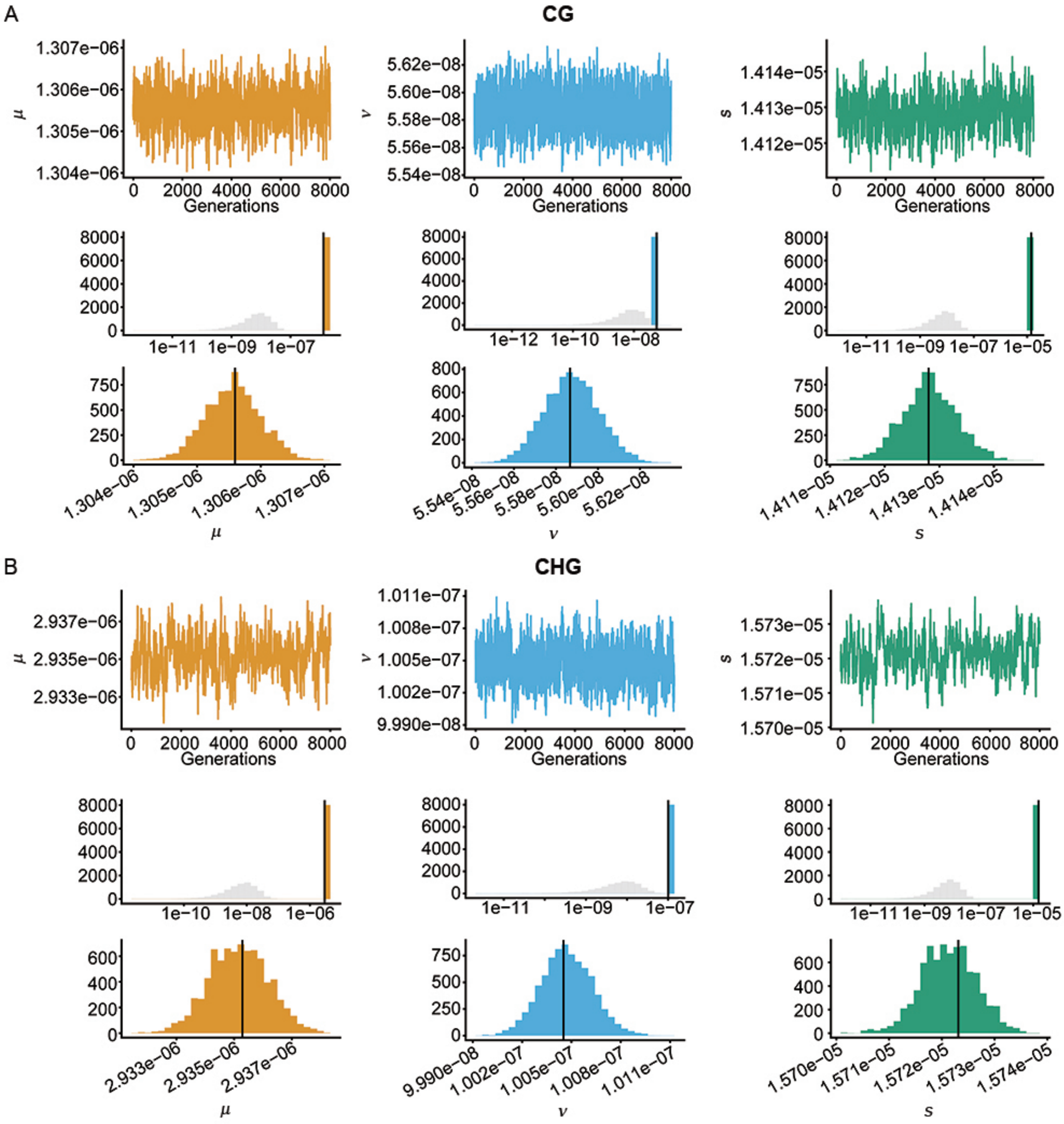

Fig. S17. MCMC tracing plots for parameters estimation. Parameters estimations using the Markov Chain Monte Carlo (MCMC) approach for forward epimutation rate (μ the left panels), backward epimutation rate (ν, the middle panels), and selection coefficient (s, the right panels) under CG (A) and CHG (B) contexts. In each plot, the top panels are the MCMC tracing plots with 25% burnin, the middle panels are the prior (grey) and posterior (colored) parameter distributions and the bottom panels are the enlarged posterior distributions.