nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19333-4 OPEN

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] & Jinliang Yang[1,2✉]

DNA methylation is a ubiquitous chromatin feature, present in 25% of cytosines in the maize genome, but variation and evolution of the methylation landscape during maize domestication remain largely unknown. Here, we leverage whole-genome sequencing (WGS) and whole-genome bisulfite sequencing (WGBS) data on populations of modern maize, landrace, and teosinte (*Zea mays* ssp. *parviglumis)* to estimate epimutation rates and selection coefficients. We find weak evidence for direct selection on DNA methylation in any context, but thousands of differentially methylated regions (DMRs) are identified population-wide that are correlated with recent selection. For two trait-associated DMRs, *vgt1*-DMR and *tb1*-DMR, HiChIP data indicate that the interactive loops between DMRs and respective downstream genes are present in B73, a modern maize line, but absent in teosinte. 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.

[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, USA. [7] Department of Evolution and Ecology, Center for Population Biology and Genome Center, University of California, Davis, CA 95616, USA. ✉email: jinliang.yang@unl.edu

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 (TE)[4,5], but broad patterns of methylation are often conserved within species[6,7]. Across plant genomes, levels of DNA methylation vary widely from euchromatin to heterochromatin, driven by the different molecular mechanisms for the establishment and maintenance of DNA methylation in CG, CHG, and CHH contexts[8,9].

DNA methylation is considered essential to suppress the activity of transposons[10], to regulate gene expression[11], and to maintain genome stability[8]. Failure to maintain patterns of DNA methylation in many cases can lead to developmental abnormalities and even lethality[12–14]. 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]. In addition, 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. As 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 (*Zea mays* ssp. *parviglumis*) near the Balsas River Valley area in Mexico about 9000 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.

In this work, we collect a set of geographically widespread Mexican landraces and a natural population of teosinte near Palmar Chico, Mexico[30], from which we generate genome and methylome sequencing data. In addition, we profile the teosinte interactome using the highly integrative chromatin immunoprecipitation (HiChIP) method. Together with the analysis from previously published genome[31], transcriptome[32], methylome[6], and interactome[33] datasets, we estimate epimutation rates and selection pressures across different timescales, investigate the DNA methylation landscape in maize and teosinte, detect differentially methylated regions (DMRs), characterize the genomic features that are related with DMRs, and functionally validate 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 (WGBS) from a panel of wild teosinte, domesticated maize landraces, and modern maize inbreds (Supplementary Data 1). 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 (Supplementary Fig. 1a). Using pseudo-references, on average about 25 million (5.6%) more methylated cytosine sites were identified than using the B73 reference (Supplementary Fig. 1b). 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 (Supplementary Fig. 2).

We found methylated cytosines in CG and CHG contexts were significantly higher in pericentromeric regions (0.54 ± 0.01 in a 1 Mb window) than in chromosome arms (0.44 ± 0.04) (Student's *t*-test P-value $< 2.2 \times 10^{-16}$) (Supplementary Fig. 3). 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 (Supplementary Fig. 4a), with ~25% of genes ($N = 6,874$) showing evidence of gene body methylation (gbm). Although the overall distribution of gbm did not differ across populations, genes with clear syntenic orthologs in Sorghum[36] exhibited little gbm (Supplementary Fig. 4b, 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 Markov Chain Monte Carlo (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 (Supplementary Fig. 5). Sensitivity test results suggested that the mSFS was insensitive to the cutoffs used for the methylation calls (Supplementary Fig. 6). As the vast majority (>98%) of CHH sites were unmethylated (Supplementary Fig. 7), we excluded CHH sites from population genetic analysis.

After testing a set of prior values, we found the initial prior rates had little impact on the posteriors, except for extremely large values (Supplementary Fig. 8), for which convergence was difficult. As 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 (Supplementary Fig. 9). 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 ~50 k[39] to ~370 k – 1 M[40]. To account for this uncertainty, we ran the models with a set of different N_e values (50 k, 100 k, 500 k, and 1 M). 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 ($v = 1.8 \times 10^{-7} - 9.0 \times 10^{-9}$ and $3.0 \times 10^{-7} - 1.5 \times 10^{-8}$) using different N_e values (Fig. 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 $N_e \times s$) stayed largely constant with values of 2.0 and 2.2 for CG and CHG tiles, respectively, 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 5 k, and downstream 5 k) were below or close to 1, and the values were above 1 for nongenic regions (i.e., 2.4 for intergenic regions and 3.5 for TE regions) (Fig. 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, whereas it was close to neutrality for ancestrally hypomethylated regions, especially for the exonic regions (Fig. 1c). In CHG contexts, selection was weak in most regions, including TE and intergenic regions, for both ancestral hyper- and hypomethylated sites.

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, whereas 91% of tiles were variable (Supplementary Fig. 5a). These variable methylation regions were 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 maize genome, respectively. To investigate whether regions of the genome exhibiting variable methylation, especially the high-frequency variable regions, are functionally relevant, we used published data from a large maize mapping population[42]. We estimated kinship matrices for single-nucleotide polymorphisms (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 (Supplementary Fig. 10a), whereas hypermethylated regions (uniformly methylated and rarely unmethylated), although accounting for 76% of the genome, contributed 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 Supplementary Fig. 10b).

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 DMRs. We employed the metilene software[44] to identify a total of 5278 DMRs (see Table 1 for numbers broken down by context and type), or about 0.08% (1.8 Mb) of the genome, including 3900 DMRs between teosinte and modern maize, 1019 between teosinte and landrace, and 359 DMRs between landrace and modern maize (Supplementary Data 2). 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]. Results suggested that methylation levels of the DMRs were largely conserved in B73 across three tissue types (Supplementary Fig. 11), consistent with the previous studies[20,46,47].

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 1330 selective sweeps between modern maize and teosinte (Fig. 2 and Supplementary Data 3, see Supplementary Fig. 12 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 (Supplementary Fig. 13, P-value < 0.001), particularly in intergenic and TE regions (Supplementary Fig. 14a). DMRs overlapping with sweeps, both hypo- and hypermethylated in maize, exhibited significantly higher allele frequency differentiation between maize and teosinte (Supplementary Fig. 14b and see Table 1 for other comparisons). We then asked whether DMRs in sweep regions were in linkage disequilibrium (LD) 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 (Supplementary Data 4).

In addition, 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 = 0.04) increased expression levels in maize compared to teosinte using published data[32]. For the 56 genes located in sweep DMRs that were

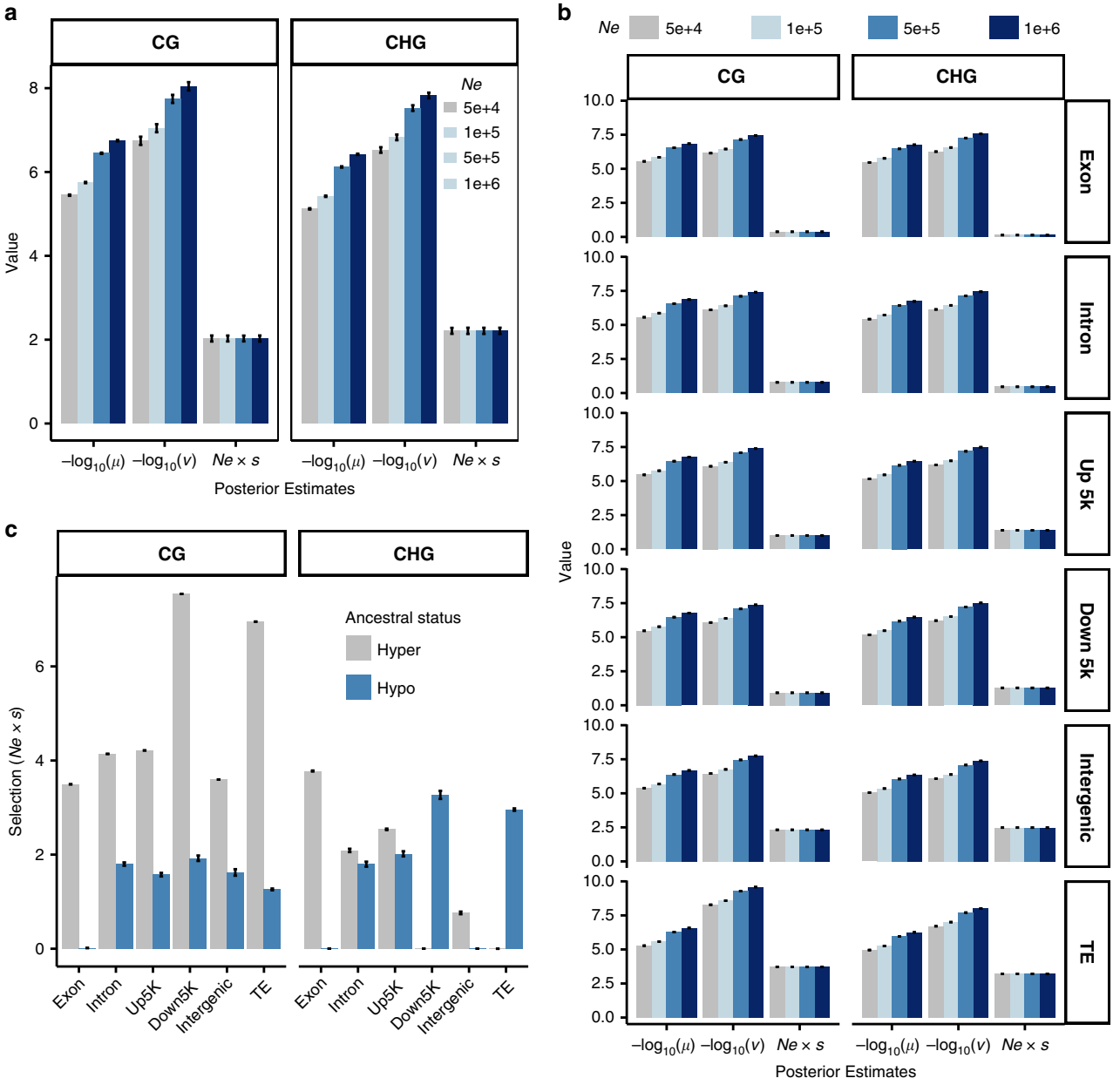

Fig. 1 Population genetic parameters inference. a Posterior estimate of mean values and standard deviations for μ, ν, and $N_e \times s$ for CG and CHG sites using four different effective population size (N_e) values. **b** Posterior estimates for different genomic features. Up 5 k, the upstream 5 kb region of a gene; Down 5 k, the downstream 5 kb 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 SDs ($N = 1600$ for each bar). Source data underlying **a**–**c** are provided as a Source Data file.

hypermethylated in maize, however, we failed to detect the significant expression differences between maize and teosinte.

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 (Fig. 3a). 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; Supplementary Fig. 15a). These CG DMRs could be mapped to $N = 229$ unique genes (Supplementary Data 5). After examining the mapping locations based on collapsed gene models, we found that DMRs were most abundant in 5′-untranslated regions (Fig. 3b), consistent with a pattern that was previously observed[55]. Using these DMR genes for a Gene Ontology (GO) analysis, we detected 15 molecular function terms that were significantly enriched (Supplementary Fig. 15b). 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[56] (permutation P-value = 0.001).

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

Comparison	Context	Type[a]	DMR	Sweep DMR[b]	Interacting DMR[c]	Interacting DMRs in sweeps[d]
Teosinte vs. Landrace	CG	Hypo in landrace	287	80 (27.8%, $P = 0.001$)*	96 (33.4%, $P = 0.001$)*	18 (18.7%, $P = 0.02$)*
		Hyper in landrace	144	36 (25%, $P = 0.001$)*	61 (42.3%, $P = 0.001$)*	6 (9.8%, $P = 6.5 \times 10^{-4}$)*
	CHG	Hypo in landrace	438	121 (27.6%, $P = 0.001$)*	115 (26.2%, $P = 0.001$)*	22 (19.1%, $P = 0.024$)*
		Hyper in landrace	150	45 (30%, $P = 0.002$)*	39 (26%, $P = 0.001$)*	8 (20.5%, $P = 0.19$)
Landrace vs. Maize	CG	Hypo in maize	143	29 (20.2%, $P = 0.078$)*	45 (31.4%, $P = 0.01$)*	9 (20%, $P = 1$)
		Hyper in maize	28	13 (46.4%, $P = 0.001$)	3 (10.7%, $P = 0.47$)	1 (33.3%, $P = 1$)
	CHG	Hypo in maize	158	36 (22.7%, $P = 0.027$)*	46 (29.1%, $P = 0.001$)*	10 (21.7%, $P = 1$)
		Hyper in maize	30	13 (43.3%, $P = 0.001$)*	5 (16.6%, $P = 0.12$)	1 (20%, $P = 0.5$)
Teosinte vs. Maize	CG	Hypo in maize	998	281 (28.1%, $P = 0.001$)*	396 (39.6%, $P = 0.001$)*	67 (16.9%, $P = 2.4 \times 10^{-10}$)*
		Hyper in maize	544	147 (27%, $P = 0.001$)*	259 (47.6%, $P = 0.001$)*	32 (12.3%, $P = 4.2 \times 10^{-13}$)*
	CHG	Hypo in maize	1,855	490 (26.4%, $P = 0.001$)*	594 (32%, $P = 0.001$)*	104 (17.5%, $P = 3.3 \times 10^{-9}$)*
		Hyper in maize	503	159 (3x1.6%, $P = 0.001$)*	124 (24.6%, $P = 0.001$)*	19 (15.3%, $P = 1.1 \times 10^{-5}$)*

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

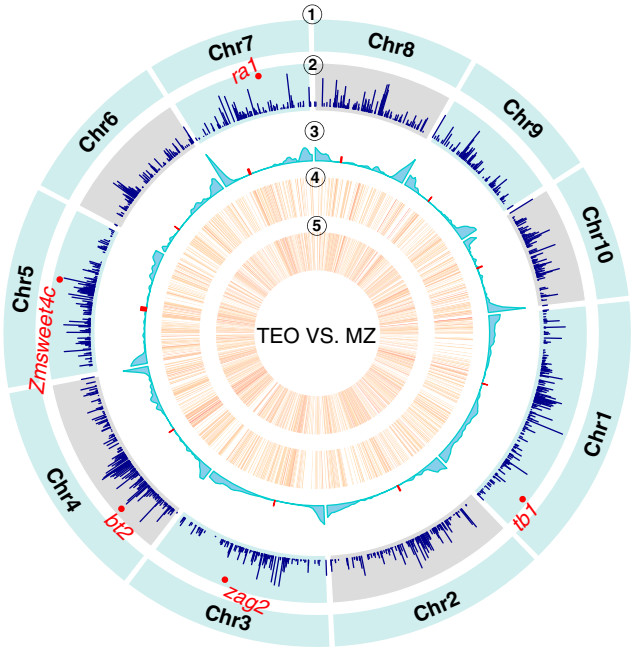

Fig. 2 Selection on differentially methylated regions. Distributions of teosinte-maize selective sweeps, DMRs, and other genomic features across ten maize chromosomes. From outer to inner circles were as follows: ① chromosome names, ② selective sweeps detected between modern maize and teosinte, ③ the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

These findings suggested a potential role for DMRs affecting regulatory regions. To investigate this possibility, we made use of recent data using chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing 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 (Fig. 3a). 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 first, second, and third levels of contacts (Supplementary Fig. 16a). Interestingly, among the 60 genes in direct contact with maize hypomethylated intergenic DMRs (Supplementary Data 6), 30 (43/60 with expression data) showed significantly (Student's paired t-test, P-value = 0.03) increased expression levels in maize compared to teosinte using published data[32]. The results were not significant for 2nd and 3rd level contacts (Supplementary Fig. 16a). We found 5/60 genes (Enrichment test P-value = 7×10^{-3}) were domestication candidate genes as reported previously[57–60]. Two of them were *Zm00001d018036* (a gene associated with cob length, P-value = 6×10^{-25}) and *Zm00001d041948* (a gene associated with shank length, P-value = 5.6×10^{-10})[57]. Further investigation of these two candidates using recently published chromatin data[61] to detect enhancer activity[62] identified H3K27ac peaks at both DMR loci (Fig. 3c and Supplementary Fig. 17a). Consistent with these enhancer signals, the expression levels of these two genes were significantly increased in maize relative to teosinte (Supplementary Fig. 16b and Supplementary Fig. 17b). Despite this functional evidence, however, 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 (Supplementary Fig. 10a). Upon closer inspection of our DMRs, we found a number of candidate flowering time genes located in sweep DMRs or interacting DMRs (Supplementary Data 7), including three genes found in both (i.e., *Zm00001d029946*, *Zm00001d015884*, and *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 (Supplementary Fig. 18) (Enrichment test P-value = 0.001). One DMR was located 40 kb upstream of *ZmRAP2.7*, a well-characterized flowering time gene, and 20 kb downstream of the *vgt1* locus, which was hypomethylated in modern maize and landrace but was hypermethylated in teosinte (Fig. 4a). A MITE transposon insertion in the *vgt1* locus is considered as the causal variant for the down regulation of

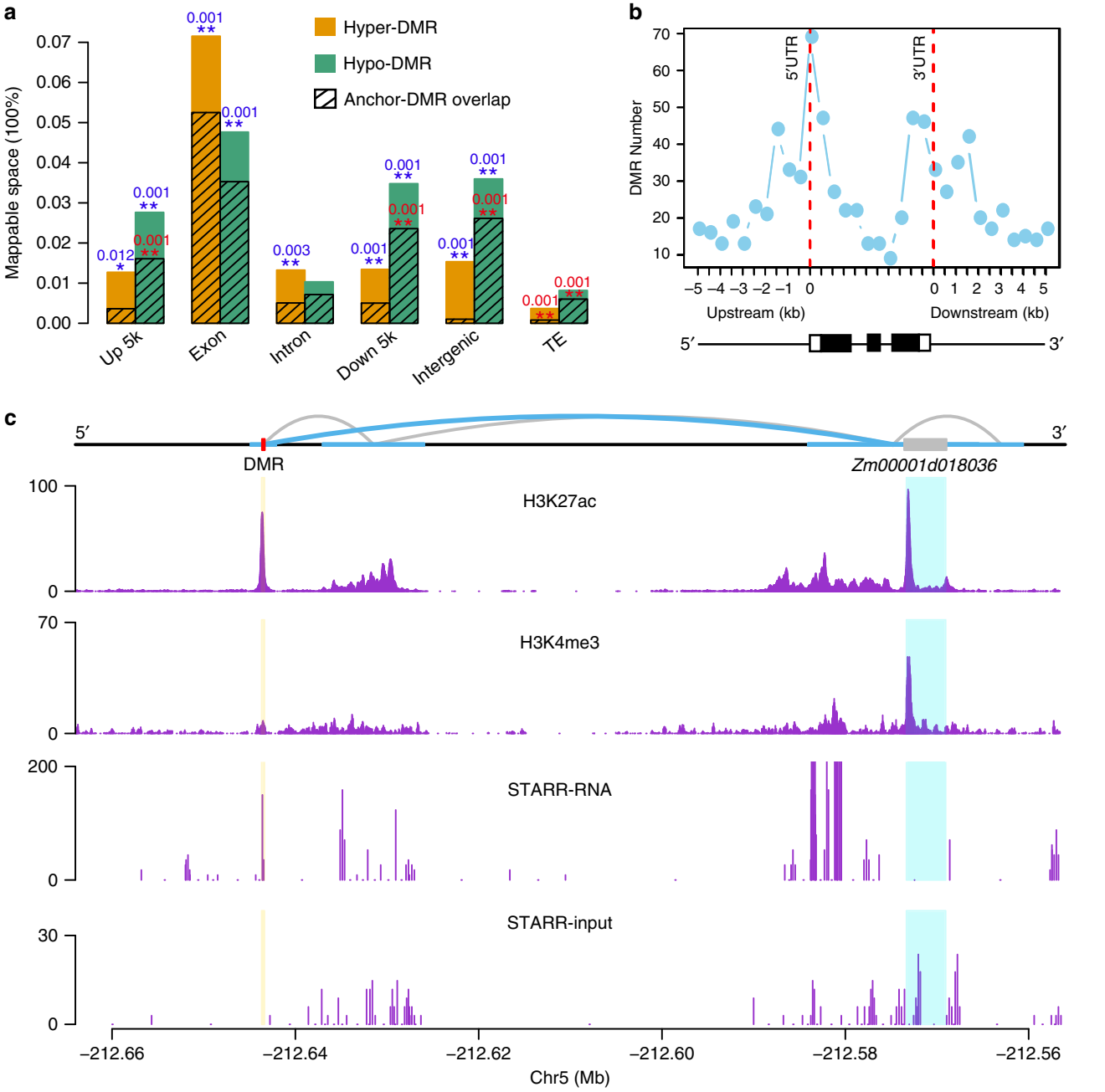

Fig. 3 Teosinte-maize CG differentially methylated regions and their associated functional features. a Breakdown of hyper-DMRs (DMR hypermethylated in maize) and hypo-DMRs (DMR hypomethylated in maize) into genomic features and their overlaps with interactive anchors using data obtained from Li et al.[33]. 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 ref. [61] 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 gray curly lines indicate other interactive contacts around the region. Horizontal thick blue lines denote the interactive anchors. Red and gray boxes indicate the DMR and gene model, respectively. Source data underlying **a**, **b** are provided as a Source Data file.

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] (Fig. 4b). In addition, we reanalyzed the maize and sorghum sequence data at the *vgt1* locus and found two conserved non-coding sequences located 1 kb downstream of the *vgt1*-DMR (Supplementary Fig. 19). To examine the interaction status in teosinte, we then generated HiChIP data for a

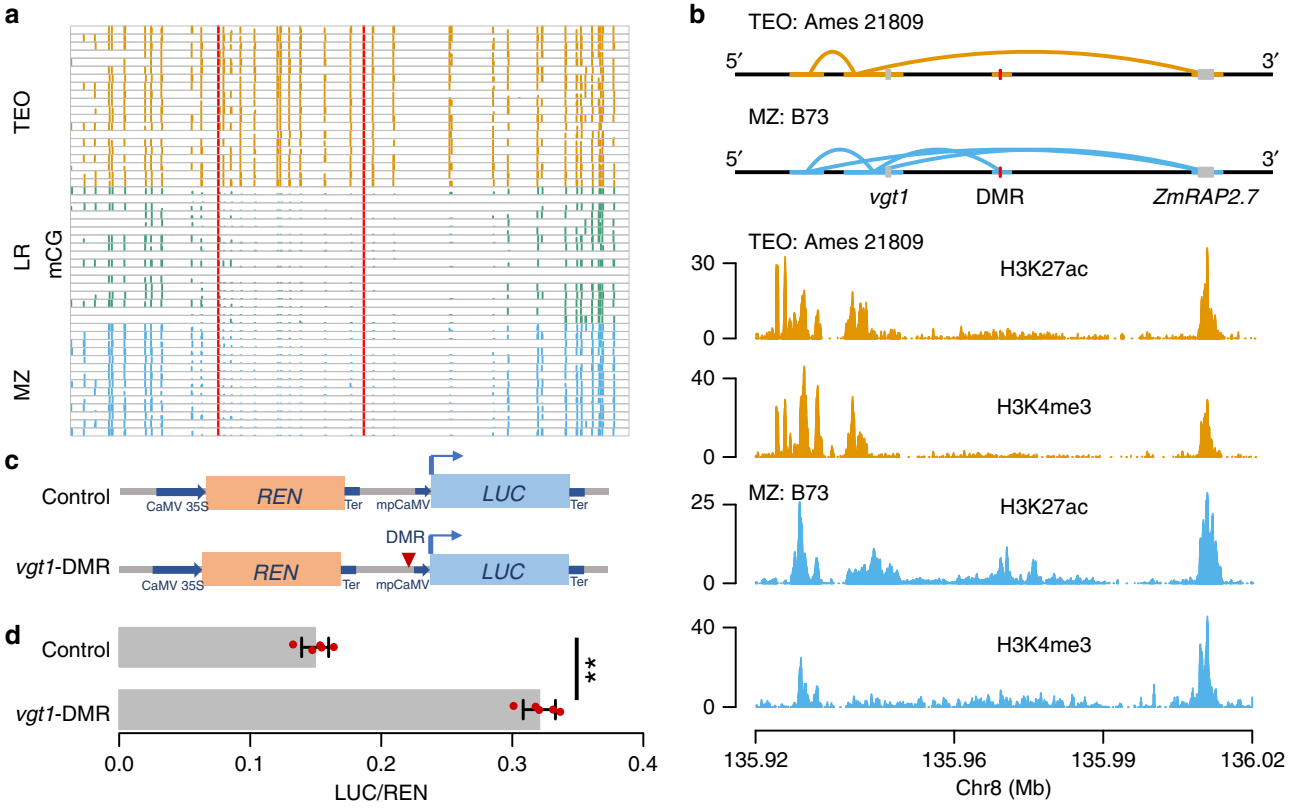

Fig. 4 Functional analysis of *vgt1*-DMR. a Levels of CG methylation around *vgt1*-DMR in maize (MZ), landrace (LR), and teosinte (TEO) populations. Vertical red lines indicate the boundaries of the *vgt1*-DMR. **b** The interactive contacts (upper panel) and colocalization with H3K27ac and H3K4me3 (lower 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 SDs. Statistical significance was determined by a two-sided *t*-test (**P-value $= 3.6 \times 10^{-8}$). Source data underlying **a**, **d** are provided as a Source Data file.

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 (Fig. 4b), suggesting that methylation at this locus might play a functional role in affecting physical interaction.

To further validate the potential enhancer activity of the 209 bp *vgt1*-DMR, we incorporated the *vgt1*-DMR sequence amplified from B73 into a vector constructed as shown in (Fig. 4c) and performed a dual-luciferase (LUC) 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 $= 2.4 \times 10^{-8}$, Fig. 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 (Fig. 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 (Supplementary Fig. 20). The physical location of the *tb1*-DMR was 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 = 0.04), 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 (Fig. 5b) of the *tb1* gene. The TE was considered as an enhancer, as shown in a transient in vivo assay[50]. Interactome data support this claim, finding physical contact between *Hopscotch* and the *tb1* gene (Fig. 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 was missing in teosinte based on our HiChIP data (Fig. 5b). By employing the circular chromosome conformation capture followed by sequencing (4C-seq) method[68], we further confirmed the absence of interaction between the *tb1*-DMR and the *tb1* gene using landrace samples showed hypermethylation at the *tb1*-DMR locus (Supplementary Fig. 21). 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 (Fig. 5b). In addition, we conducted a dual-LUC transient assay by constructing a vector similar to the *vgt1*-DMR (Fig. 4e). The results indicated that the *tb1*-DMR significantly increased the LUC/REN ratio as compared to control (Fig. 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 LD analysis using landrace genomic and methylation data segregating at this *tb1*-DMR locus (see "Methods"). As a

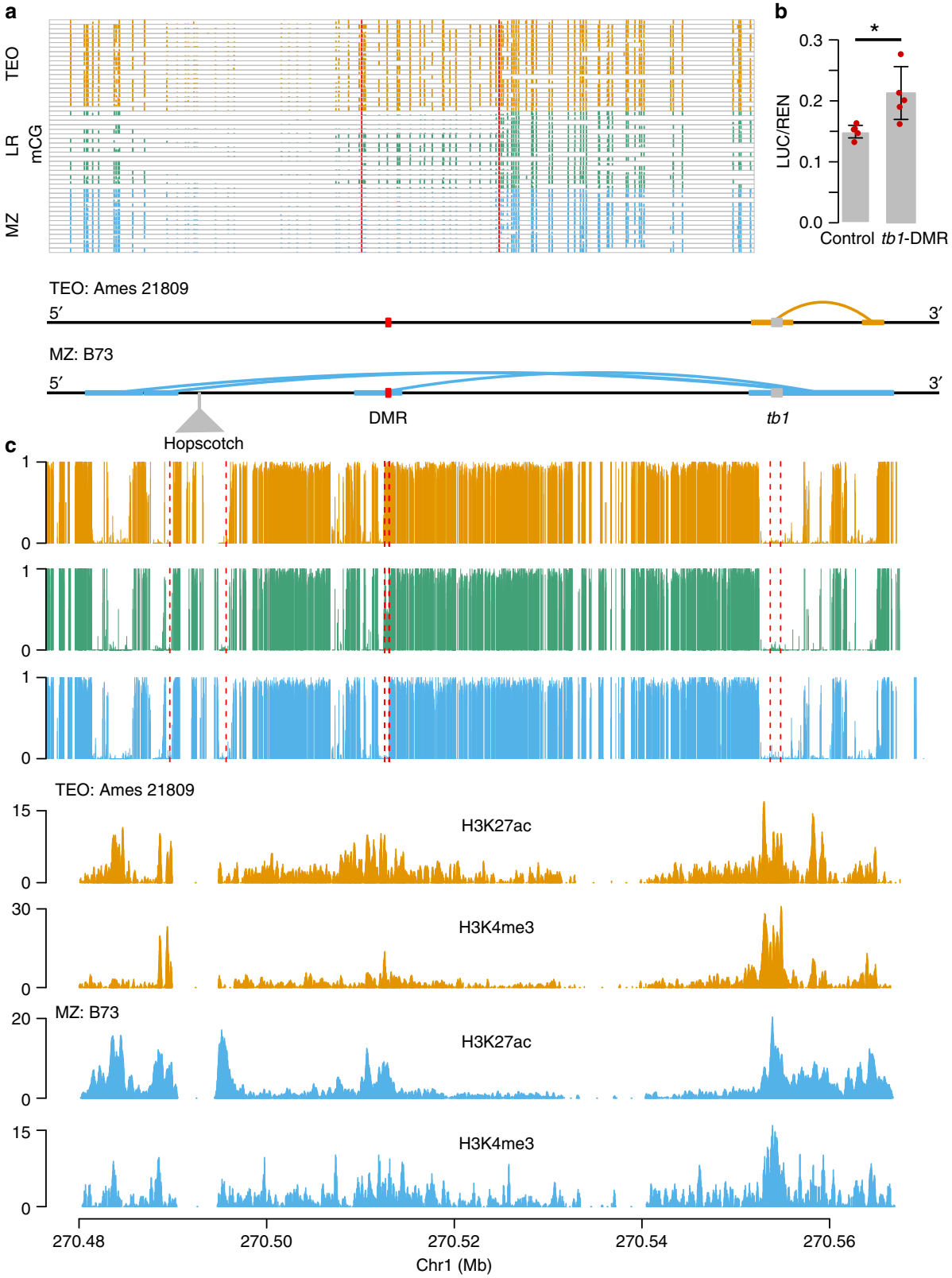

Fig. 5 A hypomethylated differentially methylated region that is upstream of *tb1* gene. a Levels of mCG for the 534-bp *tb1*-DMR in each individual methylome of the modern maize (MZ), landrace (LR), and teosinte (TEO) populations. Vertical red lines indicate the boundaries of the *tb1*-DMR. **b** Interactive 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. $N = 5$ biological replicates were performed. Error bars indicated SDs. Statistical significance was determined by a two-sided *t*-test (*P-value $= 3.4 \times 10^{-2}$). Source data underlying **a**, **c** are provided as a Source Data file.

result, we failed to detect strong LD (i.e., $R^2 > 0.1$) between the *tb1*-DMR and SNPs located at the *Hopscotch* locus (Supplementary Fig. 22), 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 (Supplementary Fig. 23a) and a two-dimensional QTL scan detected epistasis between *Hopscotch* and the *tb1*-DMR (Supplementary Fig. 23b). Further, we found that highly methylated landraces were geographically closer to the Balsas River Valley in Mexico, where maize was originally domesticated from (Supplementary Fig. 24a). As the landraces spread out from the domestication center, their CG methylation levels were gradually reduced (Supplementary Fig. 24b).

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 ten 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 DMRs. These DMRs 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 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 LD between DMRs and nearby SNPs make it difficult to assign causality, i.e., the DMRs associated with the flowering time traits may not be 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 a genome-wide association study 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 population-wide 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.

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 (Supplementary Data 1). The teosinte (*Z. 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 WGBS using Illumina HiSeq platform. In addition, 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×. 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.

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 duplicated reads, we extracted methylated cytosines using the Bismark methylation extractor and only retained sites with more than three mapped reads. The methylation level of each base pair was determined by using the number of reads supporting cytosine methylation divided by the total number of reads at each cytosine site[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

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

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

We defined a 100 bp tile 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 MCMC method (http://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} (Supplementary Fig. 8a) and lambda values of the scaled proposal distribution of 0.01, 0.05, and 0.1 (Supplementary Fig. 8b). 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.

DMRs 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 GO term analysis on selected gene lists using AgriGO2.0 with default parameters[83]. We used the significance cutoff at P-value < 0.01.

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 (one-sided 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 (15 h day time and 9 h 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 first removed chloroplast and other cell debris, resuspended it in 0.5% SDS, used 10% Triton X-100 to quench it, and then performed digestion, incorporation, and proximity ligation reactions[33]. We used two antibodies H3K4me3 (Abcam, ab8580) and H3K27ac (Abcam, ab4729) to pull down the DNA. Then, we purified DNA with the MinElute PCR Purification Kit (QIAGEN, catalog number 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.

Chromatin immunoprecipitation sequencing and HiChIP data analysis. We obtained chromatin immunoprecipitation sequencing 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 <5 kb or >2 Mb. By using the mango pipeline[89], we determined the remaining loops with three read pairs supports and the false discovery rate <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 5′-CGAA GTCTCTGAGTATGATC-3′ (forward) and 5′-GGGTTCAAAGCACCAACAG T-3′ (reverse). 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 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, rarely methylated, rarely unmethylated, and high-frequency 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 method implemented in LDAK[43].

Dual-LUC transient expression assay in maize protoplasts. To investigate the effect of DMRs on gene expression, we performed a dual-LUC 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 *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 after 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 μg plasmids into the 100 ul isolated protoplasts using polyethylene glycol (PEG) mediated transformation method[95]. After 16 h infiltration, we measured the LUC and REN activities using dual-LUC 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.

Experimental validation of the *tb1*-DMR. We performed Chop-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 EpiJET^{TM} DNA Methylation Analysis Kit (*Msp*I/*Hpa*II) (Thermo Scientific, K1441) following manufacturer's instructions. The primer sequences for PCR were 5′-ACACGCACGA AGGGTTACAG-3′ (forward) and 5′-CAGTGCTCCCTGGGTCAAA-3′ (reverse).

Statistical analyses. We performed all the statistical analyses using R software (V3.6.2, https://www.r-project.org/).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this article is available as a Supplementary Information file. The datasets and plant materials generated and analyzed during the current study are available from the corresponding author upon request. All datasets generated in this study have been uploaded to the Gene Expression Omnibus database and can be retrieved through accession number GSE145586. Source data are provided with this paper.

Code availability

The code used for the analyses are available at GitHub [https://github.com/jyanglab/msfs_teo].

Received: 25 March 2020; Accepted: 8 October 2020;
Published online: 02 November 2020

References

1. Sánchez-Romero, M. A., Cota, I. & Casadesús, J. DNA methylation in bacteria: from the methyl group to the methylome. *Curr. Opin. Microbiol.* **25**, 9–16 (2015).
2. Robertson, K. D. DNA methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).

3. Arand, J. et al. *In vivo* control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet.* **8**, e1002750 (2012).
4. Alonso, C., Perez, R., Bazaga, P. & Herrera, C. M. Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms. *Front Genet.* **6**, 4 (2015).
5. Niederhuth, C. E. et al. Widespread natural variation of DNA methylation within angiosperms. *Genome Biol.* **17**, 194 (2016).
6. Li, Q. et al. Examining the causes and consequences of context-specific differential DNA methylation in maize. *Plant Physiol.* **168**, 1262–1274 (2015).
7. Schmitz, R. et al. Patterns of population epigenomic diversity. *Nature* **495**, 193–198 (2013).
8. Zhang, H. M., Lang, Z. B. & Zhu, J. K. Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* **19**, 489–506 (2018).
9. Springer, N. M. & Schmitz, R. J. Exploiting induced and natural epigenetic variation for crop improvement. *Nat. Rev. Genet.* **18**, 563–575 (2017).
10. Deniz, Ö., Frost, J. M. & Branco, M. R. Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* **20**, 417–431 (2019).
11. Seymour, D. K. & Becker, C. The causes and consequences of DNA methylome variation in plants. *Curr. Opin. Plant Biol.* **36**, 56–63 (2017).
12. Dorweiler, J. E. et al. *mediator of paramutation1* is required for establishment and maintenance of paramutation at multiple maize loci. *Plant Cell* **12**, 2101–2118 (2000).
13. Li, Q. et al. Genetic perturbation of the maize methylome. *Plant Cell* **26**, 4602–4616 (2014).
14. Fu, F. F., Dawe, R. K. & Gent, J. I. Loss of RNA-directed DNA methylation in maize chromomethylase and DDM1-type nucleosome remodeler mutants. *Plant Cell* **30**, 1617–1627 (2018).
15. Shen, Y. T. et al. DNA methylation footprints during soybean domestication and improvement. *Genome Biol.* **19**, 1–14 (2018).
16. Hernando-Herraez, I., Garcia-Perez, R., Sharp, A. J. & Marques-Bonet, T. DNA methylation: insights into human evolution. *PLoS Genet.* **11**, e1005661 (2015).
17. Kader, F. & Ghai, M. DNA methylation-based variation between human populations. *Mol. Genet. Genomics* **292**, 5–35 (2017).
18. Manning, K. et al. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat. Genet.* **38**, 948–952 (2006).
19. Cortijo, S. et al. Mapping the epigenetic basis of complex traits. *Science* **343**, 1145–1148 (2014).
20. Eichten, S. R. et al. Epigenetic and genetic influences on DNA methylation variation in maize populations. *Plant Cell* **25**, 2783–2797 (2013).
21. Van der Graaf, A. et al. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proc. Natl Acad. Sci. USA* **112**, 6676–6681 (2015).
22. Shahryary, Y. et al. AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. *Genome Biol.* **21**, 260 (2020).
23. Charlesworth, B. & Jain, K. Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates. *Genetics* **198**, 1587–1602 (2014).
24. Vidalis, A. et al. Methylome evolution in plants. *Genome Biol.* **17**, 264 (2016).
25. Stitzer, M. C. & Ross-Ibarra, J. Maize domestication and gene interaction. *New Phytol.* **220**, 395–408 (2018).
26. Gates, D. J. et al. Single-gene resolution of locally adaptive genetic variation in Mexican maize. Preprint at https://doi.org/10.1101/706739 (2019).
27. Swarts, K. et al. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. *Science* **357**, 512–515 (2017).
28. Navarro, J. A. R. et al. A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nat. Genet.* **49**, 476–480 (2017).
29. Teixeira, J. et al. Hallauer's Tuson: a decade of selection for tropical-to-temperate phenological adaptation in maize. *Heredity* **114**, 229–240 (2015).
30. Yang, C. J. et al. The genetic architecture of teosinte catalyzed and constrained maize domestication. *Proc. Natl Acad. Sci. USA* **116**, 5643–5652 (2019).
31. Bukowski, R. et al. Construction of the third-generation *Zea mays* haplotype map. *Gigascience* **7**, gix134 (2017).
32. Lemmon, Z. H., Bukowski, R., Sun, Q. & Doebley, J. F. The role of *cis* regulatory evolution in maize domestication. *PLoS Genet.* **10**, e1004745 (2014).
33. Li, E. et al. Long-range interactions between proximal and distal regulatory regions in maize. *Nat. Commun.* **10**, 2633 (2019).
34. Wulfridge, P., Langmead, B., Feinberg, A. P. & Hansen, K. D. Choice of reference genome can introduce massive bias in bisulfite sequencing data. *Nucleic Acid Res.* 47, e117 (2019).
35. Bertioli, D. J. et al. The genome sequences of *Arachis duranensis* and Arachis ipaensis, the diploid ancestors of cultivated peanut. *Nat. Genet.* **48**, 438–446 (2016).
36. Zhang, Y. et al. Differentially regulated orthologs in sorghum and the subgenomes of maize. *Plant Cell* **29**, 1938–1951 (2017).
37. West, P. T. et al. Genomic distribution of H3K9me2 and DNA methylation in a maize genome. *PLoS ONE* **9**, e105267 (2014).
38. Benaglia, T., Chauveau, D. S., Hunter, D. R. & Young, D. S. mixtools: An R package for analyzing finite mixture models. *J. Stat. Softw.* **32**, 1–29 (2009).
39. Ross-Ibarra, J., Tenaillon, M. & Gaut, B. S. Historical divergence and gene flow in the genus *Zea*. *Genetics* **181**, 1397–1409 (2009).
40. Beissinger, T. M. et al. Recent demography drives changes in linked selection across the maize genome. *Nat. Plants* **2**, 1–7 (2016).
41. Hahn, M. W. *Molecular Population Genetics* (Sinauer Associates/Oxford Univ. Press, 2018).
42. Wallace, J. G. et al. Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet.* **10**, e1004845 (2014).
43. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
44. Jühling, F. et al. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.* **26**, 256–262 (2016).
45. Sun, Y. et al. 3D genome architecture coordinates *trans* and *cis* regulation of differentially expressed ear and tassel genes in maize. *Genome Biol.* **21**, 1–25 (2020).
46. Zhang, M. et al. Extensive, clustered parental imprinting of protein-coding and noncoding RNAs in developing maize endosperm. *Proc. Natl Acad. Sci. USA* **108**, 20042–20047 (2011).
47. Zemach, A. et al. Local DNA hypomethylation activates genes in rice endosperm. *Proc. Natl Acad. Sci. USA* **107**, 18729–18734 (2010).
48. Gardiner, L. J. et al. A genome-wide survey of DNA methylation in hexaploid wheat. *Genome Biol.* **16**, 273 (2015).
49. Song, Q., Zhang, T., Stelly, D. M. & Chen, Z. J. Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biol.* **18**, 99 (2017).
50. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* **43**, 1160–1163 (2011).
51. Zhao, D. P., Huang, Z. C., Umino, N., Hasegawa, A. & Kanamori, H. Structural heterogeneity in the megathrust zone and mechanism of the 2011 Tohoku-oki earthquake (Mw 9.0). *Geophys. Res. Lett.* **38** (2011).
52. Sosso, D. et al. Seed filling in domesticated maize and rice depends on SWEET-mediated hexose transport. *Nat. Genet* **47**, 1489 (2015).
53. Sigmon, B. & Vollbrecht, E. Evidence of selection at the *ramosa1* locus during maize domestication. *Mol. Ecol.* **19**, 1296–1311 (2010).
54. Whitt, S. R., Wilson, L. M., Tenaillon, M. I., Gaut, B. S. & Buckler, E. S. Genetic diversity and selection in the maize starch pathway. *Proc. Natl Acad. Sci. USA* **99**, 12959–12962 (2002).
55. Candaele, J. et al. Differential methylation during maize leaf growth targets developmentally regulated genes. *Plant Physiol.* **164**, 1350–1364 (2014).
56. Galli, M. et al. The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. *Nat. Commun.* **9**, 1–14 (2018).
57. Xue, S., Bradbury, P. J., Casstevens, T. & Holland, J. B. Genetic architecture of domestication-related traits in maize. *Genetics* **204**, 99–113 (2016).
58. Li, Y. X. et al. Identification of genetic variants associated with maize flowering time using an extremely large multi-genetic background population. *Plant J.* **86**, 391–402 (2016).
59. Xu, C. et al. Genome-wide association study dissects yield components associated with low-phosphorus stress tolerance in maize. *Theor. Appl. Genet.* **131**, 1699–1714 (2018).
60. Li, C. H. et al. Numerous genetic loci identified for drought tolerance in the maize nested association mapping populations. *BMC Genomics* **17**, 894 (2016).
61. Ricci, W. A. et al. Widespread long-range *cis*-regulatory elements in the maize genome. *Nat. Plants* **5**, 1237–1249 (2019).
62. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
63. Dong, Z. S. et al. A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. *PLoS ONE* **7**, e43450 (2012).
64. Salvi, S. et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc. Natl Acad. Sci. USA* **104**, 11376–11381 (2007).
65. Hufford, M. B. et al. Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
66. Rodgers-Melnick, E., Vera, D. L., Bass, H. W. & Buckler, E. S. Open chromatin reveals the functional maize genome. *Proc. Natl Acad. Sci. USA* **113**, E3177–E3184 (2016).
67. Oka, R. et al. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol.* **18**, 137 (2017).
68. Splinter, E., de Wit, E., van de Werken, H. J. G., Klous, P. & De Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. *Methods* **58**, 221–230 (2012).
69. Becker, C. et al. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**, 245–249 (2011).

70. Jiao, Y. P. et al. Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815 (2012).
71. Li, X. R. et al. Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genome Res.* **22**, 2436–2444 (2012).
72. Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular-weight plant DNA. *Nucleic Acids Res.* **8**, 4321–4325 (1980).
73. Schnable, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
74. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv:13033997 (2013).
75. Picard toolkit. http://broadinstitute.github.io/picard/ (2019).
76. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
77. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
78. Schultz, M. D., Schmitz, R. J. & Ecker, J. R. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends Genet.* **28**, 583–585 (2012).
79. Wang, H. et al. The origin of the naked grains of maize. *Nature* **436**, 714–719 (2005).
80. Tian, F., Stevens, N. M. & Buckler, E. S. Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *Proc. Natl Acad. Sci. USA* **106**, 9979–9986 (2009).
81. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
82. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
83. Tian, T. et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129 (2017).
84. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
85. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
86. Ramírez, F., Dündarm, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
87. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
88. Lareau, C. A. & Aryee, M. J. hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat. Methods* **15**, 155–156 (2018).
89. Phanstiel, D. H., Boyle, A. P., Heidari, N. & Snyder, M. P. Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics* **31**, 3092–3098 (2015).
90. Raviram, R. et al. 4C-ker: a method to reproducibly identify genome-wide interactions captured by 4C-Seq experiments. *PLoS Comput. Biol.* **12**, e1004780 (2016).
91. Yu, J. M., Holland, J. B., McMullen, M. D. & Buckler, E. S. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**, 539–551 (2008).
92. Buckler, E. S. et al. The genetic architecture of maize flowering time. *Science* **325**, 714–718 (2009).
93. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
94. Hellens, R. P. et al. Transient expression vectors for functional genomics, quantification of promoter activity and RNA silencing in plants. *Plant Methods* **1**, 13 (2005).
95. Yoo, S. D., Cho, Y. H. & Sheen, J. *Arabidopsis* mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protoc.* **2**, 1565–1572 (2007).

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 thank Mike May for help in 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

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19333-4.

Correspondence and requests for materials should be addressed to J.Y.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.