bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Evolutionary and functional genomics of DNA methylation in maize domestication and improvement

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 Agronomy 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 October 9, 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 epimutation rates and selection coefficients. We found weak evidence for direct selection on DNA methylation in any context, but thousands of differentially methylated regions (DMRs) were identified in population-wide that are correlated with recent selection. 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. 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 of cytosine residues. 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 exhibit substantial variation 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and even lethality [12–14]. Nonetheless, variation in DNA methylation has been detected both in natural plant [15] and human populations [16]. Levels of DNA methylation can be affected by genetic variation and environmental cues [17]. Additionally, heritable *de novo* epimutation — the stochastic loss or gain of DNA methylation — can occur spontaneously and has functional consequences [18, 19]. 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 [20].

In *Arabidopsis*, a multi-generational epimutation accumulation experiment [21] 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 (but see recently estimates for poplar and dandelion [22]), 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 [23] developed an analytical framework to address evolution questions for epimutations. Leveraging this theoretical framework, Vidalis et al. [24] constructed the methylome site frequency spectrum (mSFS) using worldwide *Arabidopsis* samples, but they failed to find evidence for selection 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 reveal that the dramatic morphological differences between maize and teosinte are largely due to selection of several major effect loci [25]. As maize spread across the Americas, many additional loci have played an important role in local adaptation [26]. Flowering time, a trait that directly affects plant fitness, played a major role in this local adaptation process [27–29]. Previous research, however, has focused almost entirely on DNA variation, and the contributions of methylation variation to maize domestication and adaptation remain largely elusive.

Here, we collected a set of geographically widespread Mexican landraces and a natural population of teosinte near Palmar Chico, Mexico [30], from which we generated genome and methylome sequencing data. Additionally, we profiled the teosinte interactome using HiChIP. Together with the analysis from previously published genome [31], transcriptome [32], methylome [6], and interactome [33] datasets, we estimated epimutation rates and selection pressures across different timescales, investigated the DNA methylation landscape in maize and teosinte, detected differentially methylated regions (DMRs), characterized the genomic features that are related with DMRs, and functionally validated two DMRs that are associated with adaptive traits. Our results suggest that DNA methylation genome-wide is likely only under relatively weak selection, but that methylation differences at a subset of key loci may modulate the regulation of domestication genes and 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 **methods**) that alleviated potential bias of mapping reads to a single reference genome [34] 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 [13] 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 be due to lower sequencing depth [35]. We found no significant differences between teosinte and maize as a group (**Figure S2**).

We found methylated cytosines in CG and CHG contexts were significantly higher in 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**). We calculated the average methylated CG (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* [36] exhibited little gbm (**Figure S4B-C**), consistent with previous reports [5, 37].

Genome-wide methylation is only under weak selection

As the frequency of methylation may be affected by both selection and epimutation rates, we implemented a novel MCMC approach to estimate these parameters using a population genetic model developed for highly variable loci [23]. We defined 100-bp tiles across the genome as a DNA methylation locus and categorized individual tiles as unmethylated, methylated, or heterozygous alleles for outcrossed populations (i.e., teosinte and landrace populations) and as unmethylated and methylated alleles for modern maize inbred lines (see **methods**). To determine the thresholds for methylation calls, we employed an iterative expectation maximization algorithm to fit the data [38]. We then constructed methylome site frequency spectra (mSFS) for CG and CHG sites (**Figure S5**). Sensitivity test results suggested that the mSFS was insensitive to the cutoffs used for the methylation calls (**Figure S6**). As the vast majority ($> 98\%$) of CHH sites were unmethylated (**Figure S7**), we excluded CHH sites from population genetic analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

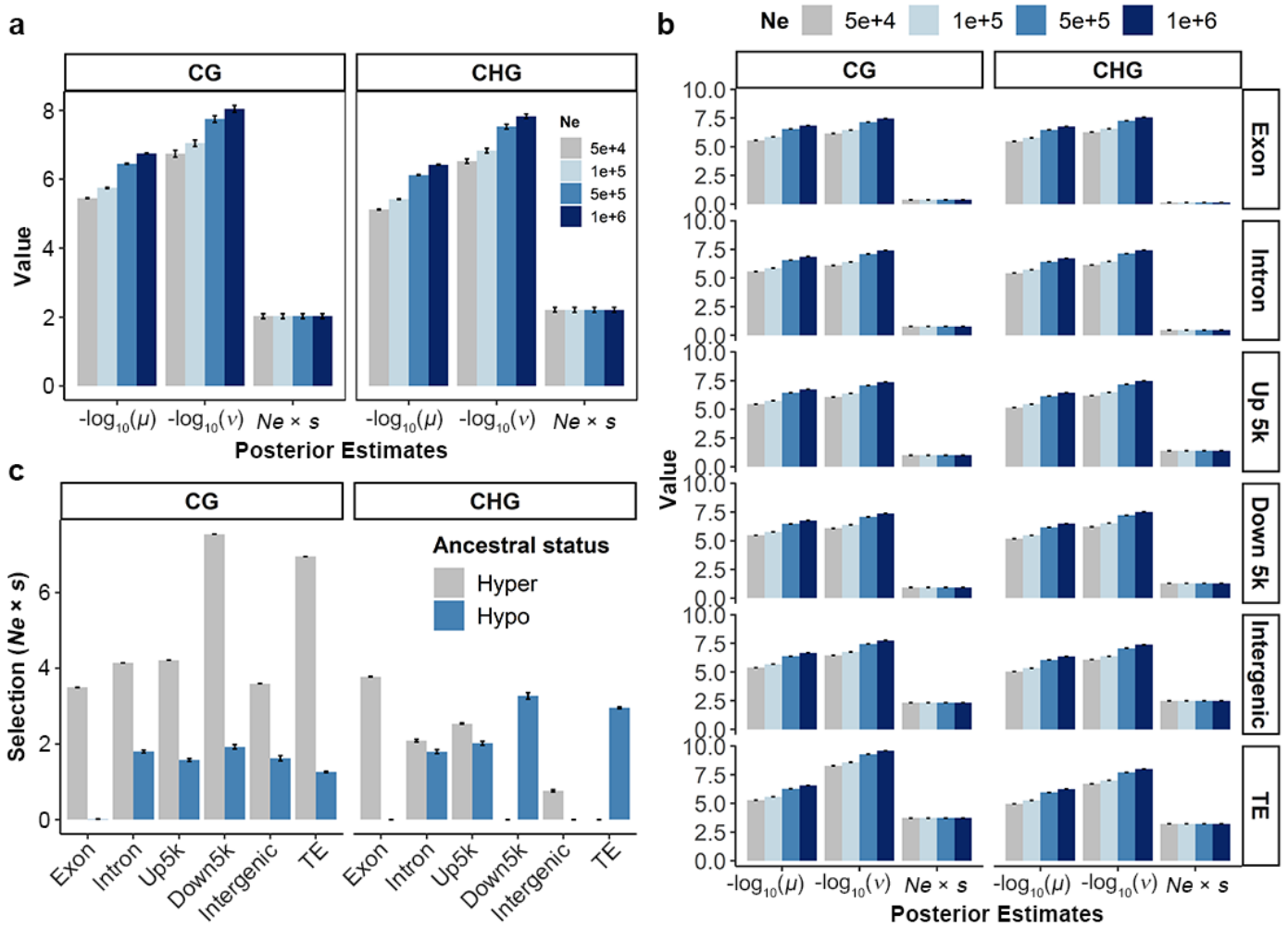

Fig. 1. Population genetic parameters inference. (**a**) Posterior estimates of mean values and standard deviations for μ, ν, and $Ne \times s$ for CG and CHG sites using four different effective population size (N_e) values. (**b**) Posterior estimates for different genomic features. Up 5k, the upstream 5k region of a gene; Down 5k, the downstream 5k region of a gene. (**c**) Posterior estimates by defining teosinte as the ancestral epiallele. Values were estimated using MCMC approach with 20% burnin (see **methods**). Error bars indicate standard deviations. Source Data underlying (**a-c**) are provided in a Source Data file. (N = 1,600 for each bar).

After testing a set of prior values, we found the initial prior rates had little impact on the posteriors, except for extremely large values (**Figure S8**), for which convergence was difficult. Because we found little difference among populations in genome-wide patterns, we estimated parameters using the combined data; estimates from individual populations were nonetheless broadly similar (**Figure S9**). Effective population size (N_e) in maize is difficult to estimate because of rapid demographic change during and post-domestication. Previous estimates of N_e in maize range from $\sim 50k$ [39] to $\sim 370k - 1M$ [40]. To account for this uncertainty, we ran the models with a set of different N_e values ($50k$, $100k$, $500k$, and $1M$). Model estimates of the epimutation rate μ for both CG ($3.6 \times 10^{-6} - 1.8 \times 10^{-7}$) and CHG ($7.6 \times 10^{-6} - 3.8 \times 10^{-7}$) sites were more than an order of magnitude higher than the backward epimutation rates ($\nu = 1.8 \times 10^{-7} - 9.0 \times 10^{-9}$ and $3.0 \times 10^{-7} - 1.5 \times 10^{-8}$) using different N_e values (**Figure 1A**), 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 for both CG and CHG tiles depended on the assumption of N_e. However, the population-scaled selection coefficient (or $Ne \times s$) stayed largely constant with values of 2.0 and 2.2 for CG and CHG tiles, indicating relatively weak selection for methylation in each context according to classical population genetic theory [41].

We then sought to test whether the population-scaled selection coefficient differs across genomic features. After fitting mSFS models separately for different genomic features, results showed that population-scaled selection coefficients in genic regions (exon, intron, upstream 5k, and downstream 5k) were below or close to 1, and the values were above 1 for non-genic regions (i.e., 2.4 for intergenic regions and 3.5 for TE regions) (**Figure 1B**), suggesting stronger selection on methylation variation outside of genes. If we consider the most common variant in teosinte as the ancestral epiallele, selection was higher in ancestrally hypermethylated regions in CG contexts, especially in TE and intergenic regions, while it was close to neutrality for ancestrally hypomethylated regions, especially for the exonic regions (**Figure 1C**). In CHG contexts, selection was weak in most regions, including TE and intergenic regions, for both ancestral hyper- and hypomethylated sites.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Regions with variable methylation contribute to phenotypic variation

Our observed CG mSFS revealed that 2% and 7% of 100-bp tiles were completely unmethylated and methylated, while 91% of tiles were variable (**Figure S5A**). These variable methylation regions can be further divided into rarely unmethylated (frequency of methylated tiles > 90%), rarely methylated (frequency of methylated tiles < 10%), and high-frequency variable regions (frequency of methylated tiles >= 10% and <= 90%), composing 69%, 2%, and 20% of the genome, respectively. To investigate whether regions of the genome exhibiting variable methylation, especially the high-frequency variable regions, are functionally relevant, we used SNP sets residing in these five regions to estimate their contribution to phenotypic variation using published data from a large maize mapping population [42]. We estimated kinship matrices for SNPs in different genomic regions and then partitioned the genetic variance for plant phenotypes using LDAK [43]. Consistent with an important functional role for genic regions and a lack of functional importance in permanently methylated regions, our results find that sites that are hypomethylated (uniformly unmethylated and rarely methylated), mainly from the genic areas, explained disproportionally larger genetic variances (**Figure S10A**). While hypermethylated regions (uniformly methylated and rarely unmethylated), although accounting for 76% of the genome, contribute only a fraction of the genetic variance for 7/23 traits. The proportion of variance explained by high-frequency sites polymorphic for methylation, ranged from 0% to 57%, with a mean value of 29%. Variance component analysis results for CHG sites were largely consistent with the results for CG sites (see **Figure S10B**).

Population level DMRs are enriched in selective sweeps

Although genome-wide selection on epimutation appears relatively weak, the observation that sites exhibiting methylation polymorphism contribute meaningfully to quantitative trait variation suggested that stronger selection could be acting at specific differentially methylated regions (DMRs). We employed the metilene software [44] to identify a total of 5,278 DMRs (see **Table 1** for numbers broken down by context and type), 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**). To check the tissue-specificity of the detected DMRs, we examined the methylation levels of these DMRs in B73 across different tissue types using published WGBS data [45]. Our results suggested that methylation levels of the DMRs were largely conserved in B73 across three tissue types (**Figure S11**), consistent with the previous studies [20, 46, 47].

Table 1. Number of differentially methylated regions (DMRs) broken down by context and type.

Comparison	Context	Type[1]	DMR NO.	Sweep DMR[2]	Interacting DMR[3]	Interacting DMRs in sweeps[4]
Teosinte vs. Landrace	CG	Hypo in landrace	287	80 (27.8%) **	96 (33.4%) **	18 (18.7%) *
		Hyper in landrace	144	36 (25%) **	61 (42.3%) **	6 (9.8%) **
	CHG	Hypo in landrace	438	121 (27.6%) **	115 (26.2%) **	22 (19.1%) *
		Hyper in landrace	150	45 (30%) **	39 (26%) **	8 (20.5%)
Landrace vs. Maize	CG	Hypo in maize	143	29 (20.2%) **	45 (31.4%) **	9 (20%)
		Hyper in maize	28	13 (46.4%)	3 (10.7%)	1 (33.3%)
	CHG	Hypo in maize	158	36 (22.7%) **	46 (29.1%) **	10 (21.7%)
		Hyper in maize	30	13 (43.3%) *	5 (16.6%)	1 (20%)
Teosinte vs. Maize	CG	Hypo in maize	998	281 (28.1%) **	396 (39.6%) **	67 (16.9%) **
		Hyper in maize	544	147 (27%) **	259 (47.6%) **	32 (12.3%) **
	CHG	Hypo in maize	1,855	490 (26.4%) **	594 (32%) **	104 (17.5%) **
		Hyper in maize	503	159 (31.6%) **	124 (24.6%) **	19 (15.3%) **

[1]Hypo and hyper indicate hypomethylated and hypermethylated regions in a given population.
[2,3]Number of DMR overlapped with selective sweeps (Sweep DMR / total DMR)[2] and involved in interactive loops (Interacting DMR / total DMR)[3]. Statistical significance was determined using one-sided permutation test (*P<0.05 and **P<0.01).
[4]Number of interacting DMR overlapped with selective sweeps (Interacting DMRs in sweeps / total interacting DMR, Chi-squared test, *P<0.05 and **P<0.01).

DNA methylation can have a number of functional consequences [15, 48, 49], 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 **methods**). We detected a total of 1,330 selective sweeps between modern maize and teosinte (**Figure 2** and **Table S3**, see **Figure S12** for results of teosinte vs. landrace and landrace vs. modern maize). Several classical domestication genes, e.g., *tb1* [50], *ZAG2* [51], *ZmSWEET4c* [52], *RA1* [53], and *BT2* [54] were among these selective signals.

We found that DMRs at CG and CHG sites were highly enriched in regions showing evidence of recent selection (**Figure S13**,

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

P-value < 0.001), particularly in intergenic and TE regions (**Figure S14A**). DMRs overlapping sweeps, both hypo- and hypermethylated in maize, exhibited significantly higher allele frequency differentiation between maize and teosinte (**Figure S14B**, see **Table 1** for other comparisons). We then asked whether DMRs in sweep regions were in linkage disequilibrium with nearby SNPs (see **methods**), as might be expected if most DMRs were the result of an underlying genetic change such as a TE insertion. Indeed, the rate of sweep DMRs in LD with local SNPs was significantly higher than expected by chance (**Table S4**).

Additionally, we detected 72 genes located in sweep DMRs (maize vs. teosinte under CG context) that were hypomethylated in maize, 24 (42/72 with expression data) of which showed significantly (Student's paired t-test, P-value < 0.05) increased expression levels in maize compared to teosinte using published data [32]. For the 56 genes located in sweep DMRs that were hypermethylated in maize, however, we failed to detect the significant expression differences between maize and teosinte.

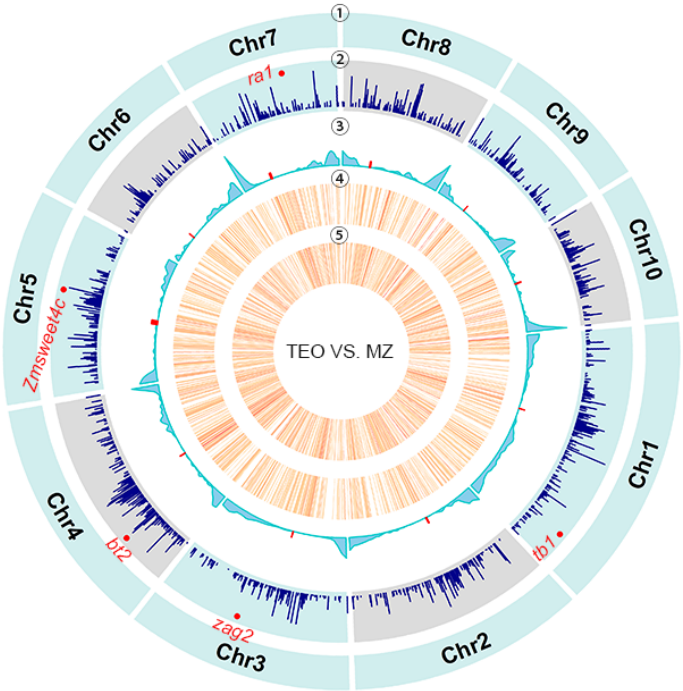

Fig. 2. Selection on differentially methylated regions. Distributions of teosinte-maize selective sweeps, DMRs and other genomic features across ten maize chromosomes. From outer to inner circles were ① Chromosome names, ② selective sweeps detected between modern maize and teosinte, ③ the recombination 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. Source Data underlying are provided in a Source Data file.

Hypomethylated regions in maize are associated with interacting loops

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 much higher than expected by chance (permutation P-values < 0.001, **Figure S15**A). These CG DMRs could be mapped to $N = 229$ unique genes (**Table S5**). After examining the mapping locations based on collapsed gene models, we found that DMRs were most abundant in at 5′ UTRs (**Figure 3**B), consistent with a pattern that was previously observed [56]. Using these DMR genes for a gene ontology (GO) analysis, we detected 15 molecular function terms that were significantly enriched (**Figure S15**B). The vast majority (14/15) of these significant terms were associated with "binding" activities, including protein, nucleoside, and ribonucleoside binding. Furthermore, we found that exonic DMRs were enriched at transcription factor binding sites identified via DAP-seq [57] (permutation P-value < 0.001).

These findings suggested a potential role for DMRs affecting regulatory regions. To investigate this possibility, we made use of recent data using long-read ChIA-PET to profile genomic regions colocalized with H3K4me3 and H3K27ac to define the interactome in maize [33]. We found that interactive anchor sequences were significantly enriched in DMRs that are hypomethylated in maize, especially in regulatory regions, including upstream 5-kb, downstream 5-kb, and intergenic regions (**Figure 3**A). We also found that DMRs located in transposable elements that were hypomethylated in maize more likely overlap with interactive anchors than expected by chance (permutation P-value < 0.001).

We hypothesized that these hypomethylated DMRs, especially intergenic DMRs overlapped with the regulatory 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 S16**A). Interestingly, among the 60 genes in direct contact with maize hypomethylated intergenic DMRs (**Table S6**), we found that 30 (43/60 with expression data) showed

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

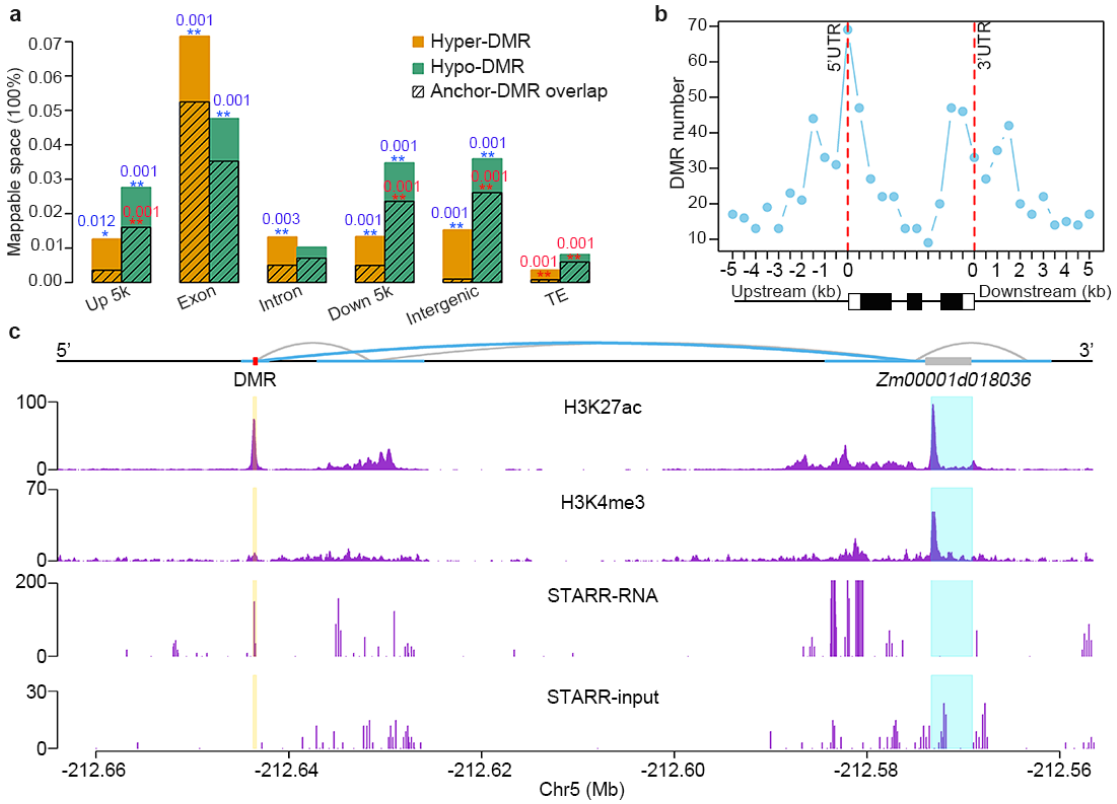

Fig. 3. Teosinte-maize CG DMRs and their associated functional features. (**a**) Breakdown of hyper-DMRs (DMR hypermethylated in maize) and hypo-DMRs (DMR hypomethylated in maize) into genomic features and their overlaps with interactive anchors using data obtained from Li et al., [33]. Blue and red stars indicated DMRs that were significantly enriched at genomic features and interaction anchors (one-sided permutation test: *P-value < 0.05, **P-value < 0.01). The numbers above the asterisks indicate the exact test P-values. (**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 [55] 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. Source Data underlying (**a**, **b**) are provided in a Source Data file.

significantly (Student's paired t-test, P-value < 0.05) increased expression levels in maize compared to teosinte using published data [32]. The results were not significant for 2nd and 3rd level contacts (**Figure S16**A). We found 5/60 genes (Enrichment test P-value < 0.01) were domestication candidate genes as reported previously [58–61]. Two of them were *Zm00001d018036* (a gene associated with cob length, P-value $= 6 \times 10^{-25}$) and *Zm00001d041948* (a gene associated with shank length, P-value $= 5.6 \times 10^{-10}$) [58]. Further investigation of these two candidates using recently published chromatin data [55] to detect enhancer activity [62] identified H3K27ac peaks at both DMR loci (**Figure 3**C and **Figure S17A**). Consistent with these enhancer signals, the expression levels of these two genes is significantly increased in maize relative to teosinte (**Figure S16**B and **Figure S17B**). Despite this functional evidence, however, we found that interacting DMRs in selective sweeps were significantly less often than expected by chance (**Table 1**).

DMRs associated with flowering time variation

Analyses above found that high-frequency regions polymorphic for methylation in our samples accounted for 15% and 17% genetic variances for two flowering time traits, days to anthesis and days to silk, respectively (**Figure S10A**). Upon closer inspection of our DMRs, we found a number of candidate flowering time genes located in sweep DMRs or interacting DMRs (**Table S7**), including three genes found in both (i.e., *Zm00001d029946*, *Zm00001d015884*, *Zm00001d025979*). We also examined several known genes in the flowering time pathway [63] and detected six DMRs located near four additional flowering time related genes (**Figure S18**) (Enrichment test P-value < 0.05).

One DMR was located 40-kb upstream of *ZmRAP2.7*, a well characterized flowering time gene, and 20-kb downstream of the *vgt1* locus, that was hypomethylated in modern maize and landrace but was hypermethylated in teosinte (**Figure 4A**). A MITE transposon insertion in the *vgt1* locus is considered as the causal variant for the down regulation of *ZmRAP2.7*, which encodes a transcription factor in the flowering time pathway [64]. We did not detect *vgt1* as a selective sweep because it is not considered a domestication or improvement candidate and our maize lines include both tropical and temperate lines [65]. We further examined LD in this regions and detected strong signals between the *vgt1*-DMR and local SNPs, suggesting that the *vgt1*-DMR is not a pure epiallele. Reanalysis of published ChIP data [33] revealed that the DMR colocalized with a H3K27ac peak and there is a physical interaction between the DMR and the *vgt1* locus in maize [33] (**Figure 4B**). Additionally, we reanalyzed the maize and sorghum sequence data at the *vgt1* locus and

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

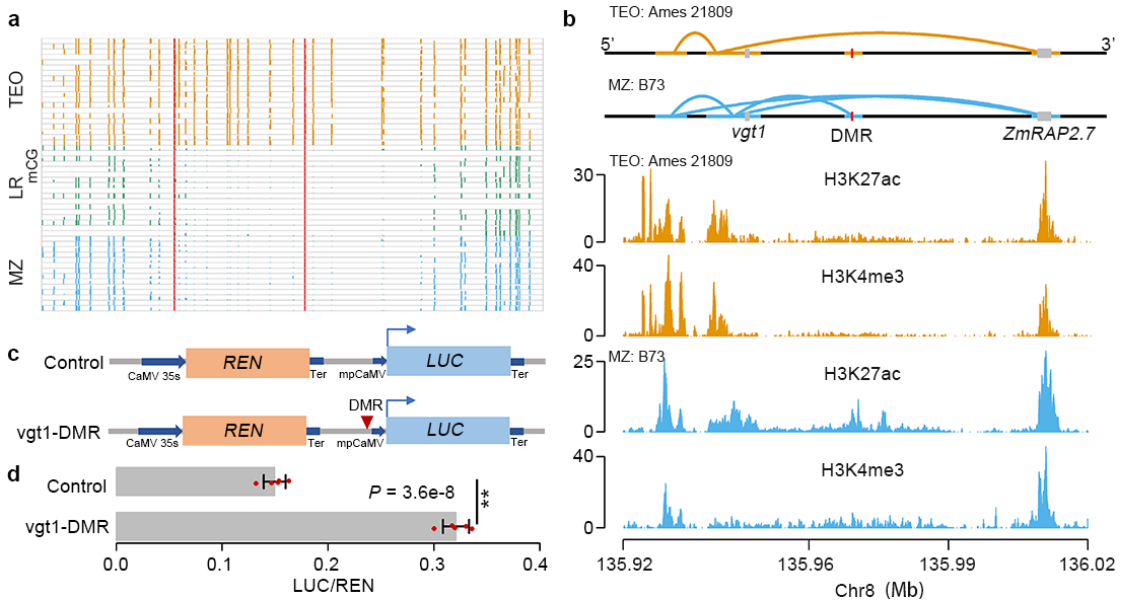

Fig. 4. Functional analysis of *vgt1*-DMR. (**a**) Levels of CG methylation around *vgt1*-DMR in maize (MZ), landrace (LR), and teosinte (TEO) populations. Vertical red lines indicate the boundaries of the *vgt1*-DMR. (**b**) The interactive contacts (upper panel) and colocalization with H3K27ac and H3K4me3 (lower panel) around *vgt1*-DMR in a maize (B73) and a teosinte (Ames 21809) samples. (**c**) The vectors constructed for functional validation of the *vgt1*-DMR using the dual-luciferase transient expression assay in maize protoplasts. (**d**) The expression ratios of LUC/REN using five biological replicates. Error bars indicated standard deviations. Statistical significance was determined by a two-sided *t*-test (**P-value = $3.6e-8$). Source Data underlying (**a**, **d**) are provided in a Source Data file.

found two conserved non-coding sequences (CNSs) located 1kb downstream of the *vgt1*-DMR (**Figure S19**). To examine the interaction status in teosinte, we then generated HiChIP data for a teosinte sample using the same tissue and antibodies (see **methods**). Although our teosinte HiChIP data identified similar peaks of H3K27ac and H3K4me3 near the region, we failed to detect a physical interaction between the *vgt1*-DMR and *vgt1* itself in teosinte (**Figure 4B**), suggesting that methylation at this locus might play a functional role in affecting physical interaction.

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 4C**) and performed a dual-luciferase transient expression assay in maize protoplasts (see **methods**). The results of the transient expression assay revealed that the maize cells harboring the DMR exhibited a significantly higher LUC and REN ratio than control (fold change= 2.2, P-value= $2.4e^{-8}$, **Figure 4D**), revealing that the DMR might act as an enhancer to activate LUC expression.

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

One of the most significant teosinte-maize CG DMRs was located 40-kb upstream of the *tb1* gene, which encodes a transcription factor acting as a repressor of axillary branching (aka tillering) phenotype [50]. This 534-bp *tb1*-DMR was hypomethylated in modern maize, hypermethylated in teosinte, and segregating in landraces (**Figure 5A**). Chop-PCR (DNA methylation-sensitive restriction endonuclease digestion followed by PCR) analysis using a modern maize (inbred line W22) and a teosinte accession (PI 8759) suggested that DNA methylation presents in both leaf and immature ear tissues in teosinte, but is absent in W22 (**Figure S20**). The physical location of the *tb1*-DMR overlapped with the MNase hypersensitive site [66] and a H3K9ac peak [67]. Phenotypic analysis of our 17 landraces indicated that the DMR was associated with the tillering (Fisher's exact test P-value < 0.05), consistent with previous observations that the hypermethylated (teosinte-like) genotypes were likely to grow tillers [50].

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 [50]. Interactome data support this claim, finding physical contact between Hopscotch and the *tb1* gene (**Figure 5B**) [33]. Direct physical contact between the *tb1*-DMR and the *tb1* gene itself in maize line B73 was also detected using ChIA-PET data [33], but this interaction is missing in teosinte based on our newly generated HiChIP data (**Figure 5B**). By employing the 4C-seq method [68], we further confirmed the absence of interaction between the *tb1*-DMR and the *tb1* gene using landrace samples showing hypermethylation at the *tb1*-DMR locus (**Figure S21**). The colocalization of *tb1*-DMR with chromatin activation marks in the region also suggested the *tb1*-DMR might act as a *cis*-acting regulatory element (**Figure 5B**). Additionally, we conducted a dual-luciferase transient assay by constructing a vector similar to the *vgt1*-DMR (**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.

To understand the correlation among these genomic components, i.e., the *tb1*-DMR, the TE insertion, and the *tb1* gene, we conducted linkage disequilibrium (LD) analysis using landrace genomic and methylation data segregating at this *tb1*-DMR locus (see **methods**). As a result, we failed to detect strong LD ($R^2 = 0.1$) between the *tb1*-DMR and SNPs located at the Hopscotch locus (**Figure S22**),

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

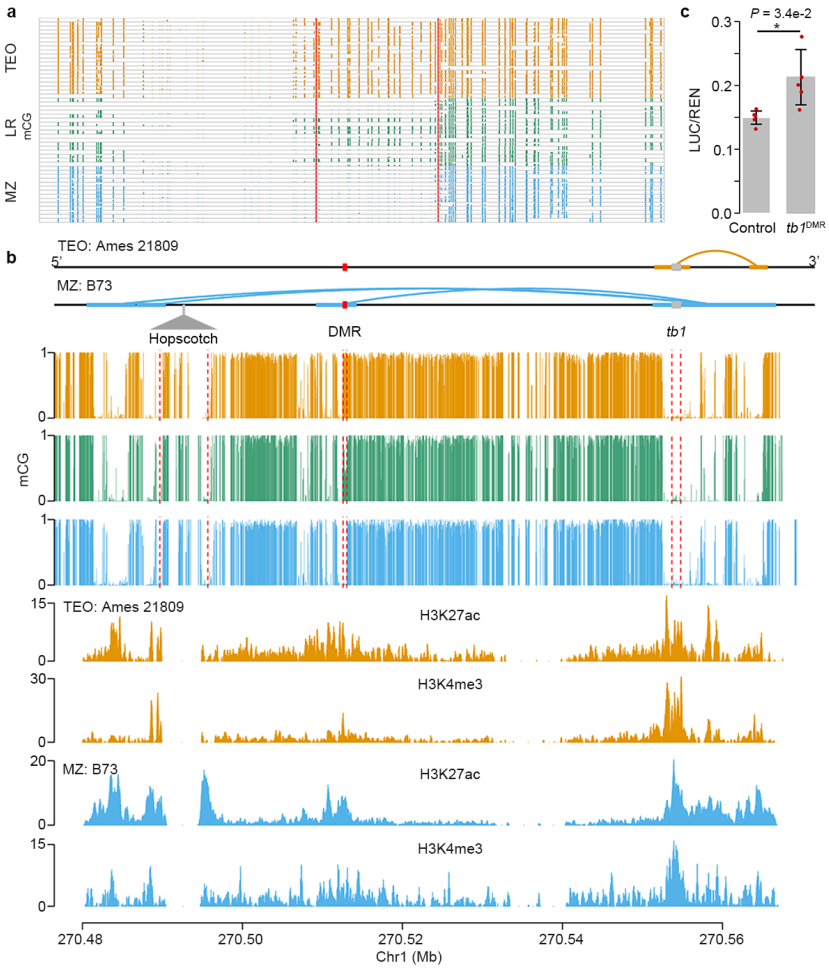

Fig. 5. A hypomethylated DMR that is upstream of *tb1* gene. (**a**) Levels of mCG for the 534-bp *tb1*-DMR in each individual methylome of the modern maize (MZ), landrace (LR), and teosinte (TEO) populations. Vertical red lines indicate the boundaries of the *tb1*-DMR. (**b**) Interactive contacts (upper 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. Five biological replicates were performed. Error bars indicated standard deviations. Statistical significance was determined by a two-sided *t*-test (**P*-value = $3.4e-2$). Source Data underlying (**a**, **c**) are provided in a Source Data file.

indicating the *tb1*-DMR might be independent from the Hopscotch locus. Reanalysis of published *tb1* mapping data [50] confirmed a significant QTL signal around the Hopscotch TE (**Figure S23A**), and a two-dimensional QTL scan detected epistasis between Hopscotch and the *tb1*-DMR (**Figure S23B**). Further, we found that highly methylated landraces were geographically closer to the Balsas River Valley in Mexico, where maize was originally domesticated from (**Figure S24A**). As the landraces spread out from the domestication center, their CG methylation levels were gradually reduced (**Figure S24B**).

DISCUSSION

In this study, we employed population genetics and statistical genomics approaches to infer the rates of epimutation and selection pressure on DNA methylation, and the extent to which SNPs located within DMRs contributed to phenotypic variation. Our results revealed that the forward epimutation rate was about 10 times larger than the backward epimutation rate. These estimates from 100-bp tiles are lower than epimutation rates estimated at nucleotides in *Arabidopsis* from epimutation accumulation experiments [69]. Even so, our estimated epimutation rates are more than an order of magnitude higher than the per-nucleotide mutation rate in maize [70].

Although population methylome modeling suggested that genome-wide DNA methylation was not under strong selection, we nonetheless show that regions harboring polymorphic methylation contribute to functionally relevant phenotypic variation. To prioritize loci likely exhibiting evolutionarily relevant methylation variation, we identified individual differentially methylated regions (DMRs). These DMR were enriched in likely functional sequence, including regulatory regions near genes, putative enhancers, and intergenic regions showing evidence of chromatin interactions. We further identified several dozen new genes that are differentially expressed between maize and teosinte, for which exonic regions directly interact with maize hypo-DMRs. We also found a strong enrichment of DMRs in regions targeted by recent positive selection. Patterns of linkage disequilibrium between DMRs and nearby

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SNPs make it difficult to assign causality, i.e., the DMRs associated with the flowering time traits maybe not the causal variants, but are consistent with the idea that many DMRs are the result of genetic changes, consistent with previous studies [7, 20]. Taken together, these results suggest that methylation might modulate physical interactions and hence likely affect gene expression. This idea fits well with previous results from GWAS that 80% of the explained variation could be attributable to trait-associated variants located in regulatory regions [71]. In total, our DMR results provide a list of candidate genes to be further tested, especially those found in selective sweeps and interacting regions. To tease apart real DMR-phenotype associations from false, future efforts should focus on genotyping the methylation status of such loci across mapping populations while modeling SNP and DMR associations with phenotypes jointly.

In addition to our genome-wide approaches that identify a large number of novel DMRs, we also conducted functional validation at two well-studied candidate loci: *vgt1* and *tb1*. In both cases, our evidence showed that methylation affects physical interactions between the gene and intergenic regulatory regions. In particular, the maize alleles having low methylation levels exhibit interactive loops and increased expression of the downstream gene compared to highly methylated alleles in teosinte.

Collectively, our results suggest a meaningful functional role for methylation variation in maize. Genome-wide variation in methylation shows signs of weak natural selection and regions exhibiting variation explain considerable phenotypic variation. We also identify a large number of DMRs, many of which overlap with signals of selection during maize domestication and improvement as well as regions of the genome important for chromatin interaction. These results suggest that further investigation of the role of methylation in affecting genome-wide patterns of chromatin interaction and gene regulation is warranted, and that naturally occurring DMRs may provide a useful source of regulatory variation for crop improvement.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 [30]. We harvested the third leaf of the teosintes and Mexican landraces at the third leaf stage for DNA extraction using a modified CTAB procedure [72]. 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 [31].

Sequencing data analysis.

The average coverage for the WGS of the 20 teosintes and 17 landraces lines was about 20 $\times$. For these WGS data, we first mapped the cleaned reads to the B73 reference genome (AGPv4) [73] using BWA-mem [74] with default parameters, and kept only uniquely mapped reads. Then we removed the duplicated reads using Picard tools [75]. We conducted SNP calling using Genome Analysis Toolkit's (GATK, version 4.1) HaplotypeCaller [76], 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 [77] 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 [78].

Population epigenetics modeling.

Spontaneous epimutation changes (i.e. gain or loss of cytosine methylation) exhibit higher rate than genomic mutation [21, 69]. 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 [23]. 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 $\alpha = 4N_e\mu$, $\beta = 4N_e\nu$, $\gamma = 2N_e s$. 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 [38]. 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 selected a set of $N_e = 50,000, 100,000, 500,000$, and $1,000,000$ [39, 40, 79, 80]. To test the prior values on the posterior distributions, we sampled μ, ν, and s from exponential proposal distributions with different prior values of 10^2, 10^4, 10^5, 10^8, and 10^{10} (**Figure S8A** and lambda values of the scaled proposal distribution of 0.01, 0.05, and 0.1 (**Figure S8B**). We ran the model using a chain length of $N = 1,000,000$ iterations with the first 20% 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 [81]. 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 [82]. 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 [44]. 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 [83]. We used the significance cutoff at P-value < 0.01.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Linkage disequilibrium (LD) analysis between DMR and local SNPs.

To test the relationship between DMRs and selective sweeps, we conducted LD analysis using SNPs located 1 kb upstream and downstream of each DMR. A DMR was determined as in LD if there are at least three SNPs displayed significant correlations with this DMR (permutation P-value < 0.01).

HiChIP sequencing library construction.

We constructed the teosinte HiChIP library according to the protocol developed by Mumbach *et al*. [84] 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 150 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 [33]. 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 [84]. 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, Cat No. 28006) and measured the DNA concentration using Qubit. To fragment and capture interactive loops, we used the Tn5 transposase kit (Vazyme, TD501) to construct the library. 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 [33] and then aligned the raw reads to B73 reference genome (AGPv4) using Bowtie2 [85]. 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 [86].

For the teosinte HiChIP sequencing data, we first aligned the raw reads to the B73 reference genome (AGPv4) using HiC-Pro [87], and then processed the valid read pairs to call interactive loops using hichipper pipeline [88] 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 [89], we determined the remaining loops with three read pairs supports and the FDR < 0.01 as the significant interactive loops.

4C-seq library construction and data analysis.

To validate the physical interaction between *tb1*-DMR and *tb1* gene, we performed 4C-seq experiments using landrace samples. We constructed the 4C-seq libraries using restriction enzymes of *Nla*III and *Dpn*II. The primer sequences for the *tb1* bait region were the same as previously described [33]. After sequencing, we aligned the reads to the B73 reference genome and then processed the uniquely mapped reads using 4C-ker program [90].

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 [91, 92]. We downloaded the phenotypic data (/iplant/home/glaubitz/RareAlleles/genomeAnnos/VCAP/phenotypes/NAM/familyCorrected), consisting of Best Linear Unbiased Predictors (BLUPs) for different traits ([42]), and imputed genotypic data (/iplant/home/glaubitz/RareAlleles/genomeAnnos/VCAP/genotypes/NAM/namrils_projected_hmp31_MAF02mnCnt2500.hmp.txt.gz) [31] from CyVerse database as described in Panzea (www.panzea.org).

In the analysis, we mapped SNPs to the invariable hypermethylated, invariable hypomethylated, and variable methylated regions. For each SNP set, we calculated an additive kinship matrix using the variance component annotation pipeline implemented in TASSEL5 [93]. 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 [43].

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 [94] 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 planted B73 in the growth chamber and kept the plants in the darkness at the temperature of about 20°C (night) and 25°C (day) to generate etiolated plants. Protoplasts were isolated from the 14-day-old leaves of B73 etiolated seedlings following the protocol [95]. Subsequently, we transformed 15 ug plasmids into the 100 ul isolated protoplasts using polyethylene glycol (PEG) mediated transformation method [95]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Experimental validation of the *tb1*-DMR.

We performed Chop-PCR (DNA methylation-sensitive restriction endonuclease digestion followed by PCR) to validate DNA methylation at *tb1*-DMR locus in different tissues of modern maize inbred line W22 and teosinte 8759. We collected the leaf tissue at the third leaf stage and immature ears of ≈5 cm in length. To evaluate the methylation level of *tb1*-DMR locus, we treated 1 μg purified genomic DNA using the EpiJETTM DNA Methylation Analysis Kit (*Msp*I/*Hpa*II) (Thermo Scientific, K1441) following manufacturer's instructions. The primer sequences for PCR were ACACGCACGAAGGGTTACAG (forward) and CAGTGCTCCCTGGGTCAAA (reverse).

Statistical analyses.

All the statistical tests were performed using R software (V3.6.2, https://www.r-project.org/).

CODE AVAILABILITY

The code used for analyses can be accessed through GitHub (https://github.com/jyanglab/msfs_teo).

ACKNOWLEDGEMENTS

J.Y. is 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. J.R.-I. is supported by NSF grant 1546719 and USDA Hatch project CA-D-PLS-2066-H. 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 Mike May for help developing the MCMC approach used here, the helpful discussion in J.R.-I.'s REHAB, and constructive suggestions from anonymous reviewers.

AUTHOR CONTRIBUTIONS

J.Y. and J.R.-I. designed this work. J.L., Q.L., N.M.S., and D.W. generated the data. H.L. and M.Z. produced the teosinte HiChIP libraries. 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 interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. J. E. Dorweiler, C. C. Carey, K. M. Kubo, J. B. Hollick, J. L. Kermicle, and V. L. Chandler, "mediator of paramutation1 is required for establishment and maintenance of paramutation at multiple maize loci," The Plant Cell **12**, 2101–2118 (2000).
13. 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).
14. 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).
15. 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).
16. I. Hernando-Herraez, R. Garcia-Perez, A. J. Sharp, and T. Marques-Bonet, "Dna methylation: insights into human evolution," PLoS genetics **11** (2015).
17. F. Kader and M. Ghai, "Dna methylation-based variation between human populations," Molecular genetics and genomics **292**, 5–35 (2017).
18. 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).
19. 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).
20. 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).
21. 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).
22. Y. Shahryary, A. Symeonidi, R. R. Hazarika, J. Denkena, T. Mubeen, B. Hofmeister, T. van Gurp, M. Colomé-Tatché, K. Verhoeven, G. Tuskan *et al.*, "Alphabeta: Computational inference of epimutation rates and spectra from high-throughput dna methylation data in plants," bioRxiv p. 862243 (2020).
23. B. Charlesworth and K. Jain, "Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates," Genetics **198**, 1587–1602 (2014).
24. A. Vidalis, D. Živković, R. Wardenaar, D. Roquis, A. Tellier, and F. Johannes, "Methylome evolution in plants," Genome biology **17**, 264 (2016).
25. M. C. Stitzer and J. Ross-Ibarra, "Maize domestication and gene interaction," New phytologist **220**, 395–408 (2018).
26. D. J. Gates, D. Runcie, G. M. Janzen, A. R. Navarro, M. Willcox, K. Sonder, S. Snodgrass, F. Rodríguez-Zapata, R. J. Sawers, and R. Rellán-Álvarez, "Single-gene resolution of locally adaptive genetic variation in mexican maize," bioRxiv p. 706739 (2019).
27. 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).
28. 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).
29. J. Teixeira, T. Weldekidan, N. De Leon, S. Flint-Garcia, J. Holland, N. Lauter, S. Murray, W. Xu, D. Hessel, A. Kleintop *et al.*, "Hallauer's tusón: a decade of selection for tropical-to-temperate phenological adaptation in maize," Heredity **114**, 229–240 (2015).
30. 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).
31. 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).
32. 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).
33. 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).
34. 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).
35. D. J. Bertioli, S. B. Cannon, L. Froenicke, G. Huang, A. D. Farmer, E. K. Cannon, X. Liu, D. Gao, J. Clevenger, S. Dash *et al.*, "The genome sequences of arachis duranensis and arachis ipaensis, the diploid ancestors of cultivated peanut," Nature genetics **48**, 438–446 (2016).
36. 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).
37. 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).
38. T. Benaglia, D. Chauveau, D. Hunter, and D. Young, "mixtools: An r package for analyzing finite mixture models," (2009).
39. J. Ross-Ibarra, M. Tenaillon, and B. S. Gaut, "Historical divergence and gene flow in the genus zea," Genetics **181**, 1399–1413 (2009).
40. T. M. Beissinger, L. Wang, K. Crosby, A. Durvasula, M. B. Hufford, and J. Ross-Ibarra, "Recent demography drives changes in linked selection across the maize genome," Nature plants **2**, 1–7 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

41. M. W. Hahn, *Molecular population genetics* (Sinauer Associates New York, 2019).
42. 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).
43. 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).
44. 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).
45. Y. Sun, L. Dong, Y. Zhang, D. Lin, W. Xu, C. Ke, L. Han, L. Deng, G. Li, D. Jackson *et al.*, "3d genome architecture coordinates trans and cis regulation of differentially expressed ear and tassel genes in maize," Genome biology **21**, 1–25 (2020).
46. M. Zhang, H. Zhao, S. Xie, J. Chen, Y. Xu, K. Wang, H. Zhao, H. Guan, X. Hu, Y. Jiao *et al.*, "Extensive, clustered parental imprinting of protein-coding and noncoding rnas in developing maize endosperm," Proceedings of the National Academy of Sciences **108**, 20042–20047 (2011).
47. A. Zemach, M. Y. Kim, P. Silva, J. A. Rodrigues, B. Dotson, M. D. Brooks, and D. Zilberman, "Local dna hypomethylation activates genes in rice endosperm," Proceedings of the National Academy of Sciences **107**, 18729–18734 (2010).
48. 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).
49. 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).
50. 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).
51. 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).
52. 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).
53. B. Sigmon and E. Vollbrecht, "Evidence of selection at the ramosa1 locus during maize domestication," Molecular Ecology **19**, 1296–1311 (2010).
54. 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).
55. 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).
56. 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).
57. 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).
58. S. Xue, P. J. Bradbury, T. Casstevens, and J. B. Holland, "Genetic architecture of domestication-related traits in maize," Genetics **204**, 99–113 (2016).
59. 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).
60. 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).
61. 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).
62. 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).
63. 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).
64. 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).
65. M. B. Hufford, X. Xu, J. Van Heerwaarden, T. Pyhäjärvi, J.-M. Chia, R. A. Cartwright, R. J. Elshire, J. C. Glaubitz, K. E. Guill, S. M. Kaeppler *et al.*, "Comparative population genomics of maize domestication and improvement," Nature genetics **44**, 808 (2012).
66. E. Rodgers-Melnick, D. L. Vera, H. W. Bass, and E. S. Buckler, "Open chromatin reveals the functional maize genome," Proceedings of the National Academy of Sciences **113**, E3177–E3184 (2016).
67. R. Oka, J. Zicola, B. Weber, S. N. Anderson, C. Hodgman, J. I. Gent, J.-J. Wesselink, N. M. Springer, H. C. Hoefsloot, F. Turck *et al.*, "Genome-wide mapping of transcriptional enhancer candidates using dna and chromatin features in maize," Genome biology **18**, 137 (2017).
68. E. Splinter, E. de Wit, H. J. van de Werken, P. Klous, and W. de Laat, "Determining long-range chromatin interactions for selected genomic sites using 4c-seq technology: from fixation to computation," Methods **58**, 221–230 (2012).
69. 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).
70. 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).
71. 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).
72. M. Murray and W. F. Thompson, "Rapid isolation of high molecular weight plant dna," Nucleic acids research **8**, 4321–4326 (1980).
73. 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).
74. H. Li, "Aligning sequence reads, clone sequences and assembly contigs with bwa-mem," arXiv preprint arXiv:1303.3997 (2013).
75. "Picard toolkit," http://broadinstitute.github.io/picard/ (2019).
76. 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).
77. B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short dna sequences to the human genome," Genome

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

biology **10**, R25 (2009).
78. 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).
79. 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).
80. 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).
81. H. Chen, N. Patterson, and D. Reich, "Population differentiation as a test for selective sweeps," Genome research **20**, 393–402 (2010).
82. 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).
83. 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).
84. 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).
85. B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," Nature methods **9**, 357 (2012).
86. 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).
87. 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).
88. C. A. Lareau and M. J. Aryee, "hichipper: a preprocessing pipeline for calling dna loops from hichip data," Nature methods **15**, 155 (2018).
89. 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).
90. R. Raviram, P. P. Rocha, C. L. Müller, E. R. Miraldi, S. Badri, Y. Fu, E. Swanzey, C. Proudhon, V. Snetkova, R. Bonneau *et al.*, "4c-ker: a method to reproducibly identify genome-wide interactions captured by 4c-seq experiments," PLoS computational biology **12**, e1004780 (2016).
91. 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).
92. 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).
93. 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).
94. 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).
95. 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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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:https://github.com/jyanglab/msfs_teo/blob/master/table/Table_S3_Selective_sweep.xlsx)

Table S4. Linkage disequilibrium (LD) analysis between DMR and local SNP. (https://github.com/jyanglab/msfs_teo/blob/master/table/Table_S4_DMR_LD.xlsx)

Table S5. The list of genes with CG teosinte-maize DMRs located at their exonic regions. (https://github.com/jyanglab/msfs_teo/blob/master/table/Table_S5_229_Exonic_DMR_Genes.xlsx)

Table S6. 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_S6_60_Genes_interactive_with_intergenic_hypo_DMR.xlsx)

Table S7. Flowering time candidate genes located at sweep and interacting DMRs. (https://github.com/jyanglab/msfs_teo/blob/master/table/Table_S7_FT_candidate_gene_at_sweep_interaction_DMR.xlsx)

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SUPPORTING FIGURES

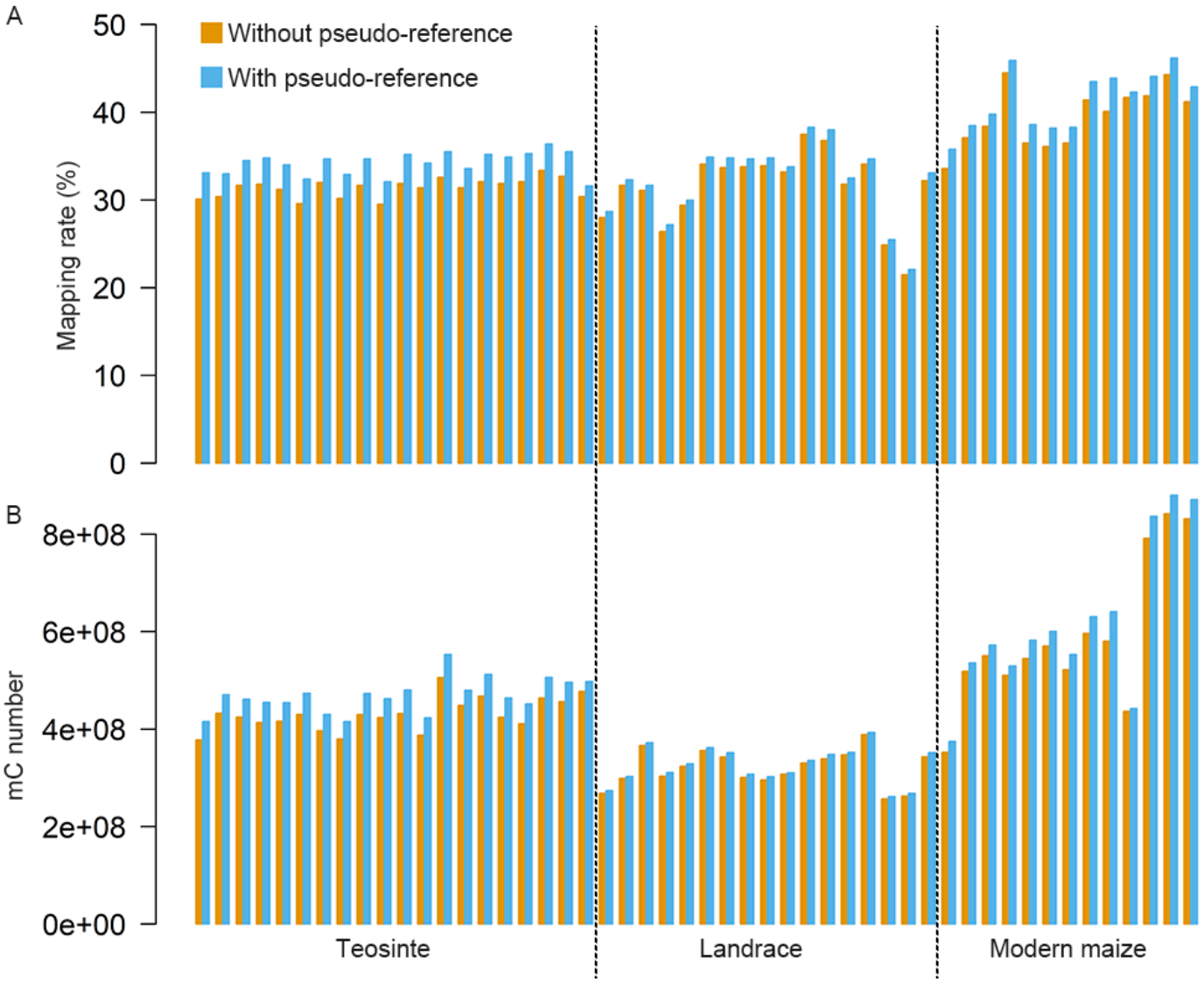

Fig. S1. Comparison of mapping rates (A) and number of methylated cytosine (mC) sites (B) with and without using pseudo-reference genome in different populations. B73 reference genome (AGPv4) was used in the analyses.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

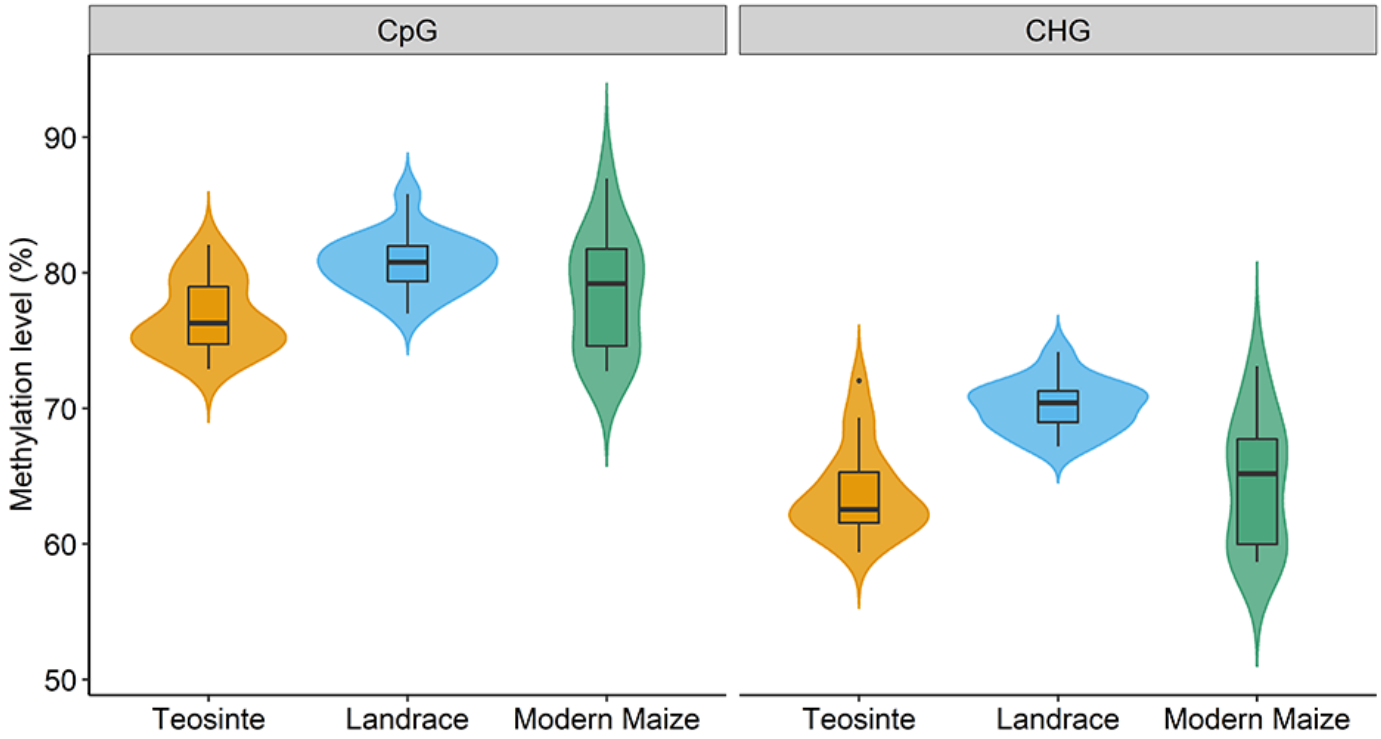

Fig. S2. Distributions of levels of DNA methylation in teosinte, landrace, and modern maize populations. Left panel denotes results from CG sites and right panel denotes results from CHG sites.

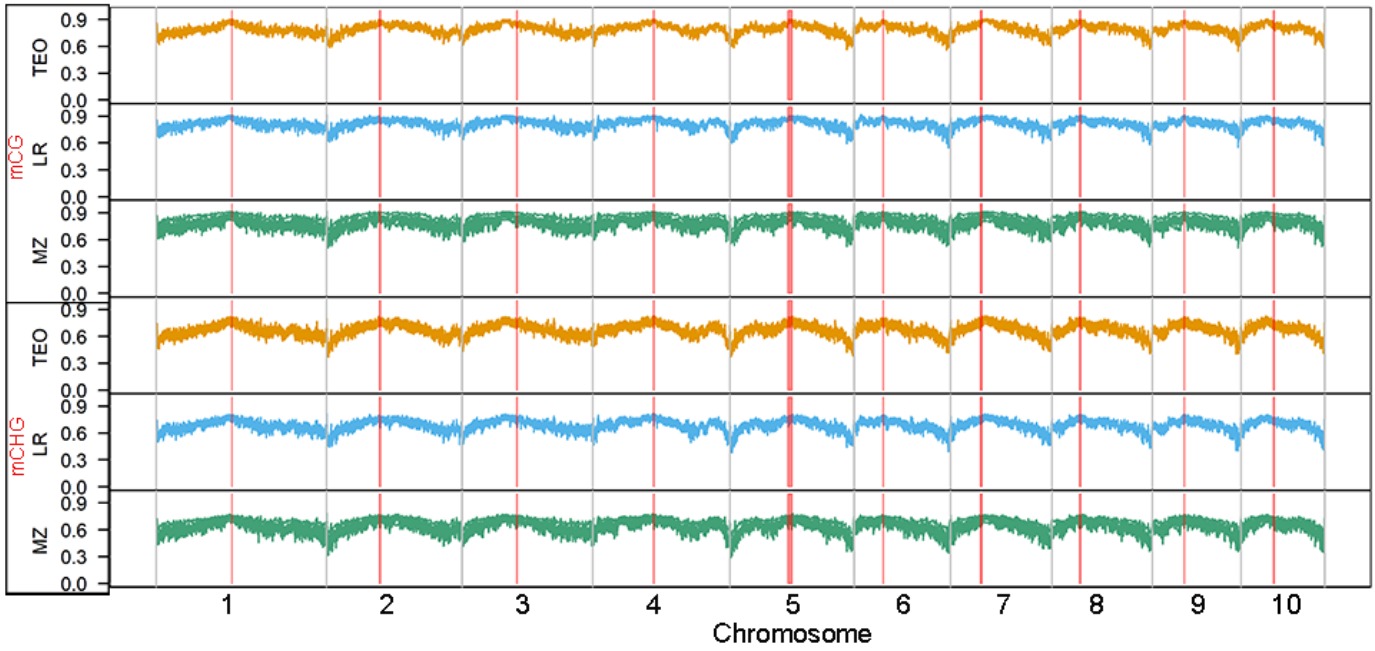

Fig. S3. Genome-wide distributions of DNA methylation across 10 maize chromosomes. TEO, LR, and MZ represent teosinte, landrace and modern maize populations, respectively. Red vertical lines indicated the pericentromeric regions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

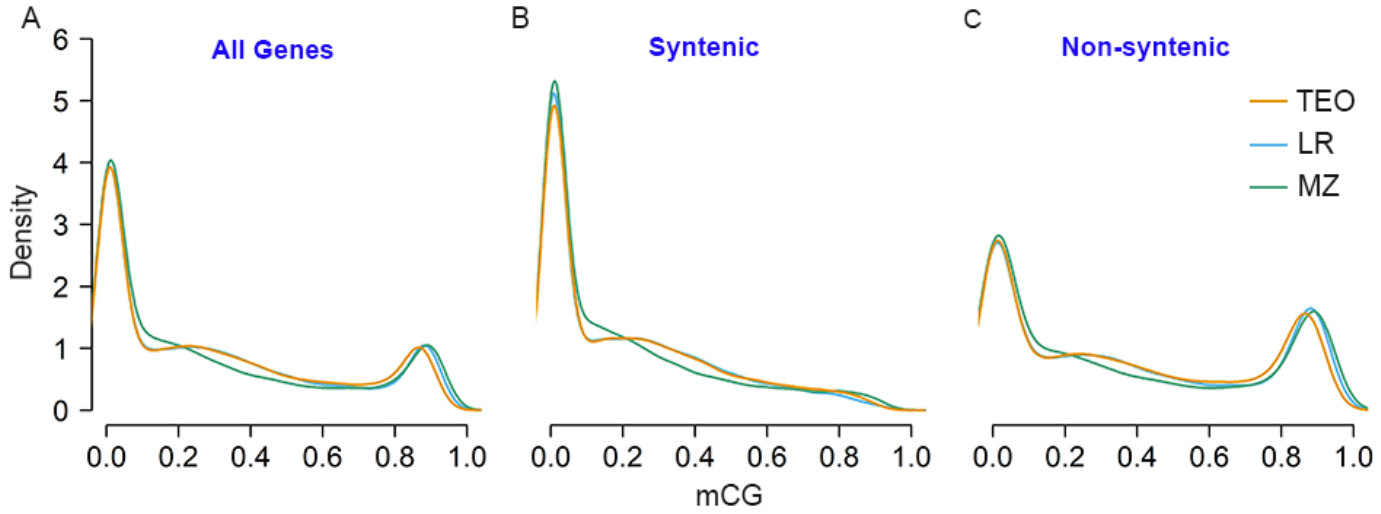

Fig. S4. Density plots of CG methylation in gene body for all the annotated maize genes (A) and for syntenic (B) and non-syntenic genes (C). TEO, LR, and MZ represent teosinte, landrace and modern maize populations, respectively. The syntenic and nonsyntenic orthologs was determined by comparing maize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

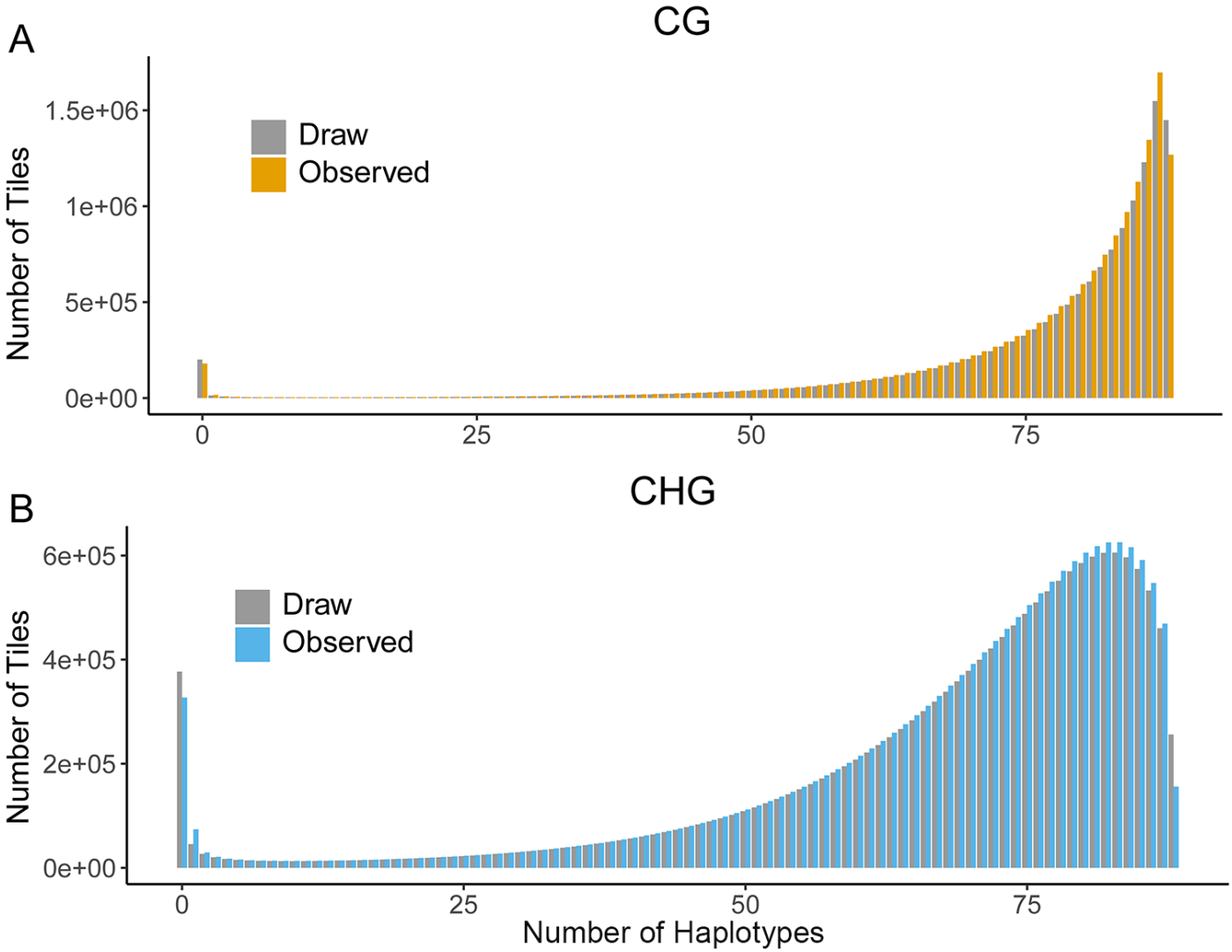

Fig. S5. Observed and posterior methylome site frequency spectra. The posterior mSFS was calculated using parameters drawn from the 1,000,000th iteration (Ne=50,000).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

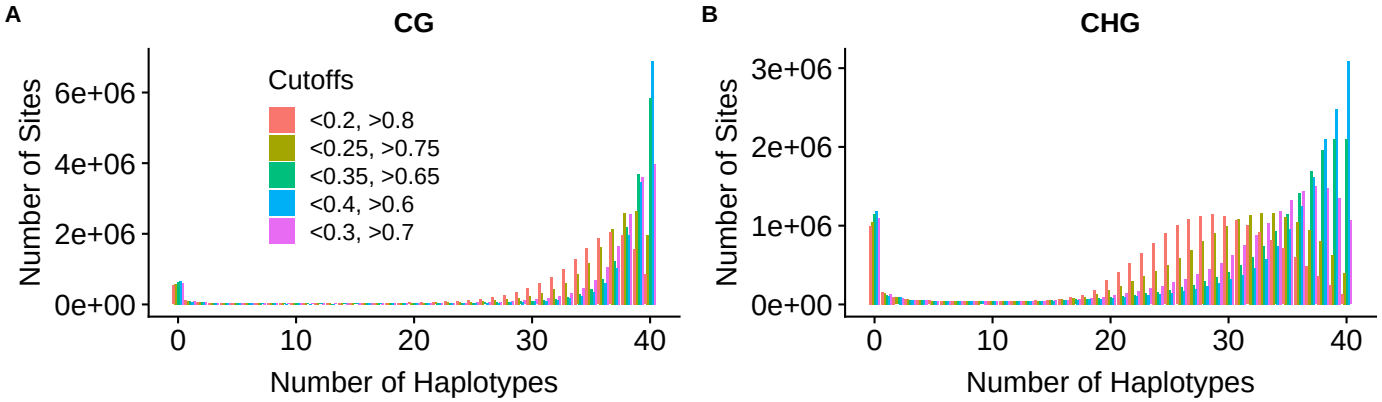

Fig. S6. Sensitivity tests using different cutoffs. Distributions of mSFS using different thresholds to determine the methylated, unmethylated, and heterozygote for the 100-bp tiles under CG (**A**) and CHG (**B**) contexts.

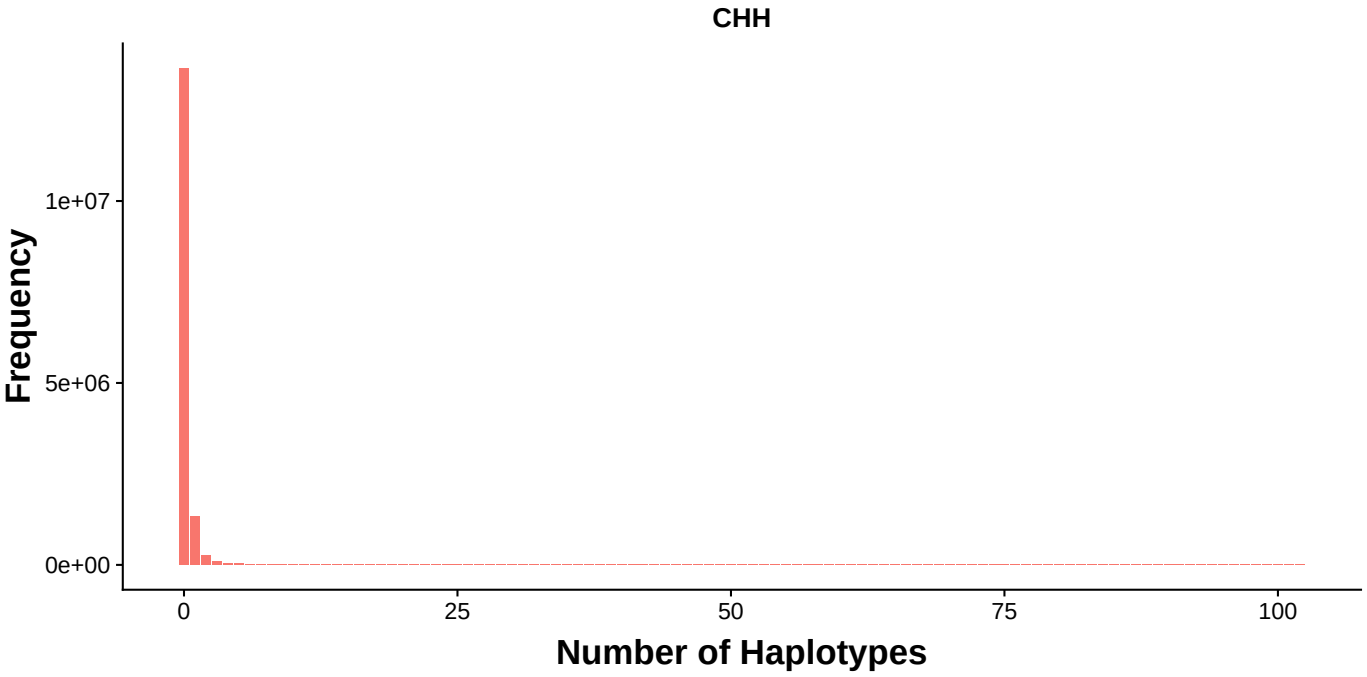

Fig. S7. Methylome site frequency spectrum under the CHH context. The distribution is highly skewed towards the unmethylated status for CHH sites.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

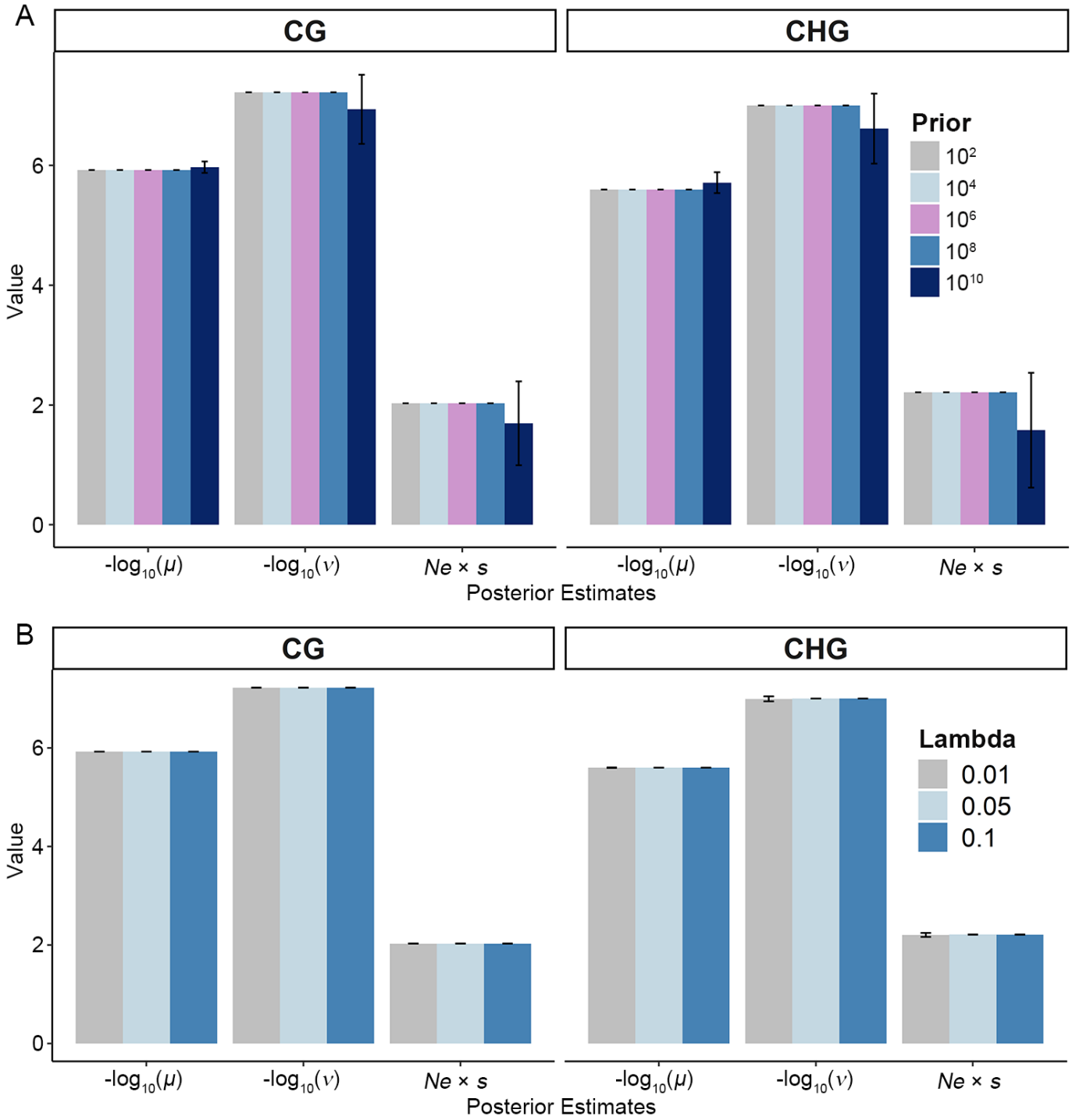

Fig. S8. The effect of prior values on posterior parameter estimations. (**A**) Prior values of 10^2, 10^4, 10^5, 10^8, and 10^{10} were used for the exponential proposal distributions. (**B**) Lambda values of the scaled proposal distribution of 0.01, 0.05, and 0.1 were used. Error bars indicate standard deviations. Data available at (N = 1,600 for each bar).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

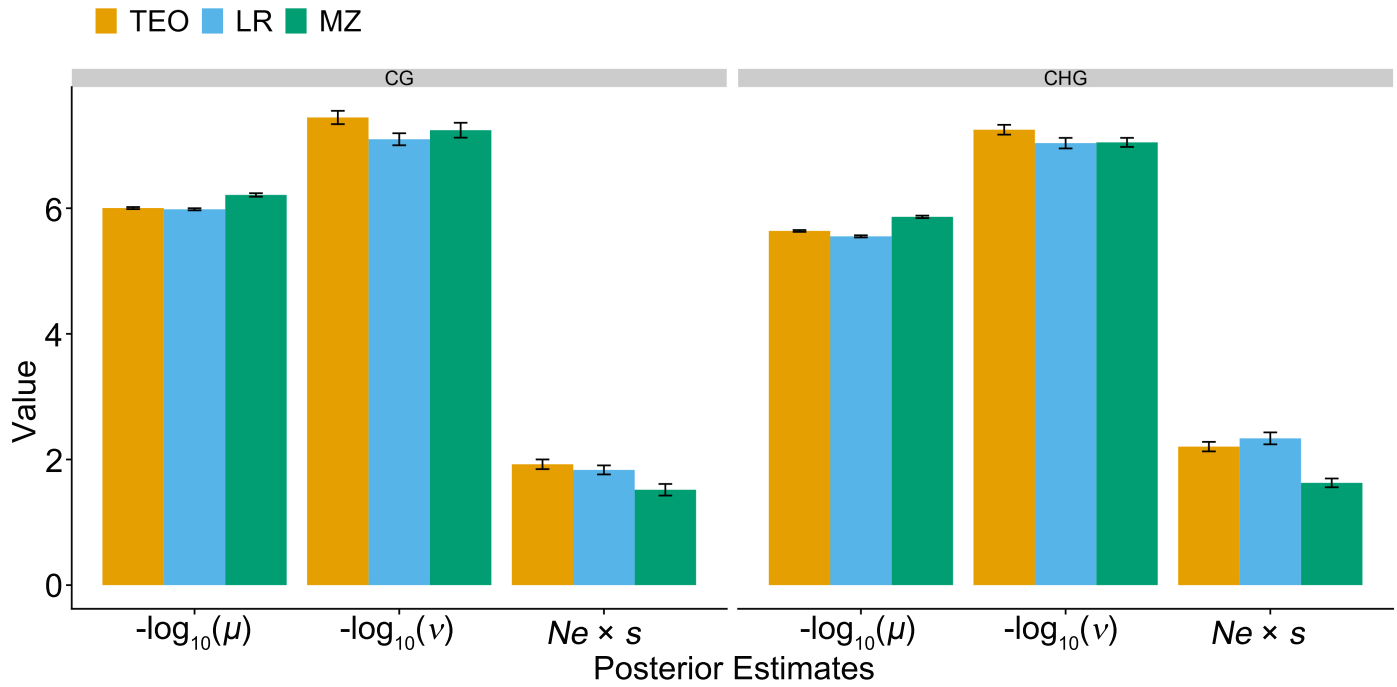

Fig. S9. Population genetic parameter inference using each individual population. Posterior estimators of mean values and standard deviations for μ, ν, and Ne $\times s$ for CG and CHG sites. Values were estimated using MCMC approach with 25% burnin. Error bars indicate standard deviations. Data available at (N = 1,600 for each bar).

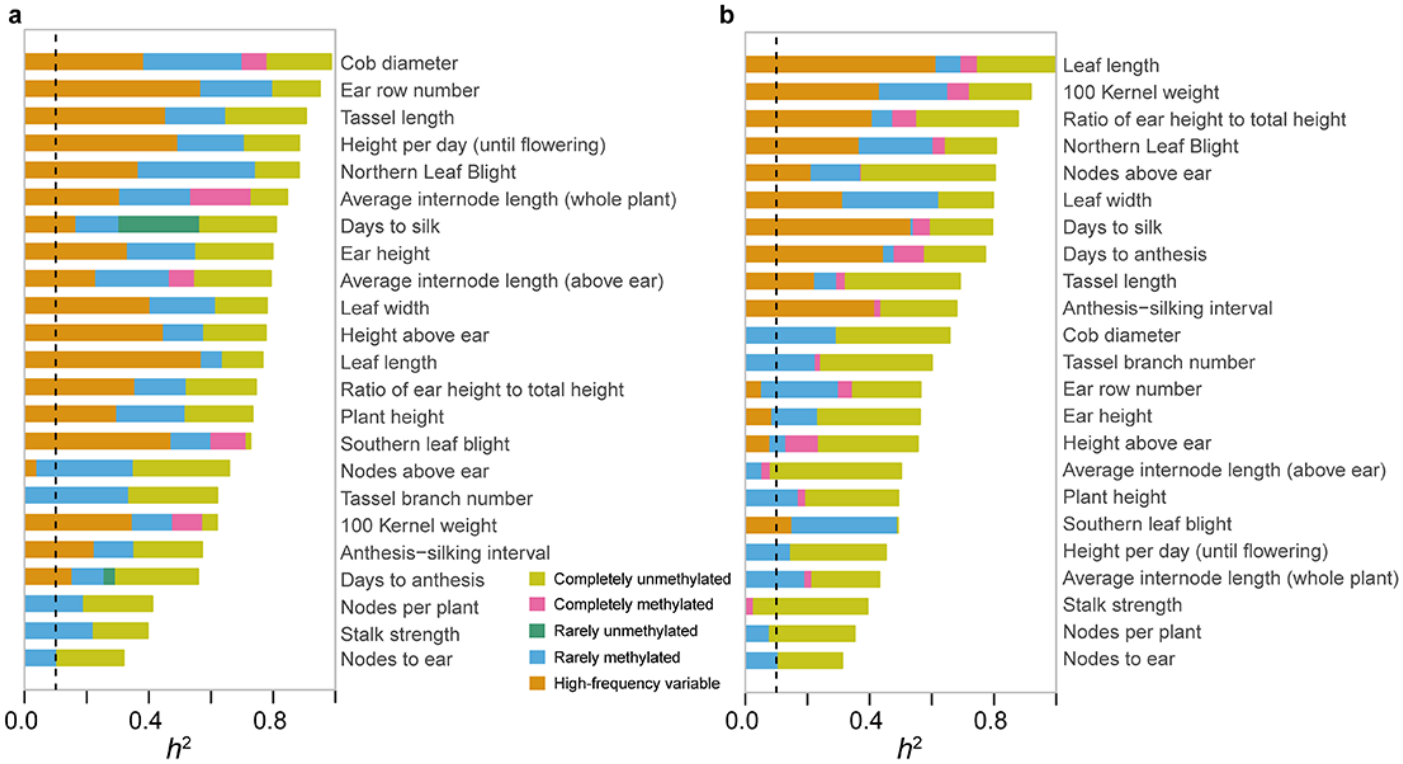

Fig. S10. Proportion of genetic variances explained by SNP subsets residing in different genomic regions with different DNA methylation status. The proportion of genetic variance explained (h^2) by different SNP subsets under CG (**A**) and CHG (**B**) contexts.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

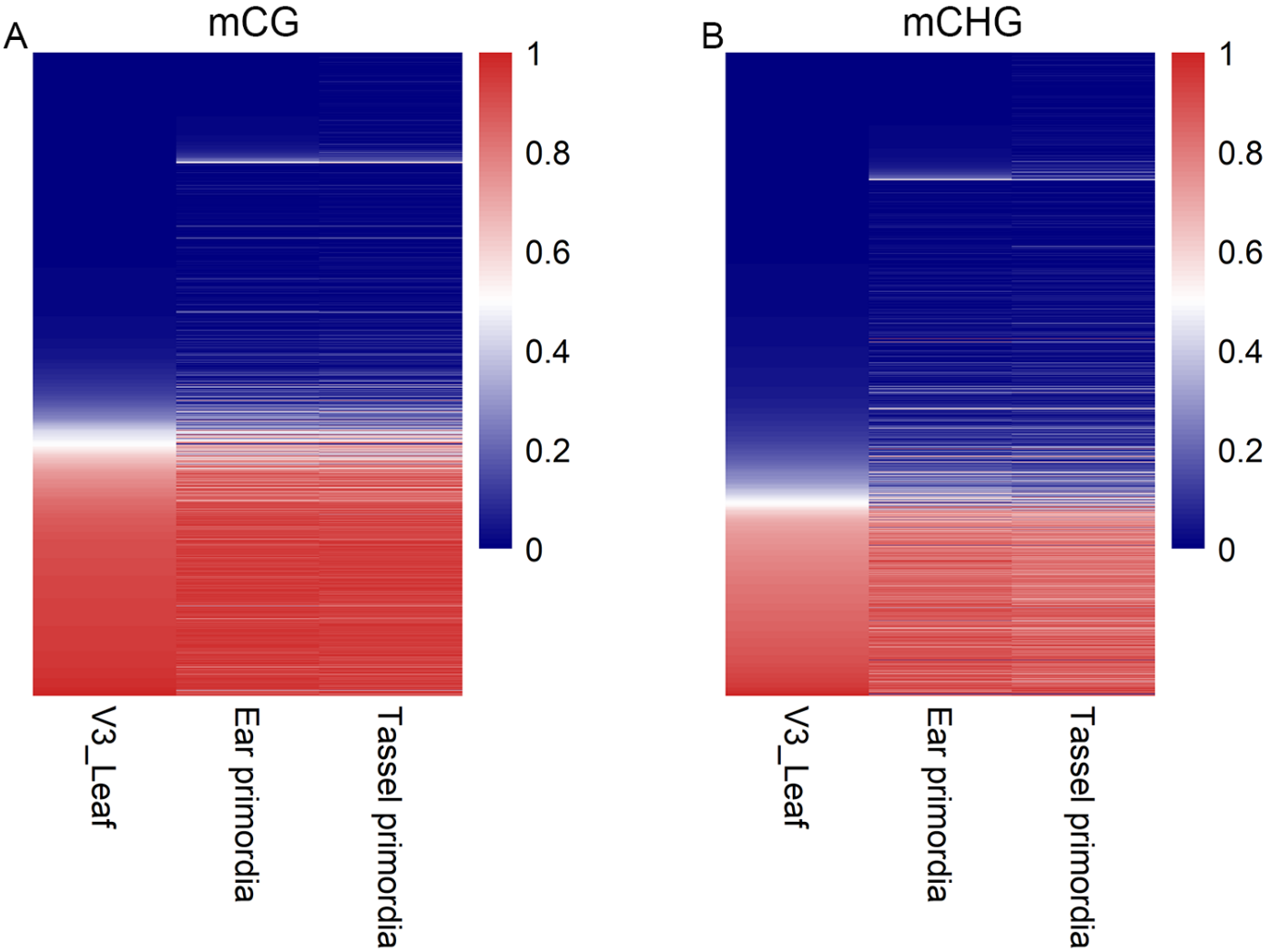

Fig. S11. The methylation patterns of the DMRs across different tissues. DNA methylation levels in CG (**A**) and CHG (**B**) of each DMR across three tissues. The colors in the heatmap indicate the high (red) or low (blue) DNA methylation levels.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

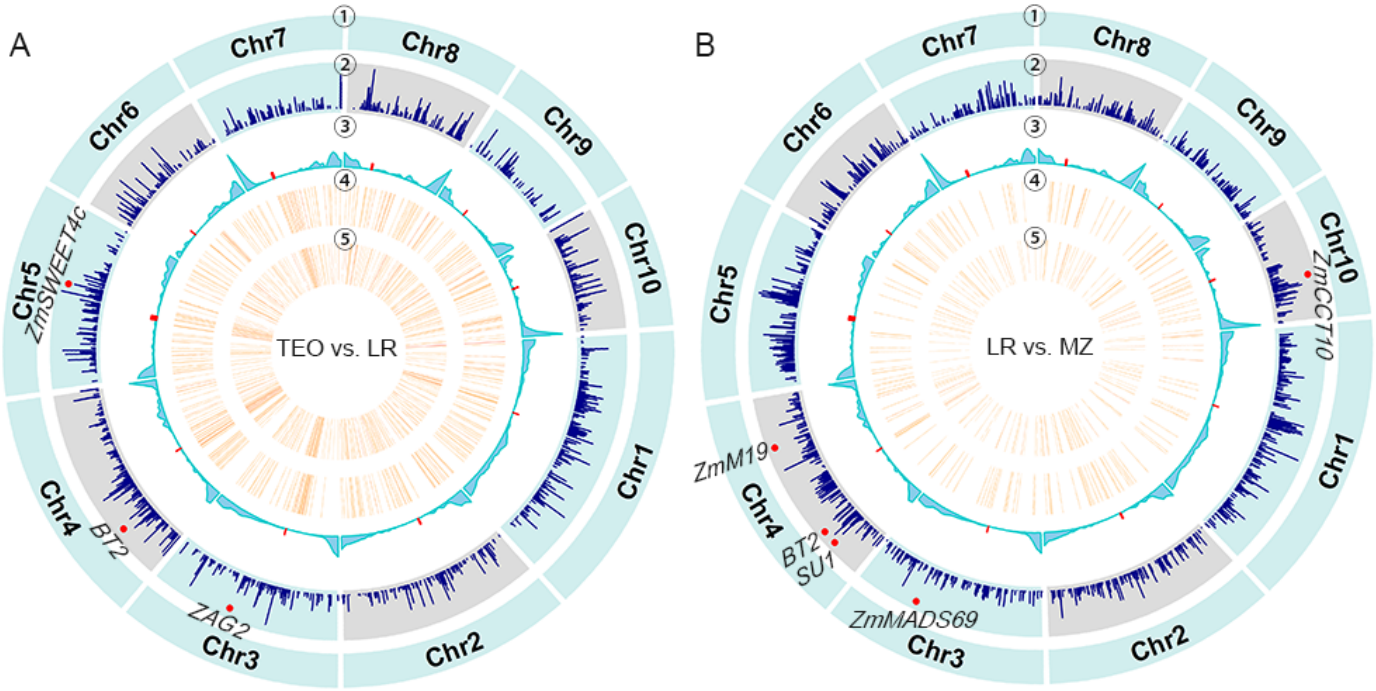

Fig. S12. Landscape of selection signals and DNA methylation variation across maize genome. Genome-wide distributions of selective sweeps, DMRs across ten maize chromosomes detected by teosinte vs. landrace (**A**) and landrace vs. modern maize (**B**). From outer to inner circles: ① chromosome names, 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

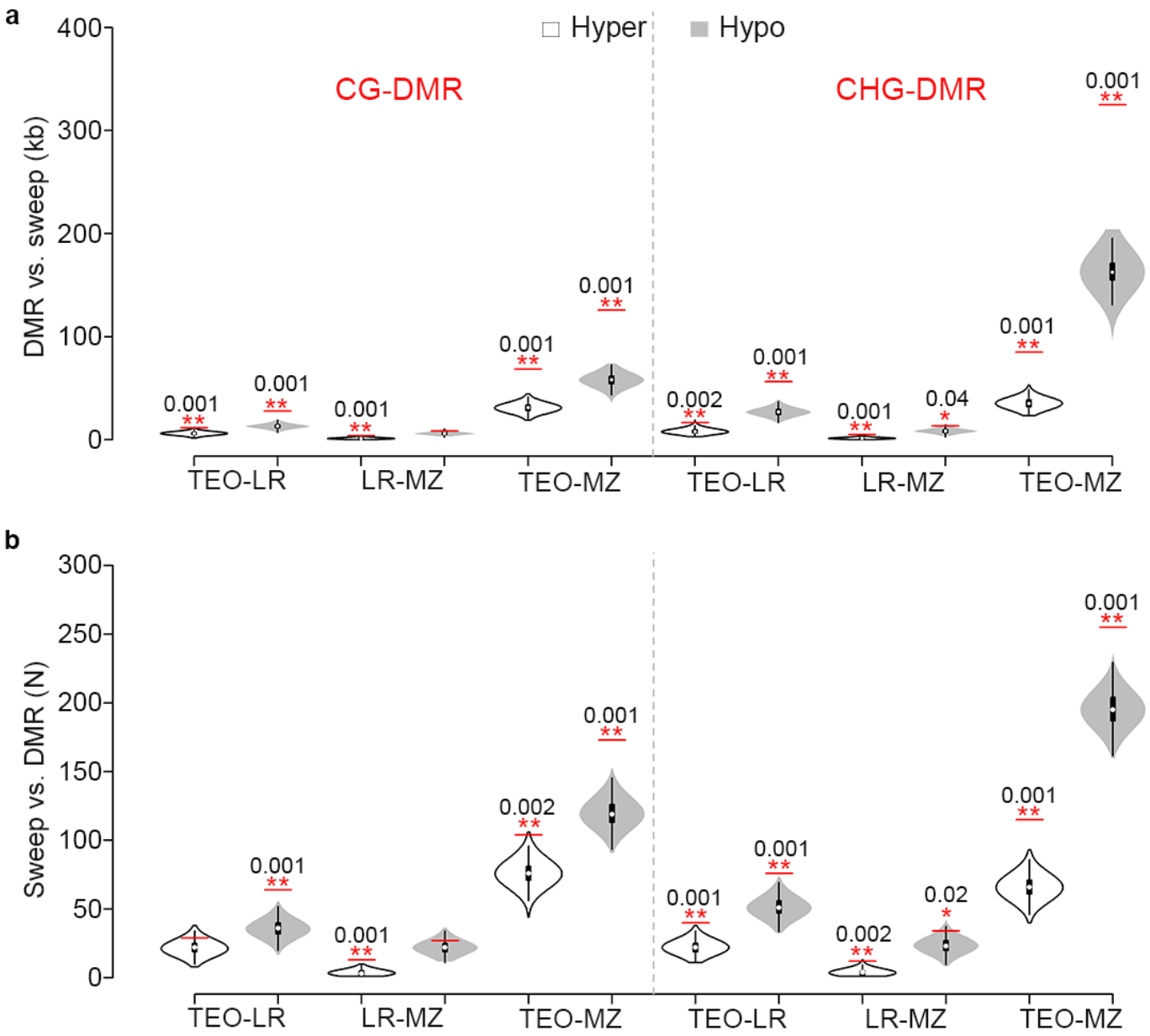

Fig. S13. Comparison between DMRs and selective sweeps. (**A**) The overlapped base-pairs between DMRs and selective sweeps. (**B**) The number of sweeps that overlapped with DMRs. Red horizontal bars indicated the observed values and violin plots showed the 1,000 one-sided permutation results using randomly selected mappable regions from the genome. Red asterisks indicated the statistical significance with one asterisk denoting *P*-value < 0.05 and two asterisks denoting *P*-value < 0.01. The numbers above the asterisks indicate the exact *P*-values. Hyper- and hypomethylation were defined based on maize.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

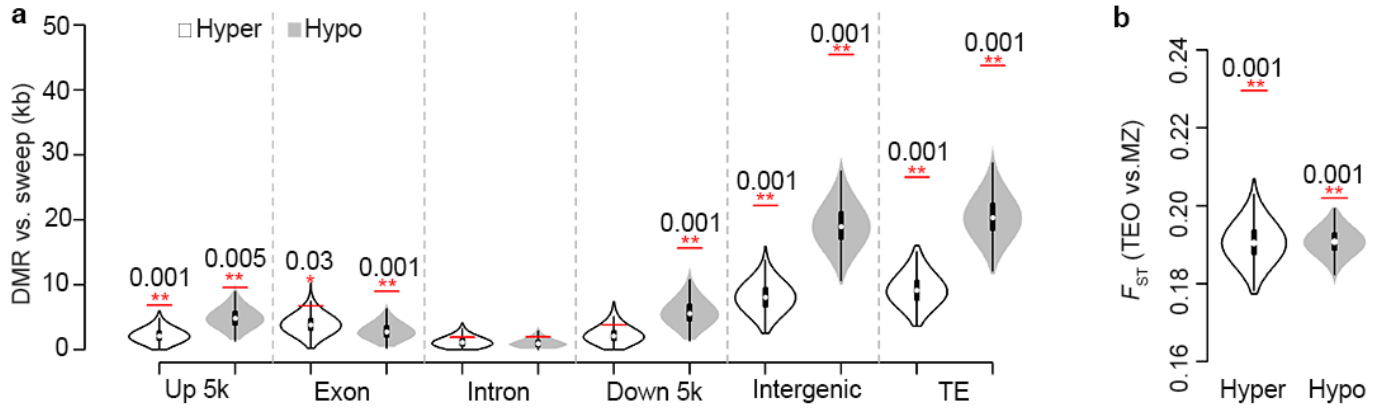

Fig. S14. Selection on differentially methylated regions. (**A**) Overlaps between teosinte-maize DMRs and selective sweeps breaking down into different genomic features. (**B**) Mean F_{ST} values of teosinte-maize DMRs that were hyper- and hypomethylated in maize. Red horizontal bars indicated the observed values and violin plots showed the 1,000 one-sided permutation results using randomly selected regions sharing the similar genomic features as the DMRs. Red asterisks indicated the statistical significance with one asterisk denoting P-value < 0.05 and two asterisks denoting P-value < 0.01. The numbers above the asterisks indicate the exact P-values.

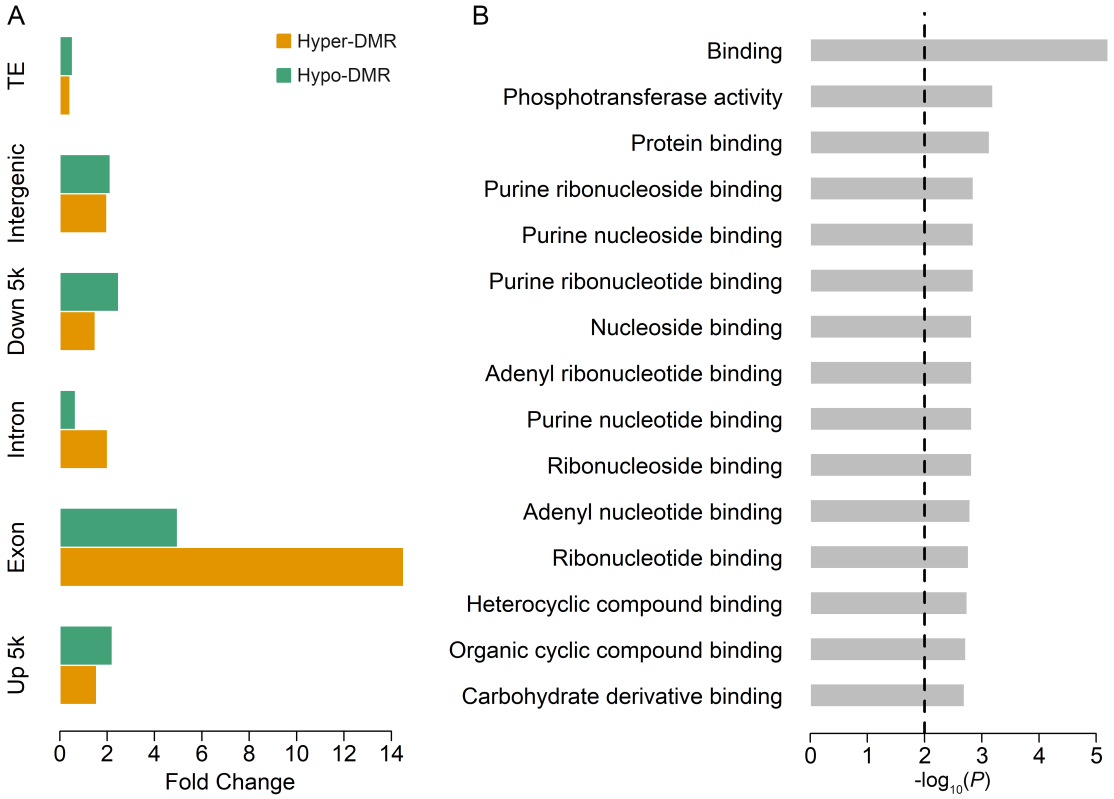

Fig. S15. Teosinte-maize CG DMRs and their associated functional features. (**A**) Fold changes of mappable DMR length relative to the mean values from 1,000 permutations. (**B**) Result of GO term enrichment test using genes exhibiting an exonic DMR. Vertical dashed line indicates the significance cutoff (Fisher's exact test, P-value=0.01).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

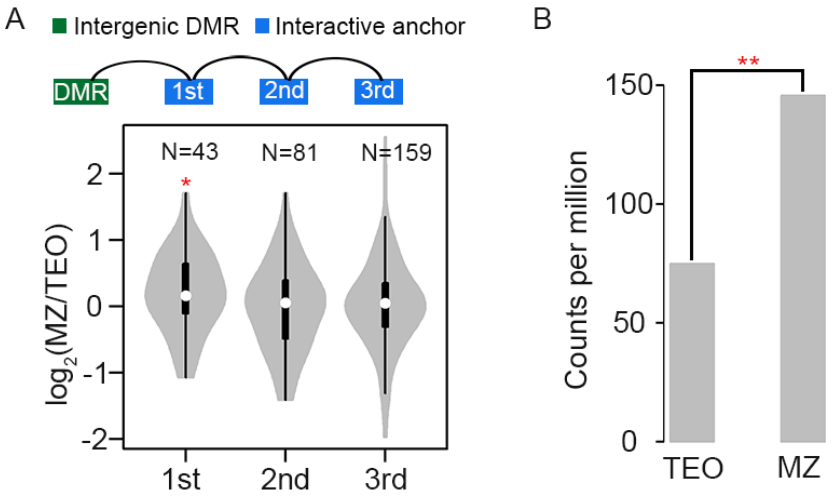

Fig. S16. Intergenic CG DMR altered downstream gene expression. (**A**) Contrast of the gene expression levels in maize relative to teosinte. In the upper panel, the schematic diagram shows the genes that involved in the 1st, 2nd, and 3rd level interactions with maize hypomethylated DMRs located in intergenic regions. Red asterisk indicates the statistically significance (two-sided paired *t*-test: *P*-value < 0.05). (**B**) Gene expression level of *Zm00001d018036* in teosinte and modern maize (Binomial test, *P*-value $= 4.6e^{-141}$).

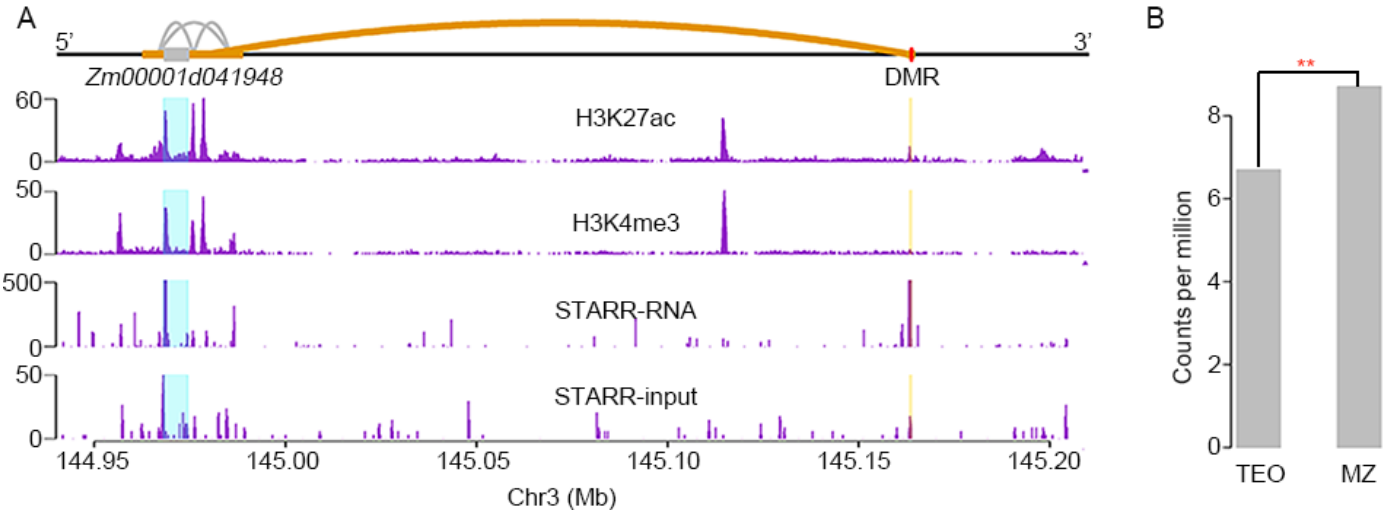

Fig. S17. Interactive loops between a DMR and a gene model *Zm0001d041948*. (**A**) Chromatin interactions (the upper panel) and ChIP-Seq profiles (the lower panels) at gene *Zm00001d041948*. Gray and red boxes indicated the physical position of the gene model and the DMR. Gray and blue lines denoted the interactive loops. (**B**) Gene expression level of *Zm00001d041948* in teosinte and modern maize (Binomial test, *P*-value $= 4.3e^{-3}$).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

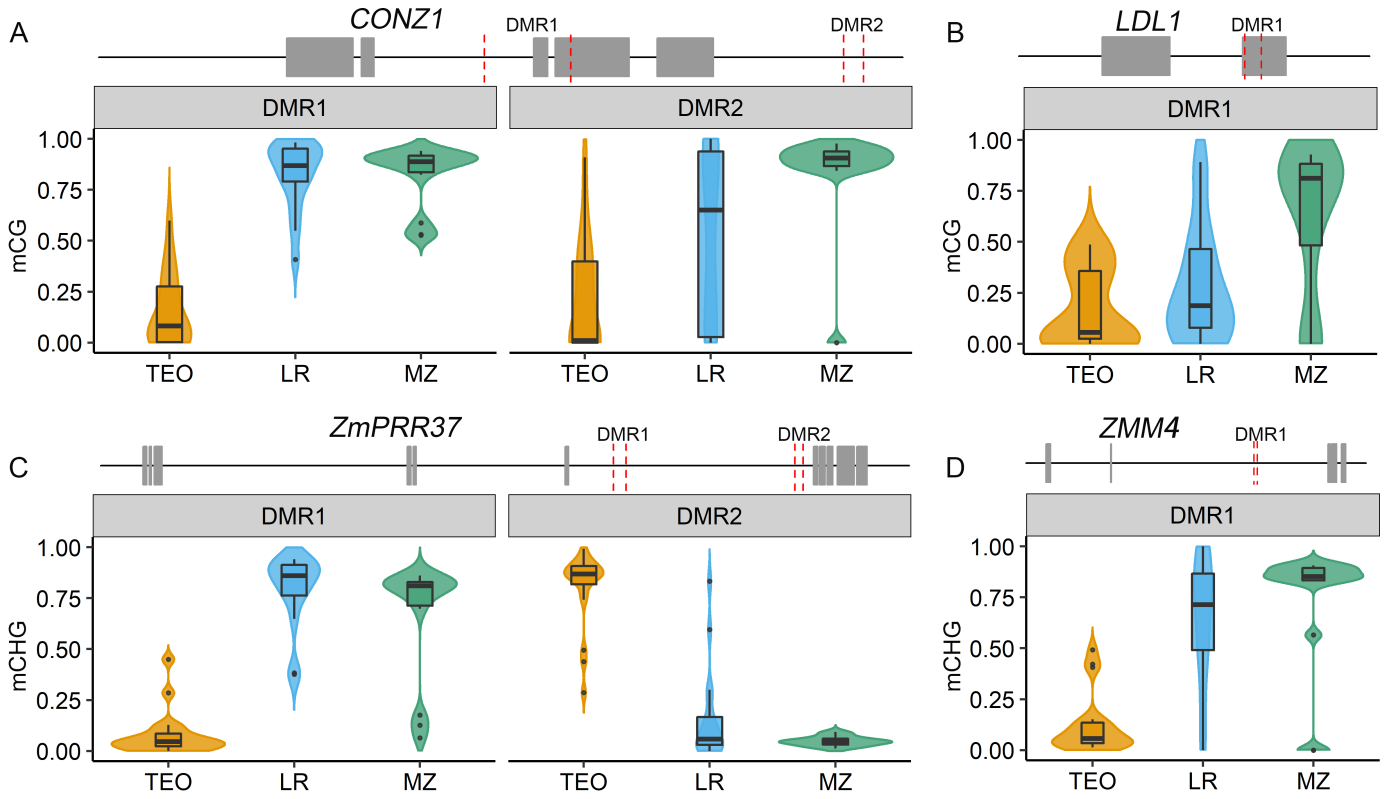

Fig. S18. Teosinte-maize DMRs located at flowering time genes. Distribution of methylation level within DMRs that located at *CONZ1* (**A**), *LDL1* (**B**), *ZmPRR37* (**C**) and *ZMM4* (**D**) in teosinte (TEO), landrace (LR), and modern maize (MZ). Two nearby vertical dashed red lines on the gene model indicated a teosinte-maize DMR.

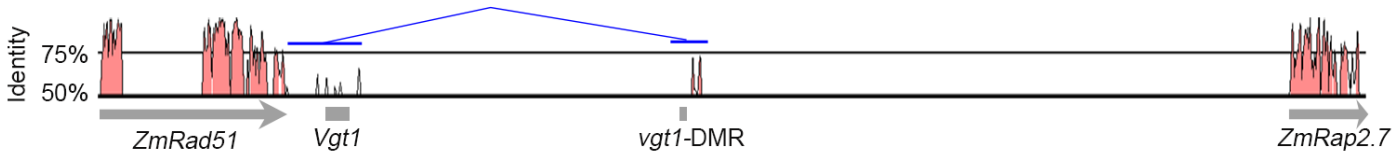

Fig. S19. CNSs between the maize and sorghum orthologous around *vgt1*-DMR. Sequence identity of the maize sequence spanning *Vgt1* and the two proximal genes with corresponding sorghum sequences. Red peaks denote CNSs identified using a window size of 100 bp. The thin blue lines indicate physical interaction between two anchor sequences (thick blue lines).

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

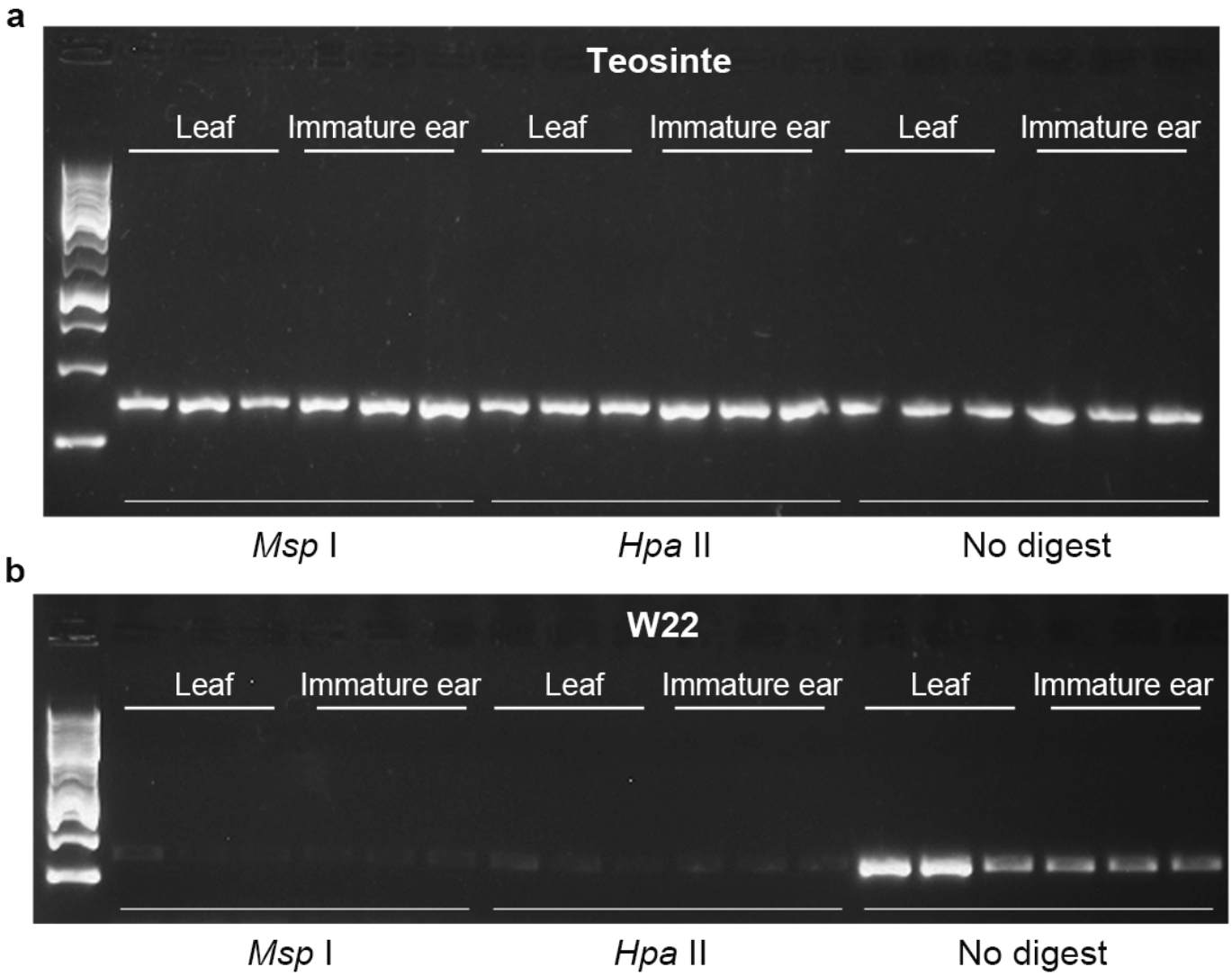

Fig. S20. Experimental validation of the *tb1*-DMR using Chop-PCR. Chop–PCR analysis of *tb1*-DMR in different tissues of W22 (**A**) and teosinte 8759 (**B**). Failure to detect a PCR product reflects the loss of DNA methylation. CG methylation was detected using HpaII and CHG methylation was detected using MspI. Three independent biological replicates are shown, each with three technical replicates. No digested DNA is used as a control.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

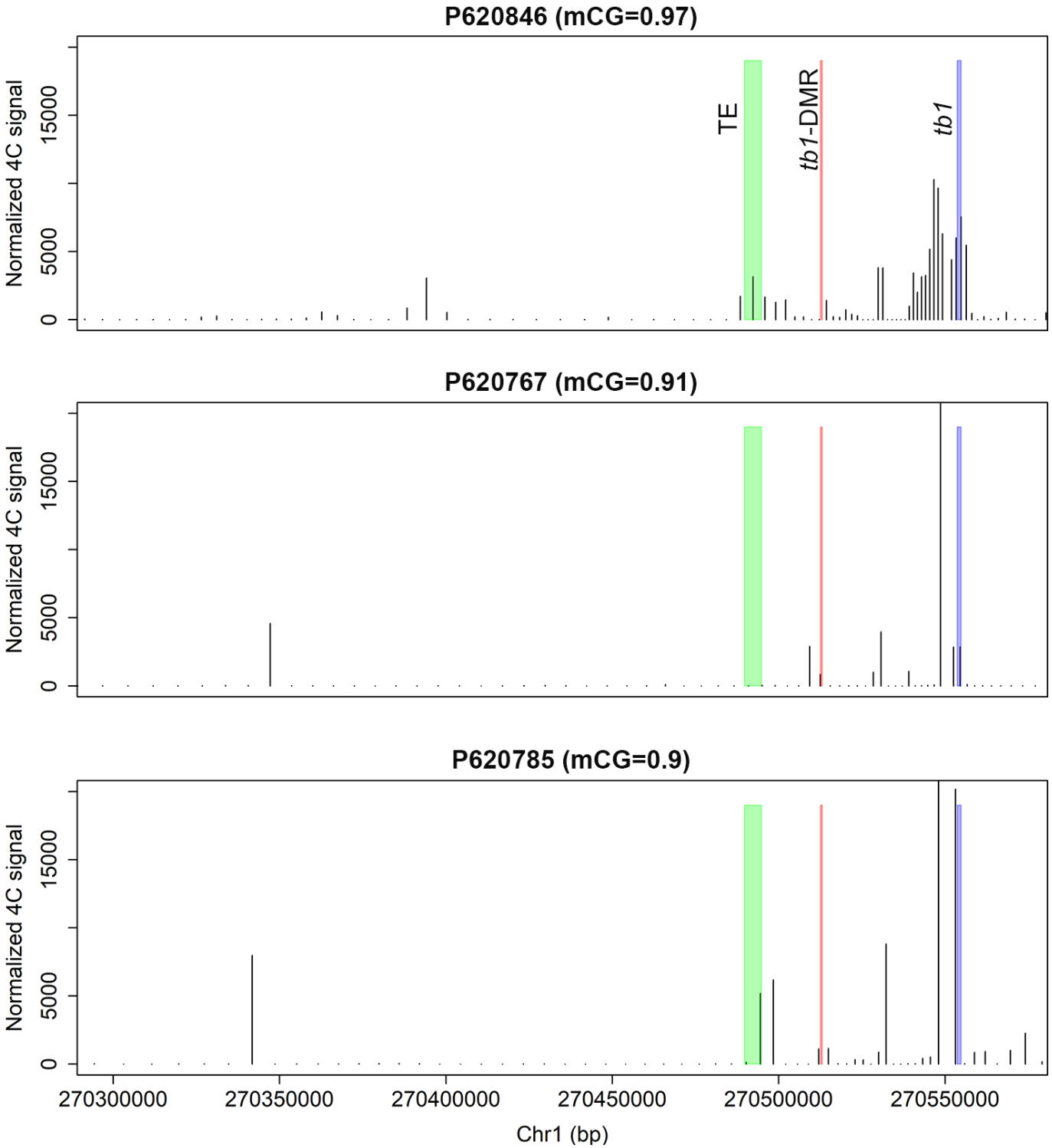

Fig. S21. The 4C-seq results of regions interacted with *tb1* gene in *tb1*-DMR hypermethylated landrace samples. The titles show the CG methylation levels of the *tb1*-DMR. The green rectangle indicates hopscotch TE; the red rectangle indicates *tb1*-DMR; and the blue rectangle indicates *tb1* gene.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

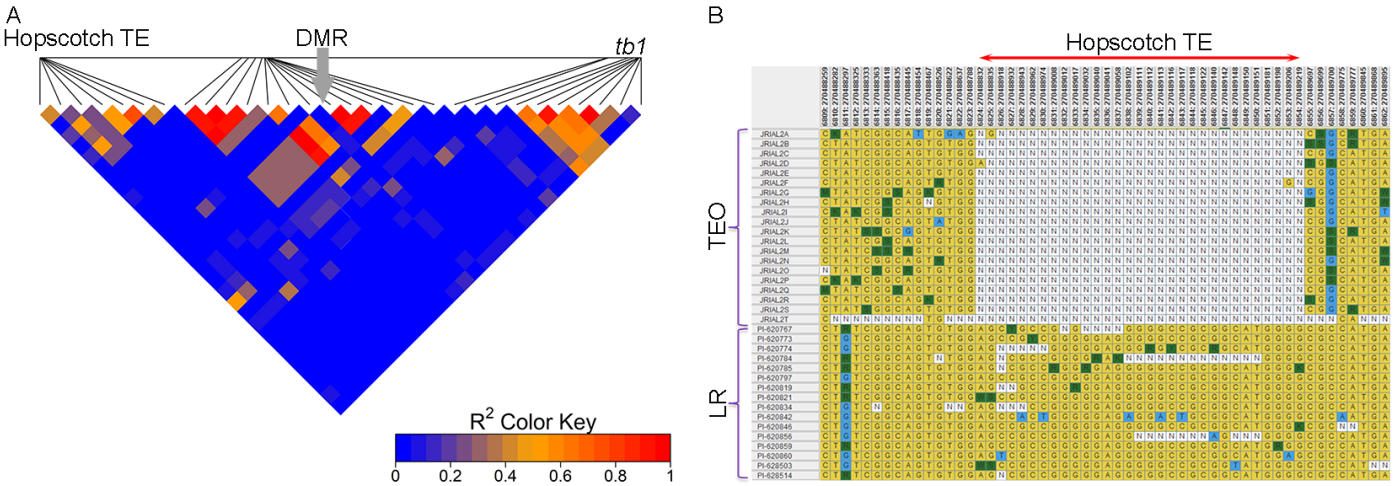

Fig. S22. The linkage disequilibrium (LD) analysis among the Hopscotch transposon, *tb1*-DMR, and *tb1* gene. (**A**) The LD heatmap using 17 landraces segregating at the tb1-DMR locus. The arrow indicated the position of the *tb1*-DMR. LD analysis was performed using SNPs (coded with 0, 1, and 2) called from the WGS data and the mCG level of the *tb1*-DMR. The horizontal black line indicates the physical positions of the SNPs. (**B**) The SNP genotypes of teosinte and landrace around the Hopscotch transposon insertion region.

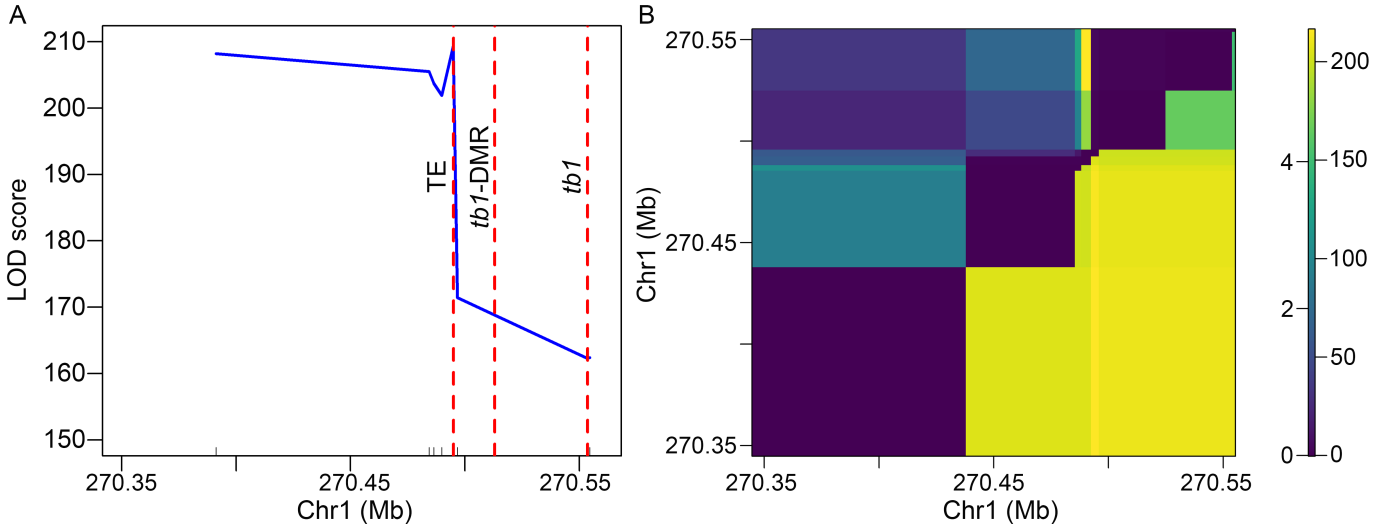

Fig. S23. The QTL scannings for the tillering phenotype around *tb1* locus. (**A**) The conventional single-QTL mapping result. Ticks above the x-axis indicated the physical positions of the markers. (**B**) Two-dimensional QTL scanning result. The lower triangle denoted the LOD scores for the joint two-locus; the upper left triangle denoted the LOD score for the epistasis of the two loci. The color scale on the right indicated LOD scores for the joint two-locus (right) and epistasis (left), separately.

bioRxiv preprint doi: https://doi.org/10.1101/2020.03.13.991117; this version posted October 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

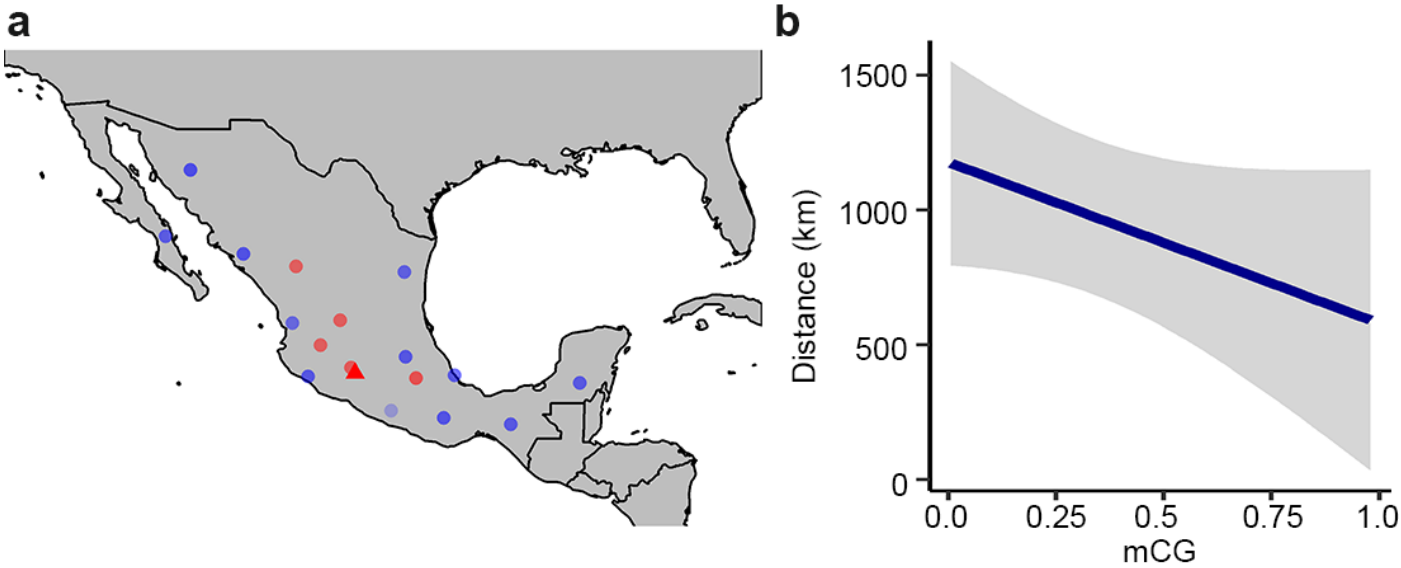

Fig. S24. Correlation analysis between geographical distributions of landrace samples and their CG methylation level at *tb1*-DMR. (**A**) Geographical distributions of the teosinte (triangle) and landrace (points) samples. Red denotes highly methylated and blue denotes lowly methylated samples in the *tb1*-DMR. (**B**) Level of mCG correlated with distance to the origin (Balsas River Valley) of maize, with grey error bands marking the 95% confidence interval.