DOI 10.15252/embj.2021110070

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

Taku Sasaki[1,4,*], Kyudo Ro[1,4], Erwann Cailleux[2], Riku Manabe[1], Grégoire Bohl-Viallefond[2], Pierre Baduel[2], Vincent Colot[2], Tetsuji Kakutani[1] and Leandro Quadrana[2,3*]

[1] Graduate School of Science, 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, 75005 Paris, France.

[3] 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, 91405 Orsay, France.

[4] Equal contribution.

* corresponding authors taku.sasaki@bs.s.u-tokyo.ac.jp and leandro.quadrana@universite-paris-saclay.fr.

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. Two such anti-silencing systems have been identified in *Arabidopsis* and were found to be mediated by the DNA-binding proteins VANC encoded by *VANDAL* transposons. Here, we show that since their origin in eudicots, anti-silencing systems have rapidly diversified 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 ten distinct VANC-induced anti-silencing systems in *Arabidopsis*. Strikingly, anti-silencing of non-autonomous *VANDALs*, which can act as reservoirs of 24nt small RNAs, is critical to prevent the demise of cognate autonomous TEs and 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.

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 and 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; Tsukahara et al. 2009; Quadrana et al. 2019; Singer, Yordan, and Martienssen 2001). 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 (Hosaka et al. 2017; Fu et al. 2013), which are abundant in the *A. thaliana* genome (Kapitonov and 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, coevolution 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 state of target TEs, potentially leading to run-away 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* induces efficient trans-hypomethylation of homologous copies (Fu et al. 2013), indicating that this population provide 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 ~10bp-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 VANDALs 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 (Colome-Tatche 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 Figure 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 Figure 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 Figure 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 *VANDALs* 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 Figure 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 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 Figure S6).

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 *VANDALs* (Appendix

Figure S6), 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 8-10bp-long motif in any 6000bp-long sequence. Given that each hypomethylated copy typically contains many short-sequence motifs (Appendix Figure S6), 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 1000bp 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 seem 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 10bp, 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 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 and 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 out 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 Figure S7) 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

Figure S7), 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 Figure S7), suggesting that this domain organization is ancestral. Conversely, DUF287 is only detected in VANC-like proteins encoded by Brassicaceae's MULEs (Fig 3E; Appendix Figure S7) 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, Monfort, and Puigdomenech 2007). Indeed, we found that this TE encodes a VANC-like protein containing both DUF1985 and Ulp1 domains (Fig 3F) and carrries 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 *VANDAL*s as traits and performed (epi)QTL mapping using the hundreds of parental DMRs that segregate in this population (Colome-Tatche 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 Figure S8), demonstrating that trans-hypomethylation in the epiRILs is determined by *ddm1*-derived, VANC-encoding *VANDAL*s. 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 Figure S8), 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 one 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 (Figure 3) within each epiQTL interval (Appendix Figure S8). In total, there are at least ten independent anti-silencing systems associated with up to ten VANC-encoding *VANDAL* copies (Appendix Figure S8). 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* (*AT1TE31190*) candidate copies identified on our epiQTL analysis (Fig 4A and B; Appendix Figure S9). 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 Figure S9). 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 constrained to short regions enriched in the motif “TGTACGTMY” (Fig 4C and E), reproducing the pattern of hypomethylation observed in the epiRILs (Fig 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 Figure S10). 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 *VANDAL*2 copy numbers by comparing WGBS coverage between VANC1-TG and control samples, but did not observe significant differences in coverage (Appendix Figure 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 (Fig 4D and Fig 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 Figure S12). 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 Figure S12), 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 24nt-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 1N1**Δ**motif copies, respectively; Fig 5D). Supporting our hypothesis, introduction of the 1N1**Δ**motif 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 Figure S13). To confirm this result genome-wide, we obtained WGBS for two independent 1N1 and 1N1**Δ**motif transgenic lines and found that all *VANDAL1/1N1/2* and *2N1* copies become systematically re-methylated following the introduction of the 1N1**Δ**motif sequence (Fig 5E and F; Appendix Figure S14). Furthermore, 1N1**Δ**motif-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 1N1**Δ**motif to induce strong silencing of related *VANDAL*s is reminiscent to the silencing activity of the natural *Mu* killer locus from maize (Slotkin, Freeling, and Lisch 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 *VANDAL*s.

DISCUSSION

Epigenetic control of TEs imposes strong selective constraints, engaging hosts and TEs in intimate co-evolutionary dynamics (Hurst and Werren 2001; Cosby, Chang, and Feschotte 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 (Schläppi, Raina, and Fedoroff 1994; Gierl, Lütticke, and Saedler 1988). Given our observations that the *A. thaliana* genome contains numeros *Spm* families showing evidence of trans-hypomethylation in the epiRILs (Appendix Figure S15), including the highly active *Spm3* (Quadrana et al. 2019; Kato, Takashima, and Kakutani 2004; Miura et al. 2001), 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 (Q. 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 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; Sheban et al. 2021; Andreev et al. 2021). 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 Figure S7), 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 shape 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 homo-polymerization 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 derive 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 24nt-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 1N1**Δ**motif sequence to induce strong and concerted epigenetic resilencing 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, Freeling, and Lisch 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, Freeling, and Lisch 2005; Burgess et al. 2020; Wang et al. 2020). However, unlike *VANDAL* killer, which does not produce long transcripts and is associated with 24nt-long siRNAs, the hairpin transcript of *Mu* killer is processed into 22nt-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 small RNAs produced from endogenous TEs by the plant specific RNA Polymerase IV (Fultz and Slotkin 2017). De novo establishment of RdDM targeting is still enigmatic, though 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 re-silenced. Under this scenario, non-autonomous *VANDAL* copies that are no longer targeted by VANCs can induce re-methylation of related *VANDAL* sequences through identity-based Pol IV-RdDM (Fultz and 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 run-away *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 (Quadrana et al. 2019; Colome-Tatche et al. 2012). All plants were grown in long-days (16 h:8h 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 *Arabidopsis thaliana* Col-0 ecotype by floral dip (Clough and 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 100nt 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 following parameters "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" (Bolger, Lohse, and Usadel 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 and 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 Figure 1D and Appendix Figure 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 (Fig 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 and 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 and 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/1kbp) density of motifs were determined by counting the number of motifs in 1000bp windows. Distance between consecutive motifs were 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 (Colome-Tatche 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 (Colome-Tatche 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 1000 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 (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, Silla-Martínez, and Gabaldón 2009). Phylogenetic tree was generated with PhyML v20160207 (Guindon and 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 250bp 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

Amino acid sequences of VANCs are predicted from RNA-seq data of *ddm1* mutants. 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 was 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 24bp-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 are 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).

ACKNOWLEDGMENTS

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.

AUTHORS'S CONTRIBUTIONS

TS, VC, TK and LQ conceived the project. TS and KR performed VANC1 transgenic experiments. TS and RM performed 1N1**Δ**motif experiments. EC extracted genomic DNA for WGBS of the epiRILs and GB processed the WGBS data. LQ analysed the WGBS data, ONT results, and performed evolutionary analyses. PB and LQ performed epiQTL mapping. TS and LQ interpreted the data. LQ wrote the paper with additional input from TS, PB, VC and TK. All the authors read and approved the paper.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

Akalin, Altuna, Matthias Kormaksson, Sheng Li, Francine E. Garrett-Bakelman, Maria E. Figueroa, Ari Melnick, and Christopher E. Mason. 2012. "MethylKit: A Comprehensive R Package for the Analysis of Genome-Wide DNA Methylation Profiles." *Genome Biology* 13 (10): R87.

Andreev, V. I., C. Yu, J. Wang, J. Schnabl, L. Tirian, and M. Gehre. 2021. "A SUMO-Dependent Regulatory Switch Connects the piRNA Pathway to the Heterochromatin Machinery in Drosophila." *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.07.27.453956.abstract.

Avsec, Žiga, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, et al. 2021. "Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax." *Nature Genetics* 53 (3): 354–66.

Baduel, Pierre, Basile Leduque, Amandine Ignace, Isabelle Gy, José Gil Jr, Olivier Loudet, Vincent Colot, and Leandro Quadrana. 2021. "Genetic and Environmental Modulation of Transposition Shapes the Evolutionary Potential of Arabidopsis Thaliana." *Genome Biology* 22 (1): 138.

Bailey, Timothy L. 2011. "DREME: Motif Discovery in Transcription Factor ChIP-Seq Data." *Bioinformatics* 27 (12): 1653–59.

Böhne, Astrid, Qingchun Zhou, Amandine Darras, Cornelia Schmidt, Manfred Schartl, Delphine Galiana-Arnoux, and Jean-Nicolas Volff. 2011. "Zisupton—A Novel Superfamily of DNA Transposable Elements Recently Active in Fish." *Molecular Biology and Evolution* 29 (2): 631–45.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer

for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.
Bonnet, Amandine, Carole Chaput, Noé Palmic, Benoit Palancade, and Pascale Lesage. 2021. "A Nuclear Pore Sub-Complex Restricts the Propagation of Ty Retrotransposons by Limiting Their Transcription." *PLoS Genetics* 17 (11): e1009889.
Burge, C., and S. Karlin. 1997. "Prediction of Complete Gene Structures in Human Genomic DNA." *Journal of Molecular Biology* 268 (1): 78–94.
Burgess, Diane, Hong Li, Meixia Zhao, Sang Yeol Kim, and Damon Lisch. 2020. "Silencing of Mutator Elements in Maize Involves Distinct Populations of Small RNAs and Distinct Patterns of DNA Methylation." *Genetics* 215 (2): 379–91.
Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. "trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses." *Bioinformatics* 25 (15): 1972–73.
Clough, S. J., and A. F. Bent. 1998. "Floral Dip: A Simplified Method for Agrobacterium-Mediated Transformation of Arabidopsis Thaliana." *The Plant Journal: For Cell and Molecular Biology* 16: 735–43.
Colome-Tatche, M., S. Cortijo, R. Wardenaar, L. Morgado, B. Lahouze, A. Sarazin, M. Etcheverry, et al. 2012. "Features of the Arabidopsis Recombination Landscape Resulting from the Combined Loss of Sequence Variation and DNA Methylation." *Proceedings of the National Academy of Sciences* 109 (40): 16240–45.
Corem, Shira, Adi Doron-Faigenboim, Ophélie Jouffroy, Florian Maumus, Tzahi Arazi, and Nicolas Bouché. 2018. "Redistribution of CHH Methylation and Small Interfering RNAs across the Genome of Tomato ddm1 Mutants." *The Plant Cell* 30 (July): tpc.00167.2018.
Cortijo, Sandra, René Wardenaar, Maria Colomé-Tatché, Arthur Gilly, Mathilde Etcheverry, Karine Labadie, Erwann Caillieux, et al. 2014. "Mapping the Epigenetic Basis of Complex Traits." *Science* 343 (6175): 1145–48.
Cosby, Rachel L., Ni-Chen Chang, and Cédric Feschotte. 2019. "Host–transposon Interactions: Conflict, Cooperation, and Cooption." *Genes & Development* 33 (17-18): 1098–1116.
Creasey, Kate M., Jixian Zhai, Filipe Borges, Frederic Van Ex, Michael Regulski, Blake C. Meyers, and Robert A. Martienssen. 2014. "miRNAs Trigger Widespread Epigenetically Activated siRNAs from Transposons in Arabidopsis." *Nature*. https://doi.org/10.1038/nature13069.
Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.
Domínguez, Marisol, Elise Dugas, Médine Benchouaia, Basile Leduque, José M. Jiménez-Gómez, Vincent Colot, and Leandro Quadrana. 2020. "The Impact of Transposable Elements on Tomato Diversity." *Nature Communications* 11 (1): 4058.
Dupeyron, Mathilde, Kumar S. Singh, Chris Bass, and Alexander Hayward. 2019. "Evolution of Mutator Transposable Elements across Eukaryotic Diversity." *Mobile DNA* 10 (March): 12.
Fultz, Dalen, and R. Keith Slotkin. 2017. "Exogenous Transposable Elements Circumvent Identity-Based Silencing, Permitting the Dissection of Expression-Dependent Silencing." *The Plant Cell* 29 (2): 360–76.
Fu, Yu, Akira Kawabe, Mathilde Etcheverry, Tasuku Ito, Atsushi Toyoda, Asao Fujiyama, Vincent Colot, Yoshiaki Tarutani, and Tetsuji Kakutani. 2013. "Mobilization of a Plant Transposon

by Expression of the Transposon-Encoded Anti-Silencing Factor." *The EMBO Journal* 32 (17): 2407–17.
Gierl, A., S. Lütticke, and H. Saedler. 1988. "TnpA Product Encoded by the Transposable Element En-1 of Zea Mays Is a DNA Binding Protein." *The EMBO Journal* 7 (13): 4045–53.
Guindon, Stéphane, and Olivier Gascuel. 2003. "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood." *Systematic Biology* 52 (5): 696–704.
Hosaka, Aoi, Raku Saito, Kazuya Takashima, Taku Sasaki, Yu Fu, Akira Kawabe, Tasuku Ito, et al. 2017. "Evolution of Sequence-Specific Anti-Silencing Systems in Arabidopsis." *Nature Communications* 8 (1): 1–10.
Hurst, G. D., and J. H. Werren. 2001. "The Role of Selfish Genetic Elements in Eukaryotic Evolution." *Nature Reviews. Genetics* 2 (8): 597–606.
Johannes, Frank, Emmanuelle Porcher, Felipe K. Teixeira, Vera Saliba-colombani, Juliette Albuisson, Fabiana Heredia, Vincent Colot, et al. 2009. "Assessing the Impact of Transgenerational Epigenetic Variation on Complex Traits." *PLoS Genetics* 5 (6): e1000530.
Johnson, Erica S. 2004. "Protein Modification by SUMO." *Annual Review of Biochemistry* 73: 355–82.
Kapitonov, V. V., and J. Jurka. 1999. "Molecular Paleontology of Transposable Elements from Arabidopsis Thaliana." *Genetica* 107 (1-3): 27–37.
Kato, Masaomi, Kazuya Takashima, and Tetsuji Kakutani. 2004. "Epigenetic Control of CACTA Transposon Mobility in Arabidopsis Thaliana." *Genetics* 168 (2): 961–69.
Krueger, Felix, and Simon R. Andrews. 2011. "Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications." *Bioinformatics* 27 (11): 1571–72.
Leeuwen, Hans van, Amparo Monfort, and Pere Puigdomenech. 2007. "Mutator-like Elements Identified in Melon, Arabidopsis and Rice Contain ULP1 Protease Domains." *Molecular Genetics and Genomics: MGG* 277 (4): 357–64.
Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100.
Li, Qing, Steven R. Eichten, Peter J. Hermanson, Virginia M. Zaunbrecher, Jawon Song, Jennifer Wendt, Heidi Rosenbaum, et al. 2014. "Genetic Perturbation of the Maize Methylome." *The Plant Cell* 26 (12): 4602–16.
Lisch, Damon. 2002. "Mutator Transposons." *Trends in Plant Science* 7 (11): 498–504.
2015. "Mutator and MULE Transposons." *Microbiology Spectrum* 3 (2): MDNA3–0032 – 2014.
Maison, Christèle, Delphine Bailly, Jean-Pierre Quivy, and Geneviève Almouzni. 2016. "The Methyltransferase Suv39h1 Links the SUMO Pathway to HP1α Marking at Pericentric Heterochromatin." *Nature Communications* 7 (July): 12224.
Marín, Ignacio. 2010. "GIN Transposons: Genetic Elements Linking Retrotransposons and Genes." *Molecular Biology and Evolution* 27 (8): 1903–11.
Miura, A., S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani. 2001. "Mobilization of Transposons by a Mutation Abolishing Full DNA Methylation in Arabidopsis." *Nature* 411 (6834): 212–14.
Ninova, Maria, Yung-Chia Ariel Chen, Baira Godneeva, Alicia K. Rogers, Yicheng Luo, Katalin

Fejes Tóth, and Alexei A. Aravin. 2020. “Su(var)2-10 and the SUMO Pathway Link piRNA-Guided Target Recognition to Chromatin Silencing.” *Molecular Cell* 77 (3): 556–70.e6.

Oberlin, Stefan, Alexis Sarazin, Clément Chevalier, Olivier Voinnet, and Arturo Marí-Ordóñez. 2017. “A Genome-Wide Transcriptome and Translatome Analysis of Arabidopsis Transposons Identifies a Unique and Conserved Genome Expression Strategy for Ty1/Copia Retroelements.” *Genome Research* 27 (9): 1549–62.

Panda, Kaushik, and R. Keith Slotkin. 2020. “Long-Read cDNA Sequencing Enables a ‘Gene-Like’ Transcript Annotation of Transposable Elements.” *The Plant Cell* 32 (9): 2687–98.

Quadrana, Leandro, Mathilde Etcheverry, Arthur Gilly, Erwann Caillieux, Mohammed-Amin Madoui, Julie Guy, Amanda Bortolini Silveira, et al. 2019. “Transposition Favors the Generation of Large Effect Mutations That May Facilitate Rapid Adaption.” *Nature Communications* 10 (1): 3421.

Rigal, Mélanie, Claude Becker, Thierry Pélissier, Romain Pogorelcnik, Jane Devos, Yoko Ikeda, Detlef Weigel, and Olivier Mathieu. 2016. “Epigenome Confrontation Triggers Immediate Reprogramming of DNA Methylation and Transposon Silencing in *Arabidopsis Thaliana* F1 Epihybrids.” *Proceedings of the National Academy of Sciences* 113 (14): E2083–92.

Robertson, Donald S. 1978. “Characterization of a Mutator System in Maize.” *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 51 (1): 21–28.

Robillard, Émilie, Arnaud Le Rouzic, Zheng Zhang, Pierre Capy, and Aurélie Hua-Van. 2016. “Experimental Evolution Reveals Hyperparasitic Interactions among Transposable Elements.” *Proceedings of the National Academy of Sciences* 113 (51): 14763–68.

Sasaki, Eriko, Taiji Kawakatsu, Joseph R. Ecker, and Magnus Nordborg. 2019. “Common Alleles of CMT2 and NRPE1 Are Major Determinants of CHH Methylation Variation in Arabidopsis Thaliana.” *PLoS Genetics* 15 (12): e1008492.

Saze, Hidetoshi, and Tetsuji Kakutani. 2007. “Heritable Epigenetic Mutation of a Transposon-Flanked Arabidopsis Gene due to Lack of the Chromatin-Remodeling Factor DDM1.” *The EMBO Journal* 26 (15): 3641–52.

Schläppi, M., R. Raina, and N. Fedoroff. 1994. “Epigenetic Regulation of the Maize Spm Transposable Element: Novel Activation of a Methylated Promoter by TnpA.” *Cell* 77 (3): 427–37.

Sheban, Daoud, Tom Shani, Roey Maor, Alejandro Aguilera-Castrejon, Nofar Mor, Bernardo Oldak, Merav D. Shmueli, et al. 2021. “SUMOylation of Linker Histone H1 Drives Chromatin Condensation and Restriction of Embryonic Cell Fate Identity.” *Molecular Cell*, December. https://doi.org/10.1016/j.molcel.2021.11.011.

Sigman, Meredith J., Kaushik Panda, Rachel Kirchner, Lauren L. McLain, Hayden Payne, John Reddy Peasari, Aman Y. Husbands, R. Keith Slotkin, and Andrea D. McCue. 2021. “An siRNA-Guided ARGONAUTE Protein Directs RNA Polymerase V to Initiate DNA Methylation.” *Nature Plants* 7 (11): 1461–74.

Singer, T., C. Yordan, and R. a. Martienssen. 2001. “Robertson’s Mutator Transposons in A. Thaliana Are Regulated by the Chromatin-Remodeling Gene Decrease in DNA Methylation (DDM1).” *Genes & Development* 15 (5): 591–602.

Slotkin, R. Keith, Michael Freeling, and Damon Lisch. 2005. “Heritable Transposon Silencing Initiated by a Naturally Occurring Transposon Inverted Duplication.” *Nature Genetics* 37

(6): 641–44.
Slotkin, R. Keith, and Robert Martienssen. 2007. “Transposable Elements and the Epigenetic Regulation of the Genome.” *Nature Reviews. Genetics* 8 (4): 272–85.
Tan, Feng, Yue Lu, Wei Jiang, Tian Wu, Ruoyu Zhang, Yu Zhao, and Dao-Xiu Zhou. 2018. “DDM1 Represses Noncoding RNA Expression and RNA-Directed DNA Methylation in Heterochromatin.” *Plant Physiology* 177 (3): 1187–97.
Tang, Alison D., Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu, and Angela N. Brooks. 2020. “Full-Length Transcript Characterization of SF3B1 Mutation in Chronic Lymphocytic Leukemia Reveals Downregulation of Retained Introns.” *Nature Communications* 11 (1): 1438.
To, Taiko Kim, Yuichiro Nishizawa, Soichi Inagaki, Yoshiaki Tarutani, Sayaka Tominaga, Atsushi Toyoda, Asao Fujiyama, Frédéric Berger, and Tetsuji Kakutani. 2020. “RNA Interference-Independent Reprogramming of DNA Methylation in Arabidopsis.” *Nature Plants* 6 (12): 1455–67.
Tsukahara, Sayuri, Akie Kobayashi, Akira Kawabe, Olivier Mathieu, Asuka Miura, and Tetsuji Kakutani. 2009. “Bursts of Retrotransposition Reproduced in Arabidopsis.” *Nature* 461 (7262): 423–26.
Wang, Dafang, Jianbo Zhang, Tao Zuo, Meixia Zhao, Damon Lisch, and Thomas Peterson. 2020. “Small RNA-Mediated De Novo Silencing of Ac/Ds Transposons Is Initiated by Alternative Transposition in Maize.” *Genetics* 215 (2): 393–406.

Figures:

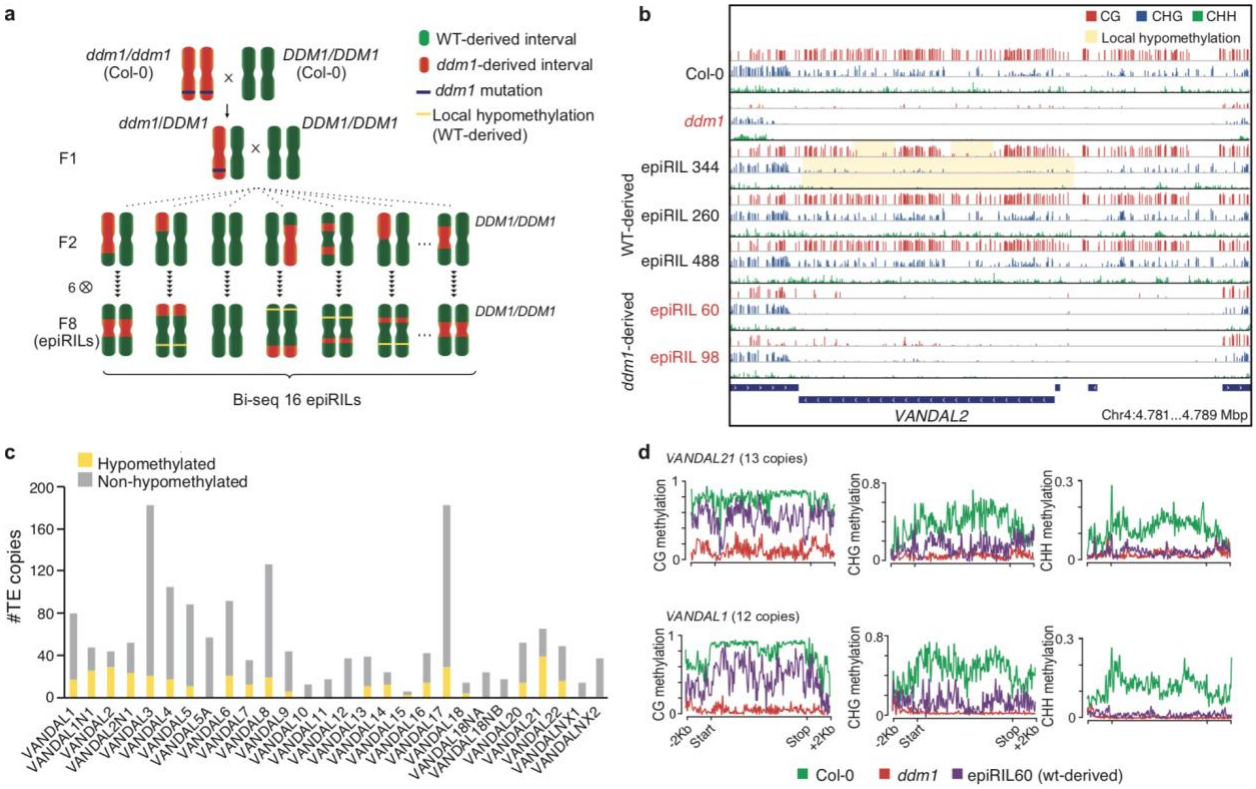

Figure 1. Most VANDAL families are trans-hypomethylated in epiRILs. a. Cartoon depicting the crossing scheme used to generate the epiRIL population. **b**. Genome browser tracks showing local hypomethylation of a WT-derived *VANDAL2* copy in Col-0, *ddm1* mutant and several epRILs that carry the WT- 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.

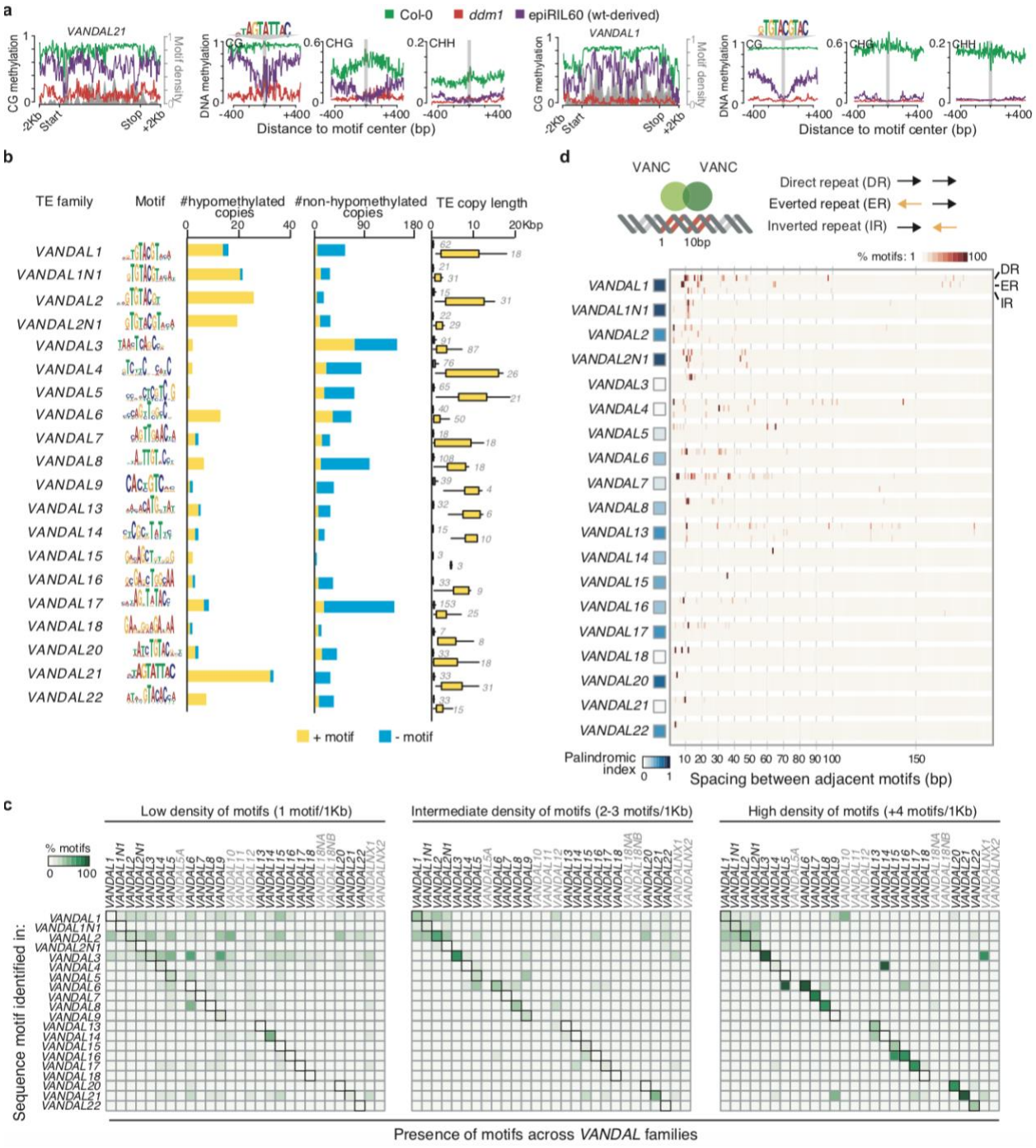

Figure 2. Sequence and syntax of motifs dictate VANC anti-silencing specificity. a. Metaplot of DNA methylation and relative motif density (arbitrary units: 1=max; 0=min) on wt-derived *VANDAL21* or *VANDAL1* copies as well as around short-sequence motifs within Col-0, *ddm1* or epiRIL60. **b**. Motif logo, number of hypomethylated and non-hypomethylated WT-derived copies and size of *VANDALs* carrying or not short sequence motifs. **c**. Proportion of isolated or increasingly clustered motifs (medium: 2-3 motifs/1Kb; high +4 motifs/1Kb) 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.

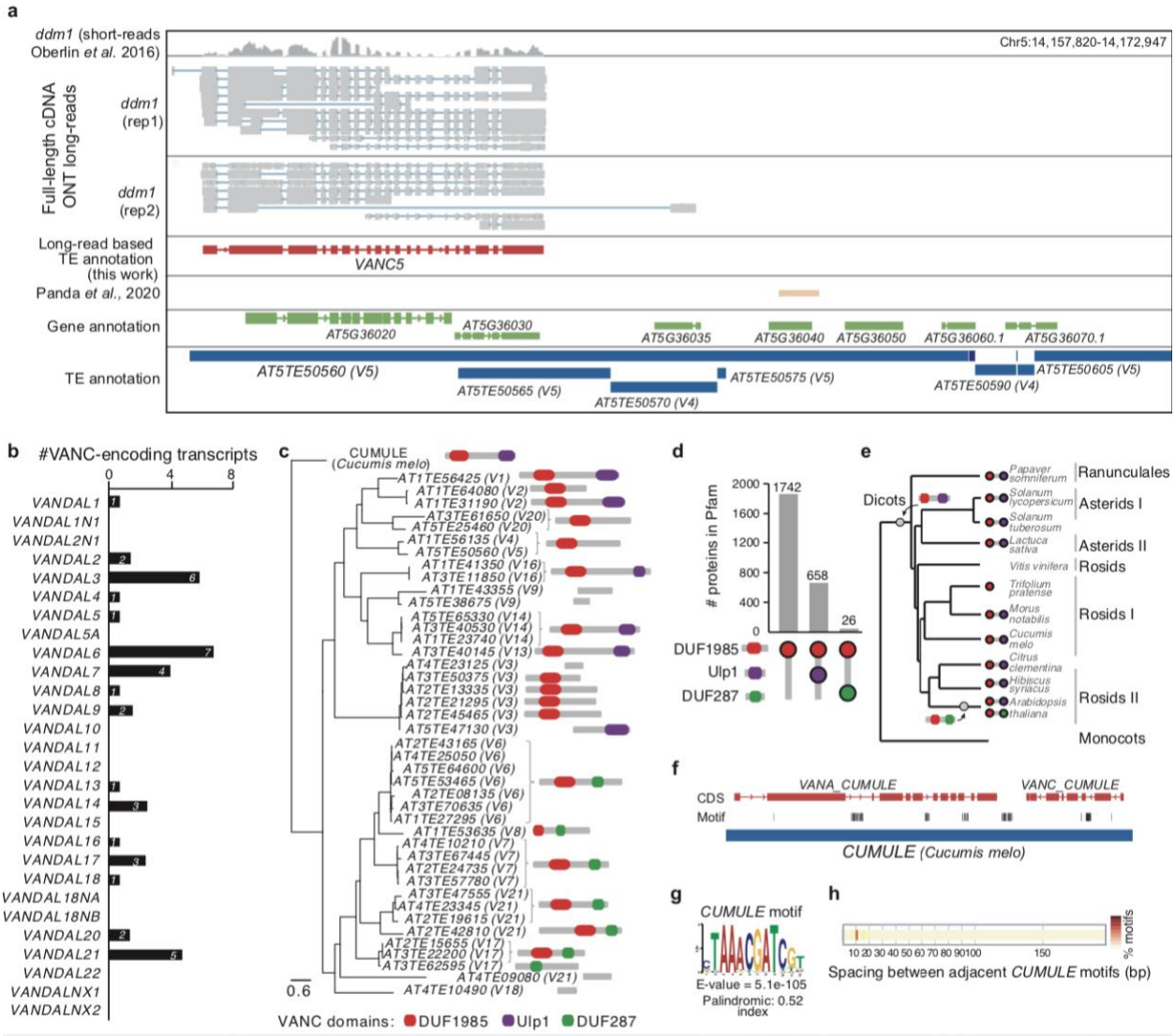

Figure 3. Diversification of VANC-dependent anti-silencing systems within and across species. a Genome Browser view of full-length cDNA nanopore reads and illumina short-reads from *ddm1* mutant plants as well as *de novo* functional annotation of TE-encoding transcripts together with TAIR10 gene and TE annotations. *VANDAL4* and *VANDAL5* copies are indicated as V4 and V5, respectively**. b**. Number of VANC-like encoding transcripts. **b**. Phylogenetic relationship among predicted VANC proteins encoded by *A. thaliana VANDAL*s 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 are 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 is also shown. **h**. Spacing between consecutive short-sequence motifs within *CUMULE*.

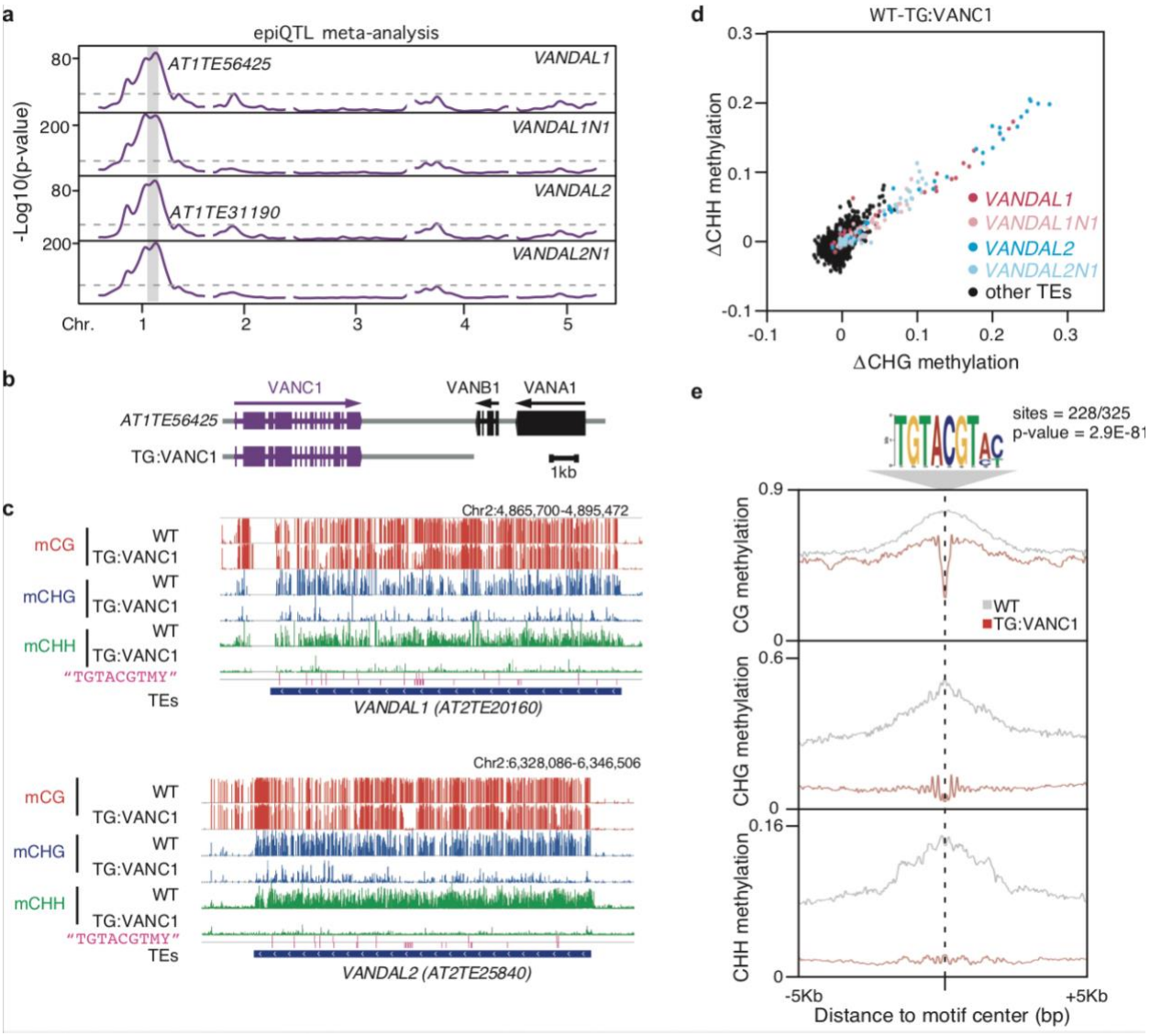

Figure 4. Ulp1-containing VANC1 induces sequence-specific hypomethylation. a. (epi)QTL mapping of VANDAL1, 1N1, 2 and 2N1 trans-hypomethylation in 105 epiRILs. The VANC-encoding VANDAL1 and VANDAL2 located within the single (epi)QTL interval are indicated in each case. **b**. Schematic diagram of structures 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).

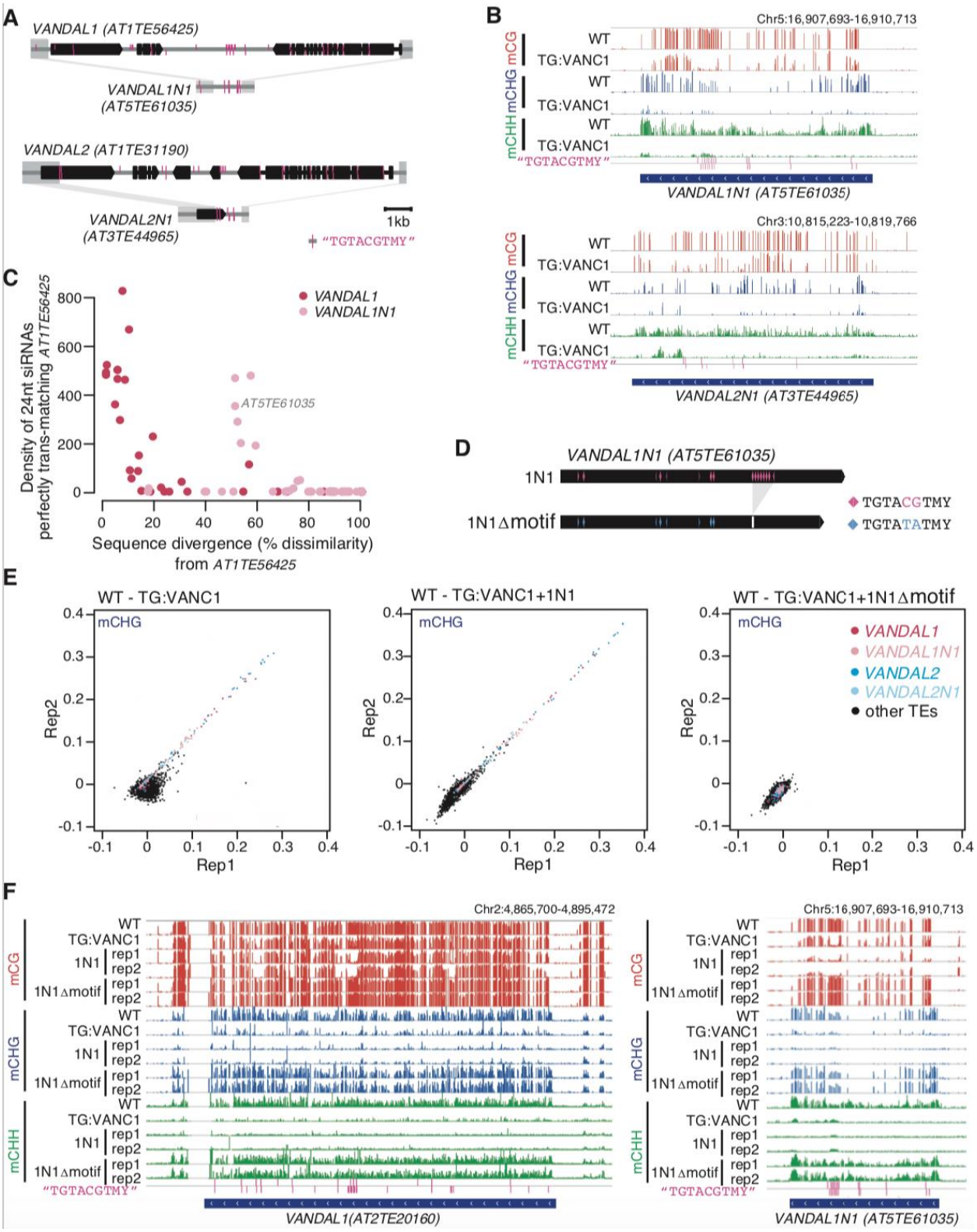

Figure 5. VANC-induced hypomethylation of non-autonomous copies prevents family-wide epigenetic resilencing. a. Schematic diagram of structures of full-length *VANDAL1* and *VANDAL2* copies and their derived non-autonomous *VANDAL1N1* and *VANDAL2N1*, respectively. Regions with high sequence

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/1000bp) of perfectly multiple-matching 24nt-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 (1N1**Δ**m). **e**. Comparison of CHG hypomethylation between replicates of plants containing VANC1, VANC1 + 1N1 or VANC1 + 1N1**Δ**m transgenes. *VANDAL1*, *1N1*, *2* and *2N1* copies are indicated by colors. **f**. Genome browser tracks showing the effect of 1N1 and 1N1**Δ**m transgenes on VANC1-induced hypomethylation over a *VANDAL1 and VANDAL1N1* copy. Methylation levels over the *VANDAL1N1* copy (*AT5TE61035*) in 1N1 samples reflects the average methylation of the endogenous and introduced copy.