A pair of transposon-derived proteins function in a histone acetyltransferase complex for active DNA demethylation

Cheng-Guo Duan[1, 2, *], Xingang Wang[2, *], Shaojun Xie[1], Li Pan[3], Daisuke Miki[1], Kai Tang[2], Chuan-Chih Hsu[3], Mingguang Lei[1], Yingli Zhong[1], Yueh-Ju Hou[2], Zhijuan Wang[2, 4], Zhengjing Zhang[1], Satendra K Mangrauthia[2, 5], Huawei Xu[2, 6], Heng Zhang[1], Brian Dilkes[3], W Andy Tao[3], Jian-Kang Zhu[1, 2]

[1]*Shanghai Center for Plant Stress Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, China;* [2]*Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907, USA;* [3]*Department of Biochemistry, Purdue University, West Lafayette, IN 47907, USA;* [4]*The State Key Laboratory of Plant Cell and Chromosome Engineering, Center for Agricultural Research Resources, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Shijiazhuang, Hebei 050022, China;* [5]*Biotechnology Section, Indian Institute of Rice Research (IIRR), Hyderabad, India;* [6]*College of Agriculture, Henan University of Science and Technology, Luoyang, Henan 471026, China*

Transposons are generally kept silent by epigenetic mechanisms including DNA methylation. Here, we identified a pair of _Harbinger_ transposon-derived proteins (HDPs), HDP1 and HDP2, as anti-silencing factors in _Arabidopsis_. _hdp1_ and _hdp2_ mutants displayed an enhanced silencing of transgenes and some transposons. Phylogenetic analyses revealed that HDP1 and HDP2 were co-domesticated from the _Harbinger_ transposon-encoded transposase and DNA-binding protein, respectively. HDP1 interacts with HDP2 in the nucleus, analogous to their transposon counterparts. Moreover, HDP1 and HDP2 are associated with IDM1, IDM2, IDM3 and MBD7 that constitute a histone acetyltransferase complex functioning in DNA demethylation. HDP2 and the methyl-DNA-binding protein MBD7 share a large set of common genomic binding sites, indicating that they jointly determine the target specificity of the histone acetyltransferase complex. Thus, our data revealed that HDP1 and HDP2 constitute a functional module that has been recruited to a histone acetyltransferase complex to prevent DNA hypermethylation and epigenetic silencing.

Keywords: DNA demethylation; transposon; domestication; histone acetylation; anti-silencing

Introduction

Transposable elements (TEs) are mobile, repetitive genetic elements that represent large portions of eukaryote genomes [1, 2]. Transposons are generally viewed as "selfish" DNA or "parasitic" elements due to the deleterious effects on host genomes. However, emerging evidence suggests that TEs play key roles in the evolution of gene function and regulation [2-4]. Moreover, an increasing number of host genes that are derived or "domesticated" from transposons have been uncovered [5]. The identification of these genes and elucidation of their

functions in host genomes are of great interest for understanding how transposons contribute to host adaptation.

In plants, TEs are regulated by epigenetic silencing mechanisms, including DNA methylation and histone modifications [6, 7]. DNA methylation plays important roles in multiple processes, including genome stability, cell responses to environmental stimuli and organ development [6-9]. DNA methylation levels and patterns are dynamic and determined by two reversible reactions: DNA methylation and demethylation [8, 10]. DNA demethylation can occur through passive or active pathways, or a combination of both. In comparison with passive DNA demethylation, specific enzymatic reactions are required for active DNA demethylation [8]. In *Arabidopsis*, active DNA demethylation is carried out by a family of bifunctional DNA glycosylases/lyases, including ROS1, Demeter (DME), DME-like 2 (DML2) and DML3. Loss-of-function mutations in ROS1 cause

DNA hypermethylation and enhanced transcriptional gene silencing (TGS) of transgenes and endogenous genes and TEs [8, 11-13]. However, how the active DNA demethylation machinery is recruited to specific genomic loci is poorly understood in plants and animals. Increased DNA Methylation 1 (IDM1), a histone acetyltransferase, is required for a subset of ROS1-mediated DNA demethylation [13]. Recently, IDM1 was shown to be in a complex with IDM2, IDM3 and a methyl-DNA-binding protein, MBD7 [13-16]. Mutations of any components in this IDM complex result in enhanced transgene silencing and DNA hypermethylation of specific genomic regions including TEs. MBD7 recognizes densely methylated CpG regions and contributes to the recruitment of IDM1 to specific genomic loci [14]. However, MBD7 alone cannot determine the target specificity of IDM1 since the complex does not associate with all genomic regions that have high DNA methylation [14]. Therefore, there are likely additional components in the IDM complex that may determine the targeting specificity of the complex jointly with MBD7.

Harbinger transposons are DNA transposons which usually encode a DDE transposase and a SANT/Myb/ trihelix domain-containing DNA-binding protein [17, 18]. Domestications of *Harbinger* transposons have been reported in mammals, *Drosophila* and *Arabidopsis* [17, 19, 20], suggesting their evolutionary importance. However, their biological functions are unclear. Here, using a forward genetic screen, we identified a pair of *Harbinger* transposon-derived anti-silencing factors, HDP1 and HDP2 in *Arabidopsis*. Loss-of-function mutations in these two genes not only triggered enhanced silencing of transgenes and some endogenous TEs, but also increased DNA methylation. Analogous to their *Harbinger* transposon counterparts, HDP1 interacts with HDP2 in the nucleus. We provide evidence that HDP1 and HDP2 are new components of the previously identified IDM histone acetyltransferase complex in which IDM1, IDM2, IDM3 and MBD7 are included. Our results suggest that HDP1 and HDP2 are required for IDM1 histone acetyltransferase activity at tested loci. Moreover, HDP2 and MBD7 share a large set of common genomic regions of chromatin association. Thus, our data revealed that HDP1 and HDP2 constitute a functional module from ancient *Harbinger* transposon which has been recruited to function in a host histone acetyltransferase complex. The HDP1 and HDP2 module is important in determining the target specificity of the histone acetyltransferase complex to facilitate DNA demethylation and to prevent epigenetic silencing.

Results

HDP1 and HDP2 prevent transcriptional gene silencing of transgenes

We previously established a transgene reporter system in *Arabidopsis* in which expression of the 35S promoter-driven *SUC2* transgene (*35S::SUC2*) causes a short-root phenotype on sucrose medium (Figure 1A). This phenotype requires the ROS1-dependent DNA demethylation pathway [21]. From an EMS mutagenesis screen, we identified three recessive mutants, *hdp1-1*, *hdp1-2* (which are allelic) and *hdp2-1*, that display normal root length and *35S::SUC2*, *35S::NPTII* (*neomycin phosphotransferase II*) transgene silencing phenotype (Figure 1A, 1B and Supplementary information, Figure S1). Map-based cloning followed by whole-genome resequencing revealed that both *hdp1* mutants had nonsense mutations in *AT1G72270* and *hdp2-1* had a nonsense mutation in *AT4G31270* (Figure 1C and 1D). The two nonsense mutations within *AT1G72270* both occur in the short transcript annotation and our RNA-seq data show transcripts only from this short form region, suggesting that the short transcript is the functional unit. Genetic complementation of *hdp1-1* and *hdp2-1* with genomic DNA, including upstream 2 kb native promoters, confirmed that the mutations in *AT1G72270* and *AT4G31270* caused the transgene silencing (Figure 1A and 1B).

To test whether DNA methylation plays a role in the transgene silencing in *hdp1* and *hdp2* mutants, we treated these mutants with the DNA methylation inhibitor 5′Aza-2′-deoxycytidine (5′Aza). 5′Aza treatment restored kanamycin resistance to *hdp1* and *hdp2* mutants (Supplementary information, Figure S1A) and increased the relative expression of transgenes compared to *35S::SUC2* plants (Supplementary information, Figure S1B), suggesting that the transgene silencing in the mutants may be caused by increased DNA methylation. Moreover, transgene silencing is likely to be at the transcriptional level, as chromatin immunoprecipitation (ChIP) assays showed that there were depletion of pol II occupancy, decreased H3K4 tri-methylation, an active histone modification, and increased H3K9 di-methylation, a repressive histone modification, at transgene promoters. 5′Aza treatment counteracted the chromatin alterations caused by *hdp1* and *hdp2* mutations (Supplementary information, Figure S1C). These results suggest that loss of HDP1 and HDP2 leads to increased DNA methylation levels, altered histone modifications and depleted Pol II at transgene promoters, resulting in decreased transcription of transgenes.

HDP1 and HDP2 prevent DNA hypermethylation and transcriptional silencing of TEs
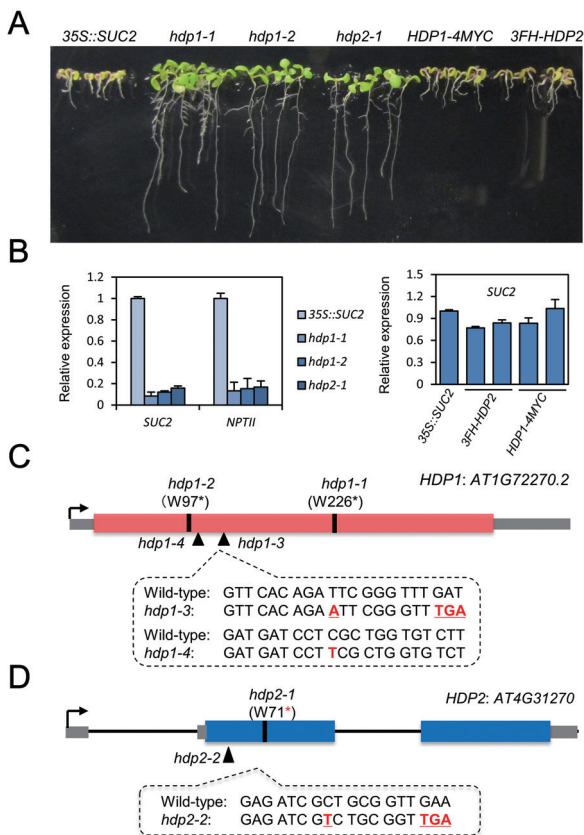
Next, we performed whole-genome bisulfite sequenc-

Figure 1 HDP1 and HDP2 prevent transcriptional silencing of transgenes. **(A)** Identification of *hdp1* and *hdp2* mutants. Introduction of *HDP1* and *HDP2* genomic DNA fully complemented the root phenotype in *hdp1-1* and *hdp2-1* mutants. **(B)** RT-qPCR showing significantly reduced transcript levels of the transgenes *SUC2* and *NPTII* in *hdp1-1* and *hdp2-1* mutants in comparison to *35S::SUC2* plants. *HDP1* and *HDP2* genomic DNA rescued the silencing of *SUC2* transgene in *hdp1-1* and *hdp2-1* mutants. Two representative transgenic lines were selected and relative expression of *SUC2* transcript was normalized to *35S::SUC2* plants. Three independent biological replicates were carried out for statistical analysis. See also Supplementary information, Figure S1. **(C)** Genomic structure and mutant alleles of *HDP1* gene. For *hdp1-1* and *hdp1-2* mutants, nucleotide substitutions of "G" to "A" lead to the changes of 226th and 97th amino acids from tryptophan (W) to stop codons. *hdp1-3* and *hdp1-4* mutants are generated by CRISPR/Cas9-mediated genome editing. Nucleotide insertions of "A" and "T" were identified in *hdp1-3* and *hdp1-4* mutants, respectively. Red box indicates the coding region of *AT1G72270* gene and gray box indicates untranslated regions. **(D)** Genomic structure and mutant alleles of *HDP2* gene. For *hdp2-1* mutant, a G-to-A substitution causes a pre-mature stop codon. An insertion of "T" was identified in *hdp2-2* mutant. Blue box indicates the coding region of *AT4G31270* gene and gray box indicates untranslated regions.

ing (WGBS) using genomic DNA from *35S::SUC2*, *hdp1-1* and *hdp2-1* seedlings. We first evaluated DNA

methylation levels at the transgene promoters. We found that DNA methylation levels at region B of the 35S promoters are increased in *hdp1-1* and *hdp2-1* relative to *35S::SUC2* plants (Figure 2A), suggesting that the transgene silencing may be caused by increased DNA methylation. We also evaluated the methylation levels of four additional endogenous ROS1 targets by DNA methylation-sensitive PCR (chop-PCR). Similar to a *ros1* mutant, DNA methylation levels at these loci were increased in *hdp1* and *hdp2* mutants relative to the *35S::SUC2* tester line (Figure 2B). To further confirm the methylation phenotype, we employed CRISPR/Cas9-based genome editing to generate *hdp1-3, hdp1-4* and *hdp2-2* mutants in the Col-0 background (Figure 1C and 1D). Chop-PCR again showed increased DNA methylation levels at the four tested loci (Figure 2B), confirming that HDP1 and HDP2 repress DNA methylation at the tested ROS1 target loci.

We then evaluated the DNA methylation change on a genome-wide level in *hdp1* and *hdp2* mutants. WGBS identified 1 243 and 2 768 hyper-differentially methylated regions (hyper-DMRs) in *hdp1-1* and *hdp2-1* mutants, respectively (Supplementary information, Table S1). Venn diagram shows that about 43.3% hyper-DMRs of *hdp1-1* overlap with that of *hdp2-1* (Figure 2C, upper panel). Moreover, box plot shows that even for those hyper-DMRs that appear specific for *hdp1-1* and *hdp2-1* in the Venn diagram, the average DNA methylation levels are still increased in *hdp2-1* and *hdp1-1* mutants compared to *35S::SUC2* plants (Figure 2C, lower panel), supporting that HDP1 and HDP2 function to prevent DNA hypermethylation of similar genomic loci. In addition, we found that most of the hyper-DMRs are mapped to TE regions (Figure 2D), suggesting that HDP1 and HDP2 may prevent TEs from silencing.

Next we profiled the transcriptomes of *hdp1-1* and *hdp2-1* by mRNA-Seq. We did not observe significant differences in gene expression between *35S::SUC2* and *hdp1-1* or *hdp2-1* plants (Supplementary information, Figure S2). TEs and repetitive sequences are the primary targets of the silencing machinery, including DNA methylation and repressive histone modifications [6, 9, 22], and are transcribed at low levels, therefore the effects of HDP1 and HDP2 dysfunction in further reducing TE expression may not be detected by mRNA-Seq. Thus, we analyzed three TE loci, *AT1TE46455, AT1TE36115* and *AT1TE35325*, which had increased DNA methylation in *hdp1-1* and *hdp2-1* mutants (Figure 2E) by RT-qPCR. We found that the expression of all three *TEs* was significantly reduced in *hdp1* and *hdp2* mutants compared to control (Figure 2E). These results suggest that HDP1 and HDP2 function in the anti-silencing of endogenous TEs
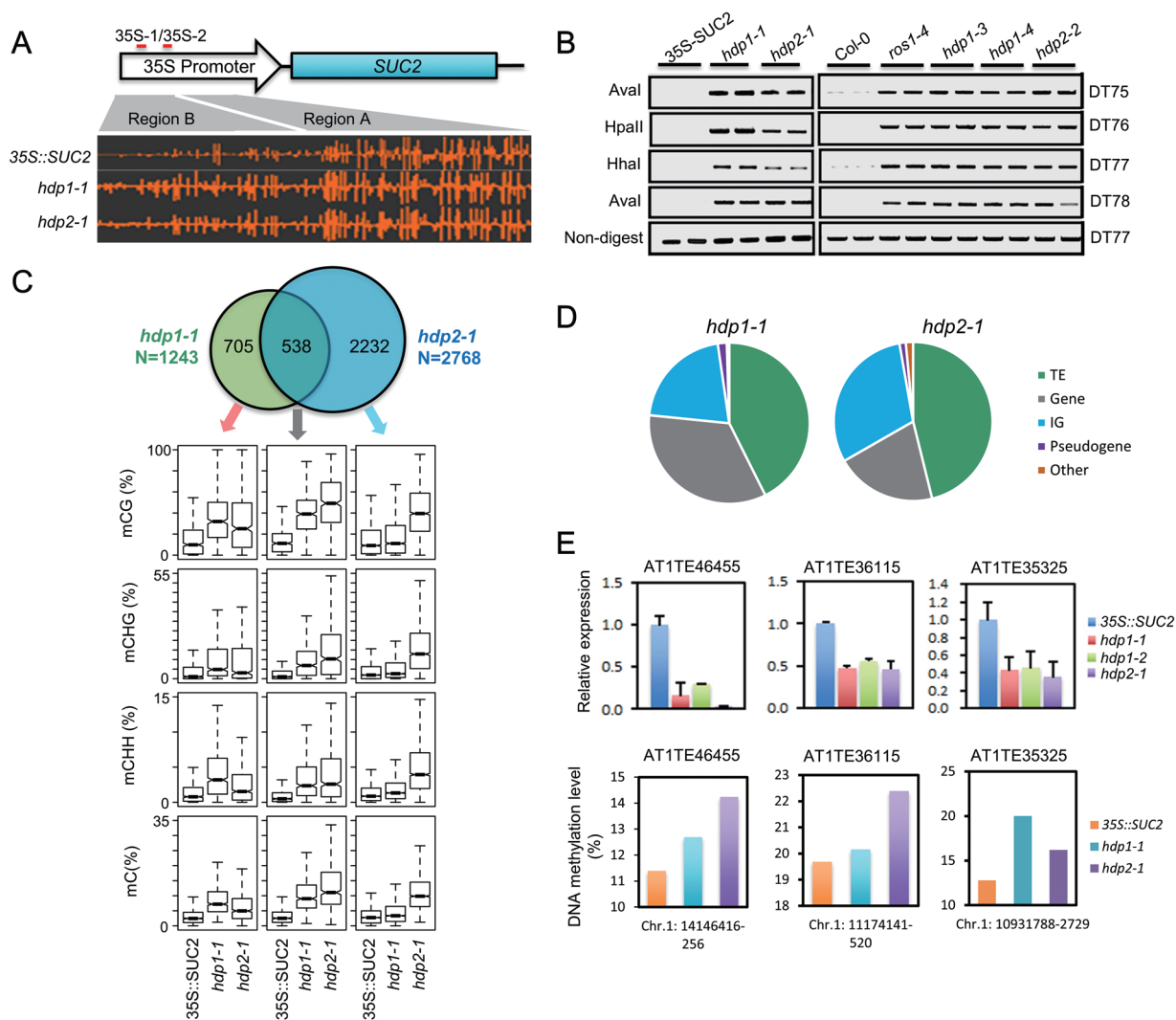
Figure 2 Dysfunctions of HDP1 and HDP2 alter DNA methylation and endogenous TE silencing. **(A)** A screenshot from Integrative Genome Browser showing DNA methylation status at the *35S::SUC2* transgene promoter region. Vertical bars on each track indicate DNA methylation levels. **(B)** Chop-PCR showing that *hdp1* and *hdp2* mutants displayed a hypermethylation phenotype in ROS1 DNA demethylation target loci [13]. Amplification of non-digested DNA served as a control. **(C)** Upper panel: Venn diagram showing the numbers of hyper-DMRs overlap between *hdp1-1* and *hdp2-1* mutants. Lower panel: Box plots displaying DNA methylation levels (CG, CHG, CHH and total C) calculated from the corresponding overlapping or unique hyper-DMRs. Increases of DNA methylation levels from any mutant compared to *35S::SUC2* wild type are statistically significant (Supplementary information, Table S1). **(D)** Hyper-DMRs annotation in *hdp1-1* and *hdp2-1* mutants. IG represents intergenic regions. **(E)** Selected endogenous TEs are repressed by *hpd1* and *hdp2* mutations. RT-qPCR results are means ± SD of three biological replicates. Relative expression was firstly normalized by the expression of *ACTIN2* and then by the expression in *35S::SUC2* wild-type plants. Therefore the relative expression of TE in *35S::SUC2* wild-type plants was defined as 1.0. The DNA methylation levels at corresponding TE regions were calculated from WGBS data. See also Supplementary information, Figure S2 and Table S1.

in plants.

HDP1 is a Harbinger transposon-derived host protein in Arabidopsis

HDP1 protein has an annotated "*Harbinger* transposase-derived nuclease domain" (InterPro accession: IPR027806) [23], suggesting that it may be a "domesticated" protein from *Harbinger* transposase (Figure 3A). The derivation of host protein-coding genes from *Harbinger* transposons has previously been reported in human, *Drosophila* and *Arabidopsis* [17, 20, 24]. A BLASTP search using the *Harbinger* transposon

database (Repbase) [25] revealed that HDP1 shares about 30% identity with some *Harbinger* transposase sequences (Supplementary information, Table S2). The *Harbinger* transposases belong to the "DDE" superfamily endonucleases, which contain a conserved catalytic triad of amino acid sequence "DDE" [26]. However, the alignment of HDP1 with selected transposases revealed that the triad is not completely conserved in HDP1 (Figure 3A, red arrows), suggesting that HDP1 probably has lost the nuclease activity. It is noteworthy that HDP1 is not a transposon remnant since, despite the homology to *Harbinger* transposase, it lacks transposon characteristics such as terminal inverted repeat (TIR) and target site duplication (TSD).

To determine the phylogenetic relationship between HDP1 and its homologous transposases, HDP1 homologous sequences retrieved from the *Harbinger* transposon database and various plant species, zebrafish and human in NCBI non-redundant (nr) protein database were used to construct a phylogenetic tree (Figure 3B). HDP1 is located in a clade consisting of sequences from all selected angiosperm species (bright green clade, *at1g72270* is HDP1), indicating that HDP1 has orthologues in these selected angiosperm species as the gene tree and species tree are consistent. The phylogenetic tree shows that several host protein clades, including the one which contains HDP1, are intermingled with clades consisting of transposase sequences (Figure 3B). These intermingled host proteins may share a common ancestor with the transposases that are located in their sister clusters. The phylogenetic tree reveals that HDP1 originated from *Harbinger* transposons and that the derivation of HDP1 had likely happened before the emergence of angiosperms.

Besides the HDP1 clade, five more orthologous clusters (clades colored in yellow, red, purple, orange and grey) containing *Arabidopsis* protein sequences and a group of non-transposon-encoded sequences are present in the phylogenetic tree (Figure 3B). Notably, the six orthologous clusters containing *Arabidopsis* sequences were not all within a single larger clade that lacks transposase sequences. Rather, two transposase clades are intermingled with the clades of orthologous angiosperm proteins. This topology suggests that the multiple *Arabidopsis* sequences were derived from independent "domestication" of *harbinger*-encoded sequences instead of solely from gene duplication after a single "domestication" event. Similarly, the *Harbinger* transposon-derived HARBI1 in human (homo_gi|767965708) is located in its own orthologous cluster which also has a sister transposase clade [27], indicating that HARBI1 in human does not share a common "domestication" event with *Harbinger*-derived proteins in plants. Collectively,

our results show that *Harbinger* transposons have been repeatedly "domesticated" into host proteins in species from different kingdoms, suggesting its functional importance during evolution.

## HDP2 is a DNA-binding protein derived from Harbinger transposons

HDP2 protein contains a SANT/Myb/trihelix DNA-binding motif (Figure 4A). Interestingly, besides the canonical transposase, *Harbinger* transposons usually have a second ORF which encodes this motif [17, 28]. The alignment of the SANT/Myb/trihelix domains from HDP2 and from *Harbinger* transposon-encoded proteins shows that three tryptophan residues critical for DNA binding are conserved between HDP2 and *Harbinger* proteins from *Arabidopsis lyrata* (Figure 4A, red stars). However, a search of Repbase with the protein sequence of HDP2 identified only one significant hit (*E*-value < 0.01; Supplementary information, Table S2). Notably, it has been suggested that the Myb-like proteins encoded by *Harbinger* transposons were significantly more divergent than the transposase [18], therefore it may be difficult to track the phylogenetic history of HDP2. Nevertheless, when we combined *Harbinger* transposon-encoded Myb-like proteins from Repbase and HDP2 homologous sequences from the NCBI nr database to construct the phylogenetic tree of HDP2, we observed intermingled clades of HDP2 homologous proteins and clades of transposon proteins (Supplementary information, Figure S3). The clade containing HDP2, like HDP1, shows agreement between genes and the taxa that they are sampled from, indicating derivation from a *Harbinger* transposon before the diversification of angiosperms.

The DNA-binding activity of HDP2 is required to prevent transgene silencing

Next, we tested the DNA-binding activity of HDP2 by electrophoretic mobility shift assay (EMSA) with recombinant maltose-binding protein (MBP)-tagged HDP2. It was reported that the *Harbinger* transposon-derived human NAIF1 protein binds to the TIR region of zebrafish *Harbinger3N_DR* transposon DNA [27]. We selected four double-stranded DNA probes in *Harbinger3N_DR* TIR region to test whether HDP2 may have DNA-binding activity. We found that MBP-HDP2 can bind to two of the DNA probes (Figure 4B and Supplementary information, Figure S4A), suggesting that HDP2 has DNA-binding activity but is not a ubiquitous DNA-binding protein. Moreover, HDP2 displayed similar DNA-binding activities to methylated and unmethylated 35S-1 DNA fragments from the *35S::SUC2* transgene promoter (Supplementary information, Figure S4B and

A

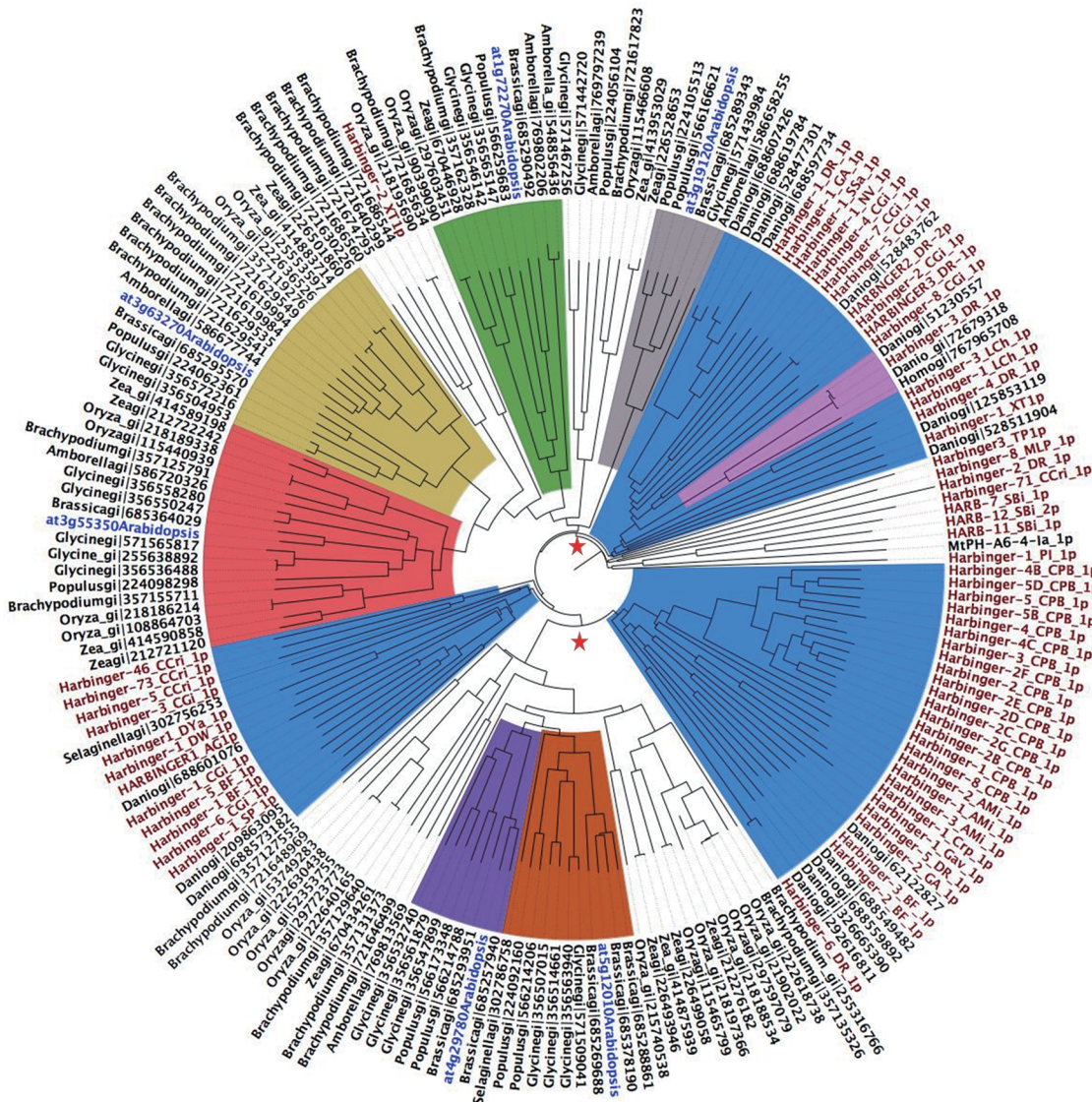*Harbinger* transposase-derived nuclease (DDE) domain
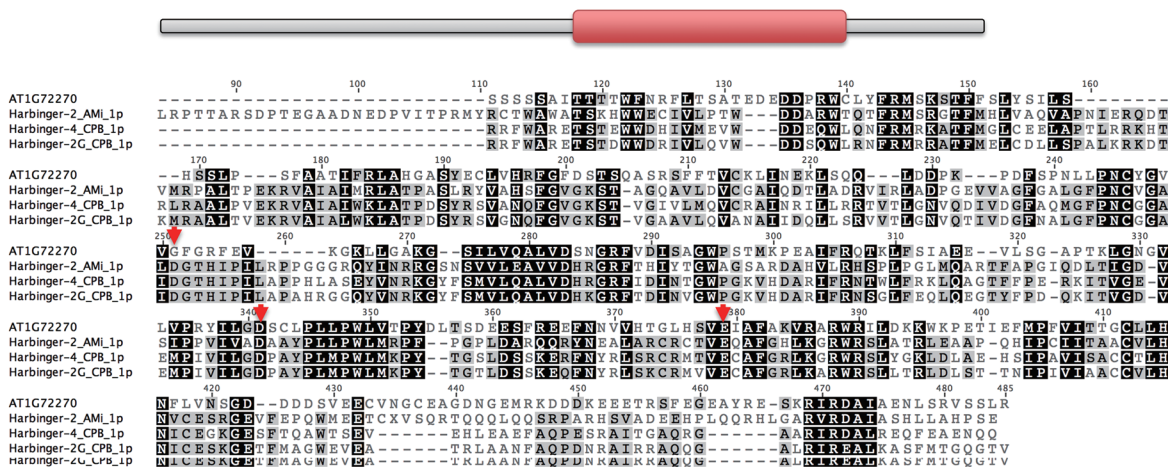
Figure 2A), suggesting that HDP2 does not discriminate the methylation state of its target DNA.

Further, we found that mutation of each of the three conserved tryptophan residues eliminated (W19R, W48R) or reduced (W71R) the DNA-binding activity of HDP2 (Figure 4C), consistent with the observation that W71 is less conserved than W19 and W48 within this DNA binding motif [29]. Thus, the DNA-binding activity of HDP2 requires these three conserved tryptophan residues. Consistent with the *in vitro* binding assays, transgenic *hdp2-1* plants expressing HDP2W19R or HDP2W48R failed to rescue the long root, transcriptional transgene silencing and DNA hypermethylation phenotypes caused by *hdp2-1* mutation (Figure 4D, 4E and Supplementary information, Figure S4C). Although HDP2W71R did recover the long root mutant phenotype to shorter roots, they were not as short as transgenic mutant plants expressing wild-type HDP2 (Figure 4D). These results indicate that the DNA-binding activity of HDP2 is required to prevent transgene silencing. In addition, we tested whether HDP2 can bind to the transgene promoters *in vivo* by performing ChIP assay in HDP2 transgenic plants. We found that wild-type and W71R mutated form of HDP2, but not the W19R and W48R mutated forms of HDP2, were enriched at the 35S promoter regions of the *SUC2*, *HPTII* and *NPTII* transgenes relative to the "no antibody" and *35S::SUC2* controls, but not at the *ACTIN7* promoter (Figure 4F and Supplementary information, Figure S4D). Thus, HDP2 binds specifically to the transgene promoter, and the DNA-binding activity is required for full function of HDP2 *in vivo*.

HDP1 and HDP2 interact in the nucleus

Transposition by the *Harbinger* transposon is facilitated by an interaction between the DNA-binding protein and the transposase [27]. The observation that both HDP1 and HDP2 may be domesticated from the *Harbinger* transposon prompted us to investigate whether a physical interaction between these two proteins has been conserved. We performed a yeast two-hybrid (Y2H) assay and found that HDP1 and HDP2 can directly interact (Figure 5A). The interaction was also confirmed by a split luciferase assay in *Arabidopsis* protoplasts (Figure 5B). To further test whether HDP1 and HDP2 interact *in vivo*, we performed a co-immunoprecipitation assay using hybrid plants generated from *HDP1-4MYC/hdp1-1* and *3FH-HDP2/hdp2-1*. Col-0 and *3FH-HDP2/hdp2-1* plants served as controls. We found that HDP1-4MYC co-immunoprecipitated with 3FH-HDP2 from hybrid plants (Figure 5C), suggesting their interaction *in vivo*. To map the domains required for the interaction, we evaluated two truncated forms of HDP1 and HDP2 in the Y2H assay. We found that the N-terminal region (amino acid 1-225) of HDP1 and the C-terminal region (amino acid 151-259) of HDP2 are required for the interaction of HDP1 and HDP2 (Supplementary information, Figure S5A). Collectively, these findings strongly suggest that HDP1 and HDP2 interact *in vivo*.

To investigate the subcellular localizations of HDP1 and HDP2, we first transiently expressed 35S promoter-driven HDP1- and HDP2-YFP fusion proteins in tobacco leaves and *Arabidopsis* protoplasts. We found that both HDP1-YFP and HDP2-YFP were localized mainly to nuclei (Figure 5D). To further confirm that their interaction occurs in the nucleus, we performed bimolecular fluorescence complementation (BiFC) in tobacco leaves. We found that YFP fluorescence was restored mostly in the nuclei expressing both HDP1-YFP[N] and HDP2-YFP[C], or both HDP1-YFP[C] and HDP2-YFP[N] (Supplementary information, Figure S5B), suggesting that HDP1 and HDP2 interact in the nucleus. We then examined co-localization of HDP1 and HDP2 at subnuclear level by immunostaining. HDP1-4MYC and 3FH-HDP2 were detected by anti-MYC and anti-FLAG antibodies, respectively. Co-staining of HDP1 and HDP2 in the same nuclei revealed many yellow foci (Figure 5E), indicating the co-localization of HDP1-4MYC and 3FH-HDP2 at the subnuclear level. Collectively, these data demonstrate that HDP1 and HDP2 physically interact in the nucleus to function in the epigenetic anti-silencing pathway.

HDP1 and HDP2 are new components of the IDM complex and are required for IDM1 histone acetyltransferase activity at tested loci

Figure 3 Phylogenetic analysis of HDP1. **(A)** Upper panel: schematic representation of HDP1 domain structure. Lower panel: MSA of HDP1 and selected *Harbinger* transposases from BLAST hits. Red arrows indicate the residues critical for nuclease enzymatic activity. **(B)** Phylogenetic tree of HDP1 homologous sequences and *Harbinger* transposases. Black nodes are HDP1 homologous sequences from BLAST search on NCBI nr database, and they are named using their genus name and gi ID in NCBI. Sequences from *Arabidopsis* are labeled in blue with gene ID. Red nodes are sequences of *Harbinger* transposases from Repbase, with their original names in Repbase. The orthologous clades of HDP1 (*at1g72270*), *at3g63270*, *at3g55350*, *at4g29780*, *at5g12010* and *at3g19120* are labeled in green, yellow, red, purple, orange and grey, respectively. Several clades of *Harbinger* transposases intermingled with host sequences are labeled in blue. Possible branching points of host proteins and transposases are labeled with red stars. See also Supplementary information, Table S2.
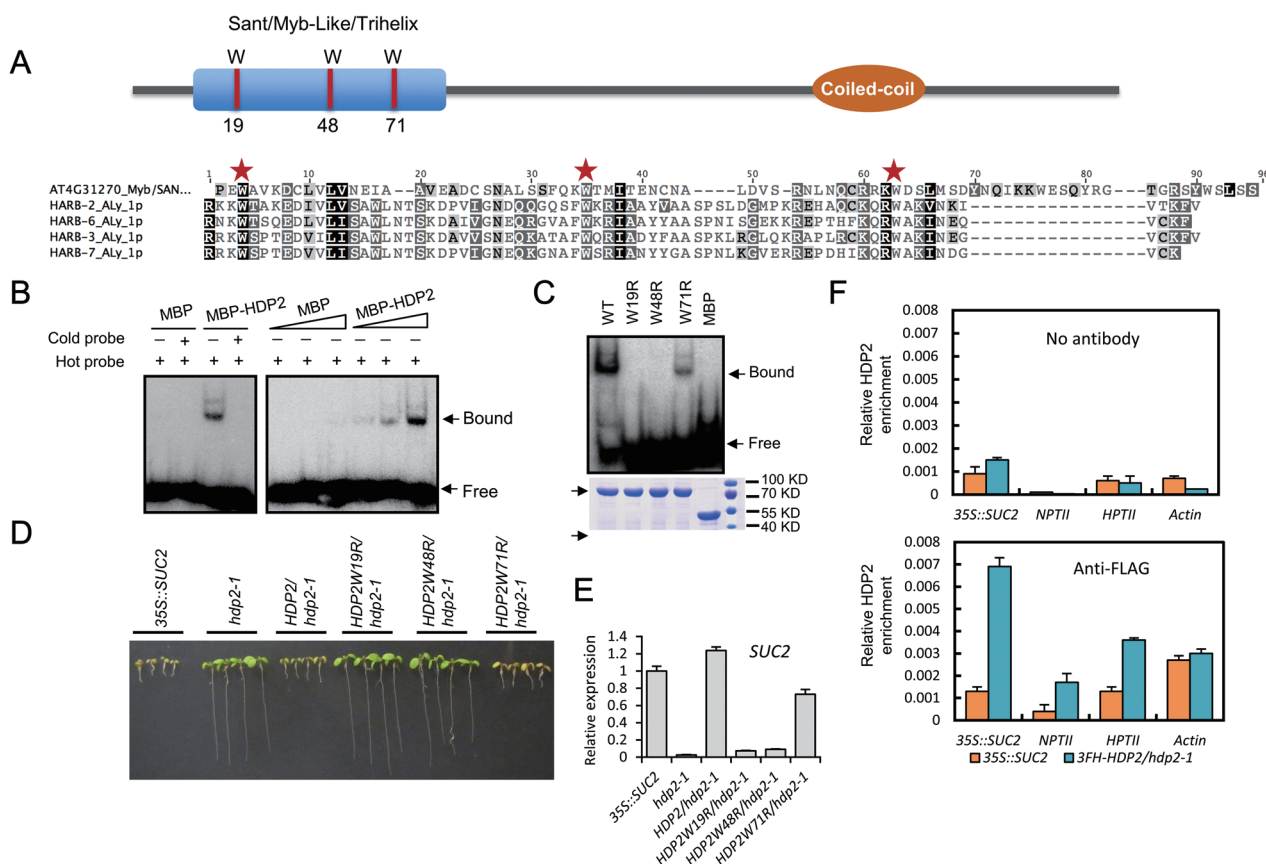
Figure 4 Domain structure of HDP2 and its DNA-binding activity. **(A)** Domain structure of HDP2 and protein sequence alignment of DNA-binding domain of HDP2 with those encoded by *Harbinger* transposons. **(B)** EMSA assay showing that MBP-HDP2 can bind to DNA *in vitro*. L1 DNA probe (see Supplementary information, Figure S4A) was used. Increasing the MBP-HDP2 concentration produced a darker mobility shifted band. MBP served as a negative control. Bound and free probes are labeled with black arrows. **(C)** Three conserved tryptophan residues are critical for DNA binding by HDP2. EMSA was performed using wild-type and mutated HDP2. Coomassie bright blue-stained SDS-PAGE gel of wild-type and mutated HDP2 with equal amounts showing that mutated proteins have the same solubility as the wild-type protein. **(D)** Complementation assay showing that HDP2W19R and HDP2W48R cannot rescue the long-root phenotype by *hdp2-1* mutation. HDP2W71R partially rescues the long-root phenotype by *hdp2-1* mutation. **(E)** RT-qPCR result showing that introduction of HDP2W19R and HDP2W48R mutated genomic sequences could not release the silencing of *SUC2* transgene. **(F)** ChIP-qPCR showing that HDP2 is enriched at the promoter regions of *SUC2*, *NPTII* and *HPTII* transgenes. Relative enrichment of HDP2 was calculated by the percentage of input DNA. No antibody was used as a parallel control. The *ACTIN* promoter was used as a target control. See also Supplementary information, Figure S3, Tables S2 and S3.

To better understand the mechanism of functions of HDP1 and HDP2, we attempted to identify HDP1- and HDP2-interacting proteins using immunoprecipitation followed by mass spectrometry (IP-MS) analysis in transgenic plants expressing native promoter-driven epitope-tagged proteins. We immunoprecipitated HDP1-4MYC and 3FH-HDP2 with anti-MYC and anti-FLAG antibodies, respectively. *35S::SUC2* plants were used as negative controls. Consistent with the interaction between HDP1 and HDP2 (Figure 5), both HDP1 and HDP2 proteins were present in their reciprocal IP products, but not in control IP (Figure 6A and Supplementary information, Table S4). Interestingly, from both HDP1 and HDP2 IP samples, but not the control, four other anti-silencing factors, including IDM1 and its associating proteins IDM2, IDM3 and MBD7, were identified. To verify the association of these six proteins, we performed IP-MS experiments with IDM1-YFP-HA, IDM2-HA, IDM3-3FLAG-3HA and MBD7-4MYC transgenic plants. The results showed that HDP1, HDP2, IDM1, IDM2, IDM3 and MBD7 were always co-purified together, with the exception that the IDM2-HA purification
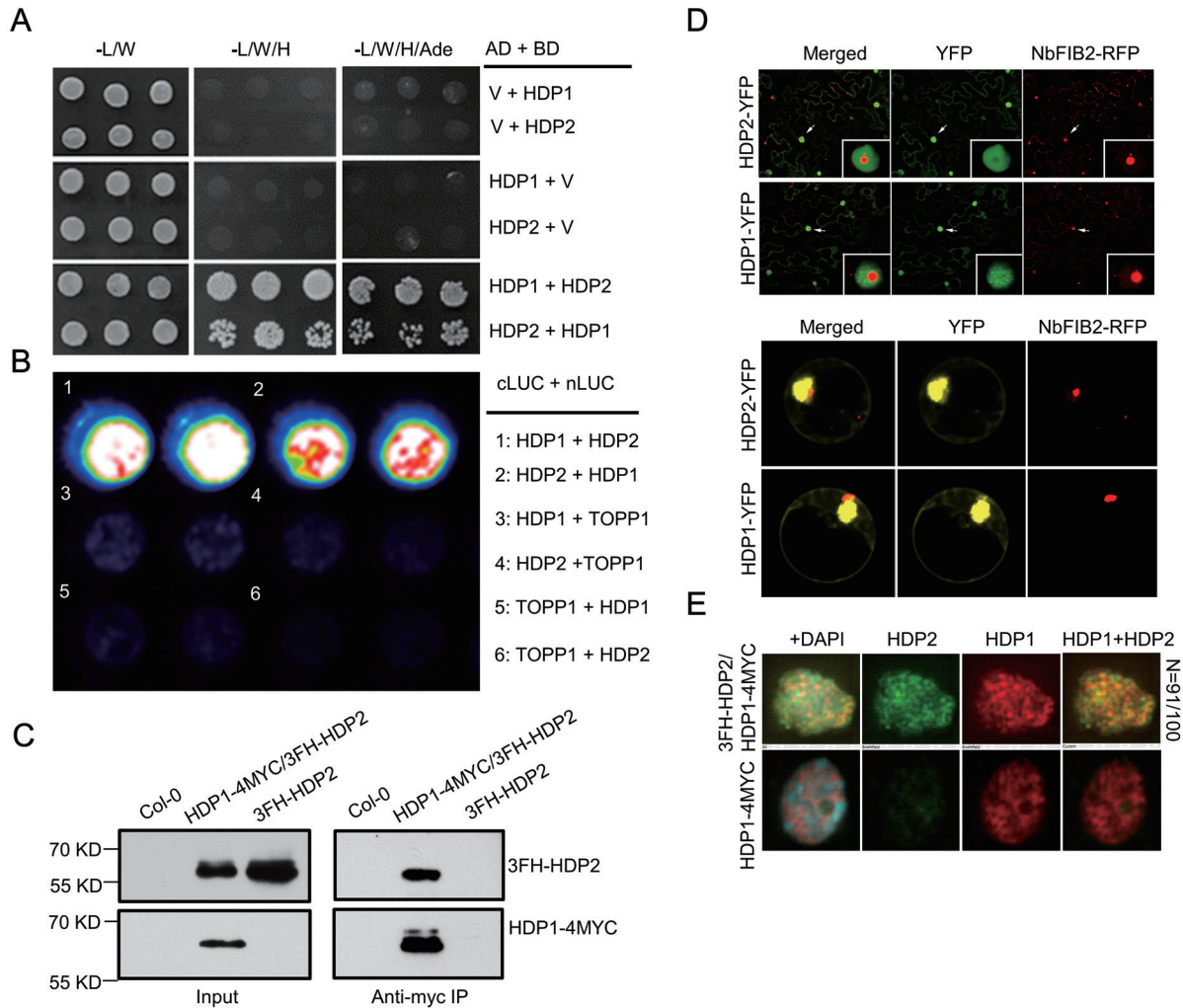
Figure 5 Physical interaction between HDP1 and HDP2. **(A)** Y2H assay showing direct interaction between HDP1 and HDP2. **(B)** Split luciferase complementation assay in *Arabidopsis* protoplasts showing the interaction between HDP1 and HDP2. TOPP1 protein was used as control protein. **(C)** HDP1-HDP2 interaction was confirmed by co-immunoprecipitation assay. Total proteins extracted from hybrid transgenic plants (*HDP1-4myc/3FH-HDP2*), *3FH-HDP2* and Col-0 plants were subjected to anti-myc immunoprecipitation, respectively. The 3FH-HDP2 and HDP1-4myc fusion proteins in input and output proteins were detected by western blotting using anti-Flag and anti-Myc antibodies, respectively. **(D)** HDP1 and HDP2 are localized to the nucleus. Full-length HDP1 and HDP2 were fused with YFP at the C-terminus under the control of 35S promoter. YFP signals were detected in tobacco leaves (upper panel) and *Arabidopsis* protoplasts (lower panel). RFP-fused NbFIB2 protein serves as nucleolus marker [40]. White arrow showing the position of nucleus zoomed in. **(E)** Immunostaining assay showing that HDP1 and HDP2 co-localize in the subnuclear level. Red and green fluorescence indicates HDP1 and HDP2, respectively. Yellow foci were observed in merged HDP1 and HDP2 fluorescence. *N* represents the ratio of nuclei showing co-localization in total examined nuclei. See also Supplementary information, Figure S4.

did not yield MBD7 (Figure 6A and Supplementary information, Table S4). The results support that the six proteins form a protein complex *in vivo*, which is referred to hereafter as the IDM complex since it contains the IDM1 histone acetyltransferase and dysfunction of any of its components causes the IDM (Increased DNA Methylation) phenotype. To further confirm the formation of IDM complex, protein gel filtration assays were performed using epitope-tagged transgenic plants. The results show that the tested components co-eluted in the same fractions (Figure 6B), supporting that a complex is formed *in vivo*. The molecular weight of the protein complex was estimated to be ~350 kDa. Considering the molecular weight of the protein tag, the estimated size is very close to the sum (~339 kDa) of the molecular weights of the six subunits at 1:1:1:1:1:1 stoichiometry.
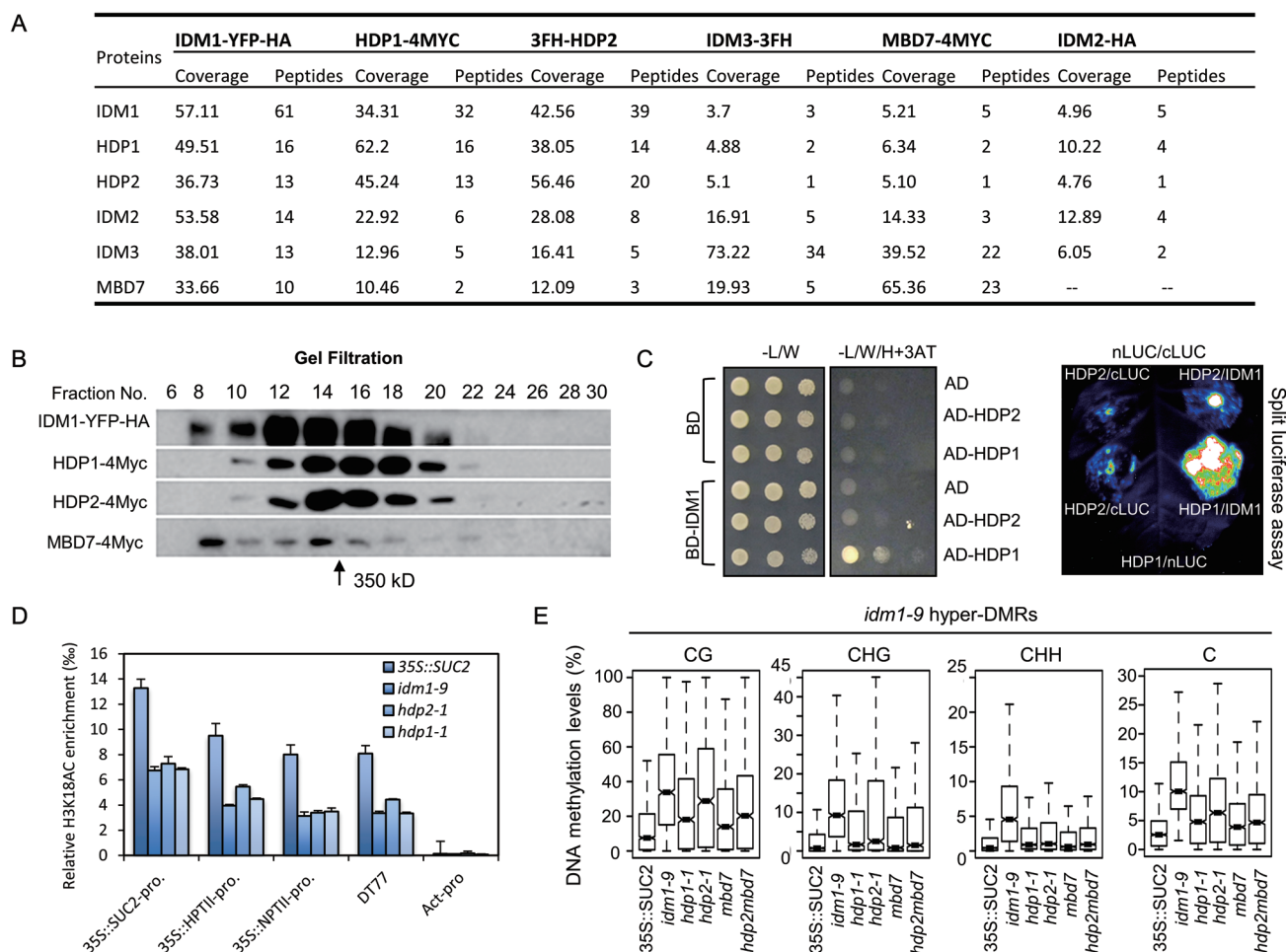
A

Proteins	IDM1-YFP-HA		HDP1-4MYC		3FH-HDP2		IDM3-3FH		MBD7-4MYC		IDM2-HA	
	Coverage	Peptides	Coverage	Peptides	Coverage	Peptides	Coverage	Peptides	Coverage	Peptides	Coverage	Peptides
IDM1	57.11	61	34.31	32	42.56	39	3.7	3	5.21	5	4.96	5
HDP1	49.51	16	62.2	16	38.05	14	4.88	2	6.34	2	10.22	4
HDP2	36.73	13	45.24	13	56.46	20	5.1	1	5.10	1	4.76	1
IDM2	53.58	14	22.92	6	28.08	8	16.91	5	14.33	3	12.89	4
IDM3	38.01	13	12.96	5	16.41	5	73.22	34	39.52	22	6.05	2
MBD7	33.66	10	10.46	2	12.09	3	19.93	5	65.36	23	--	--

Figure 6 HDP1 and HDP2 associate with IDM1, IDM2, IDM3 and MBD7. **(A)** Summary of IP-MS analyses with different bait proteins showing the association between HDP1, HDP2, IDM1, IDM2, IDM3 and MBD7. Also see Supplementary information, Table S4. **(B)** Western blotting of gel filtration assays. Western blotting showing the eluted fractions containing epitope-tagged IDM1, HDP1, HDP2 and MBD7. The arrow indicates the size of the estimated molecular weight of the protein complex. **(C)** HDP1, but not HDP2, directly interacts with IDM1 in Y2H assay and split luciferase assay in tobacco leaves. See also Supplementary information, Figure S5. **(D)** Effect of *hdp1* and *hdp2* mutations on histone H3K18AC levels. The density of H3K18AC on IDM1-dependent hyper-DMRs [13] and 35S transgene promoters was determined by ChIP-qPCR using anti-H3K18AC antibody. The ChIP signal was quantified as relative to input DNA. Two biological replicates were carried out and similar results were obtained. Shown are results from one of the biological replicates. Standard errors were calculated from three technical replicates. **(E)** Box plot showing the DNA methylation levels at different cytosine contexts calculated from the hyper-DMRs in *idm1-9* mutant. DNA methylation levels in *hdp1-1* and *hdp2-1* mutants were also increased in *idm1-9* hyper-DMRs compared to *35S::SUC2* plants. All increases are statistically significant (Supplementary information, Table S1).

To identify direct interactions within the IDM complex, we tested pairwise interactions between each component with Y2H assays. We found that HDP1, but not HDP2, can directly interact with IDM1 (Figure 6C). The N-terminal domain (amino acid residues 1-592) of IDM1 is required for its interaction with HDP1 (Supplementary information, Figure S6A). No interaction between IDM1 and truncated HDP1 proteins was observed (Supplementary information, Figure S6B), indicating that full-length HDP1 is required for its interaction with IDM1. The interaction between IDM1 and HDP1 was further verified by split luciferase assays in tobacco leaves (Figure 6C). In addition, neither HDP1 nor HDP2 showed any interaction with IDM2, IDM3 or MBD7 in either Y2H or split luciferase assays (Supplementary information, Figure S6C and S6D). Our previous report has shown that MBD7 can directly interact with IDM2 and IDM3, and IDM2 and IDM3 can directly interact with IDM1

[14]. Taken together, these results suggest that the IDM complex has two DNA-binding components, HDP2 and MBD7, and that HDP1 and IDM2/IDM3 mediate their associations with the enzymatic core IDM1, respectively.

IDM1 can acetylate histone H3 at lysine 18 (H3K18) and *idm1* mutant shows reduced H3K18Ac levels at 35S promoter regions [13-15]. We next tested whether H3K18Ac level is also affected in *hdp1* and *hdp2* mutants. ChIP-qPCR showed that the level of H3K18Ac was reduced in both *hdp1* and *hdp2* mutants at the 35S promoters and an endogenous IDM1-dependent locus tested (Figure 6D). The reduction was comparable to that in *idm1* mutant, supporting that HDP1 and HDP2 are required for the IDM1 histone acetyltransferase activity at the tested loci. Since IDM1-mediated histone acetylation is required for DNA demethylation at specific genomic regions [13, 14], we investigated whether mutations in *HDP1* and *HDP2* may cause similar DNA hypermethylation phenotypes as observed in *idm1* mutant. We performed WGBS in *hdp1-1*, *hdp2-1* and *idm1-9* mutants (all in the *35S::SUC2* background). We identified 1 243, 2 768 and 1 612 hyper-DMRs in *hdp1-1*, *hdp2-1* and *idm1-9* mutants, respectively (Supplementary information, Table S1). By intersecting the identified hyper-DMRs, we found that *idm1-9* mutant shares 416 and 562 out of its 1 612 identified hyper-DMRs with *hdp1* and *hdp2* mutants, respectively (Supplementary information, Figure S7A). We then plotted the DNA methylation levels of wild type, *hdp1*, *hdp2* and *idm1* mutants at the hyper-DMRs identified from the *idm1-9* mutant. The plots show that the DNA methylation levels in *hpd1* and *hdp2* mutants are increased compared to wild type in all cytosine contexts (CG, CHG and CHH) at the *idm1* hyper-DMRs (Figure 6E), suggesting that HDP1 and HDP2 affect DNA methylation at common genomic regions with IDM1. As both HDP2 and MBD7 are DNA-binding proteins and the increase of DNA methylation in *hdp2* mutant is less than that of *idm1* mutant, we also performed WGBS using *mbd7-1* and *hdp2-1mbd7-1* mutants to see whether HDP2 and MBD7 may have functional redundancy (Figure 6E). The results show that the *hdp2mbd7* double mutant has a similar level of DNA methylation increase to either single mutant. It is likely that there are other functionally redundant proteins to the HDPs and MBD7, and IDM1 may also function independently of the complex. Taken together, our results suggest that HDP1 and HDP2 are required for full IDM1 function *in vivo*, which is consistent with their associations in the same protein complex.

HDP2 and MBD7 bind to similar genomic regions for DNA demethylation

To determine the role of HDP2 in the IDM complex, we investigated the *in vivo* binding sites of HDP2 by performing ChIP-Seq using *3FH-HDP2/hdp2-1* transgenic plants. There were 2 354 common HDP2 binding sites identified from two biological replicates (Figure 7A). MBD7 was previously shown to bind densely methylated CpG regions and mediate the recruitment of IDM1 [14]. As HDP2 is in the same complex with MBD7, we compared the binding sites of HDP2 with those of MBD7. We found that 1 360 out of the 2 354 (58%) HDP2-enriched regions overlap with MBD7-enriched regions (Figure 7A). Similarly, 1 144 out of the 2 181 (52%) MBD7-enriched regions (common peaks from two replicates) overlap with HDP2-enriched regions (Figure 7A). In addition, we plotted ChIP signals of HDP2, MBD7 and control (ChIP signals for randomly selected regions) over MBD7 peak regions and HDP2 peak regions, respectively (Figure 7B). The distribution shows that both HDP2 and MBD7, but not the control, were enriched over the peak regions of each other, with the highest ChIP signal at the center of identified peaks. Taken together, our results show that HDP2 and MBD7 target a common set of genomic regions.

Next, we sought to determine the relationship between the roles of HDP2 and MBD7 in affecting DNA methylation and their genomic binding sites. The numbers of differentially methylated cytosines (DMCs) identified in *hdp1-3*, *hdp2-2*, *idm1-1* and *mbd7-2* mutants [14] in every 2-kb genomic regions were plotted against the corresponding ChIP signals of HDP2 and MBD7. We found that there were positive correlations between the ChIP signal and the numbers of hyper DMCs in regions where ChIP signals were high (Figure 7C). These results suggest that the enrichments of HDP2 and MBD7 at genomic regions correlate with the role of IDM complex in DNA demethylation.

Discussion

Although many transposon-derived proteins have been reported, only a few are functionally characterized. Here, we show that HDP1 and HDP2 are derived from the two proteins in *Harbinger* transposons and are present in the same complex, indicating that the components of *Harbinger* transposons can be "co-domesticated" to function in the same pathway. Although the catalytically important "DDE" triad is not conserved in HDP1, HDP2 retains the DNA binding activity. The DNA-binding protein in transposons can recognize specific sequence for transposition; however, due to the weak conservation of the SANT/Myb/trihelix domain, HDP2 may have gained new specificity beyond *Harbinger* terminal sequences. The function of animal HARBI1 and NAIF1 proteins,

Figure 7 HDP2 and MBD7 bind to similar genomic regions and the binding regions coincide with increased DNA methylation in *hdp1*, *hdp2* and *idm1* mutants. **(A)** Venn diagram showing the overlap between peaks identified from HDP2 and MBD7 ChIP-Seqs. N represents the number of peaks calculated from the overlap between two replicates of each ChIP-Seq. The number in brackets represents the percentage of overlapping peaks in total ChIP peaks for HDP2 and MBD7, respectively. **(B)** The distribution patterns of MBD7 and HDP2 ChIP signals across HDP2- and MBD7-enriched regions. For each HDP2 or MBD7 binding peak, the midpoint ("0") was first anchored. Then flanking 5-kb regions were divided into 50-bp bins. Read depth per 1 GB data were calculated for each bin and then plotted against the bins. For each identified binding peak, we randomly generated a region on the genome. The midpoint of each random region was anchored and read depth per 1 GB data over the 5-kb flanking region were calculated and plotted as the control. **(C)** Scatter plots showing the correlation between DMC density and ChIP signals. *Arabidopsis* genome was divided into 2-kb bins. The numbers of DMC identified in *hdp1-3*, *hdp2-2*, *idm1-2* and *mbd7-2* mutants in every 2-kb genomic regions were plotted against the corresponding ChIP signals of HDP2 and MBD7. ChIP signals were calculated as described in [14, 41]. Enrichment = $\log_2(8 + n1) - \log_2(8 + n2 \cdot N1/N2)$. We used n1 and n2 to represent the number of aligned ChIP and input reads in each window. N1 and N2 were the total reads in ChIP and input data, respectively. To overcome sampling noise, eight pseudo counts were added. The DMCs were defined as cytosines whose methylation level in mutant is 10% higher than that in WT plants. *X* and *Y* axis represent ChIP signals of the windows and relative DMC number in 2-kb window. Color key represents kernel density estimate of the regions on the 2-dimensional scatter plot. Horizontal dashed lines represent the average DMC numbers in all the windows.

which were also derived from the *Harbinger* transposase and SANT/Myb/trihelix domain protein, respectively, is unclear [27]. We predict that HARBI1 and NAIF1 in animals may also be involved in epigenetic regulation.

The "domestication" of *Harbinger* transposons into host proteins shows a convergent evolution pattern. Recently, another *Harbinger* transposon-derived protein, ALP1, was found to associate with *Arabidopsis* Polycomb Repressive Complex 2 (PRC2) [30]. ALP1 is present in a different orthologous clade from HDP1 in the phylogenetic tree (Figure 3B), which is consistent with their different functions and suggests that they are from independent "domestications". Previously, multiple "domestication" events of *Harbinger* transposons were also reported in *Drosophila* [19]. Our findings and others suggest that the *Harbinger* transposons have been repeatedly "co-opted" into host proteins [17, 19, 20]. This phenomenon has also been shown for the "domestication" of pogo-like transposases into CENP-B-related proteins [24, 31]. The convergent "domestication" pattern suggests their importance during evolution and that there may be additional *Harbinger* transposon-derived chromatin regulators.

Histone modifying enzymes are usually found in large protein complexes which, in addition to the enzymes, contains DNA- or chromatin-binding proteins, chaperone proteins and scaffold/hinge proteins [32]. Multiple DNA-binding proteins may cooperate to determine the target specificity of the complex [33]. Here, for the IDM complex, we propose that MBD7 and HDP2 jointly determine the genomic targeting of IDM1 (Supplementary information, Figure S7). MBD7 binds to methylated CpG sites, ensuring that the IDM complex is targeted to heavily methylated genomic region [14]. In addition, as a transcription factor-like protein, HDP2 would enable the IDM complex to target regulatory sequences. In support of this notion, we found that hyper-DMRs in *hdp2-1* mutant are enriched over promoters (1 kb upstream) of genes and TEs (Supplementary information, Table S5). Together, MBD7 and HDP2 ensure that the IDM complex is targeted only to highly methylated regulatory sequences, where active DNA demethylation is required for expression. HDP1, IDM2 and IDM3 proteins likely function to connect the DNA recognition components to the core enzyme, IDM1. However, HDP1 may also participate in helping HDP2 for target specificity or play a role in target recognition itself. In addition, IDM2 and IDM3 may also function as chaperone proteins to ensure IDM1 activity, since IDM2 and IDM3 belong to the family of α-crystallin domain proteins, most of which are small heat shock proteins that function as protein chaperones [34]. How IDM1-catalyzed histone acetylation

is connected to the recruitment of the DNA demethylase ROS1 to mediate DNA demethylation is unclear and will require future investigation.

In plants and animals, the expression and activity of TEs and other repetitive sequences are regulated by host epigenetic mechanisms, including DNA methylation and histone modifications [6]. The current work shows that transposon-derived proteins may be components of this regulatory system and this may be one possible mechanism how the epigenetic machinery primarily targets transposons. A transposon-encoded protein in maize was reported previously to also participate in DNA demethylation, although the underlying mechanism was not clear [4]. Our findings uncovered a mechanism for the regulation of TEs in which transposon-derived HDP1 and HDP2 are recruited to a host histone acetyltransferase complex to protect transposons and some endogenous genes from epigenetic silencing.

Materials and Methods

Plant materials and growth conditions

All plants were grown under a long day photoperiod (16-h light/8-h dark). For *35S::SUC2* background plants, 1/2-strength Murashige and Skoog (MS) medium containing 1% sucrose was used to observe the root phenotype and 1% glucose-containing 1/2 MS medium was used for all other experiments. For kanamycin sensitivity experiment, 50 mg/L of kanamycin was used. For DNA methylation inhibitor treatment, 5′Aza-2′-deoxycytidine (Sigma) was added into 1/2 MS medium at the concentration of 5 μM and DMSO was used as negative control. All seedlings on 1/2 MS media were photographed 14 days after germination.

The knockout mutants of *HDP1* and *HDP2* by CRISPR/Cas9 were generated as reported previously [35].

DNA methylation analysis

For DNA methylation-sensitive chop-PCR and WGBS, genomic DNA was extracted from 2-week-old seedlings using DNeasy Plant Mini Kit (Qiagen). Chop-PCR was performed as reported previously [36]. DMRs were identified according to Duan *et al.* [37]. In brief, a sliding-window approach with a 200-bp window sliding at 50-bp intervals was used to identify DMRs. Fisher's exact test was performed for methylated versus unmethylated cytosines for each context, within each window, with FDRs estimated using a Benjamini-Hochberg adjustment of Fisher's *P*-values calculated in the R environment. Windows with an FDR ≤ 0.05 were considered for further analysis, and windows within 100 bp of each other were condensed to larger regions. Regions were then adjusted to extend to DMC at each border. A cytosine was considered differentially methylated if it showed at least a 2-fold change in methylation percentage in the mutant. The regions were then filtered to include only those with at least 10 DMCs and with at least an average of a 2-fold change in methylation percentage per cytosine. The *35S::SUC2* BS-data is from Lang *et al.* [14].

Protein interaction analysis

For the Y2H assay, the full-length and truncated coding regions

of *HDP1*, *HDP2*, *IDM1*, *IDM2*, *IDM3* and *MBD7* were directly cloned into the pGADT7 (AD) and pGBKT7 (BD) vectors. For the split luciferase assay, protein coding sequences were first cloned into entry vector pENTR/D-TOPO (Invitrogen) and then into modified pEarleyGate vectors [14], which bear split luciferase coding sequence, by Gateway LR Clonase II-mediated recombination reaction (Invitrogen). For split luciferase assay, the constructs fused with split luciferase were co-transfected into *Arabidopsis* protoplast (Col-0) for overnight incubation. *Arabidopsis* protoplast was prepared as previously reported [38]. The luciferase activity was determined using CCD camera equipped with Winview software (Princeton instruments). For co-immunoprecipitation assay, the inflorescence tissues from hybrid plants crossed from *3FH-HDP2*/*hdp2-1* and *HDP1-4myc*/*hdp1-1* were collected for total protein extraction using a previously reported method [36]. Total proteins extracted from Col-0 and *3FH-HDP2*/*hdp2-1* plants were used as controls. Total proteins were then precipitated by agarose-conjugated anti-MYC antibody for 2 h at 4 °C. After washing 5 times with protein extraction buffer, the precipitated proteins were separated on 12% SDS-PAGE gel together with input proteins for immunoblot.

Multiple sequence alignment and phylogenetic tree construction

HDP1 protein sequence was used as the query for BLASTP search with NCBI nr protein database. Searched species include *Amborella trichopoda*, *Glycine max*, *Populus trichocarpa*, *Brassica rapa*, *Arabidopsis thaliana*, *Selaginella moellendorffii*, *Zea mays*, *Brachypodium distachyon*, *Oryza stiva*, *Danio rerio* and *Homo sapiens*. Top 250 protein sequences except sequences longer than 2 000 amino acids were used for protein alignment and phylogenetic tree construction. For multiple sequence alignment (MSA; Figure 3A), transposase sequences used include *Harbinge*r-2_Ami_1p, from *Alligator mississippiensis*, *Harbinger-2G_CPB_1p* and *Harbinger-4_CPB_1p*, from *Chrysemys picta bellii*. From the phylogenetic tree, we selected a subtree which includes all *Arabidopsis* protein sequences present in the tree and extracted all sequences with the subtree. The resulted list has 123 sequence entries. HDP2 protein sequence was also used as query for BLASTP search with all available *Harbinger* transposon sequences in Repbase. Top 50 hits based on *E*-value were selected together with the 123 sequences from NCBI nr database for the final phylogenetic tree construction. Command line version ClustalW [39] 2.1 was used for MSA and phylogenetic tree construction of selected protein sequence.

See Supplementary information, Data S1 for more details.

Accession numbers

The *hdp1-1* and *hdp2-1* WGBS data have been deposited to NCBI's Gene Expression Omnibus (GSE85060). The *35S::SUC2* BS-data (GSE58787) was from Lang *et al.* [14]. The WGBS data of the *hdp1* and *hdp2* mutants in Col-0 background, the ChIP-Seq and mRNA-Seq data have been deposited to NCBI's Gene Expression Omnibus (GSE89320).

Acknowledgments

Author Contributions

C-G D, XW and J-K Z designed the study. C-G D, XW, LP, DM, C-C H, ML, YZ, Y-J H, ZW, ZZ, S-K M, HX and HZ conducted the experiments. SX and KT analyzed the whole-genome sequencing data. C-G D, XW, W-A T, BD and J-K Z wrote the paper.

References

1 Britten RJ, Kohne DE. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* 1968; **161**:529-540.

2 Wicker T, Sabot F, Hua-Van A, *et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 2007; **8**:973-982.

3 Mc CB. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA* 1950; **36**:344-355.

4 Cui H, Fedoroff NV. Inducible DNA demethylation mediated by the maize suppressor-mutator transposon-encoded TnpA protein. *Plant Cell* 2002; **14**:2883-2899.

5 Sinzelle L, Izsvák, Ivics Z. Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol Life Sci* 2009; **66**:1073-1093.

6 Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 2010; **11**:204-220.

7 Matzke MA, Mosher RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet* 2014; **15**:394-408.

8 Zhu JK. Active DNA demethylation mediated by DNA glycosylases. *Annu Rev Genet* 2009; **43**:143-166.

9 He XJ, Chen T, Zhu JK. Regulation and function of DNA methylation in plants and animals. *Cell Res* 2011; **21**:442-465.

10 Wu H, Zhang Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell* 2014; **156**:45-68.

11 Gong Z, Morales-Ruiz T, Ariza RR, Roldan-Arjona T, David L, Zhu JK. ROS1, a repressor of transcriptional gene silencing in *Arabidopsis*, encodes a DNA glycosylase/lyase. *Cell* 2002; **111**:803-814.

12 Zhu J, Kapoor A, Sridhar VV, Agius F, Zhu JK. The DNA glycosylase/lyase ROS1 functions in pruning DNA methylation patterns in *Arabidopsis*. *Curr Biol* 2007; **17**:54-59.

13 Qian W, Miki D, Zhang H, *et al.* A histone acetyltransferase regulates active DNA demethylation in *Arabidopsis*. *Science* 2012; **336**:1445-1448.

14 Lang Z, Lei M, Wang X, *et al.* The methyl-CpG-binding protein MBD7 facilitates active DNA demethylation to limit DNA hyper-methylation and transcriptional gene silencing. *Mol Cell* 2015; **57**:971-983.

15 Qian W, Miki D, Lei M, *et al.* Regulation of active DNA demethylation by an α-crystallin domain protein in *Arabidopsis*. *Mol Cell* 2014; **55**:361-371.

16 Wang C, Dong X, Jin D, *et al.* Methyl-CpG-binding domain protein MBD7 is required for active DNA demethylation in *Arabidopsis*. *Plant Physiol* 2015; **167**:905-914.

17 Kapitonov VV, Jurka J. Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol* 2004; **23**:311-324.

18 Zhang X, Jiang N, Feschotte C, Wessler SR. PIF- and Pong-like transposable elements: distribution, evolution and relationship with tourist-like miniature inverted-repeat transposable elements. *Genetics* 2004; **166**:971-986.

19 Casola C, Lawing AM, Betran E, Feschotte C. PIF-like transposons are common in drosophila and have been repeatedly domesticated to generate new host genes. *Mol Biol Evol* 2007; **24**:1872-1888.

20 Hoen DR, Bureau TE. Discovery of novel genes derived from transposable elements using integrative genomic analysis. *Mol Biol Evol* 2015; **32**:1487-1506.

21 Lei M, Zhang H, Julian R, Tang K, Xie S, Zhu JK. Regulatory link between DNA methylation and active demethylation in *Arabidopsis*. *Proc Natl Acad Sci USA* 2015; **112**:3553-3557.

22 Tariq M, Paszkowski J. DNA and histone methylation in plants. *Trends Genet* 2004; **20**:244-251.

23 Jones P, Binns D, Chang HY, *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014; **30**:1236-1240.

24 Casola C, Hucks D, Feschotte C. Convergent domestication of pogo-like transposases into centromere-binding proteins in fission yeast and mammals. *Mol Biol Evol* 2008; **25**:29-41.

25 Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 2015; **6**:11.

26 Yuan YW, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci USA* 2011; **108**:7884-7889.

27 Sinzelle L, Kapitonov VV, Grzela DP, *et al*. Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes. *Proc Natl Acad Sci USA* 2008; **105**:4715-4720.

28 Jurka J, Kapitonov VV. PIFs meet tourists and harbingers: a superfamily reunion. *Proc Natl Acad Sci USA* 2001; **98**:12315-12316.

29 Kaplan-Levy RN, Brewer PB, Quon T, Smyth DR. The trihelix family of transcription factors — light, stress and development. *Trends Plant Sci* 2012; **17**:163-171.

30 Liang SC, Hartwig B, Perera P, *et al*. Kicking against the PRCs — a domesticated transposase antagonises silencing mediated by polycomb group proteins and is an accessory component of polycomb repressive complex 2. *PLoS Genet* 2015; **11**:e1005660.

31 Mateo L, Gonzalez J. Pogo-like transposases have been repeatedly domesticated into CENP-B-related proteins. *Genome Biol Evol* 2014; **6**:2008-2016.

32 DesJarlais R, Tummino PJ. Role of histone-modifying enzymes and their complexes in regulation of chromatin biology. *Biochemistry* 2016; **55**:1584-1599.

33 Siggers T, Gordan R. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res* 2014; **42**:2099-2111.

34 Scharf KD, Siddique M, Vierling E. The expanding family of *Arabidopsis thaliana* small heat stress proteins and a new family of proteins containing alpha-crystallin domains (Acd proteins). *Cell Stress Chaperones* 2001; **6**:225-237.

35 Feng Z, Mao Y, Xu N, *et al*. Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in *Arabidopsis*. *Proc Natl Acad Sci USA* 2014; **111**:4632-4637.

36 Wang X, Duan CG, Tang K, *et al*. RNA-binding protein regulates plant DNA methylation by controlling mRNA processing at the intronic heterochromatin-containing gene IBM1. *Proc Natl Acad Sci USA* 2013; **110**:15467-15472.

37 Duan CG, Zhang H, Tang K, *et al*. Specific but interdependent functions for *Arabidopsis* AGO4 and AGO6 in RNA-directed DNA methylation. *EMBO J* 2015; **34**:581-592.

38 Yoo SD, Cho YH, Sheen J. *Arabidopsis* mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat Protoc* 2007; **2**:1565-1572.

39 Larkin MA, Blackshields G, Brown NP, *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007; **23**:2947-2948.

40 Duan CG, Fang YY, Zhou BJ, *et al*. Suppression of *Arabidopsis* ARGONAUTE1-mediated slicing, transgene-induced RNA silencing, and DNA methylation by distinct domains of the cucumber mosaic virus 2b protein. *Plant Cell* 2012; **24**:259-274.

41 Baubec T, Ivanek R, Lienert F, Schubeler D. Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* 2013; **153**:480-492.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)