nature communications

Article | Article in Press

https://doi.org/10.1038/s41467-026-79325-8

Transposable element polymorphisms underlie epigenetic variation and phenotypic diversification during foxtail millet domestication

Received: 15 September 2025

Accepted: 14 October 2026

Published online: 07 November 2026

Check for updates

Wenlei Guo[1,2,3,6], Sha Tang[2,6], Yi Sui[2,6], Hui Zhang[2], Jiahao Liu[1,3], Yang Liu[1], Bin Xie[1,3], Hongyan Chang[1], Yingjie Mi[1], Jianghui Cui[4], Hui Zhi[2], Guanqing Jia[2], Xianmin Diao[2] ✉ & Jinfeng Chen[1,5] ✉

Transposable elements (TEs) generate genomic variation, yet how TE polymorphisms reshape the epigenetic landscape and contribute to phenotypic diversification during domestication remains unclear. Here, we show that TE-associated genetic variants drive heritable epigenetic changes underlying key domestication traits in foxtail millet (*Setaria italica*), one of the oldest cereal crops domesticated in China. By analyzing DNA methylomes from 60 wild and domesticated accessions, we identify 76,815 differentially methylated regions associated with extensive polymorphic TE insertions. Epigenomic variation enables the identification of a TE inventory associated with seed shattering, branching, and growth habits. Functional analyses reveal that two DNA/*PIF-Harbinger* element insertions in the 5′ UTR of the pleiotropic gene *SiGW3* modulate its expression and affect plant architecture, panicle morphology, and grain size. Our findings demonstrate that naturally occurring TE polymorphisms generate heritable epigenetic variation and contribute to phenotypic evolution during crop domestication, with implications for crop improvement.

Transposable elements (TEs) are mobile DNA sequences that constitute a major fraction of eukaryotic genomes. Active TE transposition can disrupt gene structure, alter gene expression, and reshape genomic architecture[1]. While host silencing mechanisms typically maintain TEs in a quiescent state, their mobilization can be triggered by various genetic[2–4], epigenetic[5], or environmental stimuli[6]. The genetic and epigenetic variants generated by TEs can thus be co-opted by natural or artificial selection, contributing to adaptation, development, and disease across diverse lineages[7]. TEs therefore represent potent drivers of evolutionary change.

TE transposition generates new insertions that can disrupt or misregulate protein-coding genes, potentially causing deleterious effects on the host. Therefore, TE activity is tightly controlled by the host through sophisticated epigenetic mechanisms, including DNA methylation, histone modification, and small RNA-mediated silencing[8]. In plants, most TEs are silenced through DNA methylation acting in

[1]State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, Beijing, China. [2]State Key Laboratory of Crop Gene Resources and Breeding/Key Laboratory of Grain Crop Genetic Resources Evaluation and Utilization, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. [3]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [4]North China Key Laboratory for Crop Germplasm Resources of Education Ministry, Hebei Agricultural University, Baoding, Hebei, China. [5]State Key Laboratory of Seed Innovation, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. [6]These authors contributed equally: Wenlei Guo, Sha Tang, Yi Sui. ✉e-mail: diaoxianmin@caas.cn; chenjinfeng@ioz.ac.cn

concert with small RNAs and repressive histone modifications. DNA methylation occurs in three sequence contexts in plants: CG, CHG, and CHH, where H represents A, T, or C[9]. Based on studies in Arabidopsis and rice, DNA methylation at CG sites is maintained by METHYLTRANSFERASE 1 (MET1)[9], whereas methylation at CHG and CHH contexts is catalyzed by the DNA methyltransferases CHROMOMETHYLASE 3 and 2 (CMT3 and CMT2), respectively[10]. These enzymes are recruited to regions enriched in histone H3 lysine 9 methylation (H3K9me). H3K9 methyltransferases, in turn, can be recruited to regions with CHG and CHH methylation, forming a self-reinforcing loop that maintains DNA methylation[11]. The processes described above maintain DNA methylation in plants and therefore act on sequences that have already been methylated. In contrast, de novo DNA methylation at CG, CHG, and CHH sites is established by the RNA-directed DNA methylation (RdDM) pathway, in which 24-nt siRNAs guide DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) to target TEs, predominantly at CHH sites. However, for newly introduced TEs or reactivated TEs that become derepressed in mutants defective in epigenetic pathways, non-canonical RdDM mediated by RNA polymerase II (Pol II), RNA-DEPENDENT RNA POLYMERASE 6 (RDR6), and 21-/22-nt siRNAs can initiate their silencing through post-transcriptional gene silencing[12]. Once the silenced state is established, canonical RdDM can subsequently reinforce DNA methylation, promoting long-term silencing that can be maintained across generations.

Heavily methylated TEs can influence gene regulation when located in proximity to protein-coding genes. In plants, mutations in DNA methyltransferase genes result in a range of abnormal phenotypes, including reduced seed size and vascular development, failure to flower, and spontaneous cell death[13]. However, TE-mediated mutagenesis is not always deleterious; TE insertions can sometimes confer beneficial effects on the host and be selected during adaptation to new environments. For example, TE insertions in *FLOWERING LOCUS C* (*FLC*) have been implicated in environmental adaptation in *Arabidopsis thaliana*, *Capsella rubella*, and other species in the Brassicaceae family[14,15]. Crop domestication and subsequent diversification provide excellent systems for studying adaptation driven by artificial selection, which has favored traits such as ease of harvest and high yield over just a few thousand years[16]. TEs can generate functional alleles that are subsequently selected during domestication and diversification[17,18]. A survey of 60 plant genes associated with domestication or diversification traits suggests that TE insertions likely underlie approximately 15% of these adaptive alleles in crops, such as maize, rapeseed, grape, and foxtail millet[16]. For example, in maize, a TE insertion in the *teosinte branched1 (tb1)* gene created a distal enhancer that reduced branching, a key domestication trait[19]. Consistent with this finding, a differentially methylated region (DMR) upstream of *tb1* may facilitate physical interactions between enhancers and their target genes by promoting an open chromatin state[20]. Similarly, in grape and orange, long terminal repeat (LTR) retrotransposon insertions in *Vvmby1A* and *Ruby* have been associated with variation in fruit color during breeding[21,22]. Collectively, these examples underscore the functional importance of TE-derived alleles in shaping agronomic traits during plant domestication and diversification. However, most documented cases involve individual traits in specific crops[17,23], leaving open questions about how TE polymorphisms are shaped by artificial selection during domestication, to what extent such selection reshapes the DNA methylation landscape, and how this epigenetic remodeling ultimately gives rise to favorable phenotypes.

Foxtail millet (*Setaria italica*) is one of the earliest cereal crops domesticated from *S. viridis* in the North China Plain approximately 9600–11,000 years ago[24]. It has served as an invaluable model for studying C_4 photosynthesis, crop domestication, and adaptation to semi-arid environments[25,26]. Population structure analyses have identified four major groups, W1, W2, W3, and W4, within *S. viridis* populations[27,28]. Previous studies have revealed that all Chinese *S. viridis* accessions belong to the W1 group, which is considered the direct progenitor of domesticated foxtail millet[28]. Three major genetic groups, C1, C2, and C3, have been identified in foxtail millet. The C3 lineage is globally distributed and may represent an early stage of foxtail millet domestication, whereas the C1 and C2 groups represent lineages that adapted to the high-latitude regions of northern China and the warmer climates of central and southern China[28], respectively. Artificial selection during domestication imposed a severe bottleneck on genetic diversity, resulting in a 55% reduction in nucleotide diversity (from 0.0059 to 0.0027)[29]. Domestication has also resulted in remarkable phenotypic changes in foxtail millet relative to its wild progenitor, *S. viridis*, including non-shattering seeds, reduced tiller numbers, increased grain size, delayed flowering, and reduced photoperiod sensitivity[30]. To date, candidate genes underlying seed shattering and grain size have been characterized[27,28,31]. In these studies, TE insertions were found to contribute to favorable phenotypic changes in cultivated foxtail millet. For example, an LTR insertion inactivated *Less Shattering1* (*SvLes1*), resulting in the loss of seed shattering in cultivated accessions[27]. Similarly, *Shattering 1* (*Sh1*), which encodes a protein containing zinc finger and YABBY domains, was identified as a key regulator of seed shattering in foxtail millet[32,33]. Its expression was silenced by the insertion of a DNA/*PIF-Harbinger* transposon into an exon[31]. Moreover, TE-mediated genetic variation in the heading date gene *SiPRR37* may have contributed to the adaptation of foxtail millet to northern China[34]. These studies demonstrate that TEs may have played a crucial role in driving phenotypic innovations during foxtail millet domestication and subsequent geographic expansion.

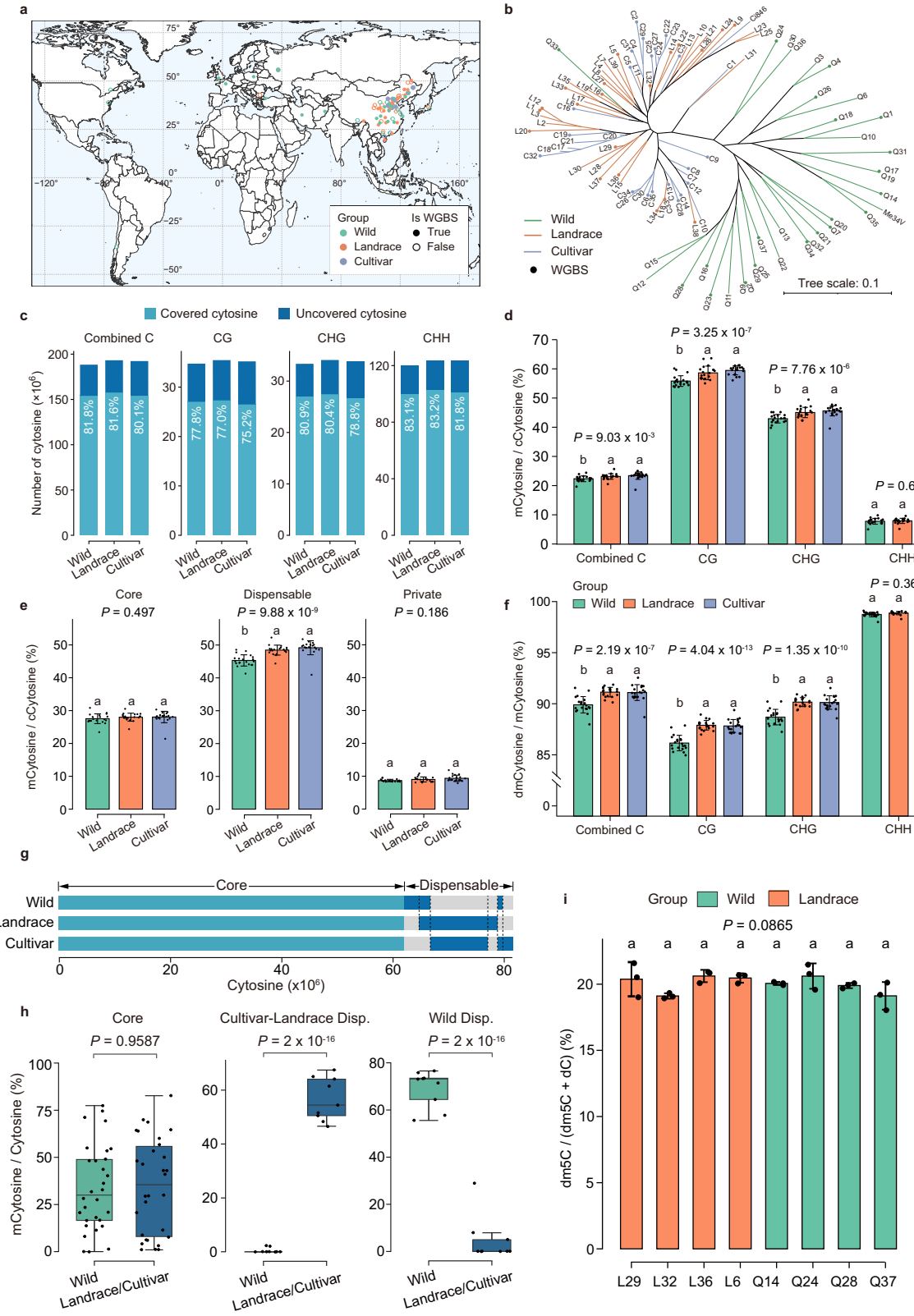
In this work, we integrate whole-genome bisulfite sequencing (WGBS) methylome, RNA-seq, and small RNA-seq data from 60 foxtail millet accessions representing wild, landrace, and cultivar groups to investigate TE diversity in foxtail millet and its contribution to the domestication of this ancient crop. By integrating these datasets with the *Setaria* pangenome, we find that DNA/*PIF-Harbinger* polymorphisms are associated with hypermethylation in dispensable regions of cultivated foxtail millet during domestication, but not during subsequent improvement. Our results suggest that extensive increase in DNA methylation may have generated numerous epigenetic alleles of agronomically important and domestication-related genes, thereby providing potential targets for artificial selection during domestication.

Results

Comparative methylome analysis reveals epigenetic shifts during domestication

To characterize genome-wide DNA methylation patterns across wild and domesticated foxtail millet, we selected 60 representative accessions, comprising 20 wild, 20 landrace, and 20 cultivar accessions, based on the phylogenetic relationships and population history of 110 accessions with available pangenome sequences[28] (Fig. 1a,b, Supplementary Fig. 1, and Supplementary Data 1). The morphological characteristics of these accessions, including plant architecture, panicle morphology, and seed size, were described in detail in a previous study (Fig. 1 of He et al.[28]), providing phenotypic context for the genetic materials analyzed here. We subsequently generated a total of 342 Gb of high-quality WGBS data from these 60 accessions, yielding an average of 42.8 ± 3.83 million reads and a mean sequencing depth of 31.6 ± 2.84 per accession (Supplementary Data 2). To evaluate DNA methylation in each accession while minimizing potential reference bias associated with the use of a single linear reference genome (Yugu1)[35,36], we aligned WGBS reads to three types of reference genomes: (i) pseudogenomes constructed from population variants, (ii) individual genome assemblies, and (iii) a graph pangenome (Supplementary Fig. 2). When using pseudogenomes or individual genome assemblies, wild accessions exhibited substantially lower unique

https://doi.org/10.1038/s41467-026-79325-8



mapping rates than their domesticated counterparts (pseudogenomes: 60.9 ± 4.4% in wild *vs.* 72.8 ± 2.8% in domesticated accessions; individual genome assemblies: 65.0 ± 7.5% in wild *vs.* 74.6 ± 3.2% in domesticated accessions; Supplementary Figs. 3g, h). This bias is likely attributable to sequence divergence between wild and domesticated groups, as well as the higher heterozygosity of wild accessions, which can compromise assembly quality and reduce mappability (Supplementary Figs. 3a–f). In contrast, alignment to the graph pangenome using MethylGrapher[37] not only eliminated this bias but also improved unique mapping rates across all three groups (90.9 ± 2.3% in wild *vs.* 91.9 ± 1.3% in domesticated accessions; Supplementary Fig. 3i and Supplementary Data 2). All subsequent methylation analyses were therefore performed using data generated through the graph pangenome approach.

Fig. 1 | DNA methylome landscapes in 60 globally collected foxtail millet accessions. a Geographic distribution of 110 accessions. Green, orange, and purple circles denote wild, landrace, and cultivar accessions, respectively. Filled and open circles represent the 60 accessions used in this study and the other 50 accessions, respectively. The geographical map was obtained from Google Maps (https://www.google.com/maps) using the R package ggmap. **b** Maximum-likelihood phylogenetic tree of 110 accessions. Branches are colored to represent different groups. Filled circles indicate the 60 accessions analyzed in this study. **c** Ratios of read-covered cytosines (cCytosines) in the CG, CHG, and CHH contexts across wild, landrace, and cultivar groups. Light and dark blue bars indicate uncovered cytosine sites and cCytosines, respectively. **d** Proportions of methylated cytosines (mCytosines) in the three contexts across the three groups. **e** Methylation levels of core, dispensable, and private cytosines across the three groups. **f** Percentages of differentially methylated cytosines (dmCytosines) in the three sequence contexts across the three groups. **g** Numbers of core and dispensable cytosines in the three groups. Light and dark blue indicate core and dispensable cytosines, respectively; gray indicates dispensable cytosines that are absent in specific groups. The table presents the number of cytosines in each dispensable category. **h** Methylation levels of 10 core and six dispensable loci specifically present in wild or landrace/cultivar accessions. Each locus was validated by bisulfite PCR in three wild and three landrace/cultivar accessions, with 7–15 sequenced clones per accession. **i** Bar plot showing total methylation levels determined by LC-MS/MS in four landrace and four wild accessions. In **d**–**f** and **i**, bars represent mean values, and error bars indicate the standard error of the mean (SEM). Each group contains 20 samples. One-way ANOVA was used to assess differences among groups, followed by LSD multiple comparisons, with Bonferroni correction. Different lowercase letters above the bar plots indicate significant differences ($P \le 0.05$). In **h**, significance was determined using a two-sided Wilcoxon rank-sum test. Box edges and centerlines represent the interquartile range and median, respectively; whiskers extend to the most extreme points within 1.5× IQR. Source data are provided as a Source Data file.

Next, we quantified DNA methylation at 26.9 ± 1.35 million CG, 27.0 ± 1.11 million CHG, and 101.3 ± 3.41 million CHH cytosine sites covered by ≥4 reads (hereafter referred to as cCytosines; Fig. 1c, Supplementary Fig. 4, and Supplementary Data 3). We observed an overall increase in DNA methylation in domesticated groups compared with their wild progenitor (ANOVA, $P = 9.03 \times 10^{-3}$ for combined C; Fig. 1d). We then classified cytosines into three categories according to their occurrence across the population: (i) Core cytosines: defined as sites covered by ≥4 reads in ≥60% of accessions across all three groups (wild, landrace, and cultivar), representing conserved genomic regions shared across the population; (ii) Dispensable cytosines: defined as sites covered by ≥ 4 reads in ≥60% of accessions within one or two groups, representing regions associated with shared structural variations within the population. (iii) Private cytosines: defined as sites covered by ≥4 reads in <60% of accessions within any group, representing highly variable regions that may have emerged recently or been lost during evolution. Our analysis revealed that dispensable cytosines exhibited significantly higher DNA methylation levels (47.6%) than core (27.9%) and private (9.04%) cytosines (Fig. 1e). This finding is consistent with our previous observation that 85.1% of structural variations in foxtail millet are associated with TE sequences[28]. The methylation level of dispensable cytosines was significantly higher in the landrace and cultivar groups than in the wild group (48.4% and 49.1% vs. 45.3%; ANOVA, $P = 9.88 \times 10^{-9}$; Fig. 1e). In contrast, no significant differences were observed in the methylation levels of core or private cytosines among the three groups (Fig. 1e). We found that domesticated accessions contained 13.8 million dispensable cytosines, compared with only 5.6 million in wild accessions (Fig. 1g). In addition, 77.2% of dispensable cytosines in domesticated accessions were associated with TEs, compared with only 55.9% in wild accessions (Supplementary Fig. 5a). In contrast, a higher proportion of gene-associated dispensable cytosines was observed in wild than in domestication accessions (12.0% vs. 2.04%), and these cytosines were enriched in genes involved in plant immune defense and oxidative stress responses (Supplementary Fig. 5b). Consistent with these observations, pairwise kinships estimated from dmCG and dmCHG alleles were strongly correlated with kinships calculated from genetic variants, including single-nucleotide polymorphisms (SNPs) and presence/absence variations (PAVs), in both dispensable and core regions (Pearson's correlation, $r = 0.73$, $P < 2.2 \times 10^{-16}$ for dmCG *vs.* PAVs and $r = 0.70$, $P < 2.2 \times 10^{-16}$ for dmCHG *vs.* PAVs; Supplementary Fig. 6). In contrast, kinship based on dmCHH alleles showed only a weak correlation with those based on PAVs in dispensable regions (Pearson's correlation, $r = -0.15$, $P < 3.77 \times 10^{-10}$). The strong correlation between epigenetic and genetic kinships indicates that the methylation variants identified here are not stochastic but are closely coupled with genetic differentiation between wild and domesticated accessions. Together with the increased methylation levels observed specifically in dispensable regions, these findings suggest that the genome-wide increase in DNA methylation in domesticated accessions is mainly driven by methylation changes within TE-associated dispensable genomic regions.

To validate DNA methylation levels at core and dispensable cytosines, we performed bisulfite Sanger sequencing of cloned polymerase chain reaction (PCR) products from 10 core and six dispensable loci, including three domestication-specific and three wild-specific dispensable loci. The sequencing results confirmed stable methylation levels at shared core cytosines and hypermethylation at group-specific cytosines (Fig. 1h). In parallel, liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis of four domesticated and four wild accessions revealed no significant difference in global methylation levels between the two groups (Fig. 1i), likely because the overall increase in DNA methylation in domestication group was relatively small (0.92%; Fig. 1d). Taken together, these results indicate that dispensable cytosines in domesticated accessions are the major contributors to methylation divergence between wild and domesticated groups.

Evolutionary dynamics of cytosine methylation variation during foxtail millet domestication

To further characterize the distribution and frequency of DNA methylation variation across the wild and domesticated foxtail millet groups, we identified differentially methylated cytosines (dmCytosines) defined as methylated cytosines (mCytosines) present in more than one but fewer than 59 of the 60 accessions. For this analysis, we included cytosines in the CG, CHG, and CHH contexts that were covered by ≥4 reads in ≥60% of accessions within at least one group, thereby capturing cytosines located on PAVs. Among the three methylation contexts, 98.9% of mCHH sites were differentially methylated (dm) among individuals, whereas the proportions of dmCG (87.3%) and dmCHG (89.7%) sites were slightly lower (Fig. 1f and Supplementary Fig. 7). We observed significantly more differentially methylated cytosines in the CG and CHG contexts in landrace and cultivar accessions than in wild accessions (ANOVA, $P = 4.04 \times 10^{-13}$ and $P = 1.35 \times 10^{-10}$; Fig. 1f and Supplementary Fig. 7). Analysis of the site frequency spectrum of differentially methylated cytosines across all three sequence contexts revealed a U-shaped distribution of dmCG and dmCHG sites in the landrace and cultivar groups (Supplementary Fig. 8a, b). Compared with the wild group, these domesticated groups exhibited an excess of both low-frequency (minor allele frequency [MAF] ≤ 0.05) and high-frequency (MAF ≥ 0.95) alleles. For dmCG sites, the proportions were 49.1% and 50.0% in the landrace and cultivar groups, respectively, compared with 40.8% in the wild group; for dmCHG sites, the corresponding proportions were 36.0% and 36.6%, compared with 29.4% in the wild group. The accumulation of high-frequency methylated cytosines in the CG and CHG contexts in domesticated foxtail millet likely reflects the combined effects of strong artificial selection and the population bottleneck, a pattern also

observed in other crops, such as rice[38,39]. In contrast, high-frequency dmCHH sites were rare across all groups (0.915%, 0.999%, and 0.417% in the landrace, cultivar, and wild group, respectively), suggesting that such variants were either rapidly lost through genetic drift or had a low probability of reaching high frequency in the population. Furthermore, low-frequency dmCG and dmCHG sites were more strongly associated with genic regions than dmCHH sites (69.8% and 74.2% vs. 39.8%; Supplementary Fig. 8c, d). By contrast, high-frequency differentially methylated cytosines were predominantly enriched in repetitive regions (73.0%, 80.7%, and 88.3% for CG, CHG, and CHH, respectively), with dmCG and dmCHG sites preferentially located in LTR retrotransposons (45.5% and 51.7%, respectively), whereas dmCHH sites were preferentially associated with DNA transposons (70.0%). These observations suggest that epimutations across all three sequence contexts are predominantly deleterious and have been subject to purifying selection, whereas the distinct genomic distributions of high-frequency variants point to context-specific evolutionary dynamics during domestication.

DMR landscape across foxtail millet domestication and improvement

Previous studies have shown that TEs account for ~60–85% of PAVs in plant genomes, including those of *A. thaliana*, broomcorn millet, and foxtail millet[28,35,40]. Given that most TEs are heavily methylated, DNA methylation analyses restricted to shared genomic regions may fail to capture variation arising from TE polymorphisms and other PAVs. We therefore analyzed DMRs located in both shared genomic regions (corresponding to core cytosines) and PAVs-associated regions (corresponding to dispensable cytosines). In total, we identified 76,815 DMRs, the vast majority of which (98.6%) were detected during domestication (landrace *vs.* wild), whereas only 1.4% were associated with the subsequent improvement phase (cultivar *vs.* landrace) (Fig. 2a, b, Supplementary Fig. 9, and Supplementary Table 1).

In contrast to the prevalence of hypomethylated DMRs reported during the domestication of other crops, such as soybean[41], maize[20,42], and rice[43], DMRs identified between domesticated foxtail millet and its wild progenitor were predominantly hypermethylated in domesticated accessions (83.6%; Supplementary Data 4). A total of 88.3% of DMRs in dispensable regions were hypermethylated, whereas comparable proportions of hypo- and hypermethylated DMRs were observed in core regions (51.9% *vs.* 48.1%). We also found that CG, CHG, and CHH DMRs exhibited distinct chromosomal distributions. CHH DMRs were concentrated in gene-rich euchromatic regions (Fig. 2a–c), whereas CG and CHG DMRs were distributed more uniformly across chromosomes (Fig. 2c and Supplementary Figs. 9a, b). Genic regions accounted for a higher proportion of CG and CHG DMRs (37.4% and 29.7%, respectively) than of CHH DMRs (20.8%) in core regions. By contrast, DMRs in dispensable regions were predominantly located in repetitive regions, with 55.6% of CG and 61.5% of CHG DMRs overlapping with retrotransposons, whereas 91.9% of CHH-DMRs overlapped with DNA transposons (Fig. 2d). Together, these results indicate that hypermethylation of these TE-associated dispensable regions contributes to the increased DNA methylation observed in domesticated foxtail millet.

To dissect the potential functional impact of these DMRs, we overlapped them with genomic regions showing signatures of artificial selection (top 5% of genetic diversity (π), F_{ST}, and *XP-CLR*) based on SNP data from the 60 accessions (Supplementary Fig. 10 and Supplementary Table 2). CHH DMRs showed the strongest enrichment in selective sweeps, with 9.45% overlapping regions of reduced π, representing a 2.15-fold enrichment over the genome-wide average of 4.40%, compared with 5.66% for CG DMRs and 5.83% for CHG DMRs. Genes associated with CG and CHG DMRs were significantly enriched in functions related to stress resistance and secondary metabolism, including disease resistance (NB-ARC, PF00931.17; hypergeometric distribution test, adjusted $P = 2.12 \times 10^{-2}$) and the triterpenoid biosynthetic process (GO:0016104; adjusted $P = 2.29 \times 10^{-3}$) (Fig. 2e and Supplementary Data 5). In contrast, genes with altered CHH methylation were enriched in developmental and energy-related processes, including *KNOX1/2* domains (PF03790 and PF03791; adjusted $P = 2.11 \times 10^{-2}$) and carbohydrate metabolism (GO:0005975; adjusted $P = 4.28 \times 10^{-3}$) (Fig. 2e and Supplementary Data 5).

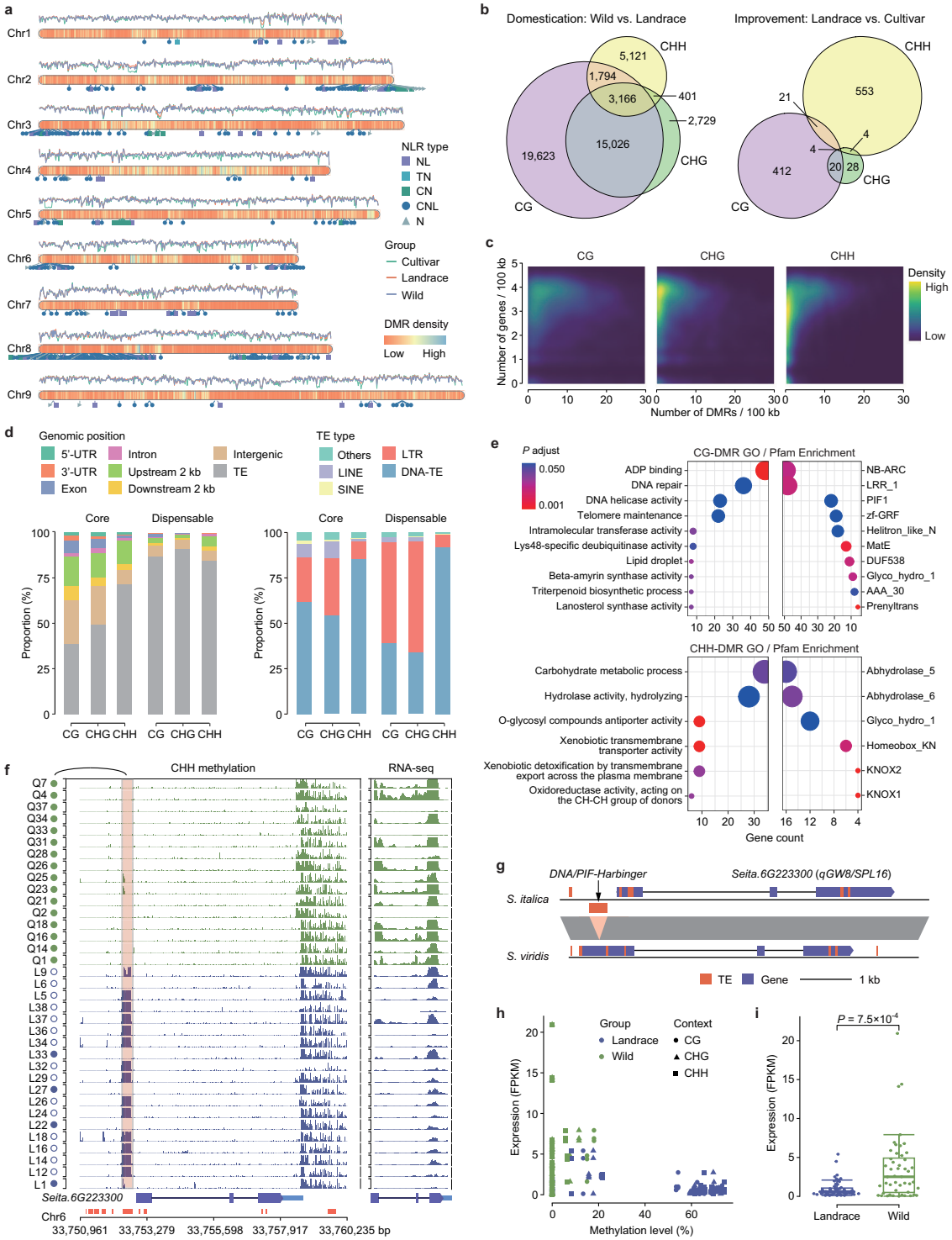
Several examples illustrate how TE-associated methylation changes may affect gene regulation. On chromosome 2, a ~171-kb region containing 14 clustered nucleotide-binding leucine-rich repeat (NLR) receptor genes exhibited extensive DNA methylation and structural variation (Supplementary Figs. 11a, b). Of the 15 DMRs identified in this region, eight overlapped with repetitive sequences spanning ≥90% of their length, and seven overlapped with PAVs spanning ≥90% of their length (Supplementary Fig. 11c). Increased CG methylation across NLR gene bodies was associated with transcriptional silencing (Supplementary Figs. 11d–f). In some cases, DMRs arose directly from PAVs: a highly methylated region present in the domesticated accession Yugu1 was absent from the wild accession A10, resulting in a hypermethylated DMR in Yugu1 (Supplementary Figs. 11g, h). Another example is *SiSPL16* (*Seita.6G223300*), the foxtail millet ortholog of rice *qGW8/SPL16*, a known regulator of grain width and yield[44,45]. We identified a CHH DMR in its 5′ UTR of *SiSPL16* that was tightly associated with a DNA/*PIF-Harbinger* insertion in domesticated accessions (Fig. 2f, g). The insertion and its associated hypermethylation were accompanied by significantly reduced *SiSPL16* expression (two-sided Wilcoxon rank-sum test, $P = 7.5 \times 10^{-4}$; Fig. 2h,i). Taken together, these results demonstrate that the marked increase in DNA methylation during foxtail millet domestication can alter the transcriptional regulation of protein-coding genes, with potential consequences for the evolution of domestication traits.

TE amplification underlies hypermethylated DMRs in domesticated foxtail millet

Having established that domesticated foxtail millet exhibits a genome-wide increase in DNA methylation, particularly in TE-associated dispensable regions, we next investigated the contribution of specific TE superfamilies to this methylation variation. To determine whether PAVs directly contribute to DMR formation, we performed a local association analysis between the methylation status of DMRs and the genotypes of PAVs within 2-kb windows (Supplementary Fig. 12a). Of the 76,815 DMRs identified, 66.5% (51,053/76,815) had at least one neighboring PAV (Supplementary Fig. 12e). Using a linear regression model, we identified 44,239 high-confidence DMR-PAV associations (false discovery rate [FDR] adjusted $P \le 0.05$ based on permutation tests). Dispensable DMRs were significantly enriched for PAV associations compared with core DMRs (chi-square test, $P < 2.2 \times 10^{-16}$; Supplementary Fig. 12f). Among PAV-associated dispensable DMRs, CHH DMRs showed the highest overlap with polymorphic TE insertions (65.0%), substantially exceeding that of CG (40.3%) and CHG (42.7%) DMRs (Supplementary Fig. 12g). Similarly, 37.6% of CG, 47.1% of CHG, and 71.9% of CHH core DMRs were located within TEs that were fixed in both domesticated and wild groups (Supplementary Fig. 12g). These results indicate that TE-associated PAVs, particularly those affecting the CHH context, are major drivers of DNA methylation differences during foxtail millet domestication.

To identify the TE superfamilies that contributed most to these DMRs, we compared the observed and expected numbers of major TE superfamilies within DMRs. Among the seven most abundant TE superfamilies (Fig. 3b), DNA/*PIF-Harbinger* transposons showed the strongest enrichment in dispensable DMRs, particularly in the CHH context (11.2-fold), whereas LTR/*Copia* and LTR/*Gypsy* retrotransposons were overrepresented in CG and CHG dispensable DMRs

https://doi.org/10.1038/s41467-026-79325-8



(1.57-fold and 1.77-fold, respectively). These TE superfamilies also exhibited distinct genomic distributions: DMRs associated with LTR retrotransposons were predominantly intergenic (80.7% for *Copia* and 86.2% for *Gypsy*; Fig. 3c), whereas those associated with DNA/*PIF-Harbinger* transposons were frequently located near or within genes (62.4% in genic regions or within 2 kb of genes; Fig. 3c), consistent with the enrichment of CHH DMRs in gene-proximal regions (Supplementary Fig. 13).

Consistent with their contribution to DMRs, these TE superfamilies have undergone substantial expansion in domesticated accessions. Landrace and cultivar genomes were significantly larger than those of wild accessions (405.74 ± 1.49 Mb and 405.31 ± 1.27 Mb vs. 396.31 ± 2.23 Mb; ANOVA, $P = 2.5 \times 10^{-4}$; Supplementary Fig. 14a–c),

Fig. 2 | Genomic distribution and functional characterization of differentially methylated regions (DMRs). a Genome-wide landscape of CHH methylation levels and DMR density. Green, orange, and purple curves represent mean CHH methylation levels in cultivar, landrace, and wild populations, respectively. Heatmaps along the chromosomes show the density of CHH DMRs. NLR genes with distinct domain combinations (C, CC domain; N, NB-ARC domain; L, LRR domain; T, TIR domain) are labeled with different shapes and colors. **b** Venn diagram of DMRs identified during domestication and improvement in the CG, CHG, and CHH contexts. **c** Correlation between gene density and DMR density (CG, CHG, and CHH) within 100-kb genomic windows. **d** Genomic distribution of core and dispensable DMRs. Regions are classified as exons, introns, 5′ UTRs, 3′ UTRs, 2-kb upstream regions, 2-kb downstream regions, TEs, and intergenic regions. TEs are further subdivided into LTR retrotransposons (LTR-TEs), DNA transposons (DNA-TEs), LINEs, SINEs, and other types. **e** GO and Pfam enrichment analyses of genes associated with CG and CHH DMRs. The x-axis shows the number of annotated genes; colors represent adjusted *P* values from the hypergeometric test. **f** Hypermethylated DMR at the *Seita.6G223300* (*SiSPL16*) locus associated with repressed gene expression. Left, CHH methylation levels in wild (green) and landrace (blue) accessions. Right, corresponding RNA-seq expression levels. Gene and TE annotations are shown below the tracks. Open and closed circles to the left of the methylation tracks indicate the presence (PAV present) and absence (PAV absent) of the structural variation, respectively. The arrow denotes the transcriptional direction. **g** Sequence comparison of the *Seita.6G223300* (*SiSPL16*) locus between *S. italica* and *S. viridis* reveals the insertion of a DNA/*PIF-Harbinger* transposon. **h** Dot plot showing the correlation between CHH methylation level and expression level (FPKM) at the *Seita.6G223300* (*SiSPL16*) locus. **i** Expression levels of *Seita.6G223300* in 16 wild and 17 landrace accessions, each with three biological replicates. Significance was determined using a two-sided Wilcoxon rank-sum test. Box edges and centerlines represent the interquartile ranges and medians, respectively; whiskers extend to the most extreme points within 1.5× IQR. Blue and green dots represent landrace and wild accessions, respectively. Source data are provided as a Source Data file.

a difference attributable to increased copy numbers of LTR/*Gypsy* (4.5–5.0 Mb or 1949–2165 additional copies), LTR/*Copia* (2.3 Mb or 1704 additional copies), and DNA/*PIF-Harbinger* elements (2.2 Mb or 5805 additional copies) (Supplementary Fig. 14d–f).

Given its strong enrichment in CHH DMRs and proximity to genes, we focused on the DNA/*PIF-Harbinger* superfamily for further analysis. Across the 60 pangenomes, these elements exhibited two predominant size peaks at ~270 bp and ~340 bp (Fig. 3d), corresponding to 10 major sequence clusters (clusters 0–9; Fig. 3e). The copy numbers of clusters 1, 2, 3, 5, 7, and 8 were significantly higher in landrace and cultivar accessions than in wild accessions (ANOVA, $P \leq 0.05$, fold change ≥1.5) (Fig. 3f), whereas cluster 4 was more abundant in the wild group (ANOVA, $P \leq 0.05$, fold change ≥ 1.5; Fig. 3e, f and Supplementary Table 3). A substantial proportion (52.7–89.1%) of elements in these clusters overlapped with dispensable DMRs that associated with TE insertions (Fig. 3g). Together, these results suggest that two distinct waves of DNA/*PIF-Harbinger* amplification accompanying foxtail millet domestication contributed substantially to the genome-wide increase in DNA methylation observed in domesticated accessions.

TE-driven siRNA waves and concurrent DNA methylation remodeling in foxtail millet

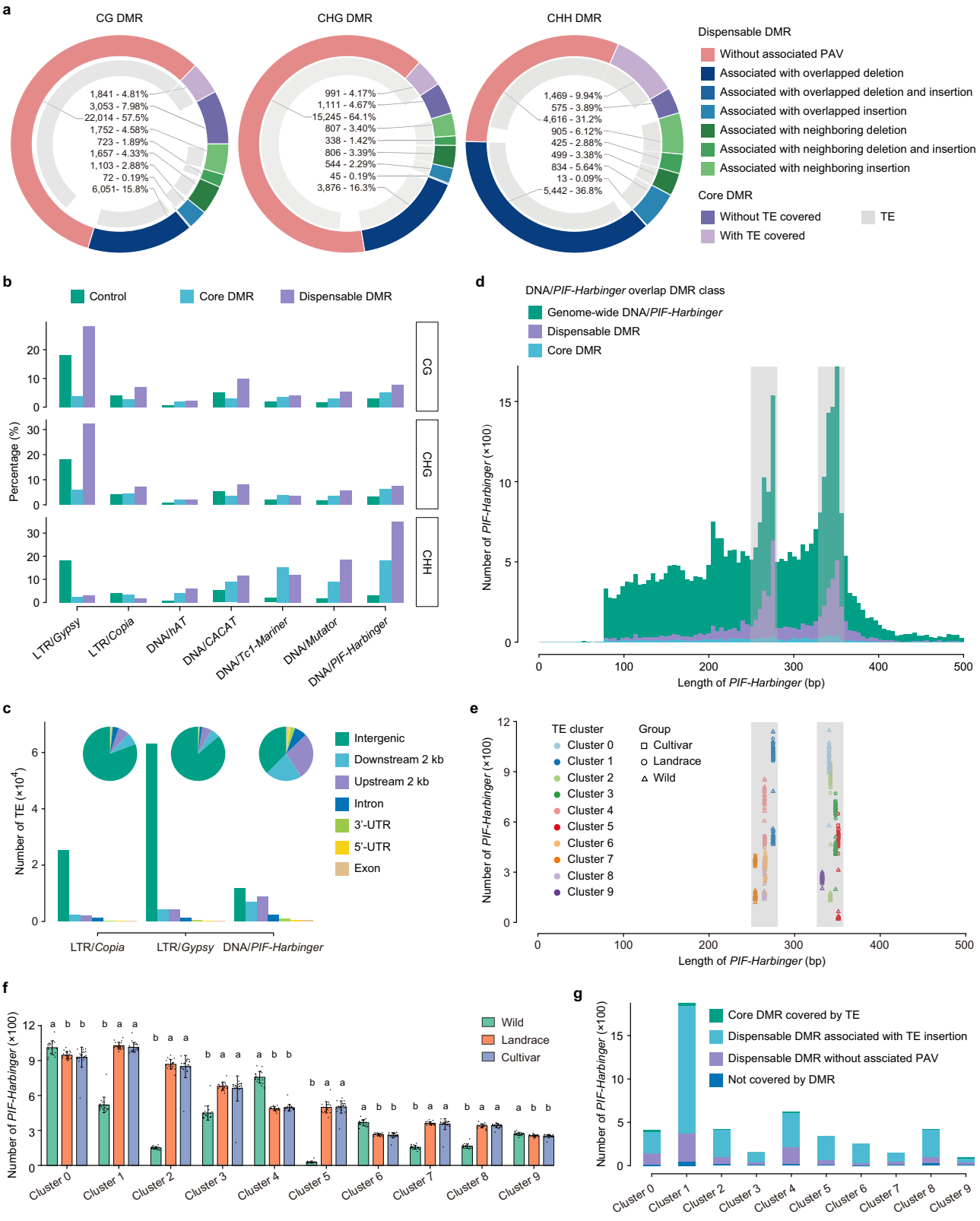
In plants, DNA methylation at miniature inverted repeat transposable elements (MITEs) is mainly guided by 24-nt siRNAs through the canonical RdDM pathway[46]. To investigate how recent TE transposition has reshaped the small RNA landscape during foxtail millet domestication, we generated small RNA libraries from nine wild and nine landrace accessions, with two biological replicates per accession. To minimize reference bias, reads were mapped to the individual genome assembly of each accession. The average alignment rate across these libraries was 80.7% ± 7.01%, with 65.7% ± 6.78% of aligned reads mapping to multiple locations (Supplementary Fig. 15a and Supplementary Data 6), suggesting that a substantial fraction of siRNAs originated from repetitive sequences. Across all accessions, 24-, 21-, and 22-nt siRNAs were the three most abundant size classes (Supplementary Fig. 15d and Supplementary Data 6). Although landrace accessions had significantly more multi-mapping reads than wild accessions (two-sided Wilcoxon rank-sum test, $P = 2.1 \times 10^{-4}$; Supplementary Fig. 15b), they had significantly fewer 24-nt siRNA reads than wild accessions (two-sided Wilcoxon rank-sum test, $P = 4.6 \times 10^{-7}$; Supplementary Fig. 15c). These findings suggest that, in addition to 24-nt siRNAs, other siRNA size classes may contribute to the increased DNA methylation observed in landrace accessions.

To compare siRNA populations more directly, we collapsed alignable reads into a set of unique-context siRNAs (u-siRNAs; Supplementary Fig. 15e). Among these, 79.2% of 24-nt u-siRNAs were derived from TEs, compared with only 44.9% of 21-/22-nt u-siRNAs (Fig. 4a). When ranked by abundance, 24-nt u-siRNAs accumulated significantly faster in landrace than in wild accessions (two-sided Wilcoxon rank-sum test, $P = 1.1 \times 10^{-5}$; Fig. 4b and Supplementary Fig. 16a), whereas 21-/22-nt u-siRNAs showed no significant difference between the two groups (Fig. 4b and Supplementary Fig. 16b). Merging all 24-nt u-siRNAs across the 36 libraries identified 101,080 differentially expressed u-siRNAs ($|FC| \geq 2$, adjusted $P \leq 0.05$), representing 18.6% of the total. Highly abundant 24-nt u-siRNAs were more likely to be upregulated in landrace accessions and exhibited higher multi-mapping rates (Fig. 4c), patterns not observed for 21-/22-nt u-siRNAs (Supplementary Fig. 16c). These contrasting patterns suggest that 24-nt and 21-/22-nt siRNAs are subject to distinct biogenesis and regulatory mechanisms.

To understand how siRNA-mediated DNA methylation was established during TE amplification, we analyzed the siRNA distribution across the 10 most abundant DNA/*PIF-Harbinger* clusters. Clusters corresponding to the 270-bp peak (clusters 1, 4, 6, 7, and 8) shared 70.3% of their 24-nt u-siRNAs and 91.1% of their 21-/22-nt u-siRNAs (Fig. 4d and Supplementary Fig. 16d), indicating extensive sequence similarity and shared siRNA production. Clusters corresponding to the 340-bp peak (clusters 0, 3, 5, 9, and 2) showed slightly lower levels of sharing u-siRNAs, with 68.9% of 24-nt and 85.0% of 21-/22-nt u-siRNAs shared among clusters. Cluster 2 was particularly divergent, sharing only 31.1% of 24-nt u-siRNAs and 58.8% of 21-/22-nt u-siRNAs with other 340-bp clusters (Fig. 4d and Supplementary Fig. 16d). Among the cluster 2-specific u-siRNAs, 48.9% of 24-nt and 85.5% of 21-/22-nt u-siRNAs mapped to multiple locations (>32 sites) (Fig. 4d and Supplementary Fig. 16d). Moreover, cluster 2 generated substantially more 21-/22-nt u-siRNAs than other 340-bp clusters (Supplementary Fig. 16f). These findings suggest that TEs within the 340-bp peak underwent rapid sequence evolution during amplification, accompanied by dynamic shifts in the siRNA populations targeting these newly evolved TE lineages.

To visualize siRNA distribution along the TE consensus sequences, we plotted read coverage across the two TE amplification peaks (Supplementary Figs. 16e, f). All 270-bp clusters showed a similar distribution pattern, with abundant 24-nt siRNAs at both ends of the TE elements and very few 21-/22-nt siRNAs in these regions. In contrast, the 340-bp clusters exhibited heterogeneous siRNA distributions, with cluster 2 uniquely enriched for 21-/22-nt siRNAs relative to 24-nt siRNAs. Cluster-specific 21-/22-nt and 24-nt siRNAs mapped predominantly to sequence regions that were divergent from those of other clusters (Fig. 4e and Supplementary Fig. 17). Phylogenetic analysis further revealed that elements corresponding to the 340-bp peak formed distinct clades, with cluster 2 elements forming a separate clade from those of other clusters (Fig. 4f), whereas elements within the 270-bp peak were more intermingled (Fig. 4g). Elements from the 340-bp clusters exhibited substantially lower DNA methylation levels than those from the 270-bp clusters, particularly in the CHH context

https://doi.org/10.1038/s41467-026-79325-8



(Fig. 4h, i). This observation, together with the enrichment of 21-/22-nt siRNAs in cluster 2, raises the possibility that these recently amplified elements are targeted by a non-canonical RdDM pathway that may silence TEs at the post-transcriptional level without fully establishing CHH methylation, a phenomenon previously described in Arabidopsis for newly inserted TEs[47,48].

TE-mediated transcriptional changes in wild and domesticated foxtail millet

To investigate how TE insertions with altered methylation influence gene expression during domestication, we first established a non-redundant set of TE-associated PAVs (TE-PAVs). Using a reciprocal overlap threshold of ≥90% between a PAV and a full-length annotated TE (Supplementary Fig. 18), we identified 41,443 TE-PAVs, representing

Fig. 3 | DMRs derived from active DNA transposon insertions. a Proportions of dispensable DMRs classified according to their association with overlapping PAVs, neighboring PAVs, or no associated PAV, and core DMRs with or without TE coverage. Outer bars represent different DMR classes; inner bars indicate whether DMRs overlap with repeat regions. **b** Expected (control) and observed proportions of DMRs associated with different TE superfamilies in the CG, CHG, and CHH contexts. Green, blue, and purple bars denote the expected (control) proportions, core DMR proportions, and dispensable DMR proportions across seven TE superfamilies, respectively. **c** Genomic distribution of DNA/*PIF-Harbinger*, LTR/*Copia*, and LTR/*Gypsy* elements. **d** Length distributions of DNA/*PIF-Harbinger* elements associated with dispensable DMRs (purple), core DMRs (blue), and genome-wide (green). **e** Copy numbers of the 10 most abundant DNA/*PIF-Harbinger* clusters in 60 accessions. The x-axis shows the mean length of each cluster; the y-axis shows the copy number per accession. Ten distinct colors represent the 10 clusters; squares, circles, and triangles represent cultivar, landrace, and wild populations, respectively. **f** Copy numbers of the 10 most abundant DNA/*PIF-Harbinger* clusters across the 60 accessions. Bars represent mean values, and error bars indicate the SEM. Significance was determined using one-way ANOVA followed by LSD multiple comparisons with Bonferroni correction. *P* values for clusters 1, 2, 4, 5, 6, 7, and 8 were $<2.2 \times 10^{-16}$; for clusters 0, 3, and 9, they were 2.5×10^{-4}, 1.3×10^{-14}, and 4.6×10^{-5}, respectively. Different lowercase letters above the bar plots indicate significant differences ($P \leq 0.05$). **g** Proportions of DMRs overlapping the 10 most abundant DNA/*PIF-Harbinger* clusters. TE positions are categorized as core DMRs (green), dispensable DMRs without associated PAVs (purple), dispensable DMRs with associated TE insertions (light blue), and regions not covered by DMRs (dark blue). Source data are provided as a Source Data file.

31.0% of all PAVs. Among these, 66.3% (27,467) were DNA/*PIF-Harbinger* elements, and 14,934 of them belong to the 10 most abundant DNA/*PIF-Harbinger* clusters (Fig. 5a). We further included 1718 DNA/*PIF-Harbinger* insertions fixed in the population, yielding in a total of 16,652 DNA/*PIF-Harbinger* elements. Hereafter, these DNA/*PIF-Harbinger* insertions are referred to as DNA/*PIF-Harbinger* insertion polymorphisms (TIPs).

TIPs were classified into three categories: low frequency (59.9%; MAF ≤ 0.05), segregating (29.8%; MAF > 0.05 and <1), and fixed (10.3%) (Fig. 5b). The 10 most abundant DNA/*PIF-Harbinger* clusters were predominantly composed of low-frequency TIPs (47.0–70.4%), with the exception of cluster 9 (14.0%) (Fig. 5c). Population distribution analysis revealed that low-frequency TIPs from clusters 0, 2, 3, and 5 (340-bp peak) and cluster 1 (270-bp peak) were predominantly found in landrace or cultivar accessions, suggesting recent amplification of these elements in domesticated groups (Fig. 5d and Supplementary Fig. 19a). In contrast, low-frequency TIPs from clusters 4, 6, 7, and 8 (270-bp peak) were enriched in the wild group (Fig. 5d and Supplementary Fig. 19a). Thus, these low-frequency TIPs represent TE families that have been recently active in either wild or domesticated groups, providing an opportunity to assess how TE polymorphisms influence gene expression during domestication.

To directly assess the transcriptional impact of these TE insertions, we performed RNA-seq on 16 wild and 17 landrace accessions, each with three biological replicates. We focused on 3,924 TIPs located within 2 kb of a gene and present in at least three accessions with the TE insertion and three without, enabling differential expression analysis between TE-present and TE-absent genotypes. This analysis identified 272 differentially expressed genes (DEGs; two-sided Wilcoxon rank-sum test; |FC| ≥ 2, adjusted $P \leq 0.05$), of which 171 showed reduced expression in the presence of a TE insertion (TE-suppressed), whereas 101 showed increased expression (TE-activated) (Fig. 5e). Genes whose expression was unaffected by nearby TIPs (non-TIP-DEGs; 3,071 genes) served as controls. TE-suppressed genes exhibited the lowest expression levels in TE-present samples, whereas TE-activated genes showed significantly lower basal expression in TE absent samples than the other two gene classes (Supplementary Fig. 19b). These findings suggest that TE-activated genes may have inherently lower expression levels and potentially function in redundant or conditional roles, such as stress responses.

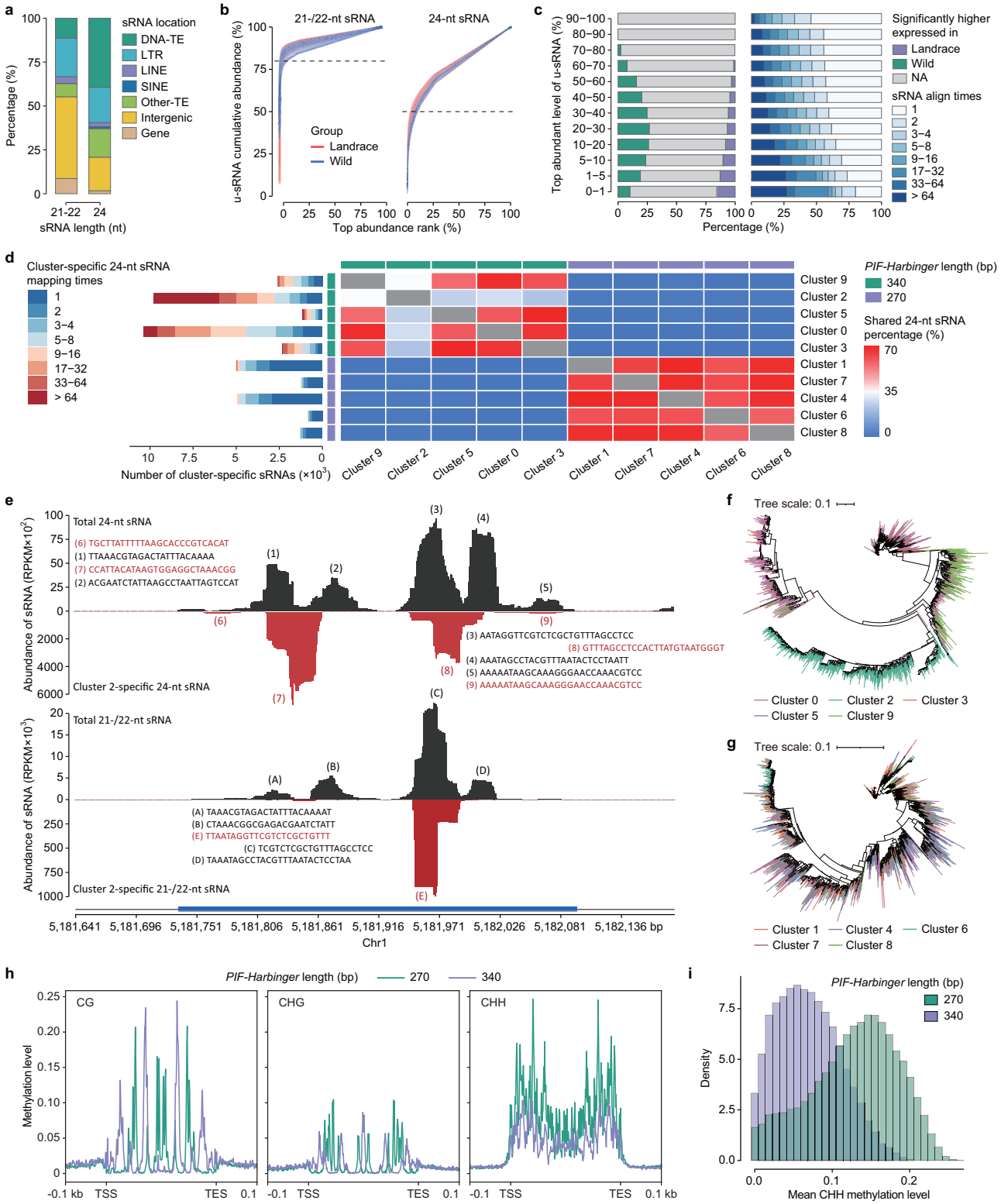
Functional enrichment analysis revealed contrasting biological roles between the two classes (Fig. 5f,g and Supplementary Data 7 and 8). TE-suppressed genes were enriched for transcription factors and abiotic stress responses, including *qGW8*|*OsSPL16*, *OsSPL18*, and *OsYABBY3*|*OsYAB4*, whereas TE-activated genes were frequently associated with immune responses, including *Xb3*, *OsGH3-5*, and *OsMSRMK2*. TE-mediated activation may involve changes in local chromatin organization, disruption of repressive elements, or the provision of novel *cis*-regulatory sequences[49,50]. We found that *Seita.9G312000* (*SiRIN13*), the foxtail millet ortholog of Arabidopsis *RPM1 Interacting Protein 13* (*RIN13*), showed a significant increase in expression from wild to landrace accessions (Supplementary Fig. 19c). This upregulation was associated with a DNA/*PIF-Harbinger* insertion in its promoter (Supplementary Figs. 19d–f). The TE insertion not only increased basal expression but also altered its gene's responsiveness to immune elicitors (Supplementary Figs. 19g, h). These findings suggest that the TE may have introduced novel regulatory elements that enhanced the gene's role in immune responses during domestication.

TE-induced DMRs associated with domestication and agronomically important traits in foxtail millet

Crops often exhibit convergent morphological changes during domestication, a phenomenon known as domestication syndrome[51]. These domestication-related traits include reduced seed shattering, increased seed number and weight, and altered flowering time[52]. Recent studies have revealed that many of these traits are controlled by parallel genetic changes, such as *Sh1* in sorghum, rice, and maize[33], and *Kernel Row Number 2* (*KRN2*) in maize and rice[53]. Given the prevalence of parallel domestication in cereals, we asked whether TE insertions have contributed to this convergent genetic basis. Examination of known domestication genes revealed that causal variants in rice are mostly not related to TE (Supplementary Data 9). In maize, 5 of 14 domestication-related variants involve TE insertions that disrupt gene structure or provide distal regulatory elements (Supplementary Data 9). Analysis of orthologs in foxtail millet revealed that 21 of 36 domestication-associated genes identified in rice and maize harbor a TE insertion within 2 kb of their foxtail millet orthologs, highlighting the potential contribution of TEs to the formation of domestication traits in this crop. By integrating published quantitative trait loci (QTLs), selection sweeps, genome-wide association study (GWAS) signals, and DMRs[28,54], we identified 710 candidate regions spanning 32.3 Mb and containing 3728 protein-coding genes. Among these genes, 1932 harbored adjacent DMRs, and 383 were differentially expressed between wild and landrace accessions. These analyses identified 59 candidate genes with known functions associated with domestication traits in other crops for further analysis (Fig. 6a, Supplementary Figs. 20 and 21, Supplementary Table 4, and Supplementary Data 10). Previously reported domestication genes in foxtail millet, including *SiLes1*[27], *SiSh1*[31], and *OsPRR37*[34,55], were successfully identified, further supporting the reliability of our approach.

Previous study reported that TE insertions at two loci, *SiLes1*[27] and *SiSh1*[31], contribute to the non-shattering phenotype in foxtail millet. Analysis of 110 *Setaria* pangenomes confirmed that a DNA/*PIF-Harbinger* insertion in the second exon of *SiSh1* is associated with reduced expression (Supplementary Figs. 22g–j). Consistent with the previous findings[31], the insertion is fixed in all landrace and cultivar accessions, suggesting that it arose as an early domestication mutation (Supplementary Figs. 21 and 22k). For *SiLes1*, an LTR/*Copia* insertion in the MY domain was previously related to the non-shattering phenotype[27]. Our analysis confirmed that this insertion suppresses

https://doi.org/10.1038/s41467-026-79325-8



SiLes1 expression (Supplementary Figs. 22a–d) and is present in 15 of 22 C1 accessions from high-latitude China and 20 of 25 C2 accessions from warmer, lower-latitude China (Supplementary Figs. 21 and 22e), suggesting that it arose during the geographic expansion of foxtail millet.

We further identified a highly divergent region on chromosome 3 containing the candidate shattering gene, *sh4*[56] (Supplementary Fig. 23). A DNA/*PIF-Harbinger* insertion in the 5′ UTR of *Seita.3G020800* (*SiSh4*) was associated with increased DNA methylation and reduced expression (Supplementary Figs. 23a–d). This insertion was present in 19 of 22 C1, 24 of 25 C2, and 3 of 5 C3 accessions (Supplementary Fig. 23e), suggesting that it may contribute to the non-shattering phenotype. Two CHH DMRs associated with DNA/*PIF-Harbinger* insertions were identified in the promoter of another shattering gene, *Seita.5G381300*, ortholog of *qSH1*[57] (Supplementary Fig. 23g). Pangenome analysis revealed four independent insertions at

Fig. 4 | sRNA landscape during the DNA/*PIF-Harbinger* amplification event. **a** Genomic locations of 21-/22-nt and 24-nt siRNA reads. Genomic locations are divided into DNA transposons (DNA-TEs), LTR retrotransposon (LTR-TEs), LINEs, SINEs, other TE classes (Other-TEs), genic regions, and intergenic regions. **b** Accumulation curves of 21-/22-nt and 24-nt siRNA reads in nine wild and nine landrace accessions. The x-axis, from left to right, represents unique sRNA reads ranked from the most abundant to the least abundant, while the y-axis represents the accumulation level of total sRNA reads in each accession. **c** Proportions of differentially expressed 24-nt unique siRNAs (DES; left) and alignment times of 24-nt unique siRNAs (right). The y-axis, from bottom to top, represents 24-nt unique siRNAs ranked from the most abundant to the least abundant. Green and purple bars represent the proportions of DES with significantly higher expression in wild and landrace accessions, respectively (left). Blue shading represents the number of alignments of siRNAs (right). **d** Heatmap of shared 24-nt siRNA proportions for the 10 most abundant DNA/*PIF-Harbinger* clusters. Purple blocks represent 270-bp clusters, whereas green blocks represent 340-bp clusters. The bar plot to the left of the heatmap indicates the number of alignments of cluster-specific 24-nt unique siRNA reads. **e** Read mapping of total (black) and cluster 2-specific (red) 21-/22-nt, and 24-nt siRNAs to a DNA/*PIF-Harbinger* cluster 2 element. **f** Phylogenetic trees of 340-bp DNA/*PIF-Harbinger* elements. Branch colors represent different clusters. **g** Phylogenetic trees of 270-bp DNA/*PIF-Harbinger* elements. Branch colors represent different clusters. **h** CG, CHG, and CHH methylation levels of 340-bp and 270-bp DNA/*PIF-Harbinger* elements across the TE body. **i**, Distribution of mean CHH methylation levels of 340-bp and 270-bp DNA/*PIF-Harbinger* elements. Source data are provided as a Source Data file.

this locus, collectively associated with a 2.71-fold reduction in expression in landrace accessions (Supplementary Figs. 23h–j). The second insertion, located in the first exon, was present in all landrace and cultivar accessions except L31 (Supplementary Figs. 21 and 23k), suggesting that it may have contributed to the transition to non-shattering early in domestication.

Ideal tiller number is a key domestication trait that influences agricultural productivity[58]. Domesticated foxtail millet typically produces a single tiller bearing one panicle. We identified several candidate tillering genes associated with DMRs. A region of high genetic divergence (F_{ST}) on chromosome 9 co-localized with QTLs for panicle number and grain number per primary branch (Supplementary Fig. 24a). Within this region, we identified orthologs of *OSH3* and *OSH1*/*KN1*. *Seita.9G109500* (*SiOSH1*/*KN1*) harbors three hypermethylated DMRs in its 5′ UTR and intron, each associated with a TE insertion (Supplementary Figs. 24b,c). The first two insertions are linked and present in all C1, C2, and C3 accessions except L15, whereas the third is absent from C3 but present in 11 of 22 C1 and 23 of 25 C2 accessions (Supplementary Fig. 24f). These insertions significantly reduced gene expression in landrace accessions (two-sided Wilcoxon rank-sum test, $P = 6.9 \times 10^{-10}$; Supplementary Figs. 24d, e). For *Seita.9G109300* (*SiOSH3*), two DNA/*PIF-Harbinger* insertions were detected in the first exon and third intron (Supplementary Figs. 24g, h). The insertion in the first exon insertion was present in all landrace and cultivar accessions except L15 and L32 (Supplementary Fig. 24j) and was associated with reduced expression (two-sided Wilcoxon rank-sum test, $P = 4.2 \times 10^{-13}$; Supplementary Fig. 24i). *Class 1 KNOTTED 1-like homeobox* (*KNOX1*) genes play key roles in axillary meristem development[59]. Reduced *OSH1* expression is associated with reduced tiller number in rice mutants[60], and *OSH1* has undergone strong selection during popcorn maize domestication[61]. The repressed expression and fixed status of these two genes in domesticated foxtail millet make them strong candidates for contributing to reduced tiller number during domestication. We further characterized another KNOX family member, *OSH6*, which showed reduced expression in domesticated accessions and harbored DNA/*PIF-Harbinger* insertions in its intron (Supplementary Fig. 25). Collectively, these findings suggest that DNA/*PIF-Harbinger* insertions, accompanied by increased DNA methylation, contributed to reduced tillering and altered panicle architecture during foxtail millet domestication.

The transition from prostrate to erect growth, controlled by *PROG1* (*PROSTRATE GROWTH 1*)[62], represents another key domestication trait. In foxtail millet, we identified a tandem duplication of the *PROG1* region, that increased its copy number to five (*SiPROG1a–e*) (Supplementary Figs. 26a, b). This region exhibited reduced nucleotide diversity and genetic differentiation (Supplementary Figs. 26c, d) and co-localized with QTLs for grain number, tiller number, and panicle weight, as well as GWAS signals for branch number.

For *SiPROG1a*, two TE insertions flanking the gene were present in all landrace and cultivar accessions and were associated with hypermethylation and reduced expression (Supplementary Figs. 26e–i). *SiPROG1b* underwent an inversion event and acquired a DNA/*CMC-EnSpm* insertion (Supplementary Fig. 27b). The inversion, which was fixed in all landrace and cultivar accessions, reversed the orientation of the gene, whereas the TE insertion converted the functional promoter into a highly methylated region (Supplementary Fig. 27a). Despite these epigenetic changes, no significant difference in expression was detected between landrace and wild accessions (two-sided Wilcoxon rank-sum test, $P = 0.91$; Supplementary Fig. 27d). Additional TE-related variants were detected at other copies, including two TE insertions at *SiPROG1c* and a deletion at *SiPROG1e*, both of which were associated with altered expression in landrace compared with wild accessions (Supplementary Figs. 27g–l). Together, these observations suggest that TE-mediated epigenetic changes at the *PROG1* locus, particularly at *SiPROG1a* and *SiPROG1b*, may have contributed to the evolution of an erect growth habit in foxtail millet, although further functional studies are required to establish causality.

SiGW3 is a pleiotropic gene that shapes plant architecture and panicle structure in foxtail millet

Genes with pleiotropic effects on agronomically important traits have been repeatedly selected during crop domestication, as documented in maize and wheat[19,63,64]. In a previous study, we showed that *SiGW3* was selected during foxtail millet domestication and was associated with increased grain size, one of the most pronounced differences between domesticated foxtail millet and its wild progenitor[28]. Here, our integrative analysis revealed that the *SiGW3* locus co-localized not only with grain size signals but also with GWAS signals for branch number, and QTLs for 1000-grain weight, grain width, leaf width, leaf length, and peduncle length (Fig. 6a,b). These findings suggest that *SiGW3* may be a pleiotropic gene contributing to multiple morphological changes during domestication.

To dissect the contribution of *SiGW3* to agronomic trait variations, we compared two previously generated overexpression (OE) lines[28] with two CRISPR/Cas9 (CR) genome edited lines generated in this study (Fig. 6c,d; CR-*SiGW3*-L1, carrying a 1-bp deletion, and CR-*SiGW3*-L2, carrying a 1-bp insertion), all in the Ci846 cultivar background. Consistent with previous findings[28], OE plants exhibited reduced grain size and 1000-grain weight compared with wild-type (WT) plants. In contrast, CR plants produced significantly shorter and wider grains and had increased 1000-grain weight (Supplementary Figs. 28f,g). These results indicate that the TE insertions do not abolish *SiGW3* function but rather downregulate its expression. Beyond grain traits, we observed significant differences in primary panicle diameter and tiller number between CR and WT plants (Fig. 6d,e), consistent with GWAS signals linking *SiGW3* to plant architecture. OE plants exhibited a larger primary panicle diameter, increased tiller number, smaller main stem diameter, shorter stature, and fewer stem nodes, resembling wild *S. viridis* and correlating with high *SiGW3* expression in the shoot apical meristem and panicle (Supplementary Figs. 29a–c). Together, these results establish *SiGW3* as a pleiotropic gene that

https://doi.org/10.1038/s41467-026-79325-8

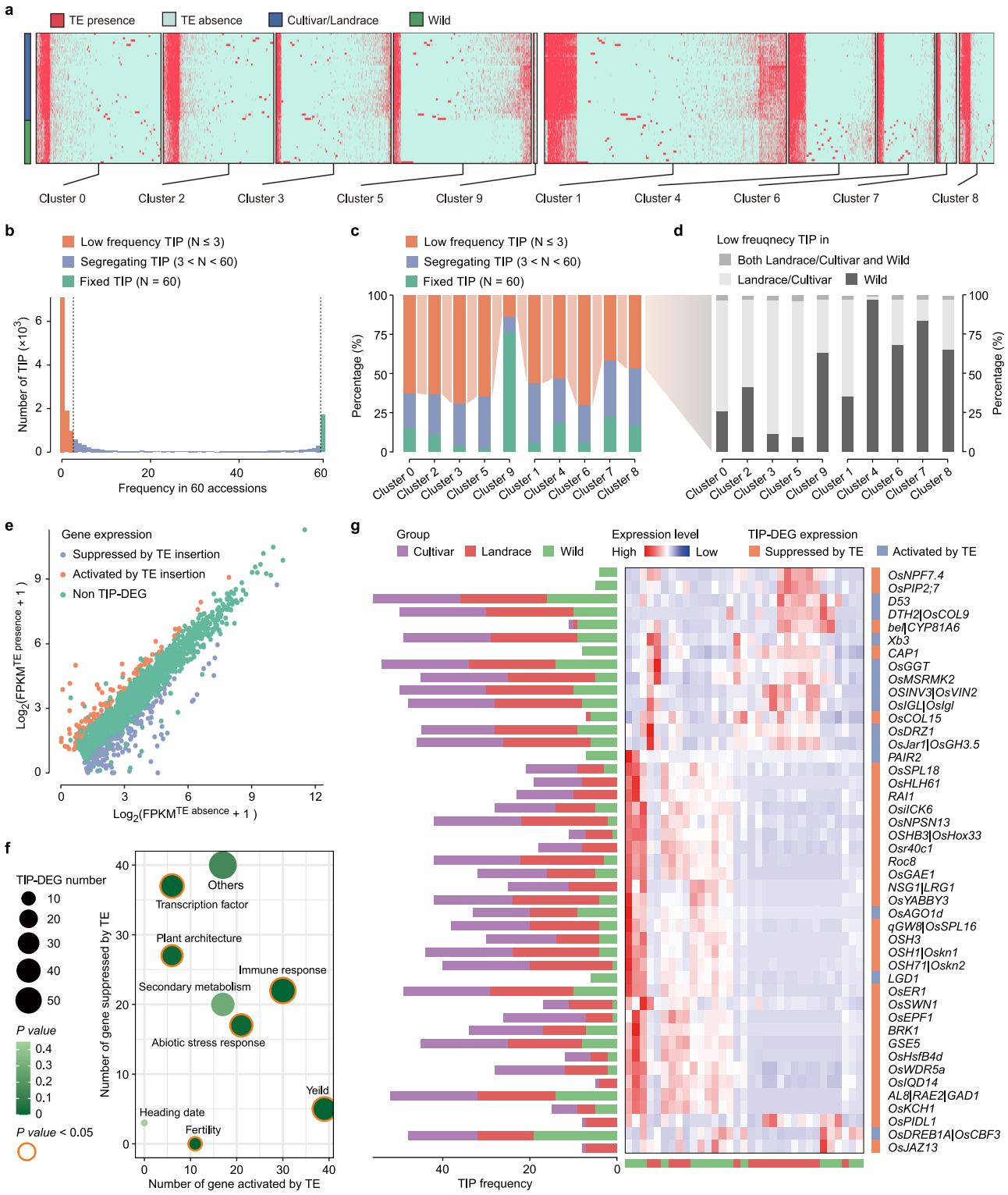
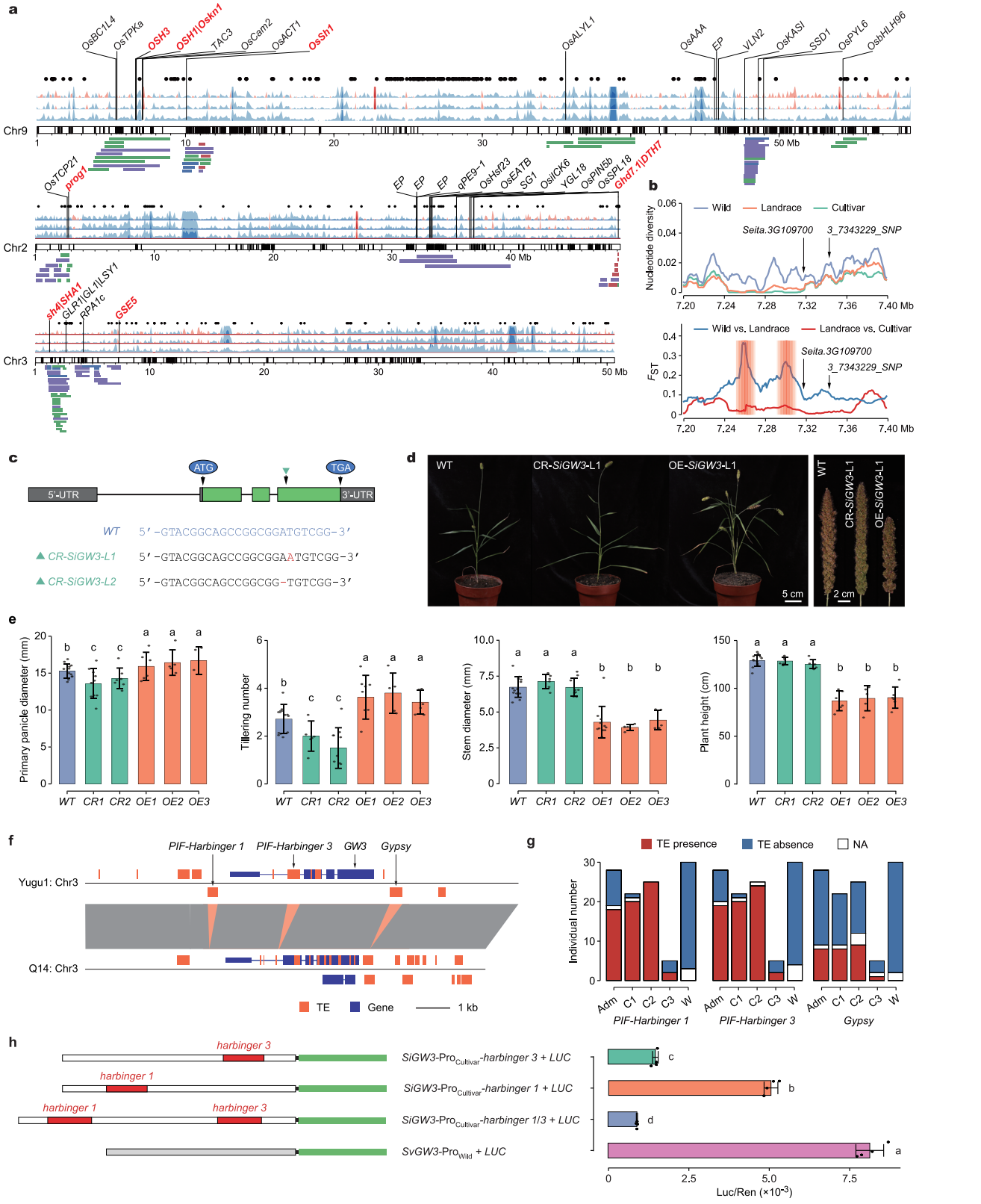


Fig. 5 | Gene expression regulated by TE insertions. a Insertion polymorphism of the 10 most abundant DNA/*PIF-Harbinger* clusters. Rows of the heatmap represent the 20 wild (green), 20 landrace (blue), and 20 cultivar (blue) accessions. Each column represents one TE insertion. Red blocks indicate the presence of a TE, whereas green blocks indicate TE absence. **b** Insertion frequencies of DNA/*PIF-Harbinger* elements across the 60 accessions. Orange, blue, and green bars represent TIPs with low-frequency TIPs ($n \leq 3$), segregating TIPs ($3 < n < 60$), and TIPs fixed in the *Setaria* population ($n = 60$), respectively. **c** Insertion frequency composition of the 10 most abundant DNA/*PIF-Harbinger* clusters. **d** Proportion of low-frequency TIPs among the 10 most abundant DNA/*PIF-Harbinger* clusters present in the wild or landrace/cultivar groups. **e** Gene expression levels in TE-absent samples (x-axis) vs. TE-present samples (y-axis). Orange, blue, and green dots represent genes with suppressed expression, activated expression, or unchanged expression (non-TIP-DEG) associated with TE insertions, respectively. **f** Putative functional differences between genes with expression activated or suppressed by TE insertions. The x-axis and y-axis represent the number of genes with expression activated and suppressed by TE insertions, respectively. Red circle borders indicate significant differences based on two-sided Pearson's chi-squared tests ($P \leq 0.05$). **g** Expression heatmap of 45 TIP-DEGs with cloned rice orthologs. The colors of the heatmap columns indicate the wild (green) and landrace (red) groups, whereas the colors of the heatmap rows indicate genes with expression activated (blue) or suppressed (orange) by TE insertions. The bar plot to the left of the heatmap indicates TE insertion frequencies in the wild (green), landrace (red), and cultivar (purple) groups. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-026-79325-8



simultaneously shapes grain morphology and plant architecture in foxtail millet.

In a previous study, we identified a 366-bp deletion ~7.2 kb upstream of *SiGW3* that co-segregated with grain size variation, suggesting that this distal sequence may repress *SiGW3* expression and thereby increase grain weight in foxtail millet[28]. Here, we identified five TE polymorphisms within 2 kb of *SiGW3* (Supplementary Figs. 28a–e) and found that two DNA/*PIF-Harbinger* insertions (*Harbinger* 1 and *Harbinger* 3) co-segregated with reduced *SiGW3* expression (Fig. 6f). These two insertions were present in all C2 accessions and in 20 of 22 C1 accessions. They were absent in L1, L20, and L3 (C3 subgroup in Heilongjiang Province, Northeast China) and in L2 (C1 subgroup in Jilin Province, Northeast China) (Fig. 6g), suggesting that their expansion

Fig. 6 | Domestication traits associated with TE insertion in foxtail millet. **a** Domestication genes affected by DMRs on chromosomes 2, 3, and 9. Horizon plots, from top to bottom, indicate CG, CHG, and CHH methylation differences between wild and cultivated populations. Black regions represent selection regions. Dots above the horizon plots indicate GWAS signals. Colored bars below the chromosomes represent QTLs associated with different phenotypes. **b** Nucleotide diversity and genetic divergence (F_{ST}) in the *SiGW3* region. **c** Two independent *SiGW3* CRISPR-edited lines. Genome-edited positions are marked by green triangles. **d** Plant architecture and panicle shape of WT (Ci846), CR-*SiGW3*-L1, and OE-*SiGW3*-L1 plants. **e**, Primary panicle diameter (n = 12, 10, 7, 6, 5, 4 for WT, CR1–2, and OE1–3, respectively), tiller number (n = 12, 6, 10, 8, 5, 6 for WT, CR1–2, and OE1–3, respectively), stem diameter (n = 13, 6, 9, 11, 5, 5 for WT, CR1–2, and OE1–3, respectively), and plant height (n = 13, 6, 6, 6, 5, 6 for WT, CR1–2, and OE1–3, respectively) of WT (Ci846), CR, and OE plants. **f** Sequence comparison of the *SiGW3* region between Yugu1 and Q14. **g** TE presence frequencies of DNA/*PIF-Harbinger* 1, DNA/*PIF-Harbinger* 3, and LTR/*Gypsy* in admixed (Adm), cultivar (C1, C2, and C3), and wild (W) groups. **h** LUC promoter assay. The white, gray, and green bars represent the cultivar *SiGW3* promoter, wild *SiGW3* promoter, and the LUC gene, respectively. Red blocks indicate *Harbinger* insertions at the *SiGW3* locus. The bar plot indicates luciferase activity from four different *SiGW3*/*SvGW3* promoter constructs (n = 4 biologically independent experiments). In **e** and **h**, bars represent mean values, and error bars indicate the SEM. One-way ANOVA was used to assess significant differences, followed by LSD multiple comparisons with Bonferroni correction. P values for the comparisons of primary panicle diameter, tiller number, stem diameter, and plant height were 5.35×10^{-3}, 1.94×10^{-7}, 1.5×10^{-12}, and 1.2×10^{-14}, respectively; the P value for luciferase activity was 6.03×10^{-14}. Different lowercase letters above the bar plots indicate significant differences ($P \le 0.05$). Source data are provided as a Source Data file.

on the *SiGW3* allele occurred after the initial spread of foxtail millet into Northeast China.

To functionally assess the effects of these insertions on gene expression, we performed luciferase (LUC) reporter assays in *Nicotiana benthamiana* (Fig. 6h). Both the *Harbinger* 1 element (*SiGW3*-$Pro_{Cultivar}$-*Harbinger* 1) and the *Harbinger* 3 element (*SiGW3*-$Pro_{Cultivar}$-*Harbinger* 3) significantly repressed luciferase expression relative to the control (*SvGW3*-Pro_{Wild}). The combination of *Harbinger* 1 and *Harbinger* 3 produced the strongest repression, exceeding the effect of either element alone. These results indicate that both *Harbinger* 1 and *Harbinger* 3 elements can repress *SiGW3* expression. Bisulfite Sanger sequencing further revealed that both elements were heavily methylated (Supplementary Fig. 30), consistent with a potential role in epigenetic silencing. Together, these findings suggest that the two DNA/*PIF-Harbinger* insertions repress *SiGW3* expression, likely through an epigenetic mechanism involving DNA methylation. The resulting pleiotropic effects on grain morphology and plant architecture likely made *SiGW3* an effective target of early domestication selection, enabling a single locus to influence multiple adaptive traits. We acknowledge that our current data cannot distinguish whether DNA methylation per se or the TE sequences themselves are required for this repression, an issue that warrants further mechanistic investigation.

Discussion

Domestication is a rapid evolutionary process accompanied by dramatic morphological and physiological changes in both plants and animals. Identifying the genetic basis of these changes has revealed causal variants underlying key domestication traits, including seed shattering (Os*Sh1* and *qSH1*)[33,57] and plant architecture (*tb1* and *PROG1*)[62,65,66]. Although early studies highlighted the role of small genetic variants, emerging evidence underscores the importance of TEs in generating both genetic and epigenetic diversity. Here, we present a comprehensive methylome atlas of wild and domesticated foxtail millet, one of the earliest crops domesticated by humans. Our analysis reveals that the substantial increase in DNA methylation during domestication is largely confined to dispensable genomic regions and is closely associated with the recent amplification of DNA/*PIF-Harbinger* MITEs. These TE-derived genetic and epigenetic variations likely facilitated artificial selection during foxtail millet domestication, as exemplified by previously characterized domestication genes controlling seed shattering (*SiLes1* and *SiSh1*) and grain size (*SiGW3*)[27,28,31]. Unlike SNP-based variation, TE amplification can rapidly generate multiple insertions, providing a rich source of genetic and epigenetic variation for selection to act upon.

DNA methylation plays a crucial role in maintaining genome stability and regulating gene transcription[67], and dynamic changes in DNA methylation can influence relatively rapid evolutionary processes such as crop domestication[68]. Previous studies in soybean[41], rice[43], and maize[20] reported that most domestication-associated DMRs were hypomethylated, suggesting that relaxed selection during domestication may lead to widespread hypomethylation[69]. However, DNA hypomethylation during domestication is not universal. In lettuce, for example, domestication was accompanied by predominantly hypermethylated DMRs (82.4%), most of which were located within TE regions, implicating TE activity in the gain in DNA methylation[70]. Our findings in foxtail millet are consistent with this model but provide a potential mechanistic explanation: the widespread presence of polymorphic DNA/*PIF-Harbinger* MITEs, which serve as a source of genetic and structural variation, introduced numerous new targets for DNA methylation, counteracting the general trend toward hypomethylation and generating abundant hypermethylated DMRs across dispensable loci. These findings suggest that polymorphic TE insertions, together with other genetic variants, can fundamentally reshape the epigenetic landscape during domestication, providing raw material for selection that differs qualitatively from that generated by point mutations.

TE-driven genomic differentiation in crops has been predominantly attributed to LTR retrotransposons[23,46], with well-documented examples in maize[20,71], *Brassica rapa*[72], and hexaploid wheat[73,74]. In foxtail millet, however, we find that DNA/*PIF-Harbinger* MITEs, rather than LTRs, represent a major source of genetic and epigenetic polymorphism. These MITEs are enriched near genes and preferentially targeted by the RdDM pathway, linking TE expansion to regulatory variation. The 270-bp and 340-bp MITE families exhibited distinct population dynamics: the former was more active in wild accessions, whereas the latter underwent amplification predominantly in domesticated accessions. Cluster 2 MITEs exhibited features characteristic of recently active elements, including high sequence homology, low CHH methylation, and enrichment of 21-/22-nt siRNAs, consistent with non-canonical RdDM targeting newly inserted TEs[48]. These observations suggest that different MITE families have undergone distinct waves of amplification, with the most recently expanded cluster 2 potentially arising from other 340-bp clusters through the accumulation of mutations and subsequently undergoing host recognition and silencing via the non-canonical RdDM pathway. The relatively low DNA methylation levels of these newly inserted MITEs may minimize their deleterious effects on nearby genes, thereby creating a buffered state that facilitates regulatory innovation.

The functional impact of TE insertions is illustrated by multiple domestication genes. The loss of seed shattering, a hallmark of domestication, is associated with TE insertions at *SiSh1* (fixed in all domesticated accessions), *SiqSH1* (present in nearly all landrace and cultivar accessions), and *SiLes1* (absent from early northern landrace accessions), suggesting a stepwise acquisition of the non-shattering phenotype during foxtail millet expansion. Beyond seed shattering, our analysis identified additional candidate genes, including *SiOSH1* and *SiOSH3* for tillering and *SiPROG1a* for erect growth, all of which harbor DNA/*PIF-Harbinger* insertions and show reduced expression in domesticated accessions. The most compelling example is *SiGW3*, a pleiotropic gene whose TE-mediated repression simultaneously

reshaped grain morphology, plant height, tillering, and panicle architecture, thereby influencing multiple domestication traits through a single locus. Its early fixation during domestication exemplifies how TE insertions can generate coordinated phenotypic shifts by rewiring key developmental regulators. These findings support a model in which TE activity may have been activated in a subset of wild accessions immediately before or during domestication, triggering to a wave of MITE insertions reminiscent of the *mPing* amplification burst in temperate japonica rice[75,76].

Whether DNA/*PIF-Harbinger* MITEs remain active in the foxtail millet population remains an open question. More than 50% of TIPs are low-frequency, and cluster 2 elements, which were recently amplified in domesticated accessions, exhibit high sequence homology and features reminiscent of active *mPing* elements in rice[75,76]. However, we have not yet identified the autonomous elements responsible for mobilizing these MITEs. Such elements may occur at very low frequencies, as in rice, where *Ping* is present in only ~6.6% of accessions[76], MITE mobilization may involve cross-mobilization by distantly related autonomous DNA transposons, as reported for rice DNA/*Mariner* elements[77]. Expanding the pangenome to include additional accessions may help uncover these elusive drivers of MITE mobilization.

A second unresolved question concerns the interplay between canonical and non-canonical RdDM pathways in shaping DNA methylation at newly amplified MITEs. Despite their recent expansion in domesticated accessions, cluster 2 elements exhibit low CHH methylation while producing abundant 21-/22-nt siRNAs, a signature of non-canonical RdDM targeting newly inserted or reactivated TEs[47]. As copy number increases, this pathway may be progressively supplanted by canonical RdDM mediated by 24-nt siRNAs[47,48], leading to more stable epigenetic silencing. The functional consequences of this transition, and whether it creates a window for regulatory variation, warrant further investigation.

Collectively, our study reveals a crucial role for TE polymorphisms in shaping epigenetic and phenotypic variation, advances our understanding of how TE diversity accelerates domestication, and identifies epialleles underlying agronomic traits in foxtail millet. More broadly, our findings raise fundamental questions about the role of TEs in crop evolution: How frequently do TE insertions coincide with domestication across different species? To what extent do TE-derived regulatory variants contribute to superior agronomic traits? How do TE-derived epigenetic markers enable plastic immune and stress responses to environmental changes? Addressing these questions will require integrated analyses of pangenomes, methylomes, and transcriptomes across diverse crops and their wild relatives, coupled with functional dissection of key TE–gene interactions.

Methods

Plant materials and growth conditions

A representative panel of 60 foxtail millet accessions (20 cultivar, 20 landrace, and 20 wild accessions) was selected from the 110 pangenome accessions (35 cultivar, 40 landrace and 35 wild accessions) with high-quality genome assemblies available[28]. Sample were selected based on the phylogenetic relationships and population structure reported in the previous study[28]. The wild group included 15 W1 accessions distributed across Eurasia that represent the ancestral population of cultivated foxtail millet[28], as well as five genetically divergent accessions (one W4 and four admixed accessions) selected to capture the broader genetic diversity of the wild population. The domesticated group comprised 12 C1, 15 C2, two C3, and 11 admixed accessions, representing the genetic diversity of the cultivated population. Samples were grown in a climate chamber at the experimental station of the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, during the 2020 growing season (June to October) under controlled conditions of 28 °C and 70% relative humidity. Young leaf tissues (14 days old) were harvested for genomic DNA and total RNA extraction, which were subsequently used for WGBS, RNA-seq, and small RNA library construction.

WGBS library construction and sequencing

Genomic DNA was used to construct WGBS libraries. Bisulfite conversion was performed using the EZ DNA Methylation Gold Kit (Zymo Research). WGBS libraries were prepared using the EpiArt DNA Methylation Library Kit for Illumina V3 (Vazyme, China) and sequenced on an Illumina NextSeq 500 platform to generate 150-bp paired-end reads. The insert sizes of the WGBS libraries were 299 ± 4.6 bp (Supplementary Data 2). To assess the bisulfite conversion rate, reads were aligned to the chloroplast genome, which is largely unmethylated in plants; thus, residual cytosines were used to estimate incomplete conversion. The average bisulfite conversion rate across all WGBS libraries was 99.31 ± 0.55% (Supplementary Data 2).

Evaluation of methylation level by LC-MS/MS

High-molecular-weight genomic DNA was extracted using the Plant Genomic DNA Kit (Tiangen, China). Purified DNA samples were digested to nucleosides using Nucleoside Digestion Mix (New England Biolabs, USA) and incubated at 37 °C overnight. The resulting hydrolysates were centrifuged at 13,000 × *g* for 10 min at room temperature, and the supernatants were analyzed by LC-MS/MS using a QTRAP 6500+ triple quadrupole/linear ion trap mass spectrometer (SCIEX, USA) coupled to a liquid chromatography system. The methylation percentage was calculated using the following formula: Methylation (%) = [5mdC/(5mdC + dC)] × 100%, where 5mdC and dC represent the molar amounts of 5-methyl-2′-deoxycytidine and 2′-deoxycytidine, respectively, determined in each DNA sample.

Preprocessing of WGBS data and evaluation of DNA methylation levels

WGBS reads were preprocessed using fastp[78] v0.23.2 to remove adapter sequences and low-quality bases with the following parameters: -z 4 -q 20 -u 30. We used three approaches to analyze the preprocessed WGBS reads: a pseudogenome approach, an individual-genome approach, and a graph pangenome approach.

The pseudogenome approach: The Syri[79] v1.7.0 tool was used to compare the 60 individual genome assemblies with the Yugu1 reference genome and identify SNPs for each accession. Pseudogenomes for the 60 accessions were then constructed by substituting the identified SNPs using in-house scripts. Bismark[80] v0.14.5 was used to align WGBS reads from each accession to its corresponding pseudogenome. The 'deduplicate_bismark' function was used to remove PCR duplicates, and 'bismark_methylation_extractor' was used to determine the methylation status of each cytosine. Cytosines covered by at least four reads (≥ 4x) were considered as covered sites. Only covered cytosine sites (cCytosine or cC) were retained for methylation analysis. Methylated cytosine sites (mCytosine or mC) were identified using a binomial test based on the bisulfite non-conversion rate estimated from the foxtail millet chloroplast genome[41,81], Benjamini–Hochberg correction was then applied to control the FDR, with adjusted $P \leq 0.05$ considered significant. Cytosine sites that were methylated in at least 1 but no more than 59 of the 60 samples were defined as differentially methylated cytosine (dmCytosine or dmC) sites, representing population-level methylation variation.

The individual genome approach: To minimize the reference bias associated with the pseudogenome approach, preprocessed WGBS reads from each accession were mapped to its corresponding individual genome assembly[28]. The software, parameters, and thresholds were the same as those used for the pseudogenome approach.

The graph pangenome approach: A graph pangenome was constructed from the genome sequences of 60 accessions

using Minigraph-Cactus v3.1.2, with the Yugu1 genome serving as the linear reference[82]. Briefly, "cactus-minigraph" was used to generate the structural variation graph with default parameters. "cactus-graphmap" and "cactus-align" were then used for base-level alignment and to anchor the genome assemblies to the graph. The final variation graph was exported in Graphical Fragment Assembly (GFA) format using 'cactus-graphmap-join' with the parameters "--filter 3 --gfa filter". We then used "methylGrapher PrepareGenome" in MethylGrapher v0.2.1[37] to convert the GFA file into fully G → A and C → T converted GFA files. The filtered WGBS reads from each accession were aligned to the converted graph genome using "methylGrapher Align", and methylated cytosines were identified using "methylGrapher MethylCall" with the "-cg_only N" parameter. Methylated cytosines were identified based on a minimum coverage of 4× and a binomial test FDR ≤ 0.05, consistent with the criteria used for the pseudogenome and individual-genome approaches. Subsequently, the coordinates and methylation calls of cytosines located in regions shared among the graph pangenome were converted to the linear Yugu1 reference genome using Ixchel. Meanwhile, insertion graph paths ≥50 bp in length were extracted from the graph pangenome using in-house scripts.

Identification of DMRs in shared genomic regions

We applied Metilene[83] to identify DMRs during domestication (landrace vs. wild) and improvement (cultivar *vs*. landrace). First, the script "metilene_input.pl" was used to generate CG, CHG, and CHH methylation ratio matrices from the MethylGrapher results. Only cytosines covered by ≥4 reads were included, while uncovered cytosines were assigned as "NA". The main program 'metilene' was then used to identify DMRs, and candidate DMRs were filtered using "metilene_output.pl". Cytosines covered by ≥4 reads in at least 60% of accessions in both comparison groups were retained for DMR calling. The following parameters were used with "metilene": "-M 300" maximum distance of 300 nt between adjacent cytosines within a DMR; "-m 8" (minimum of 8 cytosines per DMR); "-d 0.4" and "-d 0.2" (minimum methylation differences of 0.4 for CG and CHG DMRs and 0.2 for CHH DMRs); and "-f 1" (de novo mode enabled). The parameters of "metilene_output.pl" were set as "-p 0.01" (*q* value cutoff of 0.01) and "-c 8" (minimum of 8 cytosines per DMR) to filter the final DMRs.

Identification of PAV-associated DMRs

To identify DMRs associated with recent TE insertions or other PAVs, we adopted an approach similar to that described above for DMRs in shared genomic regions, based on the MethylGrapher results. Briefly, cytosines covered by ≥4× reads in ≥60% of accessions in either group were defined as PAV-associated. Uncovered cytosines were assigned as "0", representing a theoretical baseline of negligible methylation that reflects their absence in some accessions. The script 'metilene' was then used to identify PAV-associated DMRs during domestication (landrace *vs*. wild) and improvement (cultivar vs. landrace), using the same parameters described above for shared genomic regions.

Detection of selection sweeps

The SNP dataset was generated from the Syri de novo comparison between the Yugu1 reference genome and 60 pangenome sequences[79]. The SNP datasets of the 60 accessions were merged into a high-quality VCF file using bcftools (MAF ≥ 0.05 and missing rate <0.1). Selection sweeps during domestication were detected by combining *XP-CLR*[84] v1.0, $\pi_{Wild}/\pi_{Landrace}$, and the F_{ST} index. The *XP-CLR*, π, and F_{ST} values were calculated using a window size of 100 kb and a step size of 10 kb. *XP-CLR* scores were calculated using the *XP-CLR* program (https://github.com/hardingnj/xpclr). The π and F_{ST} scores were calculated using VCFtools v.0.1.13. The top 5% of *XP-CLR* score, $\pi_{Wild}/\pi_{Landrace}$, and F_{ST} values were combined to define selection sweep regions by 'bedtools merge'.

Annotation, clustering, and insertion polymorphism analysis of repetitive elements

TEs were annotated using a comprehensive protocol. Extensive de novo TE Annotator (EDTA) v1.9.6[85] was applied to each of the 60 *Setaria* pangenomes with the following parameters: "-overwrite 1 -sensitive 1 -anno 1". In parallel, RepeatMasker[86] v4.0.9 was used to annotate DNA/*PIF-Harbinger* MITEs in each genome. The RepeatMasker MITE annotations were then merged with the EDTA TE annotations to generate a consensus TE annotation set. To cluster the *Setaria* DNA/*PIF-Harbinger* elements, we integrated DNA/*PIF-Harbinger* elements from the *Setaria* pangenomes and clustered them based on the 80-80-80 principle[87] using CD-HIT[88] v4.8.1 with the following parameters: -sc 1 -c 0.8 -aL 0.8 -aS 0.8 -n 5 -M 0 -T 20. PAVs were identified in 60 *Setaria* genome assemblies using Syri[79] v1.7.0, with the Yugu1 reference genome. The coordinates of the PAVs were then compared with the annotated TE boundaries. PAVs showing ≥90% reciprocal overlap with a TE were classified as TE-derived PAVs. Population-level TE-derived PAVs were merged into a non-redundant set (TE-PAVs) using in-house scripts. DNA/*PIF-Harbinger* TE insertion polymorphisms (TIPs) were defined as a subset of TE-PAVs that had undergone recent amplification in the population and were selected for further analysis. The insertion patterns of DNA/*PIF-Harbinger* TIPs across the *Setaria* population were visualized using the R package pheatmap (https://cran.r-project.org/package=pheatmap).

DMR source analysis

For each DMR, we converted the continuous methylation level in each sample into a discrete genotype. Using hierarchical clustering (hclust with method = "single") of the methylation values across all 60 accessions for each DMR, we classified each sample as either hypermethylated (coded as 1/1) or hypomethylated (coded as 0/0). This approach captures the major methylation-state pattern of each DMR across the population. For each PAV generated by Minigraph-Cactus v3.1.2, the genotype was naturally binary, representing presence (1/1) or absence (0/0). We then performed a comprehensive association analysis between all DMRs and PAVs located within a 2-kb window. For each DMR–PAV pair, we compared the PAV genotype (presence = 1/1, absence = 0/0) with the methylation status of the DMR (1/1 = hypermethylated, 0/0 = hypomethylated) across the 60 accessions. We recorded a same-genotype match when PAV presence (1/1) corresponded to DMR hypermethylation (1/1), or PAV absence (0/0) corresponded to DMR hypomethylation (0/0). Conversely, we recorded a complementary-genotype match when PAV presence (1/1) corresponded to DMR hypomethylation (0/0), or PAV absence (0/0) corresponded to DMR hypermethylation (1/1).

We tested DMR–PAV associations using a linear regression model. For each DMR–PAV pair (d, p), we fitted the following model:

$$Y_{di} = \alpha + \sum_{k=1}^{5} \beta_k \cdot PC_{ki} + \gamma \cdot G_{pi} + \varepsilon_i 1$$

where $Y_{di} \in \{0, 1\}$ is the methylation state of DMR d in individual i (coded as 0 for hypomethylation and 1 for hypermethylation); PC_1–PC_5 are the first five principal components inferred from genome-wide markers, which capture the major axes of genetic variation, $G_{pi} \in \{0, 1\}$ is the genotype of PAV p, and $\varepsilon_i \sim N(0, \sigma^2)$ is the error term. Model parameters were estimated using ordinary least squares, and the significance of the PAV coefficient γ was assessed using a two-sided t test with the null hypothesis H_0: $\gamma = 0$. To account for multiple testing and avoid distributional assumptions, we performed 1000 genotype-shuffling permutations. For each permutation, the methylation states

of every DMR were independently permuted across individuals, and the entire analytical pipeline was rerun to construct empirical null distributions. To reduce noise from stochastic *P* value fluctuations at the small sample size ($n = 60$), we used a tier-based permutation strategy, in which *t* test *P* values were grouped into seven ordered tiers: Tier 0: [0, 0.001]; Tier 1: (0.001, 0.01); Tier 2: (0.01, 0.02); Tier 3: (0.02, 0.03); Tier 4: (0.03, 0.04); Tier 5: (0.04, 0.05); and Tier 6: >0.05. A permuted pair was counted as challenging the observed association only when its tier was strictly lower than the tier of the real-data association. Empirical *P* values were then calculated from the permutation distributions, followed by Benjamini–Hochberg FDR correction (adjusted $P \le 0.05$). For core DMRs, we classified them as TE DMRs or non-TE DMRs based on whether they overlapped with any TE regions.

RNA-seq and small RNA library construction and sequencing

Total RNA was isolated from young leaf tissues of 16 wild and 17 landrace accessions using TRIzol reagent (Thermo Fisher Scientific). RNA-seq libraries were prepared for paired-end sequencing on an MGISEQ-2000 platform, generating 150-bp paired-end reads. Small RNA was extracted from total RNA isolated from young leaf tissues of 9 wild and 9 landrace accessions using the NEBNext Small RNA Library Prep Set for Illumina. The small RNA libraries were sequenced on an Illumina NovaSeq 6000 platform.

Analysis of RNA-seq sequencing

Raw reads from 16 wild and 17 landrace accessions, each with three replicates, were filtered using fastp[78] v0.23.2 with default parameters to remove adapter sequences and low-quality bases. Clean reads were mapped to the graph pangenome using 'vg mpmap' from vg tools v1.61.0[89] and then counted using rpvg v1.0[90]. Population-level differentially expressed genes were identified using the R package DESeq2[91] with |fold change| ≥2 and adjusted $P \le 0.05$. Genes with DNA/*PIF-Harbinger* TIPs within their 2-kb flanking regions were defined as TIP genes and retained for further expression analysis. DNA/*PIF-Harbinger* TIPs were defined as a subset of TE-PAVs showing ≥90% reciprocal overlap between DNA/*PIF-Harbinger* elements and PAVs. For differential expression analysis, we considered TIP genes with at least three accessions carrying the TE and three accessions lacking it. Genes showing |fold change| ≥ 2 and adjusted $P \le 0.05$ based on the Wilcoxon rank-sum test between TE-present and TE-absent accessions were defined as TIP-differentially expressed genes (TIP-DEGs).

Analysis of sRNA-seq sequencing

sRNA-seq reads from nine landrace and nine wild accessions, each with two replicates, were filtered using fastp[78] v0.23.2 with default parameters to remove adapter sequences and low-quality reads. For each library, sRNA reads were aligned to the corresponding individual genome using bowtie[92] v1.3.1 with the parameter "-k 1". Alignable sRNA reads were then merged into a unique-context sRNA (u-sRNA) dataset and sorted by read count to plot the abundance accumulation profiles. The total u-sRNA dataset from the nine landrace and nine wild accessions was aligned to the Yugu1 reference genome using bowtie[92] v1.3.1 with the parameters "-a --best --strata" to determine the number of alignments for each u-sRNA. Population-level differentially expressed u-sRNAs were identified using the R package DESeq2[91] with |fold change| ≥ 2 and adjusted $P \le 0.05$. Deeptools[93] was used to visualize sRNA read depth across the TE bodies.

Identification of domestication candidates

To identify candidate genes associated with domestication traits, we first classified 34 QTL phenotypes into four categories: grain, panicle, seed, and plant architecture. Overlapping phenotype–QTL regions identified across two years and at two or more locations were considered domestication-associated QTL regions. DMRs within the candidate QTL regions were then examined for overlap with selection sweeps or GWAS signals. Population-level differentially expressed genes that overlapped with the filtered DMRs were identified as candidate domestication genes.

Functional enrichment analysis

InterProScan v5.52-86.0 was used to annotate GO terms for Yugu1 proteins[94], with the local mode "-dp", output format "-f tsv", and the "-goterms" flag. KEGG orthology (KO) terms were annotated using KofamScan v1.3.0 against the standard KO HMM database[95]. Pfam domains were annotated using HMMER v3.3.2 with an E-value cutoff of 1×10^{-9}. Functional enrichment analysis of candidate gene sets was performed using the R package clusterProfiler[96]. Enrichment of GO, KEGG, and Pfam terms was assessed using the hypergeometric test, with *P* values adjusted using the Benjamini–Hochberg method (pAdjustMethod = BH). Terms with an adjusted $P \le 0.05$ were considered significantly enriched.

Validation of methylation variation by bisulfite Sanger sequencing on PCR products

Genomic DNA was isolated from wild and cultivated samples using the Plant Genomic DNA Kit (Tiangen, China). Bisulfite conversion was performed using the DNA Bisulfite Conversion Kit (Tiangen, China). To validate methylation changes, we randomly selected 16 loci in foxtail millet, including 10 core loci, three dispensable loci that were hypermethylated in the landrace/cultivar group, and three dispensable loci that were hypermethylated in the wild group, respectively. We also included two DNA/*PIF-Harbinger* insertions (*Harbinger* 1 and *Harbinger* 3) at the *SiGW3* locus (Supplementary Fig. 30). Target fragments were amplified by PCR using KOD-Multi & Epi-polymerase (Toyobo, Japan). PCR products were purified using the DNA Gel Extraction Kit (Tsingke, China), and the purified fragments were cloned using the Ultra-Universal TOPO Cloning Kit (Vazyme, China). For each sample, 11–20 independent colonies were sequenced. The resulting sequences were analyzed, and CG, CHG, and CHH methylation levels were estimated using Kismeth software[97].

Functional characterization of *SiGW3*

To construct *SiGW3* CRISPR/Cas9 genome-edited plants, small-guide RNAs were designed based on the sequence context of *SiGW3* using CRISPRdirect (http://crispr.dbcls.jp/). The pYLCRISPR/Cas9-MH vector was digested with a restriction enzyme to construct a plasmid containing the sgRNA[98]. To construct overexpression plants, the full-length *SiGW3* coding sequence (CDS) was amplified from cDNA of foxtail millet (cv. Yugu1). The *SiGW3* CDS was cloned into the pCAMBIA1305.1-EGFP vector to generate the *pUbi::SiGW3*-eGFP overexpression vector. Transgenic OE and CR vectors were introduced into the *Agrobacterium* EHA105 strain and subsequently transformed into foxtail millet (Ci846) calli. The primers used for vector construction are listed in Supplementary Data 11. At least three independent transgenic lines were obtained for phenotypic evaluation and statistical analysis. Important agronomic traits, including primary panicle diameter, tiller number, branch number, stem diameter, stem node number, plant height, grain length, grain width, and 1000-grain weight, were measured[32]. Approximately 150–200 grains from each plant were randomly selected, and grain length and width were measured using grain analyzer software (SC-G Scanner, Wanshen Detection Technology Inc., Hangzhou, China).

To explore the subcellular localization of *SiGW3*, the full-length *SiGW3* coding sequence was cloned into the pCAMBIA1305-eGFP vector and introduced into the p19-containing Agrobacterium strain GV3101. The plasma membrane marker FM4-64-mCherry was co-expressed with *35S::SiGW3-GFP* in *N. benthamiana* leaves by agroinfiltration. Sections were imaged using a confocal microscope (LSM700, Zeiss, Germany), and images were analyzed and processed using ZEN Microscopy Software (Zeiss, Germany). The primers used for the

subcellular localization assay are listed in Supplementary Data 11. Plant tissues were incubated in β-glucuronidase (GUS) staining buffer containing X-Gluc at 37 °C in the dark for 12 h and then destained with 70% ethanol to visualize blue precipitates[99].

To investigate the influence of TE insertions on the *SiGW3* promoter, promoter fragments of *SiGW3* and *SvGW3* with or without LTR/*Gypsy* or DNA/*PIF-Harbinger* insertions were cloned into the pGreen0800-LUC vector. The vectors were transformed into *A. tumefaciens* strain GV3101 and infiltrated into leaves of 4-week-old *N. benthamiana* plants. Firefly luciferase activity was measured using a Luciferase Reporter Gene Assay Kit (Yeasen, 11402ES60, China). At least three biological replicates were performed for each experiment.

Phylogeny analysis

To determine the phylogenetic relationships among the 110 pangenome accessions, including 60 WGBS-sequenced accessions and 50 accessions not analyzed in this study, Syri[79] v1.7.0 was used to identify SNPs using the A10 reference genome. The SNP dataset from the 110 accessions and Yugu1 was then used to infer their evolutionary relationships using SNPhylo[100] v2018901. The phylogenetic trees of the 270-bp and 340-bp Yugu1 DNA/*PIF-Harbinger* elements were constructed using FastTree[101] v2.1.11 with default parameters.

Geographical map generation

Figures showing the geographical locations of the accessions sampled worldwide in this study were generated using the ggmap package in R v4.1.0.

Statistics

Statistical analyses and figure plotting were performed using R v4.1.0 with built-in functions and third-party R packages, including tidyverse v1.3.1, ggplot2 v3.4.3, ggpubr v0.4.0, and ggpie v0.2.5. Expression levels between two groups of accessions were compared using a two-sided Wilcoxon rank-sum test. A two-sided chi-squared test was used to assess differences in TE insertion frequencies between wild and landrace/cultivar populations using the R built-in function prop.test. One-way ANOVA was used to assess differences among clusters, followed by least significant difference (LSD) multiple comparisons with Bonferroni correction. Different lowercase letters were assigned using the R package rCompanion to indicate significant differences ($P \leq 0.05$).

Data availability

The raw sequences data, including 60 WGBS, 33 RNA-seq, and 18 small RNA-seq samples, have been deposited in the National Center for Biotechnology Information Sequence Read Archive (SRA) under BioProject PRJNA768345 [https://www.ncbi.nlm.nih.gov/bioproject/768345]. A full list of SRA accessions is listed in Supplementary Data 1. Processed RNA-seq and small RNA-seq expression data have been deposited in NCBI Gene Expression Omnibus (GEO) under accession GSE347112 and GSE347113. Source data are provided in this paper. Source data are provided with this paper.

Code availability

The custom codes used in this study has been archived in Zenodo [https://doi.org/10.5281/zenodo.21662328][102] and are also available at Github [https://github.com/Chenlab-IOZ/Foxtail_millet_methylome_code].

References

1. Huang, C. R. L., Burns, K. H. & Boeke, J. D. Active transposition in genomes. *Annu. Rev. Genet.* **46**, 651–675 (2012).
2. Guerreiro, M. P. G. Interspecific hybridization as a genomic stressor inducing mobilization of transposable elements in *Drosophila*. *Mob. Genet. Elem.* **4**, e34394 (2014).
3. Hénault, M., Marsit, S., Charron, G. & Landry, C. R. The effect of hybridization on transposable element accumulation in an undomesticated fungal species. *eLife* **9**, e60474 (2020).
4. Vicient, C. M. & Casacuberta, J. M. Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* **120**, 195–207 (2017).
5. Hirochika, H., Okamoto, H. & Kakutani, T. Silencing of retrotransposons in *Arabidopsis* and reactivation by the *ddm1* mutation. *Plant Cell* **12**, 357–368 (2000).
6. Zou, L. et al. Genetic and epigenetic reprogramming in response to internal and external cues by induced transposon mobilization in moso bamboo. *N. Phytol.* **244**, 1916–1930 (2024).
7. Modzelewski, A. J., Gan Chong, J., Wang, T. & He, L. Mammalian genome innovation through transposon domestication. *Nat. Cell Biol.* **24**, 1332–1340 (2022).
8. Fultz, D., Choudury, S. G. & Slotkin, R. K. Silencing of active transposable elements in plants. *Curr. Opin. Plant Biol.* **27**, 67–76 (2015).
9. Cokus, S. J. et al. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
10. Lindroth, A. M. et al. Requirement of *CHROMOMETHYLASE3* for maintenance of CpXpG methylation. *Science* **292**, 2077–2080 (2001).
11. Yabe, K. et al. H3K9 methylation regulates heterochromatin silencing through incoherent feedforward loops. *Sci. Adv.* **10**, eadn4149 (2024).
12. Zhang, H., Lang, Z. & Zhu, J.-K. Dynamics and function of DNA methylation in plants. *Nat. Rev. Mol. Cell Biol.* **19**, 489–506 (2018).
13. He, L. et al. DNA methylation-free *Arabidopsis* reveals crucial roles of DNA methylation in regulating gene expression and development. *Nat. Commun.* **13**, 1335 (2022).
14. Raingeval, M. et al. Retrotransposon-driven environmental regulation of *FLC* leads to adaptive response to herbicide. *Nat. Plants* **10**, 1672–1681 (2024).
15. Niu, X.-M. et al. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *Proc. Natl. Acad. Sci. Usa.* **116**, 6908–6913 (2019).
16. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
17. Lisch, D. How important are transposons for plant evolution? *Nat. Rev. Genet.* **14**, 49–61 (2013).
18. Vitte, C., Fustier, M.-A., Alix, K. & Tenaillon, M. I. The bright side of transposons in crop evolution. *Brief. Funct. Genomics* **13**, 276–295 (2014).
19. 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).
20. Xu, G. et al. Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nat. Commun.* **11**, 5539 (2020).
21. Kobayashi, S., Goto-Yamamoto, N. & Hirochika, H. Retrotransposon-induced mutations in grape skin color. *Science* **304**, 982 (2004).
22. Butelli, E. et al. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell* **24**, 1242–1255 (2012).
23. Wei, L. & Cao, X. The effect of transposable elements on phenotypic variation: insights from plants to humans. *Sci. China Life Sci.* **59**, 24–37 (2016).
24. Yang, X. et al. Early millet use in northern China. *Proc. Natl. Acad. Sci.* **109**, 3726–3730 (2012).
25. Peng, R. & Zhang, B. Foxtail millet: a new model for C4 plants. *Trends Plant Sci.* **26**, 199–201 (2021).
26. Yang, Z. et al. A mini foxtail millet with an *Arabidopsis*-like life cycle as a C4 model system. *Nat. Plants* **6**, 1167–1178 (2020).

27. Mamidi, S. et al. A genome resource for green millet *Setaria viridis* enables discovery of agronomically valuable loci. *Nat. Biotechnol.* **38**, 1203–1210 (2020).
28. He, Q. et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat. Genet.* **55**, 1232–1242 (2023).
29. Wang, C. et al. Population genetics of foxtail millet and its wild ancestor. *BMC Genet* **11**, 90 (2010).
30. Hu, H., Mauro-Herrera, M. & Doust, A. N. Domestication and improvement in the model C4 grass, *Setaria*. *Front. Plant Sci.* **9**, 719 (2018).
31. Liu, H. et al. Transposon insertion drove the loss of natural seed shattering during foxtail millet domestication. *Mol. Biol. Evol.* **39**, msac078 (2022).
32. Jia, G. et al. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat. Genet.* **45**, 957–961 (2013).
33. Lin, Z. et al. Parallel domestication of the *Shattering1* genes in cereals. *Nat. Genet.* **44**, 720–724 (2012).
34. Fukunaga, K. et al. Recombinant inbred lines and next-generation sequencing enable rapid identification of candidate genes involved in morphological and agronomic traits in foxtail millet. *Sci. Rep.* **12**, 218 (2022).
35. Igolkina, A. A. et al. A comparison of 27 *Arabidopsis thaliana* genomes and the path toward an unbiased characterization of genetic polymorphism. *Nat. Genet.* **57**, 2289–2301 (2025).
36. Bao, Z. & Weigel, D. Complexity welcome: pangenome graphs for comprehensive population genomics. *Quant. Plant Biol.* **6**, e43 (2025).
37. Zhang, W. et al. methylGrapher: genome-graph-based processing of DNA methylation data from whole genome bisulfite sequencing. *Nucleic Acids Res* **53**, gkaf028 (2025).
38. Caicedo, A. L. et al. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet* **3**, e163 (2007).
39. Liu, Q., Zhou, Y., Morrell, P. L. & Gaut, B. S. Deleterious variants in Asian rice and the potential cost of domestication. *Mol. Biol. Evol.* **3**, msw296 (2017).
40. Chen, J. et al. Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nat. Genet.* **55**, 2243–2254 (2023).
41. Shen, Y. et al. DNA methylation footprints during soybean domestication and improvement. *Genome Biol.* **19**, 128 (2018).
42. Eichten, S. R. et al. Epigenetic and genetic influences on DNA methylation variation in maize populations. *Plant Cell* **25**, 2783–2797 (2013).
43. Cao, S. et al. Asymmetric variation in DNA methylation during domestication and de-domestication of rice. *Plant Cell* **35**, 3429–3443 (2023).
44. Wang, S. et al. Control of grain size, shape and quality by *OsSPL16* in rice. *Nat. Genet.* **44**, 950–954 (2012).
45. Wang, S. et al. The *OsSPL16-GW7* regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Nat. Genet.* **47**, 949–954 (2015).
46. Song, X. & Cao, X. Transposon-mediated epigenetic regulation contributes to phenotypic diversity and environmental adaptation in rice. *Curr. Opin. Plant Biol.* **36**, 111–118 (2017).
47. Marí-Ordóñez, A. et al. Reconstructing de novo silencing of an active plant retrotransposon. *Nat. Genet.* **45**, 1029–1039 (2013).
48. Nuthikattu, S. et al. The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. *Plant Physiol.* **162**, 116–131 (2013).
49. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).
50. Horváth, V., Merenciano, M. & González, J. Revisiting the relationship between transposable elements and the eukaryotic stress response. *Trends Genet* **33**, 832–841 (2017).
51. Abbo, S. et al. Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. *Trends Plant Sci.* **19**, 351–360 (2014).
52. Chen, Q., Li, W., Tan, L. & Tian, F. Harnessing knowledge from maize and rice domestication for new crop breeding. *Mol. Plant* **14**, 9–26 (2021).
53. Chen, W. et al. Convergent selection of a WD40 protein that enhances grain yield in maize and rice. *Science* **375**, eagb7985 (2022).
54. Zhi, H. et al. Genetic control and phenotypic characterization of panicle architecture and grain yield-related traits in foxtail millet (*Setaria italica*). *Theor. Appl. Genet.* **134**, 3023–3036 (2021).
55. Li, C. et al. High-depth resequencing of 312 accessions reveals the local adaptation of foxtail millet. *Theor. Appl. Genet.* **134**, 1303–1317 (2021).
56. Li, C., Zhou, A. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936–1939 (2006).
57. Konishi, S. et al. An SNP caused loss of seed shattering during rice domestication. *Science* **312**, 1392–1396 (2006).
58. Wang, B., Smith, S. M. & Li, J. Genetic regulation of shoot architecture. *Annu. Rev. Plant Biol.* **69**, 437–468 (2018).
59. Tsuda, K., Kurata, N., Ohyanagi, H. & Hake, S. Genome-wide study of *KNOX* regulatory network reveals Brassinosteroid catabolic genes important for shoot meristem function in rice. *Plant Cell* **26**, 3488–3500 (2014).
60. Tabuchi, H. et al. *LAX PANICLE2* of rice encodes a novel nuclear protein and regulates the formation of axillary meristems. *Plant Cell* **23**, 3276–3287 (2011).
61. Fang, X. et al. Genome assembly and population genomic analysis reveal the genetic basis of popcorn evolution. *Plant Biotechnol. J.* **23**, 2911–2927 (2025).
62. Jin, J. et al. Genetic control of rice plant architecture under domestication. *Nat. Genet.* **40**, 1365–1369 (2008).
63. Zhang, Z. et al. Duplication and partitioning in evolution and function of homoeologous *Q* loci governing domestication characters in polyploid wheat. *Proc. Natl. Acad. Sci.* **108**, 18737–18742 (2011).
64. Dixon, L. E. et al. *TEOSINTE BRANCHED1* regulates inflorescence architecture and development in bread wheat (*Triticum aestivum*). *Plant Cell* **30**, 563–581 (2018).
65. Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386**, 485–488 (1997).
66. Tan, L. et al. Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* **40**, 1360–1364 (2008).
67. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
68. Becker, C. et al. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**, 245–249 (2011).
69. Björnerfeldt, S., Webster, M. T. & Vilà, C. Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome Res* **16**, 990–994 (2006).
70. Cao, S., Sawettalake, N., Li, P., Fan, S. & Shen, L. DNA methylation variations underlie lettuce domestication and divergence. *Genome Biol.* **25**, 158 (2024).
71. Clark, R. M., Wagler, T. N., Quijada, P. & Doebley, J. A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. *Nat. Genet.* **38**, 594–597 (2006).
72. Cai, X. et al. Transposable element insertion: a hidden major source of domesticated phenotypic variation in *Brassica rapa*. *Plant Biotechnol. J.* **20**, 1298–1310 (2022).

73. Kashkush, K., Feldman, M. & Levy, A. A. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* **33**, 102–106 (2003).
74. Zhang, Y. et al. Transposable elements orchestrate subgenome-convergent and -divergent transcription in common wheat. *Nat. Commun.* **13**, 6940 (2022).
75. Lu, L. et al. Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. *Proc. Natl. Acad. Sci. Usa.* **114**, E10550–E10559 (2017).
76. Chen, J. et al. Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. *Nat. Commun.* **10**, 641 (2019).
77. Yang, G., Nagel, D. H., Feschotte, C., Hancock, C. N. & Wessler, S. R. Tuned for transposition: molecular determinants underlying the hyperactivity of a *Stowaway* MITE. *Science* **325**, 1391–1394 (2009).
78. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**, e107 (2023).
79. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
80. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
81. Hofmeister, B. T., Lee, K., Rohr, N. A., Hall, D. W. & Schmitz, R. J. Stable inheritance of DNA methylation allows creation of epi-genotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome Biol.* **18**, 155 (2017).
82. Hickey, G. et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat. Biotechnol.* **42**, 663–673 (2024).
83. Jühling, F. et al. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res* **26**, 256–262 (2016).
84. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res* **20**, 393–402 (2010).
85. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
86. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **25**, 4.10.1–4.10.14 (2009).
87. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **18**, 973–982 (2007).
88. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
89. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
90. Sibbesen, J. A. et al. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nat. Methods* **20**, 239–247 (2023).
91. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
92. 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).
93. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160–W165 (2016).
94. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
95. Kanehisa, M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
96. Xu, S. et al. Using clusterProfiler to characterize multiomics data. *Nat. Protoc.* **19**, 3292–3320 (2024).
97. Gruntman, E. et al. Kismeth: analyzer of plant methylation states through bisulfite sequencing. *BMC Bioinforma.* **9**, 371 (2008).
98. Tang, S. et al. An E2-E3 pair contributes to seed size control in grain crops. *Nat. Commun.* **14**, 3091 (2023).
99. Jefferson, R. A. The GUS reporter gene system. *Nature* **342**, 837–838 (1989).
100. Lee, T.-H., Guo, H., Wang, X., Kim, C. & Paterson, A. H. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**, 162 (2014).
101. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
102. Guo, W & Chen, J. Transposable element polymorphisms underlie epigenetic variation and phenotypic diversification during foxtail millet domestication. *Zendo*. https://doi.org/10.5281/zenodo.21662328 (2026).

Acknowledgements

We are grateful to the National Crop Genebank of China (Institute of Crop Sciences, CAAS) for providing the *Setaria* germplasm used in this study. We also acknowledge the Beijing Annoroad Medical Laboratory Co., Ltd. and Igenecode Co., Ltd for sequencing services, and the computational support provided by the High-Performance Computing Platform of the State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, the Institute of Zoology, Chinese Academy of Sciences.

Author contributions

S.T., X.D., and J.C. conceived and designed the study. S.T., H. Zhang, J.L., J. Cui, H. Zhi, G.J., and Y.S. prepared the materials and performed gene function experiments. W.G., B.X., H.C., and Y.M. performed WGBS data analysis. W.G., Y.L., and Y.M. performed the DMR calling. W.G. performed mRNA and small RNA transcriptome analysis. W.G., S.T., X.D., and J.C. wrote and revised the paper.

Funding

This work was supported by the National Key R&D Program of China (2023YFD1400800 to J.C.), the National Key R&D Program of China (2021YFF1000102 to X.D.), the National Natural Science Foundation of China (32470607 to J.C.), the National Natural Science Foundation of China (32422065 to S.T.), Initiative Scientific Research Program, Institute of Zoology, Chinese Academy of Sciences (2024IOZ0106, 2023IOZ0203, 2025IOZ09, and SKLA2506 to J.C.), and the Open Research Fund Program of State Key Laboratory of Integrated Management of Pest Insects and Rodents (IPM2316 to J. Cui).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-79325-8.

Correspondence and requests for materials should be addressed to Xianmin Diao or Jinfeng Chen.

Peer review information *Nature Communications* thanks Jinliang Yang and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions 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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.