Fast co-evolution of anti-silencing systems shapes the invasiveness of Mu -like DNA transposons in eudicots

Taku Sasaki, Kyudo Ro, Erwann Caillieux, Riku Manabe, Grégoire Bohl-Viallefond, Pierre Baduel, Vincent Colot, Tetsuji Kakutani, Leandro Quadrana

Article

TRANSPARENT PROCESS

THE EMBO JOURNAL

Fast co-evolution of anti-silencing systems shapes the invasiveness of *Mu*-like DNA transposons in eudicots

Taku Sasaki[1,*,†], Kyudo Ro[1,†], Erwann Caillieux[2], Riku Manabe[1], Grégoire Bohl-Viallefond[2], Pierre Baduel[2], Vincent Colot[2], Tetsuji Kakutani[1] & Leandro Quadrana[2,**]

Abstract

Transposable elements (TEs) constitute a major threat to genome stability and are therefore typically silenced by epigenetic mechanisms. In response, some TEs have evolved counteracting systems to suppress epigenetic silencing. In the model plant *Arabidopsis thaliana*, two such anti-silencing systems have been identified and found to be mediated by the VANC DNA-binding proteins encoded by *VANDAL* transposons. Here, we show that anti-silencing systems have rapidly diversified since their origin in eudicots by gaining and losing VANC-containing domains, such as DUF1985, DUF287, and Ulp1, as well as target sequence motifs. We further demonstrate that these motifs determine anti-silencing specificity by sequence, density, and helical periodicity. Moreover, such rapid diversification yielded at least 10 distinct VANC-induced anti-silencing systems in *Arabidopsis*. Strikingly, anti-silencing of non-autonomous *VANDAL*s, which can act as reservoirs of 24-nt small RNAs, is critical to prevent the demise of cognate autonomous TEs and to ensure their propagation. Our findings illustrate how complex co-evolutionary dynamics between TEs and host suppression pathways have shaped the emergence of new epigenetic control mechanisms.

Keywords anti-silencing; epigenetics; evolution; transposable elements
Subject Categories Chromatin, Transcription & Genomics; Plant Biology
DOI 10.15252/embj.2021110070 | Received 28 October 2021 | Revised 10 February 2022 | Accepted 15 February 2022
The EMBO Journal (2022) e110070

Introduction

Transposable elements (TEs) are ubiquitous DNA sequences that move and self-propagate across genomes. Because of their potential to create large effect mutations upon transposition or by facilitating chromosomal rearrangements through recombination, TEs constitute a major threat to genome function and integrity. However, TEs are usually under tight epigenetic control, notably by DNA methylation in plants and mammals (Slotkin & Martienssen, 2007), thus limiting their mutational impact. In the model plant *Arabidopsis thaliana*, mutants deficient in the chromatin remodeler *DDM1* lose DNA methylation over most TE sequences and reactivate transcriptionally several hundreds of these, which results in increased transposition activity in few cases (Miura *et al,* 2001; Singer *et al,* 2001; Tsukahara *et al,* 2009; Quadrana *et al,* 2019). Also, many TEs transpose frequently in nature (Baduel *et al,* 2021), implying that they do occasionally evade repressive mechanisms, thus ensuring their continuous propagation.

How TEs escape epigenetic silencing remains largely unknown, except for the notable example of silencing suppression deployed by the *VANDAL21* and *VANDAL6 Mutator*-like DNA transposons (Fu *et al,* 2013; Hosaka *et al,* 2017), which are abundant in the *A. thaliana* genome (Kapitonov & Jurka, 1999). Specifically, these two TEs encode each a distinct VANC anti-silencing protein, VANC21 or VANC6, which binds to distinctive short DNA motifs accumulated in non-coding regions of cognate *VANDAL* copies, where they induce strong hypomethylation and transcriptional derepression (Hosaka *et al,* 2017). Furthermore, a *VANDAL21* copy, called *Hiun* (*Hi*), mobilized in *ddm1* (Tsukahara *et al,* 2009) and can also be mobilized in wild-type background by transgenic expression of VANC21 protein (Fu *et al,* 2013). Unlike viral suppressors, which neutralize host defense responses broadly, VANC-induced anti-silencing is highly specific, as only related TE sequences are epigenetically reactivated. Thus, co-evolution of VANC proteins and target DNA motifs may allow specific *VANDAL*s to escape epigenetic silencing and propagate through the genome while minimizing host damage (Hosaka *et al,* 2017). Nonetheless, once activated, VANC-induced anti-silencing should perpetuate the epigenetically active

1 Department of Biological Sciences, University of Tokyo, Bunkyo-ku, Tokyo, Japan
2 Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique (CNRS), Institut National de la Santé et de la Recherche Médicale (INSERM), Ecole Normale Supérieure, PSL Research University, Paris, France
*Corresponding author. Tel: +81 3 5841 4456; E-mail: taku.sasaki@bs.s.u-tokyo.ac.jp
**Corresponding author. Tel: +33 1 69 15 33 32; E-mail: leandro.quadrana@universite-paris-saclay.fr
†These authors contributed equally to this work
Present address: Institute of Plant Sciences Paris-Saclay, Centre Nationale de la Recherche Scientifique, Institut National de la Recherche Agronomique, Université Evry, Université Paris-Saclay, Orsay, France

state of target TEs, potentially leading to runaway transposition. However, wild-type genomes typically contain very few full-length *VANDAL* copies, implying that VANC-mediated anti-silencing must be interrupted at some point through a still unknown mechanism.

The *A. thaliana* genome contains 28 distinct *VANDAL* families in total. Previous analysis of F1 hybrids derived from a cross between wild type and a mutant defective in the DNA methyltransferase MET1, homolog of the mammalian DNMT1, identified lower methylation levels than the expected mid-parental values for some *VANDAL*s sequences, suggesting that they might be subjected to trans-hypomethylation (Rigal *et al,* 2016).

Using a population of *ddm1*-derived epigenetic recombinant inbred lines (epiRILs), we have previously found that transposed copies of the active *VANDAL21* copy *Hi* induce efficient trans-hypomethylation of homologous copies (Fu *et al,* 2013), indicating that this population provides a powerful tool for the systematic study of the anti-silencing factors encoded by the *VANDAL* superfamily of TEs. Here, by combining methylome data for the epiRILs, quantitative epigenetics approaches, and ectopic expression of TE-derived sequences, we have identified the complete set of active *VANDAL*-encoded anti-silencing systems in *A. thaliana*. Our results indicate that since their likely origin in the common ancestor of eudicots, VANCs and their target sequences diversified extensively. Furthermore, the *A. thaliana* VANC1-encoded anti-silencing system produced by a *VANDAL1* copy has conserved features of the most likely ancestral system, including a Ulp1 protein domain and targeting of palindromic DNA sequences. We also show that target specificity of the distinct VANCs is determined by the sequence, density and spatial arrangement of ~10-bp-long motifs, which exhibit helical periodicity and hint to a cooperative DNA binding and homodimerization of VANCs. Last, we demonstrate that non-autonomous *VANDALs*, which can serve as a reservoir of trans-matching small RNAs, are also major targets of VANC-induced anti-silencing and that impairing this targeting by removal of the short-sequence motifs triggers strong and concerted epigenetic re-silencing of cognate autonomous copies. Together, our findings revealed the complex interplay between host silencing and TEs, as well as their interactions between autonomous and non-autonomous copies that shaped the co-evolution of *VANDALs* and have potentially contributed to the emergence of novel gene control mechanisms.

Results

Extensive trans-hypomethylation of *VANDAL*s in *A. thaliana*

We set out to determine whether other *VANDAL* copies in addition to *Hi* are also subjected to trans-hypomethylation in the *ddm1*-derived epiRILs. These lines have almost identical DNA sequences but segregate many differences in DNA methylation (Colomé-Tatché *et al*, 2012) as well as few transpositionally active TEs (Quadrana *et al,* 2019). However, with the exception of *Hi*, no *VANDAL* mobilized in the epiRILs (Quadrana *et al,* 2019), enabling us to test DNA methylation levels over these TEs in the absence of confounding effects due to ongoing transposition. EpiRILs were derived from an initial cross between two isogenic individuals, one carrying a mutant allele of *DDM1* and one WT. A single F1 was then backcrossed to the WT parental line and F2 *DDM1/DDM1* progenies were propagated for six generations to generate a population of plants with mosaic epigenomes (Johannes *et al,* 2009); (Fig 1A). We obtained whole-genome bisulfite sequencing (WGBS) data for a core collection of 16 epiRILs together with siblings of the two founder plants. Overall, single-cytosine resolution methylomes confirmed the epihaplotype maps previously obtained using MeDIP and microarray-based methylomes (Appendix Fig S1). Given the crossing scheme used to derive the epiRILs (Fig 1A), around 75% of their genome on average is of WT origin and exhibit indeed WT-like methylation levels (Appendix Fig S2). Based on this property, we reasoned that putative trans-hypomethylation of *VANDAL* sequences should be readily detected in the epiRILs as local DNA methylation losses over wt-derived copies. We therefore analyzed the DNA methylation levels of wt-derived *VANDALs* (Fig 1B) and detected 244 hypomethylated copies belonging to 20 *VANDAL* families, including several *VANDAL21* sequences (Fig 1C and D). Only a small number of copies were affected per family (between 3 and 33 copies). Hypomethylation occurs both at CG and non-CG sites, but to different degrees (Fig 1D). While non-CG hypomethylation affects entire *VANDAL* sequences, CG hypomethylation is limited to short regions, resembling the sequence-specific DNA methylation loss induced by VANC21 and VANC6 (Fu *et al,* 2013; Hosaka *et al,* 2017). Importantly, hypomethylated *VANDALs* were present only in ~25% of the epiRILs that carry the corresponding wt-derived interval (Fig 1B and Appendix Fig S3). This last observation suggests that other loci, which should segregate independently of the wt-derived *VANDALs* in most cases, induce hypomethylation *in trans* and hypomethylated *VANDAL*s inherited from the *ddm1* parent are obvious candidates. Altogether, our results indicate that most *VANDAL* families may be targets of anti-silencing systems in *A. thaliana*.

Sequence and syntax of motifs determine anti-silencing specificity

Previous *in vitro* and *in vivo* studies revealed that VANC21 and VANC6, respectively, bind the short-sequence motif "YAGTATTAY" and "AGTTGTMC" (where Y can be T or C; and M can be A or C). These motifs are located within non-coding regions of *VANDAL21* and *VANDAL6* sequences, where they induce local CG hypomethylation (Fu *et al,* 2013; Hosaka *et al,* 2017). Thus, we searched in the epiRILs for short-sequence motifs overrepresented within *VANDAL21* sequences that lose CG methylation in the epiRILs and identified in this way a strong overrepresentation of the motif "YAGTATTAC" (Fig 2A). This result confirms the pattern described for VANC21-binding sites (Hosaka *et al,* 2017) with hypomethylation around short-sequence motifs extending much further at non-CG than CG sites.

Following this first confirmation of our approach, we set out to use the epiRILs to characterize the hypomethylation of all *VANDALs*. We searched for short-sequence motifs overrepresented at CG-hypomethylated regions within wt-derived *VANDAL* sequences. We detected statistically overrepresented motifs for all TE families analyzed and in all cases local CG and broad non-CG hypomethylation is observed around detected motifs (Fig 2A and B, Appendix Figs S4 and S5), reminiscent to the VANC21- and VANC6-induced loss of DNA methylation (Fu *et al,* 2013; Hosaka *et al,* 2017). Consistent with only few copies per family being trans-hypomethylated (Fig 1C), only a small fraction of *VANDALs* carries

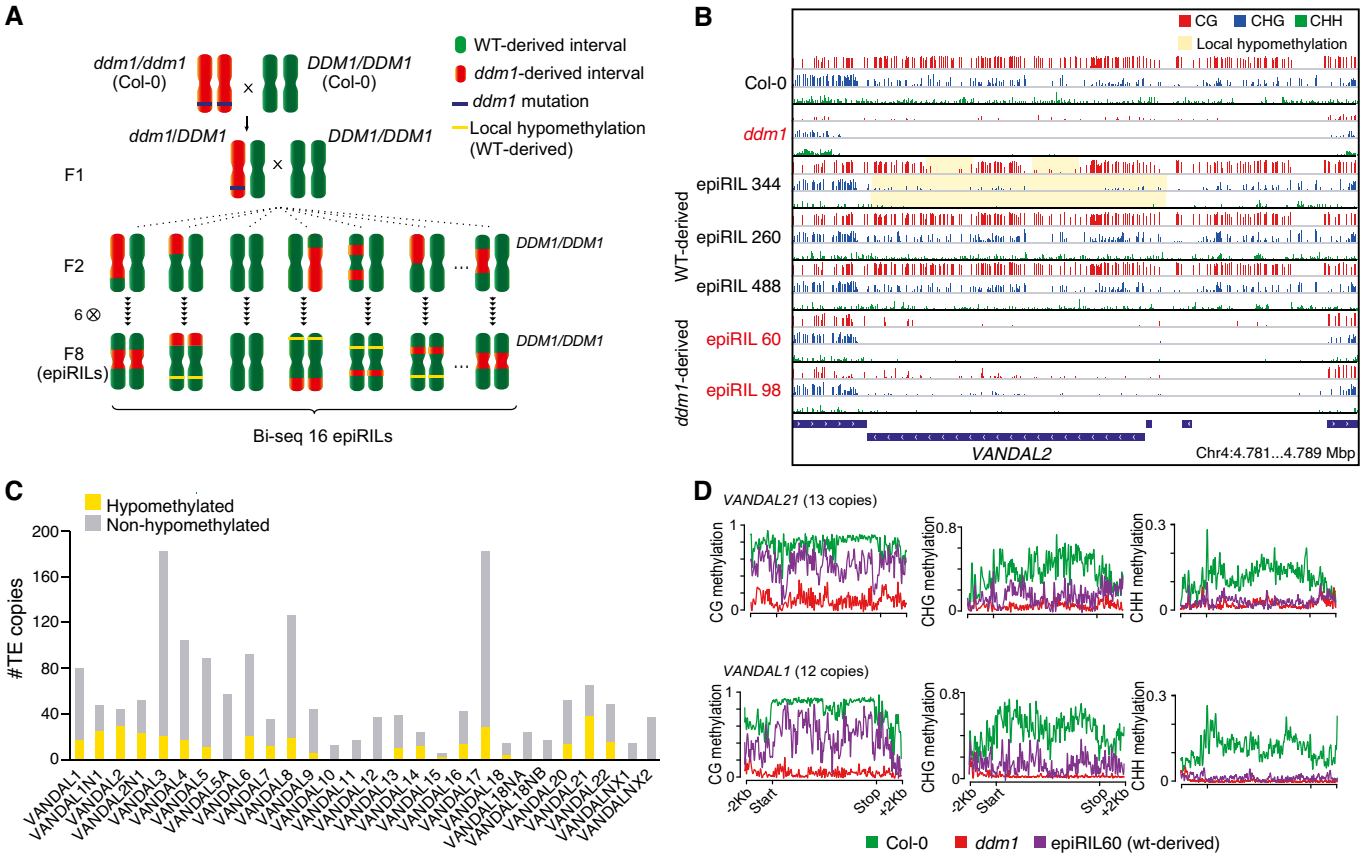

Figure 1. Most *VANDAL* families are trans-hypomethylated in epiRILs.

A Cartoon depicting the crossing scheme used to generate the epiRIL population.

B Genome browser tracks showing local hypomethylation of a WT-derived *VANDAL2* copy in Col-0, *ddm1* mutant, and several epRILs that carry the WT- or *ddm1*-derived epihaplotype at this locus (indicated on the left).

C Number of *VANDAL* copies per family as well as the number of WT-derived *VANDAL* copies hypomethylated in at least one epiRIL.

D Metaplot of DNA methylation on wt-derived *VANDAL21* or *VANDAL1* copies within Col-0, *ddm1*, or epiRIL60.

DNA-sequence motifs, and these copies typically correspond to full-length elements (Fig 2B). In fact, hypomethylated *VANDAL* sequences are much longer than non-hypomethylated ones (Appendix Fig S6A).

Despite the strong association between the presence of motifs and hypomethylation, a sizable proportion of non-hypomethylated copies do carry motifs (Fig 2B), which nonetheless accumulate at much lower density in these compared to hypomethylated *VANDAL*s (Appendix Fig S6B), suggesting that the sole presence of motifs is not sufficient for hypomethylation targeting. Furthermore, motifs enriched in specific *VANDAL* families are also detected in other families (Fig 2C), which is of course expected given the high probability ($P > 0.09$) of finding a 8- to 10-bp-long motif in any 6,000-bp-long sequence. Given that each hypomethylated copy typically contains many short-sequence motifs (Appendix Fig S6B), we reasoned that the specificity of *VANDAL* anti-silencing systems may be determined by their local density, as has been proposed for VANC21 and VANC6 (Hosaka *et al*, 2017). To test this hypothesis, we investigated the clustering of short-sequence motifs within *VANDAL* sequences. Compared to isolated motifs, which are ubiquitous among all *VANDAL*s, clusters containing four or more motifs per 1,000 bp are almost exclusively overrepresented within cognate TE families (Fig 2C), with the notable exception of *VANDAL1/1N1/2* and *2N1* families that share the same motif TGTACGTACA. In addition, motifs detected in *VANDAL5* and *VANDAL15* are also found in *VANDAL6* and *VANDAL16* copies, respectively. Notwithstanding this last result, which may be an indication that these families belong to the same anti-silencing system, local accumulation of hypomethylation motifs seems to provide an additional layer of anti-silencing specificity (Fig 2C).

Clustering of short-sequence motifs may imply that VANCs interact with DNA as homomultimers, similarly to the mode of action described for some transcription factors (Avsec *et al*, 2021). To test this possibility, we explored the spatial arrangement of short-sequence motifs (i.e., the distance between consecutive motifs organized as direct (DR), everted (ER), or inverted repeats (IR)). This analysis revealed that most hypomethylation motifs cluster as direct repeats spaced by 10 bp, which corresponds to one turn of the DNA helix. Notably, short-sequence motifs within TEs belonging to *VANDAL1/1N1/2* and *2N1* families appear to cluster indistinctly as

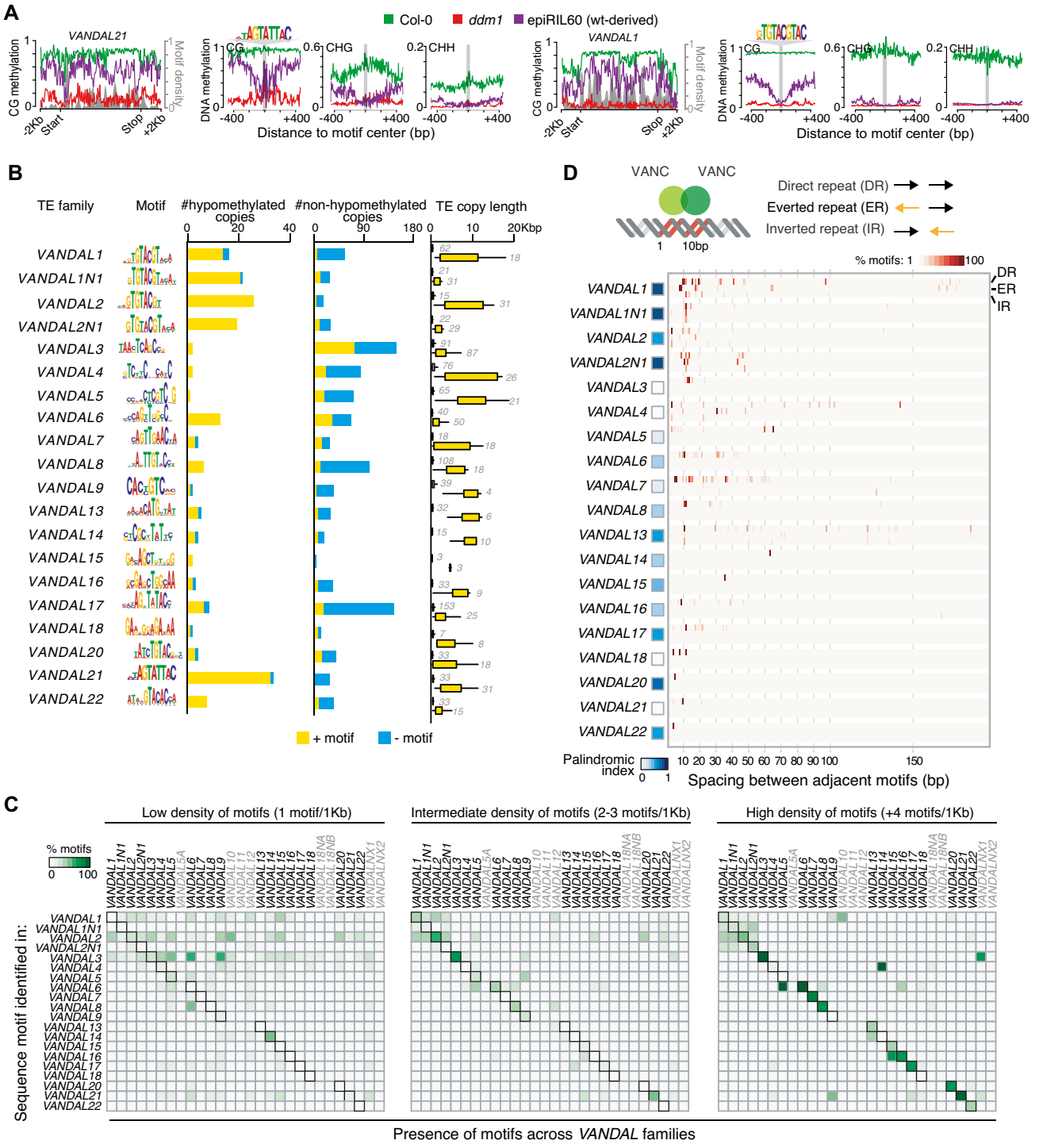

Figure 2. Sequence and syntax of motifs dictate VANC anti-silencing specificity.

A Metaplot of DNA methylation and relative motif density (arbitrary units: 1 = max; 0 = min) on wt-derived *VANDAL21* or *VANDAL1* copies as well as around short-sequence motifs within Col-0, *ddm1*, or epiRIL60. Metaplots of DNA methylation for *VANDAL1* and *VANDAL21* are from Fig 1D.

B Motif logo, number of hypomethylated and non-hypomethylated WT-derived copies, and size of *VANDALs* (in Kbp) carrying or not short-sequence motifs. For each boxplot, the lower and upper bounds of the box indicate the first (Q1) and third (Q3) quartiles, respectively, the whiskers represent data range, bounded to 1.5 * (Q3–Q1). Sample sizes are indicated in gray.

C Proportion of isolated or increasingly clustered motifs (medium: 2–3 motifs/1 Kb; high +4 motifs/1 Kb) across sequences belonging to the distinct *VANDAL* families. *VANDAL* families showing no hypomethylation in epiRILs are indicated in gray.

D Spacing between consecutive short-sequence motifs in the three possible relative orientations within specific *VANDAL* families. The proportion of palindromic sequence within each motif (i.e., Palindromic Index) is also shown.

DR, ER, or IR (Fig 2D), which is consistent with these motifs being highly palindromic (Palindromic Index 0.6–0.8). Taken together, these results establish that specificity of distinct anti-silencing systems is likely determined by the identity, syntax, and spatial organization of short-sequence motifs and that DNA-bound VANCs potentially form arrays of homopolymers with helical periodicity.

Diversification of VANC-dependent anti-silencing systems within and across species

In order to maintain the functionality of the recognition system, diversification of short motifs within *VANDALs* needs to be accompanied by parallel variations in their cognate VANC proteins. However, investigating the evolution of TE-encoding proteins is challenging because of the lack of reliable transcript annotations over TE sequences in reference genomes. To circumvent this limitation and determine the whole repertoire of *VANDAL*-encoded VANC proteins in *A. thaliana*, we carried out deep long-read Nanopore sequencing of cDNAs from *ddm1* mutant plants (see Materials and Methods). By combining this high-quality dataset with a previous "gene-like" annotation of TEs (Panda & Slotkin, 2020), we could identify 160 *VANDAL*-encoded transcripts together with their orientation, precise transcription initiation, and termination sites as well as their exon–exon boundaries (Fig 3A). *In silico* translation predicted 42 VANC-encoding genes, encompassing 16 of the 28 *VANDAL* families annotated in the reference genome and including the previously characterized *VANC21* and *VANC6* (Hosaka *et al,* 2017) (Fig 3B). The number of VANC-encoding genes varies greatly between TE families, likely reflecting differences in their coding potential. On the one hand, all the non-autonomous *VANDAL* families (*VANDAL1N1*, *VANDAL2N1*, *VANDAL5NA*, *VANDAL18NA/B*, *and VANDALNX1/2*) lack any detectable VANC-encoding transcripts. On the other hand, *VANDAL6*, *VANDAL3*, and *VANDAL21* encompass the largest number of VANC-encoding copies, supporting the notion that these families were subjected to recent amplification (Fu *et al,* 2013).

Sequence comparison of the 42 VANC proteins uncovered two main clusters, which are themselves made up of sub-groups reflecting the different *VANDAL* families (Fig 3C). Detection of conserved domains in the Pfam database using Hidden Markov Models show that most VANC-like proteins contain the domain DUF1985, either alone or in association with the Ulp1 or DUF287 domains (Fig 3C), and these combinations broadly explain the phylogenetic clustering. Remarkably, while Ulp1 domain is conserved across the tree of life (Appendix Fig S7A) and is associated with de-sumoylation activities (Johnson, 2004), DUF1985 and DUF287 have uncharacterized functions and are almost exclusively encoded by *Mu*-like elements (MULEs), named after the *Mutator* (*Mu*) DNA transposon of maize (Robertson, 1978). Placing VANC protein architectures (Fig 3D) into the plant phylogenetic tree indicated that DUF1985 alone or fused to Ulp1 preceded the radiation between rosids and asterids (Fig 3E; Appendix Fig S7B), tying the emergence of VANC proteins with that of eudicots. In addition, DUF1985 in combination with the Ulp1 domain is present across eudicot species (Appendix Fig S7A and B), suggesting that this domain organization is ancestral. Conversely, DUF287 is only detected in VANC-like proteins encoded by Brassicaceae's MULEs (Fig 3E; Appendix Fig S7A) and has no similarity with any other known protein, suggesting a recent *de novo* origin of this domain.

As the distribution, local density, and syntax of short motifs within *VANDAL* sequences is a reliable indicator of functional VANC-mediated anti-silencing in *A. thaliana* (Fig 2), we assessed whether similar motif organizations are present in the distantly related *VANDAL*-like *CUMULE* from melon (van Leeuwen *et al,* 2007). Indeed, we found that this TE encodes a VANC-like protein containing both DUF1985 and Ulp1 domains (Fig 3F) and carries outside of its coding sequences a high density of the quasi-palindromic 11 bp motif TAAACGATCGT, arranged in a one-helix-turn periodicity (Fig 3G and H). Thus, *CUMULE* likely possesses the two components of a functional VANC-induced anti-silencing system, which suggests in turn that such systems are relatively ancestral. Taken together, these results illustrate the extensive diversification of *MULE*-encoded VANC-like factors and of their targeting sequences across eudicot species.

Multiple sequence-specific anti-silencing systems coexist in *A. thaliana*

We next set out to identify the VANC-encoding copies responsible for the family-specific anti-silencing in the epiRILs. One possibility would be that hypomethylation of *VANDALs* is due to the activity of *ddm1*-derived, VANC-encoding TEs that segregate in the epiRILs. To test this hypothesis, we considered hypomethylation of wt-derived *VANDALs* as traits and performed (epi)QTL mapping using the hundreds of parental DMRs that segregate in this population (Colomé-Tatché *et al,* 2012; Cortijo *et al,* 2014). Starting with the epiQTL mapping of *VANDAL21* and *VANDAL6* hypomethylation, our approach accurately identified the full-length reference copies encoding the active VANC21 and VANC6 previously characterized (Hosaka *et al,* 2017) (Appendix Fig S8A and B), demonstrating that trans-hypomethylation in the epiRILs is determined by *ddm1*-derived, VANC-encoding *VANDALs*. Following this confirmation, we performed epiQTL mapping on the remaining *VANDAL* families with hypomethylated copies. In most cases, one or two (epi)QTL intervals were detected per family (Appendix Fig S8A), thus revealing a simple genetic architecture of anti-silencing activities overall. However, several (epi)QTL intervals were shared by various TE families, suggesting that some VANCs have broad activity. Most noticeably, one (epi)QTL interval in chromosome 1 is shared by the four families *VANDAL1/1N1/2* and *2N1* (Fig 4A), which also have similar short-sequence motifs (Fig 2B), pointing to a common anti-silencing system. Consistent with this interpretation and the simple genetic architecture of anti-silencing, we typically identified a single *VANDAL* copy expressed in *ddm1* and that encodes a full-length VANC (Fig 3) within each epiQTL interval (Appendix Fig S8A). In total, there are at least 10 independent anti-silencing systems associated with up to 10 VANC-encoding *VANDAL* copies (Appendix Fig S8A). Importantly, five of such systems involved the uncharacterized Ulp1-containing VANCs, providing a unique opportunity to investigate their anti-silencing activity experimentally.

Ulp1-containing VANC1 induces sequence-specific hypomethylation

To assess the function of Ulp1-containing VANC factors, we transformed wild-type *A. thaliana* plants with VANC-containing sequences from the *VANDAL1* (*AT1TE56425*) and *VANDAL2*

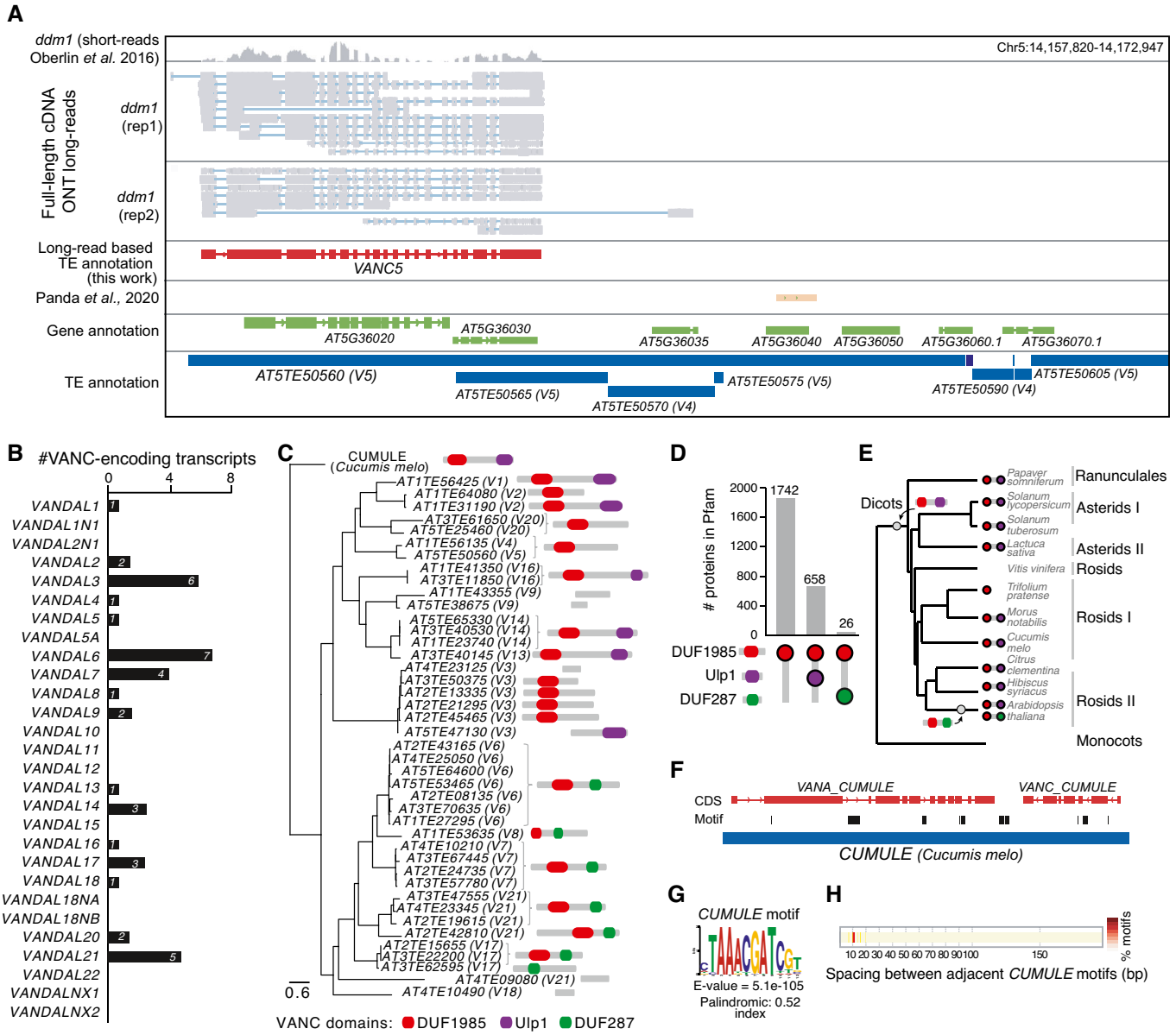

Figure 3. Diversification of VANC-dependent anti-silencing systems within and across species.

A Genome browser view of full-length cDNA nanopore reads and illumina short reads from *ddm1*-mutant plants as well as *de novo* functional annotation of TE-encoding transcripts together with TAIR10 gene and TE annotations. *VANDAL4* and *VANDAL5* copies are indicated as V4 and V5, respectively.

B Number of VANC-like encoding transcripts.

C Phylogenetic relationship among predicted VANC proteins encoded by *A. thaliana VANDALs* and *C. melo CUMULE*. Structure and presence of conserved protein domains are indicated for each VANC.

D Domain organization of VANC-like proteins in the Pfam database.

E VANC domain organizations across flowering plant species. The likely origin of Ulp1- and DUF287-containing VANCs is indicated. Representative species from different groups of eudicots are shown.

F Structure, predicted coding sequences, and localization of short-sequence motifs of *CUMULE*.

G Logo of the short-sequence motif enriched within non-coding regions of *CUMULE*. Statistical overrepresentation and Palindromic Index are also shown.

H Spacing between consecutive short-sequence motifs within *CUMULE*.

(*AT1TE31190*) candidate copies identified on our epiQTL analysis (Fig 4A and B; Appendix Fig S9A). WGBS of transformed plants revealed that the ectopic expression of VANC1, but not VANC2, is sufficient to induce strong and specific non-CG hypomethylation of both *VANDAL1* and *VANDAL2* sequences (Fig 4C and D; Appendix Fig S9B and C). Furthermore, in VANC1-transgenic plants (VANC1-TG), the loss of non-CG DNA methylation tends to affect the entire *VANDAL* sequences, while CG hypomethylation is

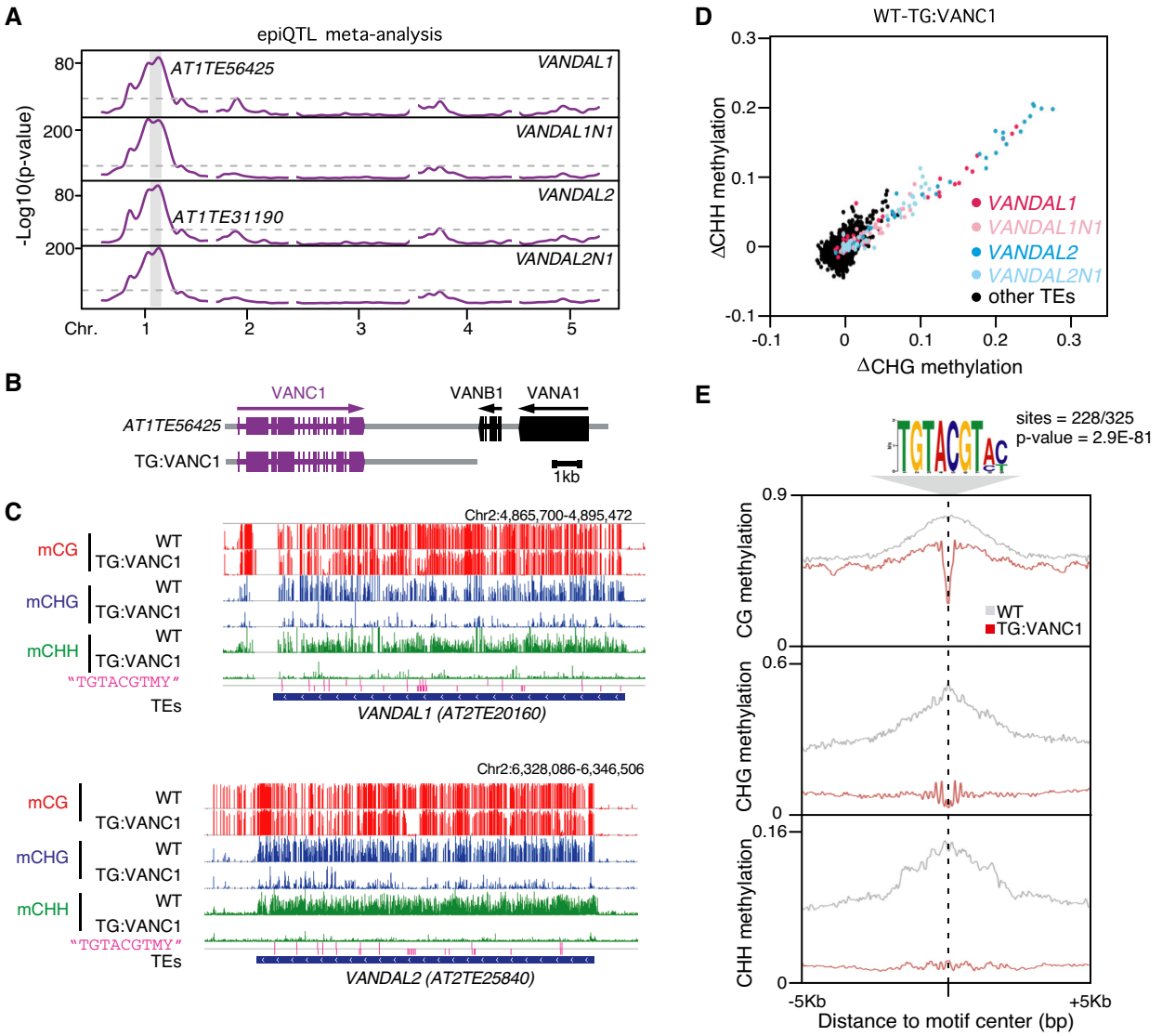

Figure 4. Ulp1-containing VANC1 induces sequence-specific hypomethylation.

A (epi)QTL mapping of *VANDAL1*, *1N1*, *2*, and *2N1* trans-hypomethylation in 105 epiRILs. The VANC-encoding *VANDAL1* and *VANDAL2* located within the single (epi)QTL interval are indicated in each case.

B Schematic diagram of structures of candidate *VANDAL1* copy and the modified transgene spanning VANC1 used. Boxes indicate exons.

C Genome browser tracks showing hypomethylation effect of VANC1 on a *VANDAL1* and *VANDAL2* copy.

D VANC1-induced DNA hypomethylation is shown for each TE at CHG sites and CHH sites. *VANDAL1*, *1N1*, *2*, and *2N1* copies are indicated by colors.

E Metaplot of DNA methylation around short-sequence motifs within wild-type (grey) or VANC1-expressing plants (red).

constrained to short regions enriched in the motif "TGTACGTMY" (Fig 4C and E), reproducing the pattern of hypomethylation observed in the epiRILs (Figs 1B and 2A). Together, these findings demonstrate that the ectopic expression of Ulp1-containing VANC1 induces strong and sequence-specific hypomethylation of related *VANDALs*.

We next tested by RT–PCR experiments whether VANC1 can induce expression of *VANDAL1/2*-encoded genes and found that *VANB* and *VANC*, but not the putative transposase *VANA*, which carries disabling mutations including a premature stop, are transcriptionally reactivated in VANC1-expressing lines (Appendix Fig S10A–C). We then assessed whether expression of *VANC1* can induce transposition of target sequences, as was previously found for VANC21 (Fu *et al*, 2013). We estimated *VANDAL1* and *VANDAL2* copy numbers by comparing WGBS coverage between VANC1-TG and control samples, but did not observe significant differences in coverage (Appendix Fig S11), consistent with the lack of expression of the putative transposase *VANA*. This last result demonstrates that VANC1-induced hypomethylation does not require and is not sufficient to trigger transposition.

Impaired VANC targeting of non-autonomous copies induces family-wide epigenetic re-silencing

Beyond the demethylation of cognate *VANDAL1* copies, the ectopic expression of *VANC1* also induces efficient demethylation of TEs belonging to the non-autonomous *VANDAL1N1* and *VANDAL2N1* families (Figs 4D and 5A). Copies belonging to these two families are short, do not produce mRNAs in *ddm1* (Fig 3B), and lack any predicted open reading frames (ORFs). Therefore, *VANDAL1N1* and *VANDAL2N1* copies must rely on factors encoded by other TEs for their amplification. To hijack the transposition machinery, such non-autonomous TEs have terminal sequences that are recognized by transposases encoded by autonomous TEs. Accordingly, terminal sequences of *VANDAL1N1* and *VANDAL2N1* copies are almost identical to the ones of *VANDAL1* and *VANDAL2*, respectively (Appendix Fig S12A and B). Notwithstanding, internal sequences of *VANDAL1N1* and *VANDAL2N1* have no sequence homology with their cognate autonomous TEs (Fig 5A), nor with any other sequence in the *A. thaliana* genome, suggesting that these non-autonomous families originated from complex sequence rearrangements. Despite their chimeric origin and the lack of sequence homology, internal regions of *VANDAL1N1* and *VANDAL2N1* copies have high densities of the VANC1 short motif "TGTACGTMY" (Fig 5A and B). However, the exact spatial organization of these motifs differs from that of *VANDAL1* and *VANDAL2* (Appendix Fig S12C), indicating that non-autonomous copies have accumulated VANC-targeting motifs anew.

Given that VANC activity is not required for transposition (Fu *et al,* 2013), and that DNA methylation of TE sequences primarily acts to repress their transcription, it is very intriguing that *VANDAL1N1* copies with no transcriptional potential are nonetheless efficiently targeted by VANC1-induced demethylation. One possibility is that the terminal sequences of non-autonomous copies could act as reservoirs of small RNAs that can then trigger the epigenetic silencing of VANC-encoding TEs *in trans* via the RNA-directed DNA methylation (RdDM) pathway. Indeed, when active VANCs are expressed, *VANDAL*-matching 24-nt small RNAs are strongly reduced (Rigal *et al,* 2016). Under this scenario, deficient hypomethylation of related *VANDAL* sequences would lead to the continuous accumulation of trans-matching small RNAs, which in turn could counteract VANC activity. Consistently, re-analysis of small RNA sequencing data obtained from wild-type inflorescences (Creasey *et al,* 2014) shows that the single VANC1-encoding copy (*AT1TE56425*) and several *VANDAL1N1* copies generate abundant perfectly multiple-matching 24-nt-long small RNAs, the density of which decreases with the DNA-sequence divergence between these TEs (Fig 5C). We thus hypothesized that VANC-induced hypomethylation of non-autonomous TEs may prevent the silencing of related autonomous *VANDALs* through identity-based RdDM. Under this scenario, impairing VANC targeting of a related *VANDAL* copy should trigger epigenetic silencing of the whole TE family. To test directly this hypothesis, we introduced in the genome of VANC1-TG plants a copy of the non-autonomous *VANDAL1N1* (*AT5TE61035*), or a modified version of it that is devoid of the short-sequence motifs required for VANC1 targeting (1N1 and 1N1Δmotif copies, respectively; Fig 5D). Supporting our hypothesis, introduction of the 1N1Δmotif sequence was sufficient to fully abolish VANC1-induced non-CG hypomethylation of the 5′ terminal region of the endogenous VANC1-encoding copy, whereas introduction of the 1N1 copy had no effect (Appendix Fig S13). To confirm this result genome wide, we obtained WGBS for two independent 1N1 and 1N1Δmotif transgenic lines and found that all *VANDAL1/1N1/2* and *2N1* copies become systematically re-methylated following the introduction of the 1N1Δmotif sequence (Fig 5E and F; Appendix Fig S14). Furthermore, 1N1Δmotif-induced DNA remethylation goes beyond the terminal regions with high sequence homology between copies (Fig 5F), implying the involvement of heterochromatin spreading and/or production of secondary small RNAs. The capacity of 1N1Δmotif to induce strong silencing of related *VANDALs* is reminiscent to the silencing activity of the natural *Mu* killer locus from maize (Slotkin *et al,* 2005), which is a non-autonomous mutated derivative of an active *Mu* transposon. Together, these results demonstrate that VANC-induced hypomethylation of non-autonomous copies is critical to avoid small RNA-mediated epigenetic re-silencing of related *VANDAL* sequences and that homologous copies lacking VANC motifs can act as *VANDAL* killers. Thus, coordinated acquisition and diversification of VANC-specific short-sequence motifs across related copies is key for the evolution of efficient anti-silencing systems and propagation of *VANDALs*.

Discussion

Epigenetic control of TEs imposes strong selective constraints, engaging hosts and TEs in intimate co-evolutionary dynamics (Hurst & Werren, 2001; Cosby *et al,* 2019). One possible outcome of these dynamics is the evolution of TEs that can escape host repression and propagate across the genome. Here, we have exploited the *ddm1* epiRIL population in combination with long-read transcriptome sequencing, quantitative genetic approaches, ectopic expression of TEs, and evolutionary analysis to identify and characterize a remarkably diverse family of VANC anti-silencing factors encoded by *VANDAL* DNA transposons specifically in eudicots.

The epiRILs were designed to investigate the epigenetic basis of phenotypic traits (Johannes *et al,* 2009; Cortijo *et al,* 2014). Our work demonstrates that this experimental population also provides a powerful system to study *VANDALs* anti-silencing, which in turn calls for its use to identify other types of TE-encoded anti-silencing systems in *A. thaliana*. Indeed, the McClintock's *Suppressor–mutator* (*Spm*) element from Maize encodes the anti-silencing factor TnpA, which can bind to, and induce hypomethylation of, cognate *Spm* copies (Gierl *et al,* 1988; Schläppi *et al,* 1994). Given our observations that the *A. thaliana* genome contains numerous *Spm* families showing evidence of trans-hypomethylation in the epiRILs (Appendix Fig S15), including the highly active *Spm3* (Miura *et al,* 2001; Kato *et al,* 2004; Quadrana *et al,* 2019), a next step would be to characterize *Spm* anti-silencing systems in this species. Furthermore, *ddm1* mutants have also been obtained in other plants, including tomato (Corem *et al,* 2018), rice (Tan *et al,* 2018), and maize (Li *et al,* 2014), thus providing useful systems to study systematically the existence of TE-encoded anti-silencing factors across plants. As a matter of fact, the putative transposase of *Mu*, MURA, was shown to demethylate the terminal inverted repeats (TIRs) of cognate transposons (Burgess *et al,* 2020), indicating that TE-encoded anti-silencing systems could be much more common than previously thought.

Our analyses indicate that VANCs likely originated in eudicots soon after their divergence from monocots and contain a characteristic N-terminal DUF1985, which in its ancestral form is typically combined with a C-terminal Ulp1 domain. This observation, together with the finding that Ulp1-containing VANC1 induces strong sequence-specific anti-silencing in *A. thaliana*, suggest that these

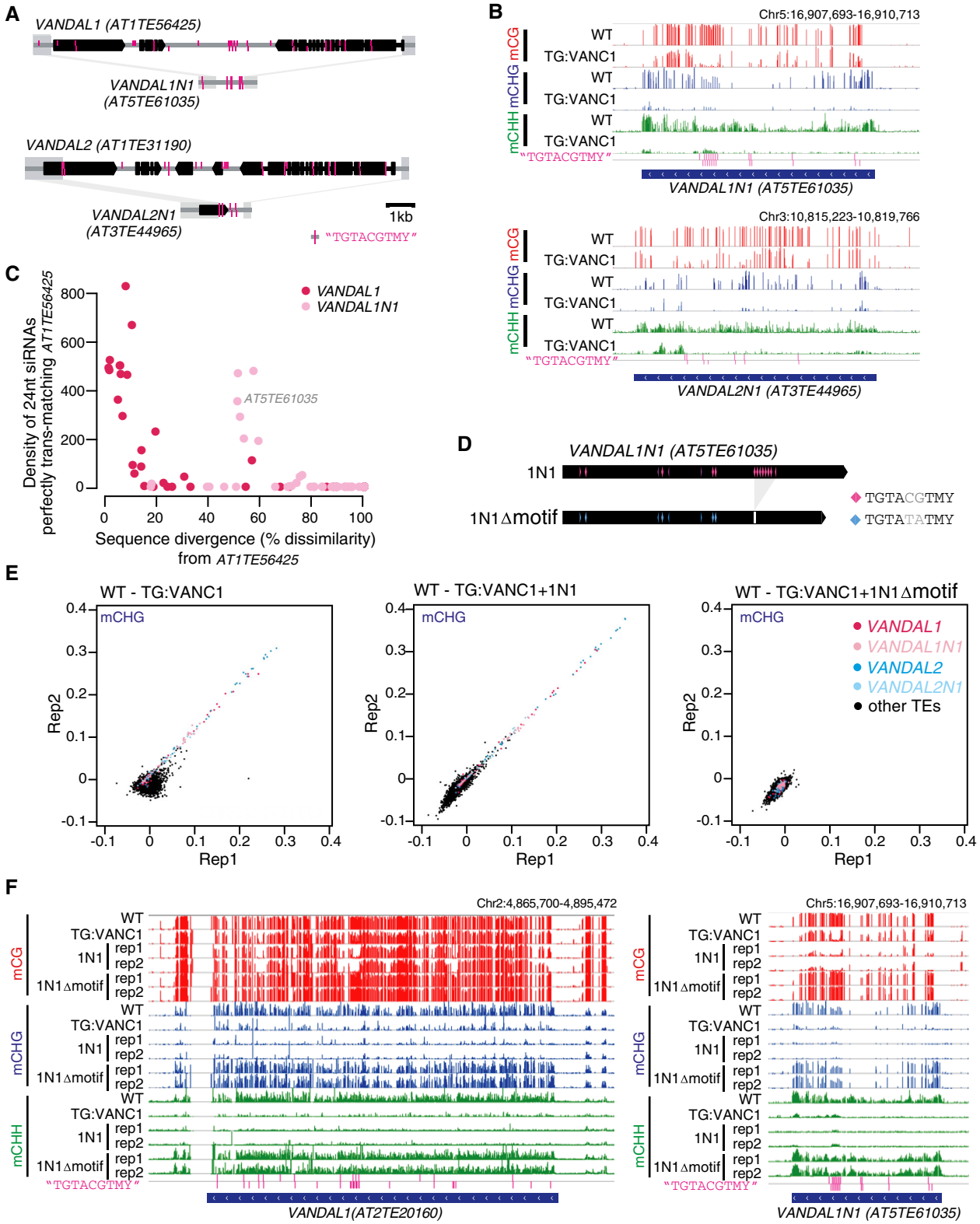

Figure 5.

Figure 5. VANC-induced hypomethylation of non-autonomous copies prevents family-wide epigenetic resilencing.

A Schematic diagram of structures of full-length *VANDAL1* and *VANDAL2* copies and their derived non-autonomous *VANDAL1N1* and *VANDAL2N1*, respectively. Regions with high sequence homology as well as the location of motifs associated with hypomethylation are indicated in grey and pink, respectively.

B Genome browser tracks showing hypomethylation effect of VANC1 on a *VANDAL1N1* and *VANDAL2N1* copy.

C Density (#siRNAs/1,000 bp) of perfectly multiple-matching 24-nt-long small RNAs and sequence divergence (% global dissimilarity) from functional VANC-encoding *VANDAL1* is shown for each *VANDAL1* and *VANDAL1N1* TE (red and pink dots, respectively).

D Schematic diagram of transgene structures of original *VANDAL1N1* (1N1) and modified version lacking all VANC1 short-sequence motifs (1N1Δmotif).

E Comparison of CHG hypomethylation between replicates of plants containing VANC1, VANC1 + 1N1, or VANC1 + 1N1Δmotif transgenes. *VANDAL1*, *1N1*, *2*, and *2N1* copies are indicated by colors.

F Genome browser tracks showing the effect of 1N1 and 1N1Δmotif transgenes on VANC1-induced hypomethylation over a *VANDAL1 and VANDAL1N1* copy. Methylation levels over the *VANDAL1N1* copy (*AT5TE61035*) in 1N1 samples reflect the average methylation of the endogenous and introduced copy.

domains were at the origin of *VANDAL*s anti-silencing systems. Incidentally, these results also reveal that VANC proteins derived from the fusion of a *de novo* originated and a captured cellular domain, DUF1985 and Ulp1, respectively. In yeast, the N-terminal and C-terminal domains of Ulp1-containing proteins are, respectively, involved in protein targeting and removal of small ubiquitin-related modifier (SUMO) (Johnson, 2004), which has been shown to repress the retrotransposon *Ty1* in this species (Bonnet *et al,* 2021) as well as to contribute to heterochromatin formation in multiple organisms (Maison *et al,* 2016; Ninova *et al,* 2020; Andreev *et al*, 2022; Sheban *et al*, 2022). Thus, it is reasonable to speculate that the N-terminal domain DUF1985 guides VANCs to target sequences while the C-terminal domain Ulp1 participates in their hypomethylation, possibly by affecting chromatin-associated SUMO. Notably, TEs encoding Ulp1-containing proteins are pervasive across the tree of life (Marín, 2010; Böhne *et al,* 2011; Lisch, 2015), suggesting that acquisition of desumoylation activities could be a recurrent evolutionary response of TEs to escape epigenetic silencing.

The deep conservation of Ulp1-containing VANCs contrasts with the many VANCs in Brassicaceae species that lack this domain and that have instead the uncharacterized DUF287, including VANC21 and VANC6. DUF287 has no sequence similarity with any other type of protein described so far, indicating that it has originated *de novo*. How these, as well as other TE-encoded orphan proteins, such as the accessory factor MURB encoded by *Mu* (Lisch, 2002), originated remains elusive. Strikingly, DUF287-containing families already account for almost half of the *VANDAL* copies in the *A. thaliana* genome, suggesting that this derived anti-silencing factor may have contributed to the invasiveness of *VANDAL*s in this group of species (Dupeyron *et al,* 2019). Because the origin of VANC-like proteins predated the radiation of eudicot species, during a time of intensive diversification and rapid evolution, it is tempting to speculate that the conflicts between host suppression pathways and TEs may have contributed to the evolution of epigenetic control systems in this remarkably diverse clade of plants. In this sense, the characteristic VANC domain DUF1985 has been recurrently captured and fused to diverse host proteins containing distinct DNA- or chromatin-binding domains (Appendix Fig S7B), likely leading to the creation of new cellular functions. Determining the precise function of VANC-containing DUF1985, Ulp1 and DUF287 domains will be key to understanding how the emergence of new TE-encoded functions shapes the evolution of TEs and host silencing mechanisms.

VANC proteins bind DNA *in vitro* and *in vivo* (Hosaka *et al,* 2017) and we provided evidence that highly specific self-recognition is determined by the sequence, density, and spatial arrangement of short motifs, which are typically spaced by 10bp within non-coding regions of *VANDAL*s. The finding that this motif syntax is conserved across distantly related *VANDAL*s points to a functional role. For instance, helical periodicity may enhance binding and polymerization of VANCs on the DNA sequence. Such homopolymerization of VANCs may shield *VANDAL* DNA sequences from DNA methyltransferases while inducing active DNA demethylation. Based on our findings, a key priority for the future will be to investigate the cooperative DNA-binding, polymerization and demethylation activity of VANC proteins.

Our study revealed that VANC-induced anti-silencing systems target autonomous as well as non-autonomous *VANDAL* families, which appear to have accumulated VANC-targeting sequence motifs anew. The latter type of TEs derives from full-length copies by truncation as well as accumulation of random mutations. To ensure propagation, these elements must hijack the transposition machinery from other TEs, leading some authors to consider non-autonomous TEs as analogous to hyperparasites (Robillard *et al,* 2016). Our findings now establish that non-autonomous copies also hijack the anti-silencing mechanisms of related *VANDAL*s to promote their own hypomethylation. This observation was initially puzzling as VANC seems to be not essential for transposition (Fu *et al,* 2013). However, we found that non-autonomous *VANDAL*s may serve as important reservoirs of multiple-matching 24-nt-long small RNAs targeting autonomous copies in a homology-dependent manner. Indeed, the density of these siRNAs decreases with the divergence between non-autonomous and autonomous *VANDAL*s. In contrast, a handful of well-spaced motifs within non-autonomous copies are sufficient to trigger VANC-induced hypomethylation and hence reduction in small RNA accumulation. Such persistent VANC targeting beyond the recognition of small RNAs would protect related *VANDAL*s (i.e., belonging to the same family) from host silencing mechanisms.

The remarkable capacity of the 1N1Δmotif sequence to induce strong and concerted epigenetic re-silencing of related *VANDAL*s is reminiscent of the silencing activity of the *Mu* killer locus, which is a naturally occurring non-autonomous derivative of the maize *Mu* transposon that express a long hairpin transcript (Slotkin *et al,* 2005). TE sequences are frequently mutated and rearranged, particularly during transposition, and it has been proposed that non-autonomous copies may be a common source of transposon silencing triggers (Slotkin *et al,* 2005; Burgess *et al,* 2020; Wang *et al,* 2020). However, unlike *VANDAL* killer, which does not produce long transcripts and is associated with 24-nt-long siRNAs, the hairpin transcript of *Mu* killer is processed into 22-nt-long small RNAs, which trigger *de novo* DNA methylation of full-length *Mu* copies (Burgess *et al,* 2020). Therefore, different transposon silencing triggers may rely on distinct molecular mechanisms. TE transgenes can

be methylated through the identity-based silencing mechanism, which is mostly dependent on 24-nt-long small RNAs produced from endogenous TEs by the plant-specific RNA Polymerase IV (Fultz & Slotkin, 2017). *De novo* establishment of RdDM targeting is still enigmatic, although recent research showed the importance of transcription by Pol II (Sigman *et al,* 2021). Conversely, reinforcement of DNA methylation can be mediated by RdDM-dependent and -independent mechanisms, which both rely on the presence of remaining epigenetic mark(s) at target loci (To *et al,* 2020). In this sense, VANCs do not erase all epigenetic marks over target TEs, as CG methylation outside motifs remains unaffected, providing the epigenetic memory that is required for resilencing. Indeed, DNA methylation of *VANDAL21* is rapidly restored when VANC21-TG is segregated apart (Fu *et al,* 2013), indicating that VANC-induced hypomethylated *VANDAL*s are in a labile epigenetic state that can be readily resilenced. Under this scenario, non-autonomous *VANDAL* copies that are no longer targeted by VANCs can induce remethylation of related *VANDAL* sequences through identity-based Pol IV-RdDM (Fultz & Slotkin, 2017). Gain and loss of VANC-targeting sequence motifs, or even the perturbation of their helical periodicity by accumulation of short indels, may happen remarkably fast due to imprecise transposition, replication slippage, unequal crossing over, and/or small-scale mutations. Therefore, the accumulation of mutations during sustained *VANDAL* proliferation is expected to eventually transform non-autonomous *VANDAL* copies from hyperparasites to killers. Such spontaneous formation of *VANDAL* killers may in turn provide an efficient self-control mechanism to limit runaway *VANDAL* proliferation, protecting genome persistence and of the TEs it contains.

To conclude, our findings reveal that the co-evolution between host silencing and TEs, as well as their interactions with hyperparasitic non-autonomous copies, shaped the diversification and invasive success of *VANDAL* TEs, with potential implications for the emergence of novel gene control mechanisms.

Materials and Methods

Plant materials

The *A. thaliana* Col-0, *ddm1-2* mutant, and the 16 epiRILs (Johannes *et al,* 2009) lines used in this work were described before (Colomé-Tatché *et al*, 2012; Quadrana *et al,* 2019). All plants were grown in long days (16 h:8 h light:dark) at 23°C. The VANC1 and VANC2 constructs were generated by amplifying genomic sequences of VANCs by PCR and cloned into pPLV01 vector double digested by *Hpa*I and *Eco*53kI using NEBuilder (NEB). The 1N1 and 1N1Δmotif constructs were cloned into *Sma*I-digested pGreenII-0179. For 1N1Δmotif, "TGTACGTMY" motifs were converted to "TGTATATMY" by PCR-based site-directed mutagenesis. Constructs were transformed into wild-type (VANC1 and VANC2) or VANC1-TG (1N1 and 1N1Δmotif) plants of *A. thaliana* Col-0 ecotype by floral dip (Clough & Bent, 1998).

Whole-genome bisulfite sequencing and DMRs detection

DNA from epiRILs was extracted using a standard CTAB protocol. Bisulfite conversion, BS-seq libraries, and sequencing (paired-end 100 nt reads) were performed by BGI Tech Solutions (Hong Kong). For WGBS of transgenic plants, bisulfite treatment and library preparation were conducted as previously described (Fu *et al,* 2013). In all cases, paired-end reads were trimmed using Trimmomatic program (version 0.33) with the following parameters "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" (Bolger *et al,* 2014). Mapping of trimmed sequences to *Arabidopsis* reference genome (TAIR10) with option "-n 1 -l 20," removal of identical reads, and counting of methylated and unmethylated cytosines were performed by Bismark ver. 0.15.0 (Krueger & Andrews, 2011). MethylKit package v0.9.4 (Akalin *et al,* 2012) was used to calculate differential CG methylation in 100 bp non-overlapping windows (DMRs) between epiRILs and wild type. Significance of calculated differences was determined using Fisher's exact test and Benjamin–Hochberg (BH) adjustment of *P*-values (FDR < 0.05) and methylation difference cutoffs of 40%. Metaplots of mCG across wt-derived VANDAL copies shown in Fig 1D and Appendix Fig S4 were performed using deeptools v3.4.0 and DNA methylation data from wt, *ddm1*, and the indicated epiRIL in each case. DNA methylation ratio in 120 bp bin for each cytosine context was calculated and compared between WT and VANC1-TG. Bins whose change in CG methylation ratio was 0.5 or more were determined as CG-hypoDMRs. Significance of decrease in DNA methylation for TEs in each cytosine context (Figs 4D and 5E) was accessed by value $(Mn/Cn - Mt/Ct)/(1/\sqrt{Cn} + 1/\sqrt{Ct})$, where Mn, Cn, Mt, and Ct are methylated cytosine (M) and total cytosine (C) counts mapped for each TEs in the non-transgenic (n) and transgenic (t) plants, respectively (Fu *et al,* 2013). Overview of the bisulfite data for the epiRILs is provided in Appendix Table S1.

Targeted bisulfite analysis

Conventional bisulfite sequencing analysis for endogenous *AT1TE56425* (*VANDAL1*) was performed as described previously (Saze & Kakutani, 2007), using primers listed in Appendix Table S2. For each sample, at least 15 clones were sequenced.

Short motif detection

DNA sequences in epiRILs CG-hypoDMRs and within annotated *VANDALs* were extracted using fastaFromBed command (bedtools) and analyzed by meme (Bailey, 2011) with the following parameters: -dna -oc -nostatus -time 18000 -maxsize 60000 -mod anr -nmotifs 3 -minw 8 -maxw 12 -revcomp. For *CUMULE* (GenBank: AY524004), DNA sequences outside coding regions, predicted using GENSCAN (Burge & Karlin, 1997), were extracted and processed as described above. Presence of specific short-sequence motifs across all *VANDAL* copies was evaluated using fimo script. Isolated, intermediate (2–3 motifs/1kbp) and high (+4 motifs/1 kbp) density of motifs were determined by counting the number of motifs in 1,000 bp windows. Distance between consecutive motifs was obtained using closestBed command (bedtools) and the frequency distribution of these distances was represented as a heatmap. Palindromic Index for motifs was calculated as the average fraction of palindromic DNA within motif instances using an in-home script (accessible at https://github.com/LeanQ/palindromes). DNA sequences in VANC1 CG-hypoDMRs overlapping *VANDAL1/2/1N1/2N1* ($N = 325$) were used for prediction of VANC1-targeted motifs by DREME script of MEME software version 4.11.0 (Bailey, 2011).

epiQTL mapping of *VANDAL*'s hypomethylation

Using methylation level based on MedIP data (Colomé-Tatché *et al*, 2012) for each wt-derived *VANDAL* copy containing CG-hypoDMRs as a trait and a total of 126 parental differentially methylated regions (DMRs) that segregate in a Mendelian fashion in 105 epiRILs (i.e., stable DMRs) as physical markers (Colomé-Tatché *et al*, 2012), we performed individuals epiQTL mappings based on the multiple QTL model (mqmsacn) from the R/qtl package. Genome-wide significance was determined empirically for each trait using 1,000 permutations of the data. LOD significance thresholds were chosen to correspond to a genome-wide false-positive rate of 5%. To summarize epiQTL results obtained for the different copies of the same *VANDAL* family, LOD scores were first transformed to *P*-values using the following function in R: pchisq(LOD*(2*log(10)), df = 1, lower.tail = FALSE)/2. Meta-analysis was calculated as previously described (Sasaki *et al,* 2019). Statistical threshold was defined as the 1% ($P = 0.01$) lowest meta-analysis *P*-values genome wide.

Full-length cDNA nanopore sequencing

Total RNA was extracted from 100 mg of rosette leaves from *ddm1-2* plants using the Nucleo-spin RNA Plant mini kit (Macherey-Nagel). Library preparation and Nanopore sequencing were performed as previously described by Domínguez *et al* (2020). Briefly, 10 ng of total RNA was amplified and converted into cDNA using SMART-Seq v4 Ultra Low Input RNA kit (Clontech). About 17 fmol of cDNA was used for library preparation using the PCR Barcoding kit (SQK-PBK004 kit, ONT) and cleaned up with 0.6× Agencourt Ampure XP beads. About 2 fmol of the purified product was amplified during 18 cycles, with a 17-min elongation step, to introduce barcodes. Samples were multiplexed in equimolar quantities to obtain 20 fmol of cDNA, and the rapid adapter ligation step was performed. Multiplexed library was loaded on an R9.4.1 flowcell (ONT) according to the manufacturer's instructions. A standard 72 h sequencing was performed on a MinION MkIB instrument. MinKNOW software (version 19.12.5) was used for sequence calling.

RT–PCR

Total RNA was extracted from seedlings of WT and VANC1-TG using TRIzol (Thermo Fisher), and treated with DNase I (Invitrogen). cDNA was synthesized using 3 μg of total RNA by SuperScript III (invitrogen). Ten times diluted cDNA was used as a template for RT–PCR. Primers used for RT–PCR were listed in Appendix Table S2.

Functional annotation of TE-encoding genes

Long reads from *ddm1* plants were mapped on the *Arabidopsis* reference genome (TAIR10) using minimap v2.11-r797 (H. Li, 2018) with the following options -ax splice -G 30k -t 12 and STAR v2.5.3a (Dobin *et al,* 2013) with the following options --outFilterMultimapNmax 50 --outFilterMatchNmin 30 --alignIntronMax 10000 --alignSJoverhangMin 3, respectively. Previously published short reads (Oberlin *et al,* 2017) were also mapped on the *Arabidopsis* reference genome (TAIR10) using STAR (Dobin *et al,* 2013). Transcript annotation was performed using the FLAIR pipeline (Tang *et al,* 2020) available at https://github.com/BrooksLabUCSC/FLAIR. First, splicing junctions based on short-read sequencing data were extracted using "junctions_from_sam.py" script and used to correct ONT long reads using "flair.py correct" script. Transcript isoforms were then detected using the "flair.py collapse" script, and transcripts supported by at least five long reads were retained. Annotated transcripts overlapping *VANDAL* elements were extracted using intersectBed and translated *in silico* using getorf vEMBOSS:6.6.0.0. Putative VANCs proteins were identified by BLAST against the functionally characterized VANC21 and VANC6 (Hosaka *et al,* 2017). Conserved protein domains were detected using HHMscan (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan) against the Pfam database.

Phylogenetic analysis

Putative VANC proteins from *A. thaliana* as well as *CUMULE* were aligned using MAFFT v7.271 and trimmed with trimAl v1.4.rev15 (Capella-Gutiérrez *et al,* 2009). Phylogenetic tree was generated with PhyML v20160207 (Guindon & Gascuel, 2003) using subtree pruning and regrafting (SPR) topological moves. Phylogenetic tree of *VANDAL1/2/1N1* and *2N1* copies was generated with Clustal OMEGA (https://www.ebi.ac.uk/Tools/msa/clustalo/) using 250 bp sequences of terminal inverted repeats (TIRs). Termini of TEs were determined by target site duplications. Sequence divergence between *AT1TE56425* and *VANDAL1* and *VANDAL1N1* copies was calculated using the command line version of Blast2seq (bl2seq), with the following parameters -p blastn -e 0.05 -D 1 -r 2 -G 5 -E 2.

Dot plot analysis

DNA sequences of non-coding regions in *VANDAL1/2/1N1/2N1* were compared by dot plot analysis. Dot plots were made with EMBOSS dotmatcher (https://www.bioinformatics.nl/cgi-bin/emboss/dotmatcher) default setting (10 window size, 23 threshold).

Small RNA analysis

Small RNA data from Col-0 wild-type inflorescence were obtained from Creasey *et al* (2014). Reads were trimmed using the Trimmomatic program (version 0.33) and first mapped on *AT1TE56425* sequence using Bowtie2 with the following parameters: --local --very-sensitive. Mapped 24-bp-long reads were then extracted using Samtools and aligned on the collection of *VANDAL1* and *VANDAL1N1* sequences, excluding *AT1TE56425*, using Bowtie2 with the following parameters: --local --very-sensitive -k 10.

Data availability

Original Scripts are available on GitHub (https://github.com/LeanQ/palindromes). WGBS of epiRILs is available at the European Nucleotide Archive (ENA) under project PRJEB47214 (https://www.ebi.ac.uk/ena/browser/view/PRJEB47214) and NCBI Gene Expression Omnibus (GEO) as GSE62206 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62206). WGBS of VANC1 transgenic lines has been deposited in the DDBJ under project PRJDB12220 (https://ddbj.nig.ac.jp/resource/bioproject/PRJDB12220).

Expanded View for this article is available online.

Acknowledgements

We thank members of the Kakutani and Colot groups and especially Raku Saito for discussions and critical reading of the manuscript. This work was supported by Japanese Society for the Promotion of Science (JSPS) KAKENHI grant numbers JP18K06348 to TS, 26221105, 15H05963, and 19H00995 to TK, Japan Science and Technology Agency (JST) CREST Grant (JPMJCR15O1 to TK), grants from the Centre National de la Recherche Scientifique (IRP SYNERTE, to LQ), the European Union Seventh Framework Programme Network of Excellence EpiGeneSys (Award 257082, to VC), and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 948674 to LQ). PB was supported by a postdoctoral fellowship (code SPF20170938626) from the Fondation pour la Recherche Médicale (FRM). Work in the Colot group is supported by Investissements d'Avenir ANR-10-LABX-54 MEMO LIFE, 506 ANR-11-IDEX-0001-02 PSL* Research University.

Author contributions

Taku Sasaki: Conceptualization; Resources; Data curation; Formal analysis; Validation; Investigation; Visualization; Methodology; Writing—review & editing. **Kyudo Ro:** Methodology. **Erwann Caillieux:** Methodology. **Riku Manabe:** Methodology. **Grégoire Bohl-Viallefond:** Data curation. **Pierre Baduel:** Formal analysis; Writing—review & editing. **Vincent Colot:** Conceptualization; Resources; Supervision; Writing—review & editing. **Tetsuji Kakutani:** Conceptualization; Resources; Supervision; Writing—review & editing. **Leandro Quadrana:** Conceptualization; Resources; Data curation; Formal analysis; Supervision; Validation; Investigation; Visualization; Methodology; Writing—original draft; Project administration.

In addition to the CRediT author contributions listed above, the contributions in detail are:

TS, VC, TK, and LQ conceived the project. TS and KR performed VANC1 transgenic experiments. TS and RM performed 1N1Δmotif experiments. EC extracted genomic DNA for WGBS of the epiRILs and GB-V processed the WGBS data. LQ analyzed the WGBS data, ONT results, and performed evolutionary analyses. PB and LQ performed epiQTL mapping. TS and LQ interpreted the data. LQ drafted the manuscript with additional input from TS, PB, VC, and TK. All the authors read and approved the manuscript.

Disclosure and competing interests statement

The authors declare that they have no conflict of interest.

References

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* 13: R87

Andreev VI, Yu C, Wang J, Schnabl J, Tirian L, Gehre M, Handler D, Duchek P, Novatchkova M, Baumgartner L *et al* (2022) Panoramix SUMOylation on chromatin connects the piRNA pathway to the cellular heterochromatin machinery. *Nat Struct Mol Biol* 29: 130 – 142

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A *et al* (2021) Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* 53: 354 – 366

Baduel P, Leduque B, Ignace A, Gy I, Gil Jr J, Loudet O, Colot V, Quadrana L (2021) Genetic and environmental modulation of transposition shapes the evolutionary potential of *Arabidopsis thaliana*. *Genome Biol* 22: 138

Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-Seq data. *Bioinformatics* 27: 1653 – 1659

Böhne A, Zhou Q, Darras A, Schmidt C, Schartl M, Galiana-Arnoux D, Volff JN (2011) *Zisupton* - a novel superfamily of DNA transposable elements recently active in fish. *Mol Biol Evol* 29: 631 – 645

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114 – 2120

Bonnet A, Chaput C, Palmic N, Palancade B, Lesage P (2021) A nuclear pore sub-complex restricts the propagation of Ty retrotransposons by limiting their transcription. *PLoS Genet* 17: e1009889

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268: 78 – 94

Burgess D, Li H, Zhao M, Kim SY, Lisch D (2020) Silencing of *Mutator* elements in maize involves distinct populations of small RNAs and distinct patterns of DNA methylation. *Genetics* 215: 379 – 391

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972 – 1973

Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* 16: 735 – 743

Colome-Tatche M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, Etcheverry M, Martin A, Feng S, Duvernois-Berthet E *et al* (2012) Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci USA* 109: 16240 – 16245

Corem S, Doron-Faigenboim A, Jouffroy O, Maumus F, Arazi T, Bouché N (2018) Redistribution of CHH methylation and small interfering RNAs across the genome of tomato *ddm1* mutants. *Plant Cell* 30: 1628 – 1644

Cortijo S, Wardenaar R, Colomé-Tatché M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury J-M, Wincker P *et al* (2014) Mapping the epigenetic basis of complex traits. *Science* 343: 1145 – 1148

Cosby RL, Chang NC, Feschotte C (2019) Host–transposon interactions: conflict, cooperation, and cooption. *Genes Dev* 33: 1098 – 1116

Creasey KM, Zhai J, Borges F, Van Ex F, Regulski M, Meyers BC, Martienssen RA (2014) miRNAs trigger widespread epigenetically activated siRNAs from transposons in *Arabidopsis*. *Nature* 508: 411 – 415

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15 – 21

Domínguez M, Dugas E, Benchouaia M, Leduque B, Jiménez-Gómez JM, Colot V, Quadrana L (2020) The impact of transposable elements on tomato diversity. *Nat Commun* 11: 4058

Dupeyron M, Singh KS, Bass C, Hayward A (2019) Evolution of *Mutator* transposable elements across eukaryotic diversity. *Mob DNA* 10: 12

Fu Y, Kawabe A, Etcheverry M, Ito T, Toyoda A, Fujiyama A, Colot V, Tarutani Y, Kakutani T (2013) Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. *EMBO J* 32: 2407 – 2417

Fultz D, Slotkin RK (2017) Exogenous transposable elements circumvent identity-based silencing, permitting the dissection of expression-dependent silencing. *Plant Cell* 29: 360 – 376

Gierl A, Lütticke S, Saedler H (1988) *TnpA* product encoded by the transposable element En-1 of *Zea Mays* is a DNA binding protein. *EMBO J* 7: 4045 – 4053

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52: 696 – 704

Hosaka A, Saito R, Takashima K, Sasaki T, Fu YU, Kawabe A, Ito T, Toyoda A, Fujiyama A, Tarutani Y *et al* (2017) Evolution of sequence-specific anti-silencing systems in *Arabidopsis*. *Nat Commun* 8: 2161

Hurst GD, Werren JH (2001) The role of selfish genetic elements in eukaryotic evolution. *Nat Rev Genet* 2: 597 – 606

Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P *et al* (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet* 5: e1000530

Johnson ES (2004) Protein modification by SUMO. *Annu Rev Biochem* 73: 355 – 382

Kapitonov VV, Jurka J (1999) Molecular paleontology of transposable elements from *Arabidopsis thaliana*. *Genetica* 107: 27 – 37

Kato M, Takashima K, Kakutani T (2004) Epigenetic control of *CACTA* transposon mobility in *Arabidopsis thaliana*. *Genetics* 168: 961 – 969

Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27: 1571 – 1572

van Leeuwen H, Monfort A, Puigdomenech P (2007) *Mutator*-like elements identified in melon, *Arabidopsis* and rice contain ULP1 protease domains. *Mol Genet Genomics* 277: 357 – 364

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094 – 3100

Li Q, Eichten SR, Hermanson PJ, Zaunbrecher VM, Song J, Wendt J, Rosenbaum H, Madzima TF, Sloan AE, Huang JI *et al* (2014) Genetic perturbation of the maize methylome. *Plant Cell* 26: 4602 – 4616

Lisch D (2002) *Mutator* transposons. *Trends Plant Sci* 7: 498 – 504

Lisch D (2015) *Mutator* and *MULE* transposons. *Microbiol Spectr* 3: MDNA3-0032-2014

Maison C, Bailly D, Quivy JP, Almouzni G (2016) The methyltransferase Suv39h1 links the SUMO pathway to HP1α marking at pericentric heterochromatin. *Nat Commun* 7: 12224

Marín I (2010) GIN transposons: genetic elements linking retrotransposons and genes. *Mol Biol Evol* 27: 1903 – 1911

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T (2001) Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature* 411: 212 – 214

Ninova M, Chen YA, Godneeva B, Rogers AK, Luo Y, Fejes Tóth K, Aravin AA (2020) Su(var)2–10 and the SUMO pathway link piRNA-guided target recognition to chromatin silencing. *Mol Cell* 77: 556 – 570.e6

Oberlin S, Sarazin A, Chevalier C, Voinnet O, Marí-Ordóñez A (2017) A genome-wide transcriptome and translatome analysis of *Arabidopsis* transposons identifies a unique and conserved genome expression strategy for *Ty1/Copia* retroelements. *Genome Res* 27: 1549 – 1562

Panda K, Slotkin RK (2020) Long-read cDNA sequencing enables a "gene-like" transcript annotation of transposable elements. *Plant Cell* 32: 2687 – 2698

Quadrana L, Etcheverry M, Gilly A, Caillieux E, Madoui M-A, Guy J, Bortolini Silveira A, Engelen S, Baillet V, Wincker P *et al* (2019) Transposition favors the generation of large effect mutations that may facilitate rapid adaption. *Nat Commun* 10: 3421

Rigal M, Becker C, Pélissier T, Pogorelcnik R, Devos J, Ikeda Y, Weigel D, Mathieu O (2016) Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in *Arabidopsis thaliana* F1 epihybrids. *Proc Natl Acad Sci USA* 113: E2083 – 2092

Robertson DS (1978) Characterization of a mutator system in maize. *Fundam Mol Mech Mutag* 51: 21 – 28

Robillard É, Le Rouzic A, Zhang Z, Capy P, Hua-Van A (2016) Experimental evolution reveals hyperparasitic interactions among transposable elements. *Proc Natl Acad Sci USA* 113: 14763 – 14768

Sasaki E, Kawakatsu T, Ecker JR, Nordborg M (2019) Common alleles of *CMT2* and *NRPE1* are major determinants of CHH methylation variation in *Arabidopsis thaliana*. *PLoS Genet* 15: e1008492

Saze H, Kakutani T (2007) Heritable epigenetic mutation of a transposon-flanked *Arabidopsis* gene due to lack of the chromatin-remodeling factor DDM1. *EMBO J* 26: 3641 – 3652

Schläppi M, Raina R, Fedoroff N (1994) Epigenetic regulation of the maize *Spm* transposable element: novel activation of a methylated promoter by TnpA. *Cell* 77: 427 – 437

Sheban D, Shani T, Maor R, Aguilera-Castrejon A, Mor N, Oldak B, Shmueli MD, Eisenberg-Lerner A, Bayerl J, Hebert J *et al* (2022) SUMOylation of linker histone H1 drives chromatin condensation and restriction of embryonic cell fate identity. *Mol Cell* 82: 106 – 122.e9

Sigman MJ, Panda K, Kirchner R, McLain LL, Payne H, Peasari JR, Husbands AY, Slotkin RK, McCue AD (2021) An siRNA-guided ARGONAUTE protein directs RNA polymerase V to initiate DNA methylation. *Nat Plants* 7: 1461 – 1474

Singer T, Yordan C, Martienssen RA (2001) Robertson's *Mutator* transposons in *A. thaliana* are regulated by the chromatin-remodeling gene *Decrease in DNA Methylation* (*DDM1*). *Genes Dev* 15: 591 – 602

Slotkin RK, Freeling M, Lisch D (2005) Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat Genet* 37: 641 – 644

Slotkin RK, Martienssen R (2007) Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* 8: 272 – 285

Tan F, Lu Y, Jiang W, Wu T, Zhang R, Zhao Y, Zhou DX (2018) DDM1 represses noncoding RNA expression and RNA-directed DNA methylation in heterochromatin. *Plant Physiol* 177: 1187 – 1197

Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* 11: 1438

To TK, Nishizawa Y, Inagaki S, Tarutani Y, Tominaga S, Toyoda A, Fujiyama A, Berger F, Kakutani T (2020) RNA interference-independent reprogramming of DNA methylation in *Arabidopsis*. *Nat Plants* 6: 1455 – 1467

Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T (2009) Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* 461: 423 – 426

Wang D, Zhang J, Zuo T, Zhao M, Lisch D, Peterson T (2020) Small RNA-mediated *de novo* silencing of *Ac/Ds* transposons is initiated by alternative transposition in maize. *Genetics* 215: 393 – 406

