A domesticated Harbinger transposase forms a complex with HDA6 and promotes histone H3 deacetylation at genes but not TEs in Arabidopsis

Citation for published version:

Digital Object Identifier (DOI):
10.1111/jipb.13108

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Journal of integrative plant biology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
A domesticated Harbinger transposase forms a complex with HDA6 and promotes histone H3 deacetylation at genes but not TEs in Arabidopsis

Xishi Zhou†, Junna He2,3†, Christos N. Velanis4†, Yiwang Zhu1†, Yuhan He†, Kai Tang3,5, Mingku Zhu3,6, Lisa Graser4,7, Erica deLeau4, Xingang Wang2†, Lingrui Zhang3, W. Andy Tao8, Justin Goodrich4*, Jian-Kang Zhu5* and Cui-Jun Zhang1,5*

ABSTRACT

In eukaryotes, histone acetylation is a major modification on histone N-terminal tails that is tightly connected to transcriptional activation. HDA6 is a histone deacetylase involved in the transcriptional regulation of genes and transposable elements (TEs) in Arabidopsis thaliana. HDA6 has been shown to participate in several complexes in plants, including a conserved SIN3 complex. Here, we uncover a novel protein complex containing HDA6, several Harbinger transposon-derived proteins (HHP1, SANT1, SANT2, SANT3, and SANT4), and MBD domain-containing proteins (MBD1, MBD2, and MBD4). We show that mutations of all four SANT genes in the sant-null mutant cause increased expression of the flowering repressors FLC, MAF4, and MAF5, resulting in a late flowering phenotype. Transcriptome deep sequencing reveals that while the SANT proteins and HDA6 regulate the expression of largely overlapping sets of genes, TE silencing is unaffected in sant-null mutants. Our global histone H3 acetylation profiling shows that SANT proteins and HDA6 modulate gene expression through deacetylation. Collectively, our findings suggest that Harbinger transposon-derived SANT domain-containing proteins are required for histone deacetylation and flowering time control in plants.

Keywords: deacetylation, flowering, Harbinger, HDA6, histone acetylation, protein complex, SANT

INTRODUCTION

In eukaryotic cells, histones and DNA are packed and ordered into highly structured units called nucleosomes. Nucleosomal histones are subject to multiple posttranslational modifications, such as methylation, acetylation, phosphorylation, and ubiquitination, which play critical roles in various biological processes that include transcriptional regulation, chromosome packaging, and repair of DNA damage (Bannister and Kouzarides, 2011; Zhao et al., 2019; Shim et al., 2020). Acetylation of histones occurs at specific lysine residues on their N-terminal tails and is associated with transcriptional activation (Struhl, 1998). Histone acetyltransferases (HATs) and histone deacetylases (HDAs or HDACs) regulate histone acetylation status by adding or removing acetylation marks, respectively (Marmorstein and Earley et al., 2010; Velanis et al., 2020). Histone acetylation is crucial for controlling gene transcription.

The Arabidopsis (Arabidopsis thaliana) genome encodes 18 annotated HDACs that fall into three families based on sequence: the homologs of yeast (Saccharomyces cerevisiae) REDUCED POTASSIUM DEPENDENCY 3 (RPD3), the plant-specific HISTONE DEACETYLASE 2 (HD2), and the Sir2-like family (Hollender and Liu, 2008; Chen et al., 2019; Yruela et al., 2021). HD6A, an RPD3-type HDAC, has been extensively studied in Arabidopsis. HD6A is involved in various biological processes such as transcriptional gene silencing (Probst et al., 2004; Earley et al., 2010; To et al., 2011; Liu et al., 2012; Yu et al., 2017; Yang et al., 2020), responses to drought and salt stress (Chen et al., 2010; Kim et al., 2017), expression of circadian clock genes (Hung et al., 2018), mRNA polyadenylation (Lin et al., 2020), and flowering (Wu et al., 2008; Yu et al., 2011; Ning et al., 2019). The hda6 mutant shows late flowering phenotype due to the upregulation and hyperacetylation of the flowering repressor loci FLOWERING LOCUS C (FLC), MAIZE AFFECTING FLOWERING 4 (MAF4), and MAF5 (Wu et al., 2008; Yu et al., 2011; Ning et al., 2019). In Arabidopsis, FVE and FLOWERING LOCUS D (FLD) interact with HDA6 to repress FLC, MAF4 and MAF5 expression through histone deacetylation and demethylation (Sun et al., 2011; Yu et al., 2012). Recently, HISTONE DEACETYLATION COMPLEX 1 (HDC1), SIN3-LIKE proteins (SNLs), and MULTICOPY SUPPRESSOR OF IRA 1 (MSI1) were reported to be shared subunits of the HDA6 and HDA19 complexes (Ning et al., 2019). Compared to hda6 mutants, which flower later in both long days (LD) and short days (SD), the hda19, hdc1, sni2 sni3 sni4, and msi1 mutants showed delayed flowering time in LD only (Steinbach and Hennig, 2014; Ning et al., 2019), indicating that HDC1, SNLs, and MSI1 may not participate in the HDA6-mediated control of flowering. Notably, transcriptome analysis indicated that HDA6 is involved in the transcriptional regulation of many genes and transposable elements (TEs) (Yu et al., 2016). It was later reported that HDA6 interacts with the histone H3 lysine 9 methyltransferases SU(VAR)3-9 HOMOLOG 4 (SU(VH)4), SU(VH)5, and SU(VH)6 to co-regulate the silencing of TEs (Yu et al., 2017). However, relatively little is known about how HDA6 specifically regulates gene transcription.

Transposable elements are repetitive sequences that constitute a large part of plant and animal genomes. Although usually considered as selfish or parasitic, TEs are important sources of emerging new genes (Sinzel et al., 2009; Bourque et al., 2018). PIF/Harbinger class transposons usually encode two proteins: a nuclease and a SANT/myb/trihelix domain-containing DNA binding protein (Kapitonov and Jurka, 2004; Zhang et al., 2004). Interaction between these two proteins is required to reconstitute transposase activity and effective transposition of PIF/Harbinger elements (Sinzel et al., 2008; Hancock et al., 2010). Several genes that originated by domestication of PIF/Harbinger transposases have been reported to play critical roles in regulating gene expression (Liang et al., 2015; Duan et al., 2017; Velanis et al., 2020). For instance, Arabidopsis HARBINGER DERIVED PROTEIN 1 (HDP1) and HDP2, a pair of anti-silencing factors derived from Harbinger transposons, can interact with one another and also associate with INCREASED DNA METHYLATION 1 (IDM1), IDM2, IDM3 as well as METHYL-CPG-BINDING DOMAIN 7 (MBD7) to form a histone acetyltransferase complex that prevents hypermethylation and silencing of specific loci (Duan et al., 2017). ANTAGONIST OF LIKE HETEROCHROMATIN PROTEIN 1 (ALP1) and ALP2 are another pair of domesticated Harbinger-derived proteins that antagonize silencing mediated by Polycistr group (F1) proteins (Liang et al., 2015; Velanis et al., 2020). ALP1 and ALP2 associate with MSI1, a core component of the H3K27me3 histone methyltransferase complex Polycomb repressive complex 2 (PRC2), and may displace the accessory components EMBRYONIC FLOWER 1 (EMF1) and LIKE HETEROCHROMATIN PROTEIN 1 (LHP1) to form a variant PRC2 complex (Velanis et al., 2020). Phylogenetic analysis indicates that there are at least four other domesticated Harbinger-related nucleases in Arabidopsis (Duan et al., 2017). However, their partners are unknown and it is unclear whether they are also part of histone modifying complexes.

In this study, by using immunoprecipitation combined with mass spectrometry (IP-MS), we found that HDA6 associates with Harbinger transposon-derived proteins (HHP1, SANT1, SANT2, SANT3 and SANT4), and MBD proteins (MBD1, MBD2, and MBD4) in Arabidopsis. Split luciferase assays in Nicotiana benthamiana leaves showed that HHP1 and MBD interact directly with HDA6 and SANT proteins. However, HDA6 and SANT proteins did not interact with one another directly, suggesting that their association is mediated by their interactions with HHP1 and MBD proteins. Mutations of all four Arabidopsis SANT genes increased the expression of the key floral repressors FLC, MAF4, and MAF5, resulting in a late flowering phenotype similar to that of hda6 mutants. However, the hhp1 single mutant and mbd1/2/4 triple mutant had normal flowering times. Transcriptome analysis in the wild type and various mutant backgrounds revealed that HDA6, HHP1, MBD, and SANT proteins regulate the expression of largely overlapping groups of...
Journal of Integrative Plant Biology

genes involved in biotic and abiotic stress responses, with inactivation of HDA6 or SANT proteins having the strongest phenotypes at the molecular level. Furthermore, genome-wide histone H3 acetylation profiling revealed that SANT1, SANT2, SANT3, and SANT4 play important parts in mediating deacetylation of histone H3. Taken together, our data uncover a critical role for Harbinger transposon-derived SANT domain-containing proteins in histone deacetylation and flowering time control in plants.

RESULTS

HHP1 has a function distinct from that of its paralog ALP1

We previously reported that the Arabidopsis genome contains one paralog for ALP1, At3g53540 (Liang et al., 2015), which we refer to hereafter as HDA6-associated Harbinger transposon-derived protein 1 (HHP1) (see below). We obtained a mutant allele, hhp1-1 (Salk_122829), with a T-DNA insertion in the second exon predicted to introduce a premature stop codon in the HHP1 coding sequence. The hhp1-1 mutant did not show any obvious morphological abnormalities when grown under standard conditions (Figure S1A). Given the sequence similarity between ALP1 and HHP1, we postulated that the two genes might function redundantly. However, the hhp1-1 alp1-1 double mutant exhibited the same phenotype as the alp1-1 single mutant (Liang et al., 2015; Figure S1A). Mutation of alp1 partially suppressed the early flowering and leaf curling phenotypes observed in the PcG mutant lhp1, whereas hhp1-1 lhp1-3 double mutant plants were phenotypically similar to the lhp1-3 single mutant (Figure S1B, C). In addition, the morphology and flowering time of the lhp1-3 alp1-1 triple mutant were comparable to those of the lhp1-3 alp1-1 double mutant (Figure S1B, C). Together, these results indicated that HHP1 is unlikely to function like ALP1 as a component of the PRC2 complex.

Domesticated transposase regulates histone deacetylation

Harbinger transposon-derived proteins co-purify with HDA6 and MBD proteins

HDP1 and ALP1 are Harbinger transposase-derived proteins and function as components of two distinct chromatin modifying complexes (Liang et al., 2015; Duan et al., 2017; Velanis et al., 2020). Since HHP1 is also a Harbinger transposon-derived protein (Liang et al., 2015), we next undertook a proteomics approach to test whether HHP1, like ALP1 and HDP1, is part of a chromatin modifying complex. To determine the function of HHP1, we performed immunoprecipitation followed by mass spectrometry (IP-MS) with transgenic plants overexpressing a translational fusion between HHP1 and the green fluorescent protein (HHP1-GFP). We identified HDA6, three MBD proteins (MBD1, MBD2, and MBD4), and four closely related SANT domain-containing proteins among the HHP1-GFP immunoprecipitated products in samples from the transgenic plants, but not the controls (Table 1). Based on its association with HDA6, we designated the ALP1 paralog HDA6-associated Harbinger transposon-derived protein 1 (HHP1); we named the four SANT domain-containing proteins SANT1, SANT2, SANT3, and SANT4 (Table 1). To confirm the association of these proteins, we performed a reciprocal IP-MS experiment using HDA6-Flag transgenic plants. In agreement with the initial IP-MS results, HHP1, SANT proteins (SANT1, SANT2, SANT3, and SANT4) and MBD proteins (MBD1 and MBD4) co-purified with HDA6, confirming that HDA6 and HHP1 interact with the same set of proteins (Table 1).

As an independent test of interaction between these proteins, we performed split luciferase assays in N. benthamiana leaves. We established that HHP1 and MBD proteins (MBD1, MBD2, and MBD4) form homodimers and heterodimers (Figure 1A, B and Figure S2A, B). Similarly, HHP1 and MBD proteins (MBD1, MBD2, and MBD4) interacted with HDA6 and SANT3 and more weakly with SANT1, SANT2, and SANT4 (Figure 1A, B and Figure S2A, B). No interactions were detected between HDA6 and SANT proteins (SANT1, SANT2, SANT3, and SANT4) (Figure 1C and Figure S2C). These

Table 1. List of proteins co-immunoprecipitating with HHP1 and HDA6. The interacting proteins were affinity-purified and analyzed by mass spectrometry

<table>
<thead>
<tr>
<th>Accession</th>
<th>Protein</th>
<th>HHP1-IP (Rep. 1)</th>
<th>HHP1-IP (Rep. 2)</th>
<th>HHP1-IP (Rep. 3)</th>
<th>HDA6-IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT5G63110</td>
<td>HDA6</td>
<td>62.63</td>
<td>18</td>
<td>53.29</td>
<td>17</td>
</tr>
<tr>
<td>AT3G55350</td>
<td>HHP1</td>
<td>74.63</td>
<td>27</td>
<td>74.88</td>
<td>27</td>
</tr>
<tr>
<td>AT2G47820</td>
<td>SANT3</td>
<td>39.01</td>
<td>27</td>
<td>36.52</td>
<td>26</td>
</tr>
<tr>
<td>AT1G50500</td>
<td>SANT4</td>
<td>23.06</td>
<td>19</td>
<td>32.79</td>
<td>23</td>
</tr>
<tr>
<td>AT1G90050</td>
<td>SANT1</td>
<td>28.71</td>
<td>9</td>
<td>42.14</td>
<td>17</td>
</tr>
<tr>
<td>AT1G90040</td>
<td>SANT2</td>
<td>23.27</td>
<td>7</td>
<td>37.43</td>
<td>12</td>
</tr>
<tr>
<td>AT3G63030</td>
<td>MBD4</td>
<td>51.08</td>
<td>8</td>
<td>30.11</td>
<td>5</td>
</tr>
<tr>
<td>AT4G22745</td>
<td>MBD1</td>
<td>47.55</td>
<td>8</td>
<td>19.12</td>
<td>3</td>
</tr>
<tr>
<td>AT5G35330</td>
<td>MBD2</td>
<td>37.50</td>
<td>6</td>
<td>50.00</td>
<td>7</td>
</tr>
</tbody>
</table>

www.jipb.net | Month 2021 | Volume 00 | Issue 00 | 1-13 | 3
results suggested that HHP1 and/or the MBD proteins may act as scaffolds for the association with the HDA6 and SANT proteins in one or more complexes.

In both published examples of domesticated Harbinger transposases in Arabidopsis, the nuclease-derived component (HPD1 or ALP1) retained its cooperation with the protein derived from the DNA-binding domain (HPD2 or ALP2) (Duan et al., 2017; Velanis et al., 2020). The ancestral Harbinger DNA-binding component shows limited similarity to SANT/Myb/trihelix motif proteins (Kapitonov and Jurka, 2004; Duan et al., 2017). As the SANT proteins identified here not only bear a SANT motif (Figure S3A) but also, in most cases (SANT2, SANT3, SANT4), can interact directly with the domesticated Harbinger-related nuclease HHP1, we hypothesized that HHP1 and SANTs arose via the co-domestication of ancestral Harbinger transposase nuclease and DNA binding components. In agreement with this hypothesis, protein sequence alignments revealed conservation across the SANT/Myb/trihelix domains of the four SANT proteins and the DNA-binding component of Harbinger transposases (Figure S4). In addition, HHP1 lacked the DDE motif required for nuclease activity (Li et al., 2015), while the SANT proteins lacked two of the three conserved tryptophan residues necessary for DNA binding activity (Figure S4), consistent with domestication and acquisition of novel functions.

**SANT1, SANT2, SANT3, and SANT4 promote flowering in Arabidopsis**

Because SANT1/2/3/4, HHP1, and MBD1/2/4 physically associate with HDA6, we asked whether any of these proteins share common functions with HDA6 in Arabidopsis. BLAST analysis revealed that SANT1, SANT2, SANT3, and SANT4 are more closely related to each other than to any other Arabidopsis protein with a SANT domain (Figure S3B), indicating that they may function redundantly. We first obtained a mutant allele (Salk_031353) harbouring a T-DNA insertion in the promoter of SANT3 (Figure S5A). To explore redundancy between SANT family members, we generated higher-order sant mutants in the Salk_031353 background using a multiplex CRISPR/Cas9 gene editing system (Zhang et al., 2016) with guide RNAs that targeted all four genes (Figure S5A). Through genotyping, we first identified a higher order mutant bearing large deletions in SANT1, SANT2, and SANT4 (Figure S5B). However, subsequent RT-PCR analysis showed that the expression of SANT3 was reduced but not eliminated by the Salk_031353 T-DNA insertion and we therefore termed this quadruple mutant sant-weak (Figure S5C). After screening further we identified sant-null, containing an additional frameshift mutation causing early translation termination in SANT3, as well as the large deletions in SANT1, SANT2, and SANT4 (Figure S5B). When grown under long day conditions, the sant-weak mutant flowered slightly later than Col-0 (Figure S6), while the sant-null mutant showed an obviously late flowering phenotype, but less severe than that of the hda6 mutant (Figure 2A, B). To test if HHP1 and MBDs are also required for normal flowering in Arabidopsis, we obtained a likely null hhp1 mutant allele (hhp1-2) via CRISPR/Cas9 genome editing and a mbd1/2/4 triple mutant (Figure S7). However, both the hhp1-2 single mutant and the mbd1/2/4 triple mutant flowered at the same time as Col-0 (Figure 2A, B), arguing against the involvement of these genes in flowering time control in Arabidopsis.
Domesticated transposase regulates histone deacetylation

SANT1/2/3/4 and HDA6 co-regulate the expression of flowering repressors

Several studies have shown that the late flowering phenotype of hda6 is the result of increased expression of the flowering repressors FLC, MAF4, and MAF5 (Wu et al., 2008; Yu et al., 2011; Ning et al., 2019). To further study the function of SANT1, SANT2, SANT3, SANT4, HHP1, MBD1, MBD2, and MBD4 in transcriptional regulation of genes related to flowering, we performed transcriptome deep sequencing (RNA-seq) of 12-day-old Col-0, sant-null, hhp1-2, mbd1/2/4, and hda6 seedlings grown in long days. A visual inspection of RNA-seq reads mapping to each locus in the Integrated Genome Viewer (IGV) browser showed that FLC, MAF4, and MAF5 are upregulated strongly in the sant-null and hda6 mutants, and to a lesser extent in hhp1-2 and mbd1/2/4, relative to Col-0 seedlings (Figure 3A–C). RT-qPCR analyses confirmed that all three genes are upregulated in the sant-null and hda6 mutants, and much more modestly so in hhp1-2 and mbd1/2/4 (Figure 3D). These results indicated that SANT1/2/3/4 and HDA6 contribute to the regulation of flowering time in long days and suggested that they may act together to promote flowering.

SANT1/2/3/4 and HDA6 co-regulate global gene expression

HDA6 is a histone deacetylase that functions in transcriptional repression. We therefore tested whether SANT1/2/3/4, HHP1, and MBD1/2/4 shared functions with HDA6 in transcriptional regulation of the Arabidopsis genome. We identified 924 and 1,047 genes that were upregulated in the hda6 and sant-null mutants, respectively, relative to Col-0 samples,
Figure 4. SANT1/2/3/4 and HDA6 co-regulate the expression of genes but not TEs
(A) Number of differentially expressed genes (left) and TEs (right) in hhp1-2, mbd1/2/4, sant-null, and hda6 mutants relative to Col-0 (P < 0.05 and a two-fold cutoff were used). (B) Venn diagrams showing the overlap of upregulated genes (left) and TEs (right) in hhp1-2, mbd1/2/4, sant-null, and hda6 mutants. (C) Heatmap illustrating the transcript levels of hda6-mediated upregulated genes (left) and TEs (right) in hda6 and other indicated mutants. The color scale indicates normalized FPKM values. (D) Boxplot of transcript levels for hda6-mediated upregulated genes (left) and TEs (right) in hda6 and indicated mutants. Each box represents the log2 (FPKM + 1 values of mutants/FPKM + 1 values of Col-0).

and another 239 and 317 that were downregulated, consistent with roles of HDA6 and the SANT proteins in repressing gene expression (Figure 4A; Dataset S1). However, we detected fewer (207 and 357) upregulated genes in the hhp1-2 and mbd1/2/4 mutants, respectively, relative to Col-0 (Figure 4A; Dataset S1). The relatively modest transcriptional changes in these two mutants were consistent with their weaker effects on FLC, MAF4 and MAF5 expression and flowering time relative to the hda6 and sant-null mutant.

Of the 924 upregulated genes in hda6, 511 (55%) genes were also upregulated in the sant-null mutant (Figure 4B; Figure S8). Heatmap and box plot analysis showed that most of the up-regulated genes in hda6 also appeared up-regulated in the sant-null mutant (Figure 4C, D), suggesting that SANT1/2/3/4 and HDA6 have similar functions. Consistently, we also found that most of the up-regulated genes in the sant-null mutant appeared upregulated in hda6 (Figure S9). Although relatively fewer genes were upregulated in hhp1-2 or in mbd1/2/4 than in hda6, a high proportion (about 50%) of them were also upregulated in hda6; the extent of this overlap was statistically significant, as determined by a hypergeometric distribution (Figure S8). Gene ontology (GO) term enrichment analyses revealed that the genes up-regulated in sant-null and hda6 mutants are involved in stress responses (Figure S10). HDA6 has been reported to be required for transcriptional silencing of TEs (Yu et al., 2017). Indeed, 178 TEs were upregulated in the hda6 mutant relative to Col-0 in our RNA-seq data (Figure 4A; Data S2), whereas only 31, 16, and 21 TEs were upregulated in the sant-null, hhp1-2, and mbd1/2/4 mutants, respectively (Figure 4A; Dataset S2). The expression of most of the TEs upregulated in hda6 was not affected in the sant-null, hhp1-2, and mbd1/2/4 mutant backgrounds (Figure 4B-D). These results suggested that even though SANT1/2/3/4, HHP1, and MBD1/2/4 form a complex with HDA6, they are not involved in maintaining TE silencing at heterochromatin. Collectively, these results indicated that SANT1/2/3/4 and HDA6 co-regulate the expression of genes, but not that of TEs.

SANT proteins are required for histone H3 deacetylation
Previous studies revealed that HDA6 is an RPD3-type HDAC critical for histone H3 deacetylation of target sites. The association of SANT proteins with HDA6 prompted us to examine their contribution to histone H3 deacetylation in vivo. We investigated the global landscape of histone H3 acetylation in wild-type, sant-null and hda6 mutant seedlings by chromatin immunoprecipitation followed by sequencing (ChiP-seq). Based on a principal component analysis of three replicates per genotype, the results of the ChiP-seq experiment were consistent and reproducible (Figure 5A). We identified 784 regions with higher levels of histone H3 acetylation in the hda6 mutant (Dataset S3). Compared to wild type, 2,751 regions showed significantly higher levels of histone H3 acetylation in the sant-null mutant (Dataset S3). It is surprising that more regions are affected in sant-null than in
**Domesticated transposase regulates histone deacetylation**

**Figure 5.** SANT1/2/3/4 and HDA6 co-regulate histone acetylation

(A) Principal component analysis of H3KAc-ChIP-seq data. Each dot represents one replicate of the indicated ChIP-seq sample. (B) Heatmap representation of H3KAc levels for upregulated genes identified by RNA-seq in sant-null (left) and hda6 (right). The color scale indicates normalized RPKM values. DEG, differentially expressed gene. (C) Boxplot of H3KAc levels for upregulated genes identified by RNA-seq in sant-null (left) and hda6 (right). Each box represents the RPKM values (reads per kilobase per million mapped reads). (D) Snapshots of the genome browser illustrating expression and H3KAc levels at three representative loci in Col-0, sant-null, and hda6.

**DISCUSSION**

Transposable elements are becoming increasingly recognized as major contributors to genetic diversity and adaptive evolutionary innovations (Oliver et al., 2013; Hosaka and Kakutani, 2018), either as a direct consequence of insertional transposition (Henaff et al., 2014; Makarevitch et al., 2015) or through the neofunctionalization of TE-encoded transposases as host genes following their molecular domestication (Bundock and Hooykaas, 2005; Jangam et al.,...
Domesticated transposase regulates histone deacetylation

Our study adds to the growing body of evidence supporting the recruitment of domesticated transposases by the host core epigenetic machinery as a recurrent theme in the regulation of developmental and/or environmental responses. We show that five co-domesticated proteins derived from the *Harbinger* transposon, HHP1 and SANT1–SANT4, form one or more novel histone deacetylase complexes with HDA6. Moreover, we demonstrate that these complexes are important for regulating the expression of common target genes via histone deacetylation.

### HHP1 is in histone deacetylase complexes with HDA6, MBD and SANT proteins

Our reciprocal IP-MS experiments indicate that HHP1 and HDA6 interact with one another and associate with the MBD1, MBD2, MBD4, and SANT1–SANT4 proteins. These results are also supported by split luciferase assays that show numerous direct interactions between these proteins, including HHP1 with HDA6, as well as MBD1, MBD2, and MBD4 and SANT2, SANT3, and SANT4. However, HDA6 did not interact directly with the SANT proteins, suggesting that the MBD and HHP1 proteins mediate their interaction. Until the biochemical purification of the entire HDA6-containing complex(es), it is unclear whether the four SANT proteins and the three MBD proteins form one or more complexes with HHP1 and HDA6. Given that SANT1–SANT4 are closely related, as are MBD1, MBD2, and MBD4 (Zemach and Grafi, 2003), redundancy between individual SANT or MBD protein is likely, raising the possibility that various HDA6/HHP1/MBD/SANT subcomplexes may exist within cells but differ in their constituent SANT or MBD member. If these interactions are biologically relevant, a prediction is that inactivation of the different components of the complex(es), as with the *hda6* mutant or higher-order mutants in the case of *MBD1/2/4* or *SANT1/2/34* genes, will result in overlapping phenotypes. Our transcriptome analysis of 12-day-old wild-type and mutant seedlings strongly supported this hypothesis. In particular, the overlap between genes misregulated in the mutants affecting the different components was statistically significant and much higher than would be expected by chance (Figure S8). However, there are differences in phenotypic severity, with *sant-null* and *hda6* having a much greater effect on the transcriptome than *hhp1* or *mbd1/2/4* (Figure 4). The HHP1 and MBD proteins are presumably less critical for the function of the complex(es) than HDA6 and SANT1–SANT4. The split luciferase assays indicate that HHP1 and MBD proteins can each interact with both HDA6 and SANT proteins, so one possibility is that inactivation of both HHP1 and the MBD proteins is necessary to fully disrupt the complexes. Although the MBD domain was identified based on its binding to methylated CG dinucleotides (mCG) (Meehan et al., 1998), a previous study suggests that *Arabidopsis* MBD1, MBD2, and MBD4, unlike MBD5, MBD6, and MBD7, does not bind mCG (Zemach and Grafi, 2003). Proteins with MBD domains are frequently associated with histone deacetylases, both in animals and in plants (Nan et al., 1998; Ng et al., 1999; Zemach and Grafi, 2003). Thus, MBD1/2/4 may act as scaffold proteins involved in recruiting HDA6 and SANT1–SANT 4 rather than binding to mCG.

### Preferential recruitment of domesticated *Harbinger* transposase to chromatin-modifying complexes

The two components of *Harbinger* transposases are typically domesticated as an interacting pair, as with ALP1/ALP2 and HDP1/HDP2 in plants or Harb1/Naif in vertebrates (Sinzelle et al., 2008; Liang et al., 2015; Duan et al., 2017; Velanis et al., 2020). The sequence similarity between SANT1–SANT4 and the SANT/Myb/trihelix proteins of *Harbingers*, together with the ability of HHP1 and SANT2, SANT3, and SANT4 to interact directly, indicate that HHP1 and SANT1–SANT4 likely arose through co-domestication of ancestral *Harbinger*-encoded transposases. As observed previously (Duan et al., 2017; Velanis et al., 2020), the nuclease component tends to diverge less and is more easily identifiable than the SANT/myb/trihelix component of the complexes. Neither retain their original functions — HHP1 lacks the DDE catalytic triad necessary for nuclease activity, and SANT1–SANT4 lack the three tryptophan residues needed for DNA binding. Strikingly, in all three cases where domesticated *Harbinger* transposases have been functionally characterized in *Arabidopsis*, they appear to have been repurposed as components of different chromatin-modifying complexes: ALP1 and ALP2 are part of a Polycomb H3K27me3 histone methyltransferase complex, HDP1 and HDP2 are in a histone acetyltransferase complex, and HHP1 and SANT1–SANT4 in a histone deacetylase complex (Figure 6). One distinguishing feature of *Harbinger* TE’s that may contribute to this status is their preferential insertion close to genes (Jiang et al., 2003; Zhang et al., 2004; Hancock et al., 2011), which may have favored the association of the host epigenetic machinery with *Harbinger* transposases, initially to recruit chromatin modifiers to the new TE and limit their effects on adjacent host gene activity.

### The SANT complex influences histone deacetylation and flowering

The histone acetylation profiles and differentially expressed genes identified among the various genotypes tested here suggest that HDA6 and SANT proteins act together to repress common targets via histone deacetylation. Notably, the *sant-null* mutant and *hda6* mutant, and to a lesser extent the *hhp1* and *mbd1/2/4* mutants, affect gene expression similarly, whereas only the *hda6* mutant deregulates TE expression. We hypothesize that HDA6 mediates TE expression by participating in different complexes, such as those containing the H3K9me2 histone methyltransferases SUVH4, SUVH5, and SUVH6 (Yu et al., 2017). The SANT complex instead promotes flowering, by repressing the flowering repressors FLC, MAF4, and MAF5. Furthermore, GO term enrichment analysis reveals that a significant fraction of the genes that are differentially expressed in *sant-null* and *hda6*
Domesticated transposase regulates histone deacetylation harboring a 16-bp deletion (Figure S7A), was also generated by CRISPR/Cas9-mediated editing, using the pHEE401 vector (plasmid #71286, Addgene) and a sgRNA with the sequence GUGUAGCUJiGGGCiGUAGG. The higher-order sant mutants, sant-weak and sant-null, both in the Salk_031353 background, were generated in the sant3 (Salk_031353) mutant background via multiplex CRISPR/Cas9-mediated editing (Zhang et al., 2016) using a combination of three sgRNA pairs that targeted the four SANT genes (Figure S5). The primers used for genotyping and for sgRNA production are listed in Dataset S5.

The 35 S::HHP1-GFP construct was generated by Gateway recombination between an entry clone containing the HHP1 coding sequence and the pGWB5 destination vector. The construct was introduced into Col-0 background by Agrobacterium (Agrobacterium tumefaciens)-mediated transformation via the floral dipping method. The complementation lines of sant-null were generated by cloning the SANT1, SANT3, and SANT4 genomic coding regions and regulatory sequences into a modified pCAMBIA1305 vector, and subsequent Agrobacterium-mediated transformation of sant-null mutants via the floral dipping method. The primers used for plasmids construction are listed in Dataset S5.

Arabidopsis seeds were grown on half-strength Murashige and Skoog (MS) medium under a long-day photoperiod (16 h light, 22°C/8 h darkness, 20°C), and then transferred to soil in growth chambers with the same conditions. To measure flowering time, the number of rosette leaves and number of days were recorded when the inflorescence stem reached 1 cm.

**Immunoprecipitation and mass spectrometry (MS)**

Flower tissue (3 g) collected from transgenic or wild-type plants was used for affinity purification. Tissue was ground in liquid nitrogen and homogenized in 15 mL of lysis buffer (50 mM Tris, pH 7.6, 150 mM NaCl, 5 mM MgCl2, 10% glycerol, 0.1% NP-40, 0.5 mM dithiothreitol [DTT], 1 mM PMSF, protease inhibitor cocktail 1 tab/50 mL, Roche) for 10 min at 4°C. Following centrifugation, the supernatant was incubated with 10 μg of anti-Flag (Sigma, F1804) antibody or anti-GFP antibody (Invitrogen, A-11122) and 100 μL of Dynabeads Protein G (Invitrogen, 10003D) for 3 h at 4°C with agitation. Beads were then washed three times with lysis buffer and three times with wash buffer (150 mM NaCl, 50 mM Tris–HCl pH 8.0, 5 mM MgCl2). Immunoprecipitated proteins were subjected to liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis as previously described (Zhang et al., 2018).

**Materials and methods**

**Plant materials and growth conditions**

*Arabidopsis thaliana* accession Columbia-0 (Col-0) was used as the wild type. The T-DNA insertion lines *hhp1-1* (Salk_122829), *mbd1* (Salk_025352), *mbd4* (Salk_042834), and *sant3* (Salk_031353), as well as the EMS *hda6* allele (*axe1-5*), were obtained from the Arabidopsis Biological Resource Center. The *mbd2* mutant was generated via CRISPR/Cas9 genome editing, resulting in a single-nucleotide insertion in the first exon of *MBD2* (Figure S7B). The *mbd1*/*mbd2*/*mbd4* triple mutant was generated by genetic crossing of the *mbd1*, *mbd2*, and *mbd4* single mutants. The *hhp1-2* mutant, mutants take part in defense and abiotic stress responses, hinting that the novel histone deacetylase complex(es) might be crucial for coordinating a plethora of immunity and/or adaptive physiological responses.

**MATERIALS AND METHODS**

**Plant materials and growth conditions**

*Arabidopsis thaliana* accession Columbia-0 (Col-0) was used as the wild type. The T-DNA insertion lines *hhp1-1* (Salk_122829), *mbd1* (Salk_025352), *mbd4* (Salk_042834), and *sant3* (Salk_031353), as well as the EMS *hda6* allele (*axe1-5*), were obtained from the Arabidopsis Biological Resource Center. The *mbd2* mutant was generated via CRISPR/Cas9 genome editing, resulting in a single-nucleotide insertion in the first exon of *MBD2* (Figure S7B). The *mbd1*/*mbd2*/*mbd4* triple mutant was generated by genetic crossing of the *mbd1*, *mbd2*, and *mbd4* single mutants. The *hhp1-2* mutant,
Domesticated transposase regulates histone deacetylation

buffer (10 mM MES, pH 5.6, 10 mM MgCl2, and 100 μM aceto-syringone) to a final cell density of OD600 = 0.8. Equal volumes of resuspended cells were mixed in different combinations and incubated at room temperature for 2 h, before infiltration into the abaxial surface of Nicotiana benthamiana leaves. After infiltration, plants were returned to normal growth conditions for 48 h. Luciferase activity was detected with a luminescence imaging system (Princeton Instrument). The primers used for related plasmid construction are listed in Dataset S5.

**Transcriptome deep sequencing (RNA-seq) and data analysis**

We collected triplicate samples for the wild type, hhp1-2, mbd1/2/4, sant-null, and hda6 for RNA-seq analysis. Total RNA was extracted from 12-day-old seedlings grown under long day conditions, with TRIzol reagent (Invitrogen) and then sent to Novogene Corporation (Beijing, China) for library construction and sequencing. Libraries were generated according to the manufacturer’s instructions with the NEBNext UltraTM RNA Library Prep Kit for Illumina (NEB, USA) and sequenced on an Illumina NovaSeq 6000 platform. After removing adapter sequences and low-quality reads, clean reads were mapped to the TAIR10 Arabidopsis genome using TopHat2 with the parameter “--2-b-very-sensitive” (Kim et al., 2013). Mapped reads were visualized with Integrative Genomics Viewer (IGV) from the BAM files. Differential expression of genes and transposable elements was determined using Cuffdiff (|log2 fold change| > 1) within Cufflinks (Trapnell et al., 2013). Heatmaps and boxplots of differentially expressed genes and TEs were generated in R. Gene ontology (GO) enrichment analysis was performed using the R package clusterProfiler in Bioconductor (Yu et al., 2012). The raw reads were deposited at the Gene Expression Omnibus (GEO) database (accession number: GSE167288).

**Analyses of transcript levels by RT-qPCR**

Total RNA from 12-day-old Arabidopsis seedlings grown on half-strength MS medium was extracted with TRIzol reagent (Invitrogen) and treated with DNase I (Takara) to remove contaminating genomic DNA. Total RNA was then reverse-transcribed into cDNA with PrimeScript RT reagent Kit (Takara, RR047A). Real-time quantitative PCR was performed on an Applied Biosystems 7500 Real-Time PCR System using ChamQ Universal SYBR qPCR Master Mix (Vazyme). ACTIN7 was used as reference. Three biological replicates were performed for real-time PCR. The primers sequences can be found in Dataset S5.

**Chromatin immunoprecipitation followed by sequencing (ChIP-seq) and data analysis**

Chromatin immunoprecipitation (ChIP) was performed as previously described, with minor modifications (Zhu et al., 2012). Briefly, 2 g of 12-day-old seedlings grown on half-strength MS medium was collected and fixed in 1% formaldehyde for 15 min and then ground into powder in liquid nitrogen. Nuclei were isolated and chromatin was sheared into 200- to 500-bp fragments by sonication with a Bioruptor (Diagenode). After centrifugation, the sonicated chromatin was incubated with anti-acetylated Histone H3 antibody (Merck Millipore, 06-599) overnight and then incubated with Dynabeads Protein G (Invitrogen, 10003D) for 2 h with agitation at 4°C. The precipitated chromatin was washed and eluted with elution buffer (0.5% SDS and 0.1 M NaHCO3) at room temperature, then concentrated via phenol–chloroform extraction and ethanol precipitation.

Immunoprecipitated DNA from three biological replicates were sent to Novogene Corporation (Beijing, China) for library construction and sequencing (Illumina NovaSeq, 6000; 150-bp pair-end reads). Adapter sequences and low-quality reads were removed from the raw data, after which clean reads were mapped to the TAIR10 Arabidopsis genome using Bowtie2 with default parameters; PCR duplicates were removed by using the Sambamba program (Langmead and Salzberg, 2012; Tarasov et al., 2015). Enriched ChIP peaks were identified with the MACS2 program and differentially enriched peaks were determined by the R package DiffBind in Bioconductor (FDR < 0.05; log2 (fold change) > 1). The read count of each gene region was calculated with the “intersect” command from the BEDTools suite (Quinlan, 2014). The H3KAc level of each gene is given as reads per kilobase per million (RPKM). The heatmap and box-plots of H3 acetylation levels in upregulated genes were generated in R. The raw H3Ac ChIP-seq data were deposited at the GEO database (accession number: GSE167288).

**ACKNOWLEDGEMENTS**

We thank the Arabidopsis Biological Resource Center for providing Arabidopsis mutant lines used in this study. This work was supported by grants from the National Key Research and Development Program of China (2020YFE0202300), the Central Public-interest Scientific Institution Basal Research Fund, the BBRC under the Grant Reference BB/P008569/1 to J.G. C.N.V. and E.dL., and an Erasmus plus training award to L.G.

**AUTHOR CONTRIBUTIONS**


**REFERENCES**

Domesticated transpose regulates histone deacylation


Domesticated transposase regulates histone deacetylation


SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article: http://onlineibrary.wiley.com doi/10.1111/j.1365-3108_suppinfo

Figure S1. Phenotypic and genetic analysis of hhp1 mutants
(A) Representative photographs of 53-day-old plants grown in short days. (B) Representative photographs of 39-day-old plants grown in short days.
(C) Summary of flowering time in short days, expressed as rosette leaf number. Different letters above bars indicate significant differences (two-tailed Student’s t-test, P < 0.05). Data are shown as means ± standard error of the mean. Scale bars, 1 cm.

Figure S2. Exploration of the interaction between HHP1, HDA6, MB1, MB2, and SANT1, SANT2, SANT3, and SANT4, as determined by split luciferase assays
(A) Luciferase complementation imaging (LCI) assay showing that MB2 interacts strongly with HHP1, MB1, MB2, MB4, and HDA6 and weakly with SANT1–SANT3. (B) LCI assay showing that MB4 interacts strongly with HHP1, MB1, MB2, MB4, HDA6, and SANT3 and weakly with SANT1, SANT2, and SANT4. (C–F) LCI assays showing that SANT1, SANT2, and SANT4 do not interact with one another.

Figure S3. Domain structure and phylogenetic analysis of SANT proteins
(A) Domain structure of SANT proteins. The domain architecture of each protein was determined from the SMART (Simple Modular Architecture Research Tool) database. Magenta denotes low-complexity regions. (B) Phylogenetic tree of the domesticated transposases ALP2 and HDP2, SANT protein, and Harbinger transposase DNA binding proteins.

Figure S4. SANT proteins share similarity with Harbinger transposase DNA binding proteins
Alignment of protein sequences for SANT1, SANT2, SANT3, and SANT4 transposase DNA binding proteins. The three tryptophan residues that are highly conserved among Harbinger DNA binding proteins are highlighted in red boxes.

Figure S5. Genotyping analysis of sant mutants
(A) Schematic diagrams of SANT1, SANT2, SANT3, and SANT4 loci. Dark blue and black boxes represent untranslated regions and exons, respectively, and black lines indicate introns. The black and orange arrows indicate the locations of PCR primers used in genotyping analysis and sgRNAs used for multiplex CRISPR/Cas9 editing, respectively. The same two sgRNAs targeted both SANT1 and SANT2, which share very high sequence similarity. Black triangle indicates T-DNA insertion site of Salk_031353. (B) RT-PCR analysis of SANT3 gene expression in Col-0 and the sant-weak and sant-null mutants. ACTIN7 was included as reference. (C) Genotyping analysis of SANT1, SANT2, and SANT4 in Col-0, sant-weak, and sant-null. (D) Electropherogram of the SANT3 gene region in Col-0, sant-weak, and sant-null targeted by genome editing. The red dashed box indicates the 1st insertion in sant-null compared to Col-0. (E) Relative transcript levels of At1g50830, At1g71890, and At5g54095, as determined by RT-qPCR in Col-0, the sant-null mutant, and transgenic plants carrying the SANT1, SANT2, or SANT4 genes introduced in the sant-null background. Error bars indicate the standard deviations of three technical repeats.

Figure S6. Flowering phenotypes of Col-0 and sant-weak and sant-null mutants grown in long-day conditions
Days to bolting (left) and number of rosette leaves upon bolting (right) for Col-0, sant-weak, and sant-null plants grown in long-day conditions. At least
20 plants were scored for each line. The y-axes denote days to bolting (left) and rosette leaf number (right). Each dot represents one plant. Different letters indicate statistically significant differences (ANOVA; Tukey HSD’s multiple range test, P < 0.05).

Figure S7. Genotyping analysis of hhp1-2 single mutant and mbd1/2/4 triple mutants
(A) Electropherogram of HHP1 in Col-0 and the hhp1-2 mutant over the region targeted by genome editing. (B) Electropherogram of MBD2 in Col-0 and the mbd1/2/4 mutant over the region targeted by genome editing. The red dashed box indicates the 1-bp insertion in the mutant compared to wild type. (C) Schematic diagrams of the MBD1 and MBD4 loci indicating the positions of T-DNA insertion lines. Gray and black boxes represent untranslated regions and exons, respectively, and black lines indicate introns. The arrows indicate PCR primers used for genotyping. (D) Genotyping analysis of mbd1 and mbd4 T-DNA mutant alleles.

Figure S8. Venn diagrams showing the overlap of up-regulated (P-value<0.05 and a twofold cutoff was used) genes identified in hda6, hhp1-2, mbd1/2/4 and sant-null. The P-values were calculated using the hypergeometric distribution.

Figure S9. Transcriptome analysis of genes upregulated in the sant-null mutant
(A) Heatmap representation of normalized expression in Col-0 and the indicated mutants of genes that were upregulated in sant-null. The color scale indicates normalized FPKM values. (B) Boxplot showing the expression in the indicated mutants of genes that were upregulated in sant-null. Each box represents log2(FPKM+1 values of mutants/FPKM+1 values of Col-0).

Figure S10. Gene Ontology (GO) term enrichment analysis of upregulated genes in hda6 (left) and sant-null (right). The top 20 most enriched GO biological processes are shown as a bubble chart. The gene number (count) and p.adjust are given.

Figure S11. Heatmap representation (A) and boxplot (B) of H3KAc levels for regions with higher levels of histone H3 acetylation in sant-null (left) and hda6 (right). The color scale indicates normalized FPKM values.

Figure S12. The histone H3 acetylation levels at FLC, MAF4, and MAF5 as determined by ChIP-seq. Snapshots of the genome browser and bar graphs illustrate H3KAc levels at FLC, MAF4, and MAF5 in Col-0, sant-null, and hda6. The results are from three replicates.

Dataset 1. RNA transcript levels of differentially expressed protein coding genes in hhp1-2
Dataset 2. RNA transcript levels of differentially expressed TEs in hhp1-2
Dataset 3. List of higher levels of H3Ac peaks in sant-null mutant
Dataset 4. H3KAc levels of hda6-mediated up-regulated genes
Dataset 5. List of primers used in this study