Studies on anti-silencing mechanisms of non-TIR transposons in Arabidopsis

Hosaka, Aoi
Doctor of Philosophy

Department of Genetics
School of Life Science
SOKENDAI (The Graduate University for Advanced Studies)

Table of Contents

Summary

Transposable elements (TEs) are mobile genetic elements parasitizing the host genome. Since TEs are potentially harmful to the host, they are silenced by epigenetic mechanisms such as DNA methylation. In mutants deficient for maintaining DNA methylation, various developmental defects are induced due to insertions of TEs into genic regions.

On the other hand, some TEs counteract the host defense. An Arabidopsis transposable element named *Hiun* (*Hi*) is methylated and silent in wild type. When *Hi* is experimentally introduced into wild type plants as a transgene, endogenous *Hi* copy shows drastic decrease of DNA methylation, transcriptional activation, and mobilization. These anti-silencing effects are mediated by an ORF encoded in *Hi*, which is called VANC. Importantly, the *Hi* transgene induces the loss of DNA methylation specifically on VANDAL21 members but not on other TEs, which is in contrast to the global effects of counter-defense systems of infectious parasites that disrupt host surveillance machinery itself. Such a specific, rather than global, counter-defense would facilitate proliferation of this TE, with minimum deleterious effects to the host.

Although the sequence-specific anti-silencing by VANDAL21 seems reasonable biologically, the mechanism and evolution are enigmatic. Important questions would be (i) how the VANC functions in a sequence-specific manner and (ii) the evolutional dynamics of the sequence-specific anti-silencing system. To address these questions, I first confirmed that the VANC is sufficient for inducing sequence-specific loss of DNA methylation. To determine genome-wide distribution of the VANC protein, I performed ChIP-seq analyses against the VANC protein and found

that VANC predominantly binds to the non-coding regions of VANDAL21 TEs. I speculated that VANC may recognize specific DNA sequences and searched for motifs statistically overrepresented in VANC binding regions at VANDAL21 TEs. Among 7 identified motifs, “YAGTATTAY” (Y=T or C) motif was most commonly observed at the VANC binding regions. Furthermore, the motif was highly accumulated in VANDAL21 TEs, while it was rarely found in TEs of other VANDAL families. *In vitro* analyses suggested that the VANC protein is a double-stranded DNA-binding protein with preference binding to the motif sequence. Although the motif also exists outside of VANDAL21 TEs, they are unlikely targeted by VANC, indicating that the presence of the motif is not sufficient to explain preferential binding of the VANC. A very specific feature of the VANC binding regions compared to other regions with the motif is that the density of the motifs was much higher inside of VANDAL21 TEs. This suggests that the motif density is important for preferential binding of the VANC at VANDAL21 TEs *in vivo*.

To know how VANC localization affects DNA methylation, I examined the relationship between the VANC localization and DNA methylation and found that the VANC localization may trigger local loss of DNA methylation.

Importantly, I found that the accumulation of the motif at the VANC binding regions is mediated by the formation of tandem repeats. Comparative analyses of these sequences showed that the structures of tandem repeats are unstable, most likely due to expansion and contraction of the repeats. Furthermore, the turnovers of DNA sequences at the VANC binding regions have frequently occurred by repetitive sequence homogenizations of tandem repeats during the evolution of VANDAL TEs indicating

the rapid evolution of the VANC binding regions. The presence of tandem repeats may be advantageous for the evolution of VANDAL TEs, because they can efficiently amplify the VANC recognition motifs to increase their affinity and specificity to the VANC proteins, and can efficiently replace binding motifs of ancestral VANC proteins with those of a mutated VANC with efficient functions. In fact, VANC genes are known to evolve rapidly under positive selections. Taken together, I propose that the sequence-specific anti-silencing system evolves rapidly under co-evolution of VANC genes and their binding regions.

Introduction

Transposable elements (TEs) constitute significant portions of genomes in vertebrates and plants (Wicker *et al.,* 2007; Feschotte and Pritham, 2007; Tenaillon *et al.,* 2010). TEs are classified into two types according to their transposition manner; copy-and-paste (Class I) or cut-and-paste (Class II) (Wicker *et al.,* 2007). Since TE movement is mutagenic, most of them are silenced by epigenetic mechanisms, such as DNA methylation. In the TEs of plants, both CG and non-CG contexts of cytosine can be methylated (Cokus *et al.,* 2008; Lister *et al.,* 2008) and both of them can function for silencing the TEs (Law and Jacobsen, 2010).

Intriguingly, some TEs have developed mechanisms to counteract the defense systems in the host. For example, McClintock's *Suppressor-mutator* (*Spm*) element in maize encodes a protein TnpA, which induces loss of DNA methylation at a promoter region of *Spm* and reactivates it *in trans* (Schlappi *et al.,* 1994; Cui and Fedoroff, 2002). In the case of *Mutator* (*Mu*), another well-characterized TE in maize, an autonomously mobile copy named *MuDR* also reactivates silent copies of *Mu* (Brown and Sundaresan, 1992; Lisch *et al.,* 1995; 1999). *MuDR* contains two genes, *mudrA* and *mudrB*, the former encoding a transposase responsible for the loss of DNA methylation and transposition of Mu TEs (Eisen *et al.,* 1994; Lisch, 2002). TEs similar to the maize *Mu* are widespread in eukaryotes and they are referred to as *Mu*-like elements (MULEs) (Jiang *et al.,* 2004). ORFs related to *mudrA* are generally found in autonomous copies of these MULEs.

While all MULEs in maize contain conserved ~220bp TIRs, some MULEs in *Arabidopsis* genomes lack any recognizable TIRs and are classified as non-TIR-MULEs

(Le *et al.,* 2000; Yu *et al.,* 2000). Phylogenic analyses indicate that non-TIR-MULEs in *Arabidopsis* genomes form large families and they are derived from TIR-MULEs and proliferated recently. They generally possess several ORFs in addition to an ORF encoding *Mutator*-like transposase domain. In *Arabidopsis thaliana*, a group of non-TIR-MULEs called VANDAL21 transposes in background of DNA methylation deficient mutants (Tsukahara *et al.,* 2009). An autonomous copy of VANDAL21/AT2TE42810, referring to as *Hiun* (*Hi*), encodes three ORFs; VANA, which is a putative transposase, VANB, and VANC (Fu *et al.,* 2013). Importantly, a transgene of VANC induced hypomethylation, transcriptional activation, and excision of endogenous *Hi*, indicating that VANC is a novel anti-silencing factor. Furthermore, full-length of *Hi* transgene induces the loss of DNA methylation specifically in *Hi* and other VANDAL21 members (Fu *et al.,* 2013). These observations indicate that *Hi* harbors sequence-specific anti-silencing system, which is in contrast to the viral counter-defense systems that globally interrupt host surveillance (Boualem *et al.,* 2016). Because TEs cannot be horizontally transferred, reduction of host fitness by disruption of the host surveillance system would be deleterious for survival of TEs. However, it is unknown how the sequence-specificity is established. It is particularly interesting that even with high target specificity of the anti-silencing system, non-CG methylation is reduced in the entire region of VANDAL21 TEs, which are more than 8kb in length. Another important question is the evolutionary dynamics of the sequence-specific anti-silencing system.

In this thesis, I first confirmed that VANC is sufficient for inducing sequence-specific loss of DNA methylation at VANDAL21 TEs. ChIP-seq analyses of

the VANC protein revealed that VANC exclusively binds to non-coding regions of VANDAL21 TEs, and is associated with the local loss of DNA methylation. Importantly, VANC recognizes short DNA motifs enriched in non-coding regions of VANDAL21 TEs. I also found that the motif at non-coding regions is amplified by the formation of tandem repeats, and repetitive sequence homogenization processes by expansion and contraction of repeats have contributed for the rapid turnover of the motifs. These observations suggest that tandem repeats at the non-coding regions of VANDAL21 TEs are advantageous for rapid co-evolution between the VANC genes and VANC-targeted non-coding regions. I will further discuss the molecular basis and the evolutional dynamics of sequence-specific anti-silencing system.

Results

VANC is sufficient to induce sequence-specific loss of DNA methylation

It has previously been shown that a transgene of full-length *Hi* induces hypomethylation specifically on VANDAL21 copies (Fu *et al.,* 2013). To examine whether the VANC is involved in target specification, I analyzed DNA methylation profiles of ΔAB transgenic plants which lacks VANA and VANB transcription units in the Hi transgene, as shown in Figure 1, by whole-genome bisulfite sequencing (BS-seq). Both ΔAB and *Hi* transgenic plants showed very similar DNA methylation profiles, suggesting that VANC is sufficient for inducing the sequence-specific loss of DNA methylation at VANDAL21 loci (Figure 2, Supplemental Figure 1). Both in ΔAB and *Hi* transgenic plants, non-CG methylation was reduced in the entire region at VANDAL21 TEs, whereas effects on CG sites tend to local (Figure 2a).

Generation of anti-VANC polyclonal antibody

To analyze the function of VANC protein, I generated anti-VANC polyclonal antibody. Firstly, 6xHis-tagged full-length VANC (AT2G23480) protein was expressed in bacteria under induction by arabinose. Expression was confirmed by both CBB staining and western blotting (Supplemental Figure 2a, 2b). The protein was purified by nickel column and was used for immunization against rabbits (Supplemental Figure 2c).

To verify the specificity of immunized antiserum, nuclear proteins were extracted from both WT and ΔAB transgenic plants and used for western blotting (Supplemental Figure 2d, 2e). Clear signals were observed from the soluble fraction of nuclear extracts of ΔAB transgenic plants, but not in WT plants. Besides, the size of the

highest band was similar to that of the recombinant 6xHis-VANC protein expressed in bacteria. Taken together, immunized antiserum can specifically detect the VANC protein *in vivo*. I also confirmed that immunized antiserum can be used for immunoprecipitation (Supplemental Figure 2d, 2e). Additional lower-sized signals were also observed from the nuclear extracts and immunoprecipitated samples of VANC transgenic plants, possibly representing partially degraded VANC protein.

Predominant localization of the VANC protein at non-coding regions of VANDAL21 TEs

I speculated that sequence-specificity of the VANC protein is established by direct binding on their target DNA sequences. To examine this, I performed ChIP-seq using anti-FLAG and anti-VANC antibodies against FLAG-VANC and ΔAB transgenic plants, respectively (Figure 1).

Genome-wide views of FLAG-VANC distribution showed several strong signals (shown by arrows in Figure 3). 21 out of 29 loci with the strong signal correspond to the VANDAL21 copies. ChIP-seq with the anti-VANC antibody showed similar distribution of the signals (Supplemental Figure 3)

Next, I observed localizing patterns of FLAG-VANC signals within each of the VANDAL21 TEs. Interestingly, the signals were accumulated on the non-coding regions, such as upstream of the genes, intergenic regions, and introns (Figure 4a).

VANC targets lineage-specific short motifs accumulated in non-coding regions of VANDAL21 copies

To understand how VANC determines its targets, I searched for sequences statistically overrepresented in VANC binding regions at VANDAL21 TEs. (N=89; Table 1). Among 7 identified sequences, "YAGTATTAY" (Y=T, or C) was most commonly observed (Table 1). I analyzed the distribution of these 7 motifs around VANC binding regions and found that the "YAGTATTAY" motif was most highly associated with that (Figure 4b). Besides, the "YAGTATTAY" motif was accumulated at the non-coding regions of VANDAL21 TEs, which is consistent with FLAG-VANC binding patterns (Figure 4a, Supplemental Figure 6). Interestingly, the "YAGTATTAY" motif was highly enriched only in TEs of VANDAL21 (Figure 4c). Furthermore, number of the motif in a TE was correlated with FLAG-VANC accumulation and loss of non-CG methylation induced by the ΔAB transgene (Figure 4d-f). These results suggest that the "YAGTATTAY" motif is the most likely to be the major target of VANC protein. However, in addition to the motif within VANDAL21, many motifs are located outside of VANDAL21 TEs (3679 sites of 3953 in total) in *A. thaliana* genome, and are not likely targeted by VANC, indicating that the presence of the motif is not sufficient to explain preferential binding of the VANC protein. Since many of FLAG-VANC binding regions at VANDAL21 TEs seem to contain several motifs with the same orientation, I suspected that the motif density might be important for the VANC localization. Indeed, the regions with more than 3 motifs in the same orientation in 1kb were predominantly found in VANDAL21 TEs (Figure 4g). This suggests that VANC recognizes the "YAGTATTAY" motif and the density is important for the specific binding of VANC protein.

Direct binding of VANC to the "YAGTATTAY" motif was examined by

EMSA (Electro Mobility Shift Assay). GST-tagged recombinant VANC protein (GST-VANC) purified from bacteria was incubated with two types of radiolabeled dsDNA probes; one is the act2 probe that was the 50bp dsDNA sequence derived from ACT2/AT3G18780 used as a control, and Motif-plus probe in which 9bp sequence of the act2 probe was converted into "CAGTATTAC" which is one the "YAGTATTAY" variants (Figure 5a, 5b, Supplemental Table 1).

When the probe was incubated with GST-VANC, Motif-Plus probe showed a distinctive shift, whereas act2 probe was also shifted to smeared signals without discrete bands (Figure 5a), suggesting that VANC protein has a specific binding mode to dsDNA when "CAGTATTAC" sequence is present in the probe. The addition of the unlabeled Motif-Plus outcompeted the interaction of the labeled Motif-Plus with GST-VANC, but the unlabeled act2 lacking the motif did not (Figure 5b). This result indicates the preferential binding of GST-VANC to the "CAGTATTAC" sequence *in vitro*.

In vitro analyses showed that VANC directly recognized at least "CAGTATTAC" motif. Notably, GST-VANC binds to dsDNA containing the motif but not to ssDNA derived from the dsDNA probe, indicating that VANC is a dsDNA-specific binding protein (Figure 5c).

VANC localization is tightly associated with the loss of DNA methylation

Next I examined the effect of VANC localization on DNA methylation. Efficacies of VANC on DNA methylation differ among contexts; while non-CG methylation was reduced in the entire regions of TEs, reduction of CG methylation was

limited (Figure 2c). To see the effects of VANC localization on DNA methylation more directly, I investigated DNA methylation profiles in 10kb segments centered on VANC localized regions (Figure 6). All contexts of DNA methylation were drastically decreased in VANC localized regions indicating that VANC binding is associated with the local loss of DNA methylation. Consistently, VANC localization at VANDAL21 TEs was associated with the local loss of DNA methylation (Supplemental Figure 6).

I also asked whether the VANC enrichment on VANDAL21 TEs is associated with transcriptional activation by analyzing RNA-seq datasets of ΔAB transgenic plants. As shown (Figure 7), not all of VANC-bound TEs were activated their transcription, suggesting that the VANC localization is not sufficient for inducing transcription. Regardless of the activation of transcription, VANC localization at VANDAL21 TEs was associated with local loss of DNA methylation, implying that the loss of DNA methylation is directly triggered by VANC, rather than the consequence of transcriptional activation (Supplemental Figure 6; i.e. AT2TE20140, AT3TE52540, and AT4TE16990).

Rapid accumulation of VANC recognition motifs by the formation of tandem repeats

VANDAL21 family can be divided into two subgroups according to sequences of VANA transposase core domain (Fu *et al.,* 2013). Hereafter I call these subgroups as VANDAL21_1, to which *Hi* belongs, and VANDAL21_2 (V21_1 and V21_2 in Figure 8, also see Supplemental Figure 4). While the "YAGTATTAC" motif (hereafter I call C-type motif) was observed in both subfamilies, the "YAGTATTAT"

motif (hereafter I call T-type motif) was accumulated only in VANDAL21_2 in *A. thaliana*, but not in VANDAL21_2 in *A. lyrata* (Figure 8). This result indicates that the differential accumulation of the "YAGTATTAY" variants reflects phylogenic relationship. Furthermore, the previous report showed that, while ORFs encoding transposase domain are relatively conserved among VANDAL families, other ORFs and non-coding regions are highly diverged (Supplemental Figure 5; Fu *et al.,* 2013). These observations suggest that such rapid gain and loss of the motif is associated with the divergence of the non-coding regions.

I wondered how the motifs accumulate so rapidly. Sequence comparisons of upstream regions of VANA ORF, where the motifs are arrayed, showed that the accumulation of C-type motif was associated with the formation of tandem repeats (Figure 9a). The same feature was also observed at introns of VANB (Figure 9c) and VANC (Figure9e), suggesting that the motif is amplified by the formation of tandem repeats. Tandem repeats are known to be unstable (Gemayel *et al.,* 2010). Indeed, comparisons of tandem repeats in each subfamily showed that repeat sequences vary with many insertions or deletions, in addition to point mutations (Figure 9b, 9d, 9f). Furthermore, sequences among subfamilies were highly diverged due to repetitive expansion and contraction of repeats (Figure 9a-d). These observations suggest that expansion and contraction of repeats during evolution of VANDAL TEs have driven the rapid gain and loss of the VANC recognition motifs.

I wondered whether the formation of tandem repeats is specific to VANDAL TEs. To address the question, I searched for tandem repeats within TEs of various families. Interestingly, most of superfamilies of Class II TEs possess certain amount of

tandem repeats including MuDR superfamily to which VANDAL TEs belong, while Class I TEs have much less numbers of tandem repeats in *Arabidopsis thaliana* (Figure 18). In addition, TEs in 11 species across the kingdoms were used for the analysis. Although average numbers and lengths of tandem repeats in each superfamily varies among species, generally Class II type TEs contain more tandem repeats than Class I TEs, suggesting that the formation of tandem repeats is not specific to VANDAL TEs, but rather very conserved nature of Class II type TEs.

Discussion

Genome-wide distribution of VANC protein

Here I defined the genome-wide distribution of VANC. It predominantly localizes at non-coding regions of VANDAL21 TEs. Both genomic localization and *in vitro* analyses suggest that VANC directly recognizes the "YAGTATTAY" DNA motif. Besides, density of the "YAGTATTAY" motif appears to be important for preferential binding of VANC, since motif dense regions are exclusively found at VANDAL21 TEs. Similary, some transposase-binding motifs are found in repetitive sequences at subterminal regions of TEs and are required for stable binding (Gierl *et al.,* 1988; Bravi-Angel *et al.,* 1995; Becker and Kunze, 1997; Raina *et al.,* 1998; Mack and Crawford, 2001; Hashida *et al.,* 2006). This suggests that the accumulation of the motifs for TE-encoded genes by tandem repeats for target specification is the general strategy of Class II type TEs. This may reflect the result that Class II TEs generally have more tandem repeats than Class I TEs.

Then, how does the densely arrayed motif contribute to the target specificity of VANC? Another anti-silencing factor of Spm TE, referred as TnpA, is known to recognize 11bp sequences accumulated on subterminal regions of the transposon, and intermolecular interactions between DNA-bound TnpA proteins are essential for stable binding on subterminal regions (Schläppi *et al.,* 1998). Analogous to TnpA, VANC may employ similar mechanisms for specific binding at non-coding regions of VANDAL21 where the motif is densely arrayed.

Although VANC was also accumulated at upstream regions of VANDAL17, there was no "YAGTATTAY" motif (Supplemental Figure 6; AT2TE11015,

AT2TE15655, AT3TE22200, and AT3TE62595). This suggests that VANC may recognize other motifs that I could not detect in this study.

In vitro analyses revealed that VANC specifically recognizes dsDNA. VANC does not have any known functional domains. It would be informative to analyze biochemical properties and three-dimensional structures of VANC proteins to understand the molecular basis of sequence-specific anti-silencing system.

Relationship between VANC localization and hypomethylation

I showed that the VANC localization was associated with the loss of DNA methylation While demethylation on CG sites was limited on the VANC binding regions, demethylation on non-CG was often spread over the entire region of the VANDAL TEs. This may be explained by the differences of DNA methylation pathways. CG methylation pattern in mother cell is inherited to daughter cells during replication by maintaining DNA methylation states by DNA methyltransferase, MET1. On the other hand, regulation of non-CG methylation pathway involves signal amplification steps. For example, plant-specific RNA polymerase IV transcribes on methylated regions to facilitate 24-nt of siRNA production, which guides effector complex including DNA methyltransferase, DRM2 to the targets. To maintain non-CG methylation, self-reinforcing loop of DNA methylation and histone H3K9 methylation is required (Matzke and Mosher, 2014). Once loss of DNA methylation is induced by VANC, amplification steps may be disrupted, resulting in spread of hypomethylation. A big remaining question is how the anti-silencing achieved. One possible pathway could be that a VANC primarily function as a transcription activator and the transcription

induces the loss of DNA methylation. However, VANC induced loss of CG methylation in some of the VANDAL21 copies without detectable transcriptional activation, suggesting that the primary function of the VANC is not transcriptional activation but trigger of demethylation. Then, how does the VANC localization affect DNA methylation? There are three possible models; (1) the VANC actively removes DNA methylation with its enzymatic activity, (2) VANC recruits DNA demethylases to the target regions, or (3) VANC-binding prevents from DNA methylation. Further study will be necessary to answer the question.

Co-evolution of VANC proteins and non-coding sequences mediated by tandem repeats

Dot-plot analyses of non-coding regions demonstrated that rapid accumulation of the VANC recognition motifs is associated with the formation of tandem repeats. Another important point is the fast evolution of non-coding regions. Tandem repeats are structurally unstable due to the expansion or contraction of repeats induced by either unequal crossing over during recombination or strand-slippage replication (Gemayel *et al.,* 2010). Indeed, repeat numbers vary at some VANC targeted-regions. Furthermore, sequences in these regions evolve rapidly, presumably due to the repetitive homogenization processes by expansion of repeats. The formation of tandem repeats would be beneficial not only to amplify the VANC recognition motif to increase affinity or specificity of VANC, but also to efficiently replace the binding motif of an ancestral VANC with that of a mutated VANC if the mutated VANC has more favorable functions, such as higher anti-silencing efficiency and target specificity. Therefore, the

tandem repeats would be advantageous for rapid co-evolution with the VANC genes. In fact, VANC-related genes are also known to evolve rapidly under positive selections (Fu *et al.,* 2013). As a consequence, this would lead to differentiation of VANDAL families. This is somewhat similar to evolution dynamics of centromere. Centromere forms large tandemly arrayed sequence referred as centromeric-satellites (Henikoff 2001). The sequences of centromeric-satellites are highly diverged even in closely related species, due to repetitive homogenization by expansion of centromeric-satellites (Henikoff 2001; Hall *et al.,* 2003). Proteins of centromere components that bind to centromeric-satellites are also known to evolve rapidly to specifically recognize the sequences and that presumably reflects the co-evolution between centromere-satellites and the components (Malik and Henikoff, 2001; Talbert *et al.,* 2004; Dawe and Henikoff, 2006).

Concluding Remarks

In this thesis, I found that the VANC binds to non-coding regions of VANDAL21 TEs by recognizing the specific short DNA motif and triggers local hypomethylation. This is the first report that revealed genome-wide distribution of the TE-encoded anti-silencing factor. Furthermore, rapid co-evolution of VANC genes and the non-coding regions is mediated by the repetitive formation of tandem repeats. These findings provide novel insights into evolutional aspects of TEs for their survival in the host genome.

Materials and Methods

Plant Materials

Arabidopsis thaliana strain Columbia-0 (Col-0) was used as "wild type". Transgenic lines with full length Hi and ΔAB Hi in pPZP2H-lac are described previously (Fu *et al.,* 2013). The FLAG tagged VANC construct was generated by two steps: (i) For generating VANC construct with 3x FLAG tag in C-terminus, 3xFLAG sequence and linear ΔAB Hi in pBluescript II SK (-) are generated by PCR and they are combined by In-Fusion HD cloning kit (Takara); (ii) The FLAG-tagged VANC sequence was PCR amplified and cloned in sma l-digested pPZP2H-lac vector by In-Fusion HD cloning kit.

Primers used for constructions and sequencing are listed at Supplemental Table 1. The sequence of 3xFLAG is

"GACTACAAAGACGATGACGACAAGGATTATAAGGATGACGATGATAAAGACTATAAAGATGATGATGACAAA".

Coomassie brilliant blue staining and Western blotting

Protein extract was on 5-20% gradient acrylamide gel (ATTO). CBB stain one (Nacalai) was used for CBB staining. For western blotting, the gel, hybond-P PVDF membrane (GE), and filter papers were equilibrated in Transfer buffer (25mM Tris, 192mM Glycine, 10% Methanol, 0.05% SDS: sodium dodecyl sulfate). Trans-blot SD semi-dry transfer cell (Bio-Rad) was used for transfer of proteins to the membrane. After transfer, the membrane was immuno-hybridized with following steps; blocking in TBS-T buffer (50mM Tris-HCl, 150mM NaCl, 0.05% Tween 20) containing 0.5% of ECL blocking

agent (GE) for 1 hour, washed with TBS-T buffer 3 times, incubated in TBS-T buffer with primary antibody for 1 hour, washed with TBS-T buffer 3times, incubated in TBS-T buffer with HRP-labeled secondary antibody for 1 hour, and washed with TBS-T buffer 3times. ECL prime Western Blotting Detection Reagents (GE) was used to induce Chemiluminescence. Signals were analyzed by LAS4000mini (GE). To detect VANC protein in vivo, Can Get Signal solutions (TOYOBO) were used during immuno-hybridization, instead of TBS-T buffer.

Generation of anti-VANC polyclonal antibody

Total RNA was isolated from *Hi* transgenic plants by PureLink Plant RNA Reagent (Thermo Fisher Scientific). 1ug of RNA was used for cDNA synthesis with AMV ver3.0 (Takara). VANC cDNA was amplified by PrimeSTAR GXL (98℃ 10sec 60℃ 15sec 68℃ 3min, 30cycles) and A-tailed by ExTaq (Takara). The cDNA was TA-cloned into pGEM T-easy vector by Mighty Mix (Takara). Cloned full-length of VANC cDNA was amplified with AttB1 and AttB2 sequence added primers by PrimeSTAR GXL. The PCR fragment was cloned into pDEST17 vector by one-tube BP and LR Gateway reaction system following manufacture's protocol (Thermo Fisher Scientific). The pDEST17 vector containing VANC cDNA was transformed to E. coli of BL21-Al strain. Cells were pre-cultured for 8 hours in 5ml of LB liquid medium and the 0.5ml of the culture was inoculated in 25ml of LB liquid medium. After 3 hours of incubation, expression of 6xHis-tagged VANC protein (6xHis-VANC) was induced for 3 hours by adding up to 0.2% of L-arabinose. All culture steps were performed at 37°C. Cells were harvested by centrifugation at 6,000xG for 10 minutes. Cells were lysed, and

the insoluble fraction containing 6xHis-VANC was purified by Bugbuster Master mix (Millipore) following manufacture's protocol. The purified insoluble fraction was solubilized in denaturing Binding buffer (6M Urea, 30mM Imidazole, 1X PBS buffer). 6xHis-VANC was captured by HisTrap Ni sepharose column (GE), and eluted with Elution buffer (6M Urea, 200mM Imidazole, 1X PBS buffer). Purified 6xHis-VANC was used for immunizing rabbits (MBL).

EMSA assay

cDNA of VANC was cloned into pDEST15 vector as shown above. The vector was transformed into was transformed to E. coli of BL21-Al strain. Cells were pre-cultured for 8 hours in 5ml of LB liquid medium and the 1ml of the culture was inoculated in 100ml of LB liquid medium. After 3 hours of incubation at 37°C, the culture was incubated at 25°C for 1hour. Expression of VANC protein GST-tagged on N-terminus (GST-VANC) was induced for 24 hours at 25°C by adding up to 0.2% of L-arabinose. Cells were harvested and lysed by Bugbuster master mix. GST-VANC was purified by GSTrap HP (GE) following manufacture's protocol. Protein concentration was quantified by NanoDROP 2000 (Thermo). To radiolabel dsDNA probes, firstly 10uM of 49bp forward and complement oligonucleotides were mixed and heat denatured at 95° C for 5minutes. To anneal the oligonucleotides Temperature was decreased 5° C in every 5 minutes until become 50° C. 1uM of annealed oligonucleotides, 1U of T4PNK, 1xT4PNK buffer, and 0.5MBq of [α-32P] ATP was incubated for 1 hour at 37°C in 10ul of reaction solution. Radiolabeled oligonucleotides were column purified by using MicroSpin G-25 Columns (GE). 5ng of GST-VANC, 0.1uM of labeled

dsDNA, and 2ul of modified GRA buffer (Tris-HCl (ph 7.5) 0.15M, NaCl 0.6M, MgCl2 0.03M, Triton-X, 0.4%, Glycerol 40%, and 5mM DTT) was incubated at 4°C for 30 minutes in 20ul of a reaction solution (Hashida *et al.,* 2006). The reaction solution was electrophoresed on non-denaturing 5-20% gradient PAGE gel (ATTO) with 0.5xTBE buffer by 30mA. RI signals were detected by FLA-9000 (FUJI).

Library Preparation and High throughput sequencing

ChIP-seq

Approximately 5.0g of matured rosette leaves were fixed with by formaldehyde and ChIP was performed as previously described (Ito et al. 2015) by using antiserum of 6xHis-VANC immunized rabbit, and anti-FLAG antibody produced in rabbit (F7425 SIGMA). 1.2 ng of input and ChIP DNA was used for library construction by KAPA hyper prep kit (Kapa Biosystems) following manufacture's protocol. The libraries were amplified by 15 cycles of PCR by KAPA Hifi-PCR solution, and sequenced either by Miseq as 74bp of paired-end reads or Hiseq 4000 as 50bp of single-end reads.

BS-seq

Matured rosette leaves of WT and VANC transgenic plants were used for DNA extraction. Bisulfite treatments and library preparations were performed as previously described (Fu *et al.,* 2013).

RNA-seq

RNA was extracted from matured rosette leaves of WT, and ΔAB transgenic plants by PureLink Plant RNA Reagent (Thermo Fisher Scientific) and sent to Takara-bio Biomedical center (Takara) for library preparation and sequencing.

Bioinformatic Analyses

For BS-seq, paired-end reads were qualified using Trimmomatic-0.33 software with following options "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" (Bolger *et al.,* 2014). Qualified reads were mapped by "bismark" command of bismark (0.14.3) software with following options "-n 1 -l 20". PCR duplicates were removed from mapped bam files by "deduplicate_bismark" command (Krueger F & Andrews, 2011). Base-resolution of read counts of methylated and unmethylated cytosines were obtained as CX_reports files by "bismark_methylation_extractor" command with following options "--bedGraph --CX --cytosine_report".

For ChIP-seq, reads were mapped by Bowtie (0.12.8) with "-n 2 -M 1" and "-X 1000" options for single-end and paired-end reads, respectively (Langmead *et al.,* 2009). Resulting sam files were converted into bam files and sorted by SAMtools (0.1.18) (Handsaker *et al.,* 2009). To identify peaks of FLAG-VANC, sorted bam files of anti-FLAG immunoprecipitated sample of WT and FLAG-VANC transgenic plants were compared by MACS2 (2.1.0) "callpeak" command with following options "-g 135000000 -B -q 0.01" (Zhang *et al.,* 2008). DNA sequences of VANC enriched regions defined by MACS2 in VANDAL21 TEs were extracted. Short motifs were searched by a DREME script of MEME software (4.11.0) under default parameters except maximum core width was set as 9 (Bailey, 2011).

For RNA-seq, paired-end reads were mapped by tophat (1.4.0) with default parameters (Trapnell *et al.,* 2009). Reads mapped on specific regions were counted by "coverage"

command of BEDtools (2.16.2) (Quinlan & Hal 2010). Custom Perl scripts were used for downstream analyses. These datasets were visualized on IGV genome browser (Lander *et al.,* 2011). TAIR10 annotation was used for all sequence analyses. Estimation of phylogeny of VANDAL families was performed as described previously (Fu *et al.,* 2013) except sequences were aligned by MUSCLE algorithm and ClustalX (2.0.12) was used for construct neighbor-joining tree (Edgar 2004; Thompson *et al.,* 2002; Bateman, 2007; Saitou & Nei, 1987). Alignments of non-coding regions were performed by MUSCLE algorithm. For identification of tandem repeats in TEs, TE sequences in eukaryotic genomes were obtained from Repbase (21.05), and Tandem Repeats Finder (4.09) was used with following parameters" 2 5 7 80 10 50 500" (Jurka *et al.,* 2005; Benson, 1999).

References

Bailey, T.L. (2011). DREME: Motif discovery in transcription factor ChIP-seq data. Bioinformatics 27, 1653–1659.

Bateman, A. (2007). ClustalW and ClustalX version 2.0. Bioinformatics 21, 2947–2948.

Becker, H.A., and Kunze, R. (1997). Maize Activator transposase has a bipartite DNA binding domain that recognizes subterminal sequences and the terminal inverted repeats. Mol. Gen. Genet. 254, 219-230.

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.

Bolger AM, Lohse M & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–20

Boualem, A., Dogimont, C., and Bendahmane, A. (2016). The battle for survival between viruses and their host plants. Curr Opin Virol 17, 32–38.

Bravo-Angel, A.M., Becker, H.A., Kunze, R., Hohn, B., and Shen, W.H. (1995). The binding motifs for Ac transposase are absolutely required for excision of Ds1 in maize. Mol. Gen. Genet. 248, 527-534.

Brown, J., and Sundaresan, V. (1992). Genetic study of the loss and restoration of Mutator transposon activity in maize: evidence against dominant-negative regulator associated with loss of activity. Genetics 130, 889–898.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452, 215–219.

Cui, H., and Fedoroff, N.V. (2002). Inducible DNA demethylation mediated by the maize Suppressor-mutator transposon-encoded TnpA protein. Plant Cell 14, 2883–2899.

Dawe, R.K., and Henikoff, S. (2006). Centromeres put epigenetics in the driver's seat. Trends Biochem. Sci. 31, 662–669.

Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Eisen, J.A., Benito, M.I., and Walbot, V. (1994). Sequence similarity of putative transposases links the maize Mutator autonomous element and a group of bacterial insertion sequences. Nucleic Acids Res. 22, 2634–2636.

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

Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. 44, 445–477.

Gierl, a, Lütticke, S., and 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.

Hall, S.E., Kettler, G., and Preuss, D. (2003). Centromere satellites from Arabidopsis populations: Maintenance of conserved and variable domains. Genome Res. 13, 195–205.

Hashida, S.-N., Uchiyama, T., Martin, C., Kishima, Y., Sano, Y., and Mikami, T. (2006). The Temperature-Dependent Change in Methylation of the Antirrhinum Transposon Tam3 Is Controlled by the Activity of Its Transposase. Plant Cell 18, 104–118.

Henikoff, S. (2001). The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. Science 80. 293, 1098–1102.

Ito, T., Tarutani, Y., To, T.K., Kassam, M., Duvernois-Berthet, E., Cortijo, S., Takashima, K., Saze, H., Toyoda, A., Fujiyama, A., et al. (2015). Genome-wide negative feedback drives transgenerational DNA methylation dynamics in Arabidopsis. PLoS Genet. 11, e1005154.

James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. (2011) Integrative Genomics Viewer. Nature Biotechnology 29, 24–26

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. (2004). Pack-MULE transposable elements mediate gene evolution in plants. Nature 431, 569–573

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.

Krueger F & Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27: 1571–2

Langmead B, Trapnell C, Pop M & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. 11, 204–220.

Le, Q.H., Wright, S., Yu, Z., and Bureau, T. (2000). Transposon diversity in Arabidopsis thaliana. Proc. Natl. Acad. Sci. U.S.A. 97, 7376–7381.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-9

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

Lisch, D., Chomet, P., and Freeling, M. (1995). Genetic characterization of the Mutator system in maize: behavior and regulation of Mu transposons in a minimal line. Genetics 139, 1777–1796.

Lisch, D., Girard, L., Donlin, M., and Freeling, M. (1999). Functional analysis of deletion derivatives of the maize transposon MuDR delineates roles for the MURA and MURB proteins. Genetics 151, 331–341.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, a H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133, 523–536.

Mack, a M., and Crawford, N.M. (2001). The Arabidopsis TAG1 transposase has an N-terminal zinc finger DNA binding domain that recognizes distinct subterminal motifs. Plant Cell 13, 2319-2331.

Malik, H.S., and Henikoff, S. (2001). Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics 157, 1293–1298.

Matzke, M.A., and Mosher, R.A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat. Rev. Genet. 15, 394–408.

Quinlan AR and Hall IM. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841-2

Raina, R., Schlappi, M., Karunanandaa, B., Elhofy, A., and Fedoroff, N. (1998). Concerted formation of macromolecular Suppressor-mutator transposition complexes. Proc Natl Acad Sci U S A 95, 8526-8531.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

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

Talbert, P.B., Bryson, T.D., and Henikoff, S. (2004). Adaptive evolution of centromere proteins in plants and animals. J Biol 3, 18.

Tenaillon, M.I., Hollister, J.D., and Gaut, B.S. (2010). A triptych of the evolution of plant transposable elements. Trends Plant Sci. 15, 471–478.

Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinformatics Chapter 2, Unit 2.3.

Trapnell C, Pachter L & Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–11

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

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8, 973–982.

Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41, 331–368.

Yu, Z., Wright, S.I., and Bureau, T.E. (2000). Mutator-like elements in Arabidopsis thaliana. Structure, diversity and evolution. Genetics 156, 2019–2031.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, R137.

Acknowledgements

First of all, I appreciate my supervisor, Dr. Tetsuji Kakutani. I thank Dr. Taku Sasaki, Raku Saito, and Kazuya Takashima for kindly providing me NGS datasets for analyses and gave me advices. I thank the members of Progress committee, Dr. Ken-ichi Nonomura, Dr. Takuji Iwasato, Dr. Kazuhiro Maeshima, and Dr. Hitoshi Sawa, and former Progress committee, Dr Takehiko Kobayashi, and Dr Tatsuo Fukagawa for critical comments and encouragements. I thank all lab members in Kakutani Laboratory.

Figure Legends

Figure 1. *Hiun*-derived transgenes used in this study.

Schematic diagram of structures of *Hiun* and *Hiun*-derived transgenes used in this study. Black lines, Grey boxes, and grey lines indicate intergenic regions, coding regions, and introns, respectively. ΔAB lacks VANB and VANA ORFs as previously described (Fu *et al.,* 2013). FLAG-VANC contains genomic region of VANC ORF and its flanking regions. 3xFLAG sequence was fused to C-terminus of VANC ORF (see materials and methods).

Figure 2. VANC is sufficient for inducing sequence-specific loss of DNA methylation.

(a) DNA methylation profiles of a ΔAB transgenic plant (ΔAB), a full length *Hiun* transgenic plant (*Hi*), and their control wild type plants (WT_1, WT_2). Datasets of WT_1 and ΔAB were sequenced in this study, and datasets of WT_2 and *Hi_* were derived from a previous study (Fu *et al.*, 2013). Left and right ends of each TE are shown as broken lines. Each point represents proportion of methylated cytosine for a sliding window with seven fractions after separating each TE, including half-length of the TE for left and right flanking regions, for 200 fractions as previously described (Fu *et al.,* 2013). VANDAL21 copies showed decrease of DNA methylation in both ΔAB and full length *Hiun*. *CACTA2* (AT1TE42210) is shown as a negative control.

(b, c) Comparison of decrease in TE DNA methylation between full-length *Hiun* and

ΔAB transgenic plants at CHG sites (b) and CHH sites (c). The significance of decrease in DNA methylation was assessed by the value (Mn/Cn – Mt/Ct)/(1/√‾Cn + 1/√‾Ct), where Mn, Cn, Mt and Ct are methylated cytosine (M) and total cytosine (C) counts mapped for each TE in the non-transgenic (n) and transgenic (t) plants, respectively (Fu *et al.,* 2013). TEs more than 1kb long are plotted (N=5842). VANDAL21 copies are colored red.

Figure 3. VANC was localized mainly on VANDAL21 loci.

A genome-wide views showing enrichment of FLAG-VANC signal. In each 10kb, FLAG-VANC enrichment was calculated by log-scaled values of (IPt/INt) /(IPn/INn), where [IP] and [IN] are read counts for IP and input reads in FLAG-VANC transgenic [t] and non-transgenic [n] plants, respectively. Positions where FLAG-VANC is enriched more than 2 folds compared to input are pointed by arrows. 21 out of these 29 loci correspond to VANDAL21 loci (red arrows). Details for each of all these loci are shown as magnified views in supplemental Figure 6.

Figure 4. VANC signals were found in non-coding regions of VANDAL21 TEs that have specific short motifs.

(a) Coverage of FLAG-VANC signals on three of the VANDAL21 loci (arrow number 6, 12, and 18 in Figure 3). Black boxes, grey boxes, and grey lines indicate VANDAL21 TEs, coding regions, and introns, respectively. Position of the C-type

("YAGTATTAC") motif and the T-type ("YAGTATTAT") motif are shown as green and orange bars, respectively. Bars on top and beneath the grey line indicate that the motifs exist on forward and reverse strand, respectively. Distribution of the motif and the coverage of FLAG-VANC signals of other loci are shown in Supplementary Figure 6. (b) Distributions of motifs around VANC binding regions at VANDAL21 TEs (N=89). Numbers of the motifs found in each 50bp are summed and shown in Y-axis. Broken line corresponds to summits in the VANC binding regions.

(c) The "YAGTATTAY" motif is specifically accumulated in VANDAL21 TEs. Numbers of the motifs are counted in each TEs longer than 1kb. Distribution of the numbers in each VANDAL21 families and the other TEs are shown as boxplot. (d-f) Scatter plots comparing the "YAGTATTAY" motif numbers in each TE with VANC enrichment (d) and effects on DNA methylation at CHG sites (e) and CHH sites (f). Enrichment of FLAG-VANC was accessed by value (IPt/INt)/(IPw/INw) where [IP] and [IN] are read counts for immunoprecipitated and input read counts in FLAG-VANC transgenic [t] and wild type [w] plants, respectively. Effects on DNA methylation are assessed by the value (Mn/Cn – Mt/Ct), where Mn, Cn, Mt and Ct are methylated cytosine (M) and total cytosine (C) counts mapped for each TE in the non-transgenic (n) and transgenic (t) plants, respectively. TEs more than 1kb long are plotted (N=5842). VANDAL21 copies are colored red. (g) Motif dense regions are predominantly located inside of VANDAL21 TEs. The number of the motif around 1kb window in each motif sites was counted. For each number of the motif, the numbers and the percentages of the motif sites located within VANDAL21 TEs are shown as blue and red bars, respectively.

Figure 5. Binding mode of VANC protein to dsDNA *in vitro.*

(a) EMSA showing that binding mode of VANC protein to DNA depends on presence of "CAGTATTAC" sequence. GST-tagged full-length of VANC protein purified from bacteria was used. Adding 0.1mM DTT into GRA buffer slightly reduced smear signals. (b) Competitor assay with unlabeled dsDNA sequences. Only motif-containing dsDNA effectively outcompetes visible interaction between GST-VANC and labeled dsDNA. (c) EMSA showing that GST-VANC specifically binds to dsDNA but not to either forward strand (ss-Fw) or reverse strand (ss-Rv) of single-stranded DNA. Probe structures are illustrated. Stars indicate radiolabeled 5' ends.

Figure 6. VANC binding was associated with the local loss of DNA methylation.

50bp-binned average DNA methylation profiles with 150bp-sliding window of WT and ΔAB transgenic plants on 10kb regions around summits of VANC binding regions within the VANDAL21 loci (N=89). For each 150bp, Proportion of DNA methylation was assessed by counts of methylated cytosine divided by total counts of cytosine at CG sites (a), CHG sites (b) and CHH sites (c).

Figure 7. Relationship between VANC localization and transcription.

Scatter plots comparing VANC enrichment and transcriptional activation of

VANDAL21 TEs by expressing VANC transgene. Assessment of VANC enrichment was described in Figure 4.

Expression level of VANDAL21 TEs were assessed by RPKM (Read Per kilobase Million mapped reads) normalized read counts on VANDAL21 TEs of ΔAB transgenic plants (a, c) and WT plants(b, d).

Figure 8. Rapid turnovers of VANC recognition motifs during the evolution of VANDAL TEs.

Phylogenic relationship among VANDAL21 and related VANDAL families within the genomes of *A. thaliana* and *A. lyrata*. *A. lyrata*-specific lineages are shown by red lines. The phylogenic relationship was estimated by Neighbor-joining method using transposase core domain. Numbers of C-type and T-type motifs are shown for each TE copy.

Figure 9. Formation of tandem repeats is associated with rapid accumulation of the VANC recognition motifs.

(a) VANDAL21 copy sequences are compared in upstream regions of VANA transcription units. Compared nine sequences are comprised of, VANDAL21 sequences in genomes of *A. lyrata* and *A. thaliana* (Figure 8). In this dot plot, regions with 10bp exact match are shown by dots. Presence of C-type and T-type motifs are shown as green and orange dots, respectively. (b) Alignment of sequences of VANDAL21_1

(N=5), VANDAL21_2 in *A. lyrata* (N=6), and VANDAL21_2 in *A. thaliana* (N=5). Positions of C-type motif and T-type motif are colored in green and orange, respectively. Sequences pointed by black triangles are those used for the dot-plot in (a). (c) Intronic regions of VANC are compared in six VANDAL21 copies in *A. thaliana*. The format is as shown in (a). (d) Alignment of sequences shown in (c). (e) Intronic regions of VANB are compared in four VANDAL21 copies in *A. thaliana*. The format is as shown in (a). (f) Alignment of sequences shown in (e).

Figure 10. Tandem repeats in eukaryotic TEs.

Average number of tandem repeats in major superfamilies of 12 eukaryotic TE references derived from RepBase21.05. Tandem repeats were identified by Tandem Repeats Finder (4.09) and classified by their lengths of repeat units. Class I and Class II type TEs were indicated as blue and pink lines, respectively.

Motif[a]	sites[b]	P-value[c]	E-value[d]
YAGTATTAY	**69**	**9.3e-24**	**2.5e-19**
AGTATTYC	36	3.9e-10	9.9e-6
GWAATACC	24	9.9e-9	2.5e-4
AAACAAAS	27	1.2e-8	0.0003
CAATATYA	32	6.6e-8	0.0016
AATCAMAAT	20	2.9e-7	0.007
GTACTMGTA	18	1.5e-6	0.036

Table 1. List of significantly overrepresented motifs in VANC binding regions at VANDAL21 TEs (N=89)
a: Overrepresented motif
b: Sequences matching the motif
c: The p-value of Fisher's Exact Test for enrichment of the motif in the sequences
d: The motif p-value times the number of candidate motifs tested

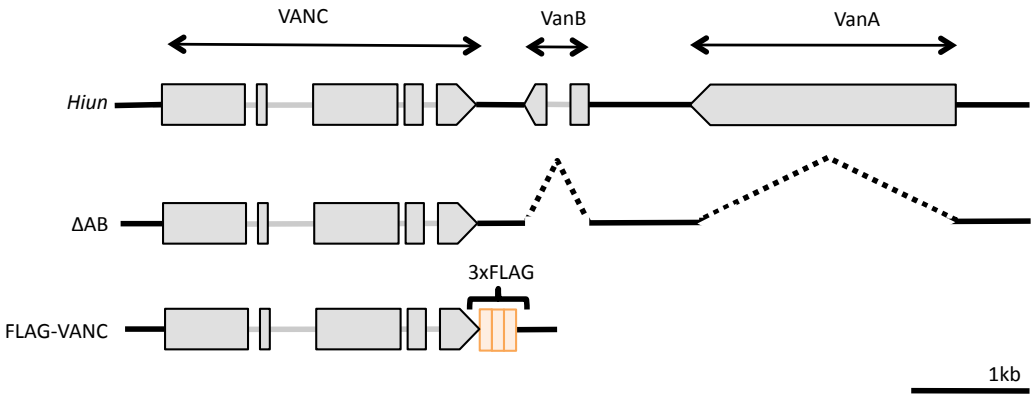


Figure 1. *Hiun*-derived transgenes used in this study

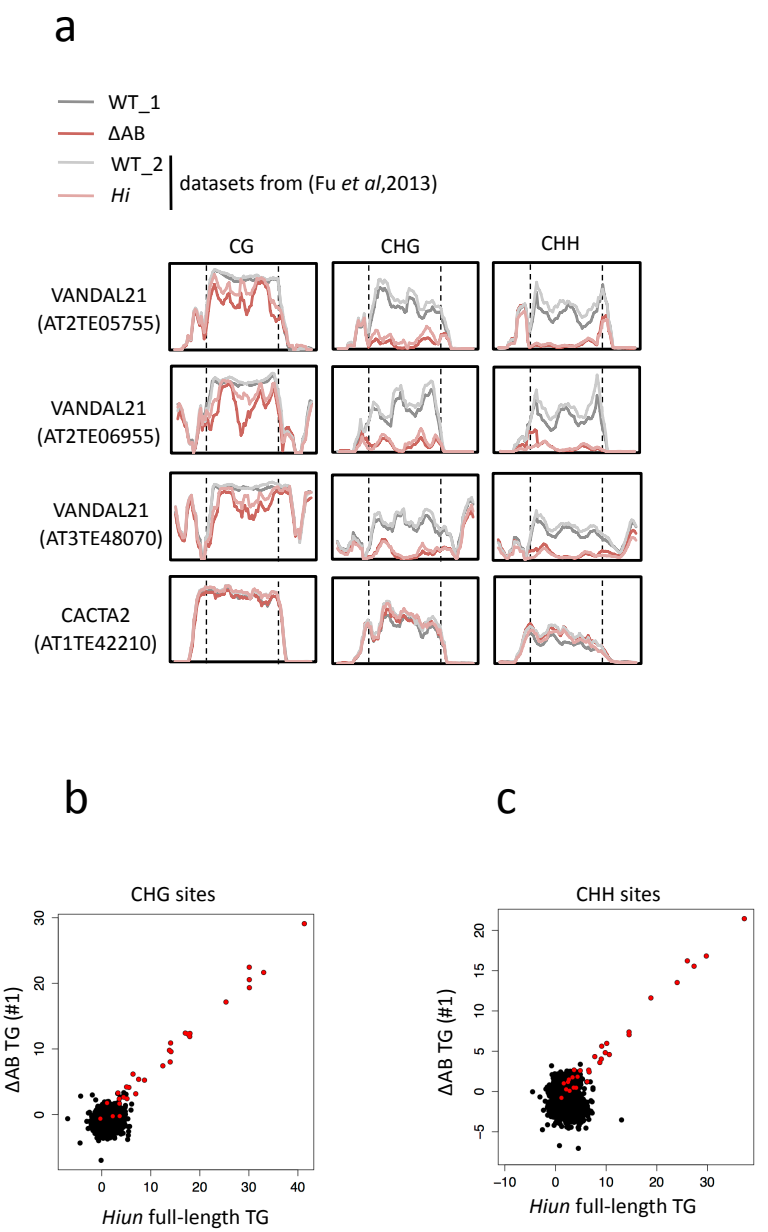


Figure 2. VANC is sufficient for inducing sequence-specific loss of DNA methylation

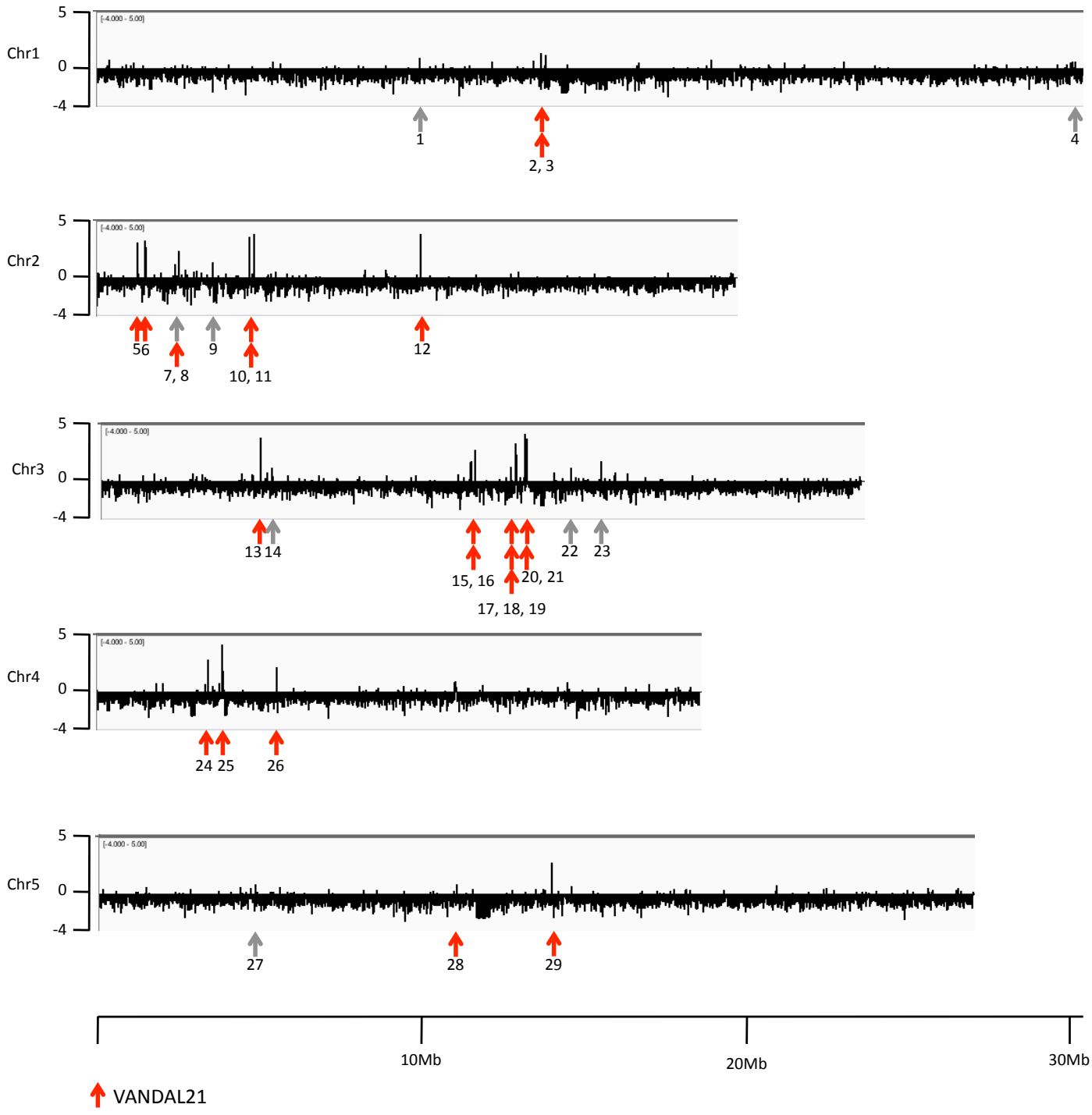


Figure 3. VANC was localized mainly on VANDAL21 loci

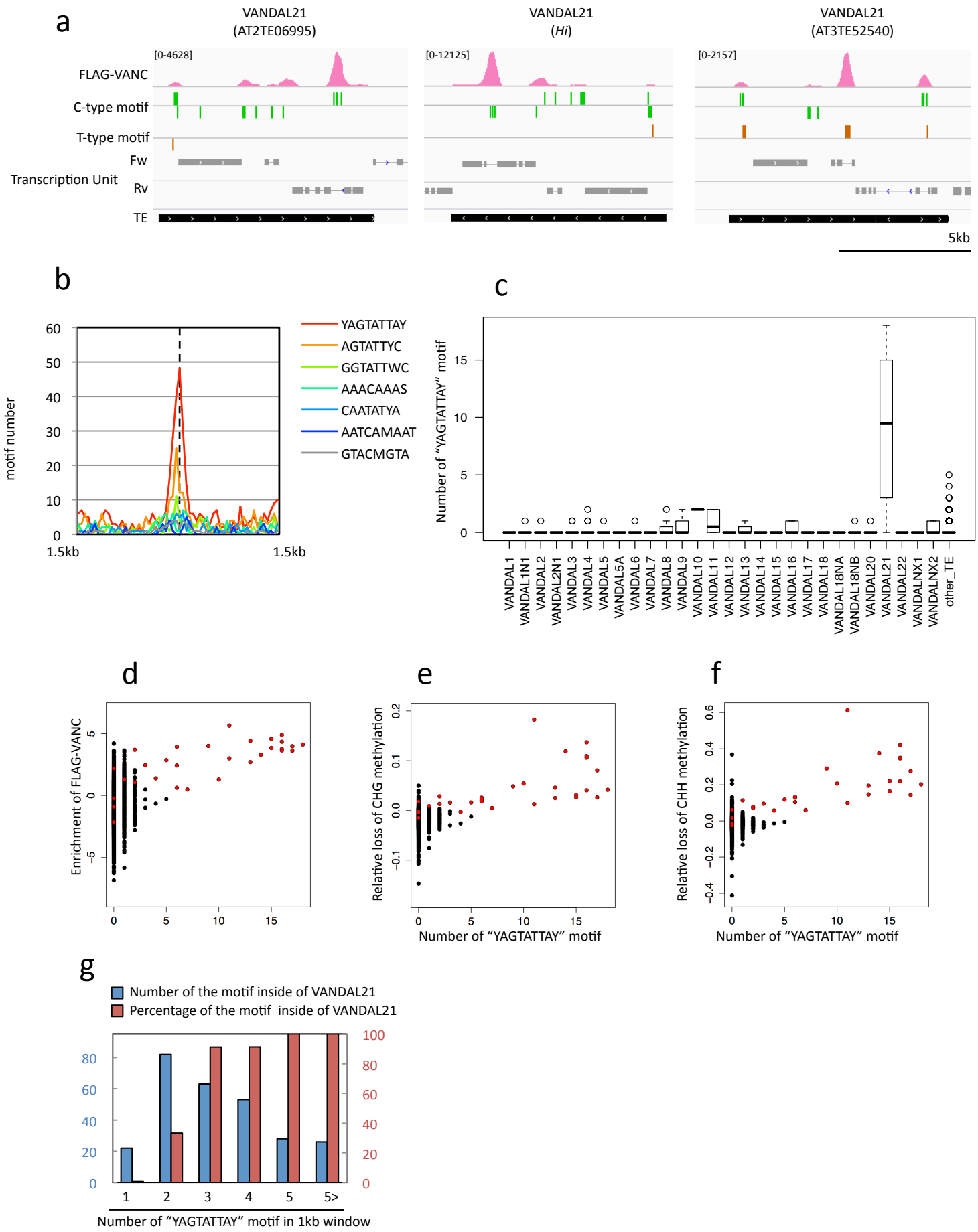


Figure 4. VANC signal was found in non-coding regions of VANDAL21 TEs, and those regions have specific short motifs.

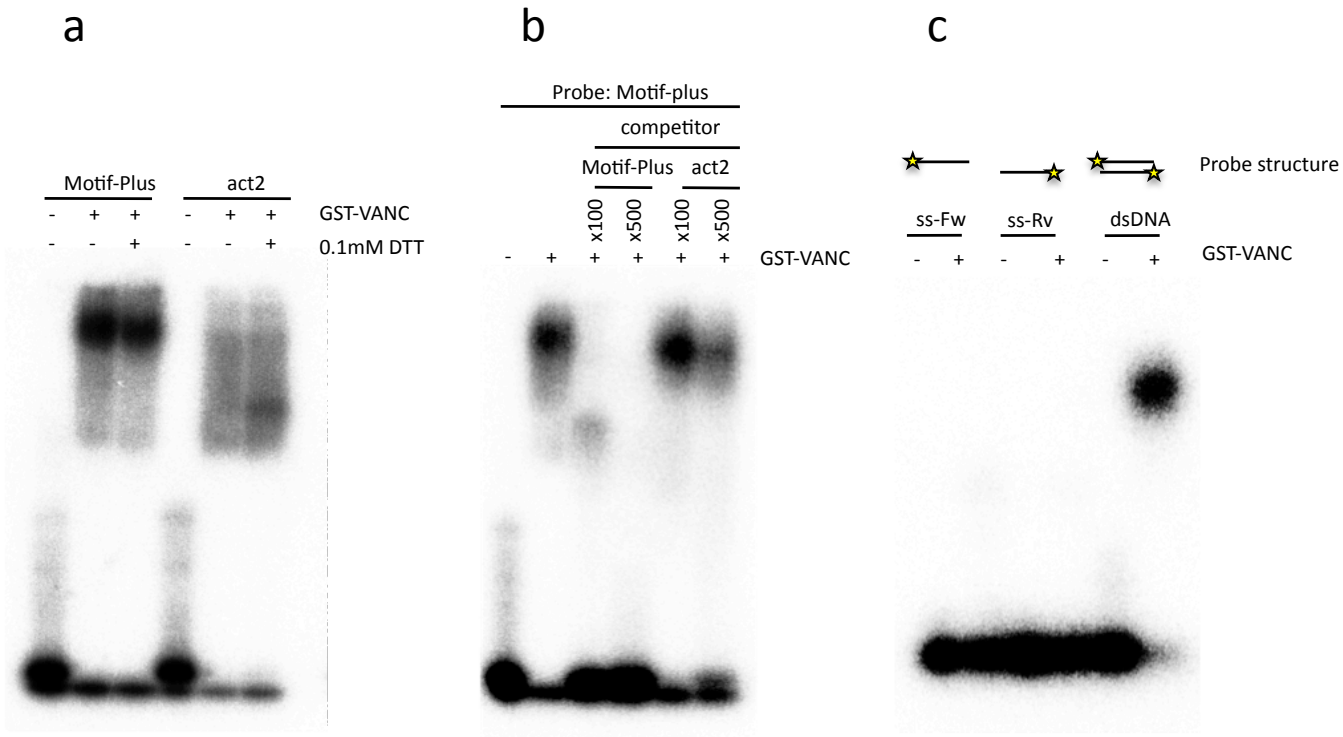


Figure 5. Binding mode of VANC protein to dsDNA *in vitro*

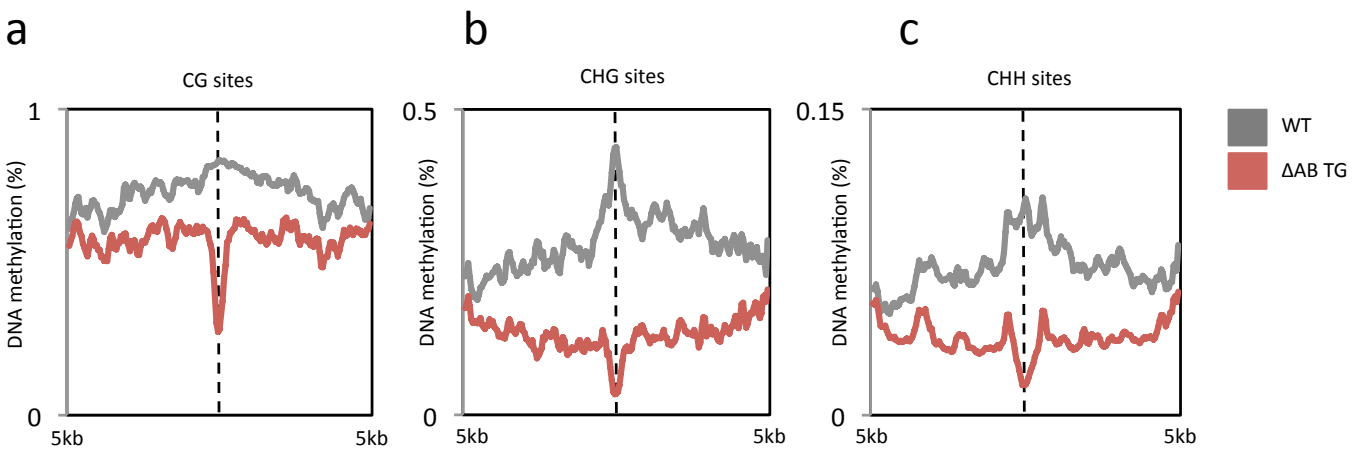


Figure 6. VANC binding was associated with local loss of DNA methylation

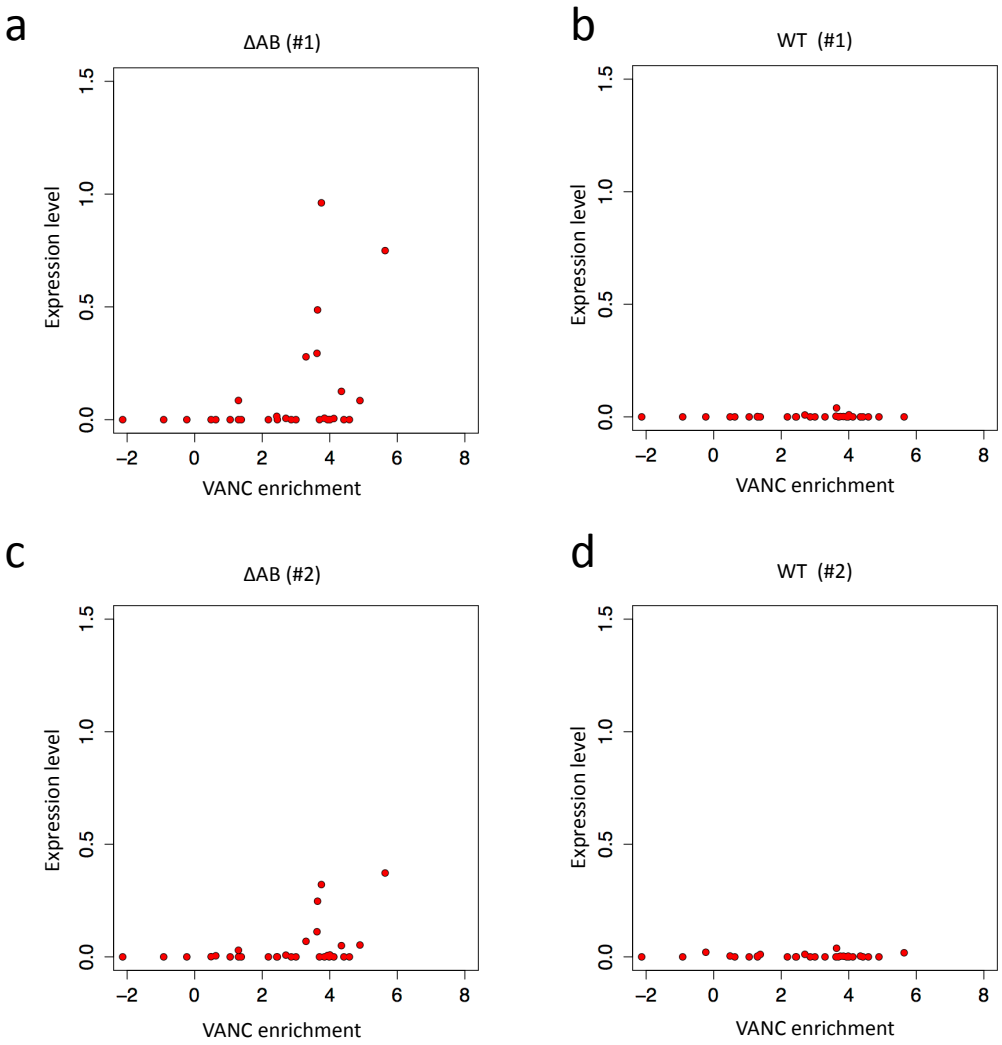


Figure 7. Relationship between VANC localization and transcription

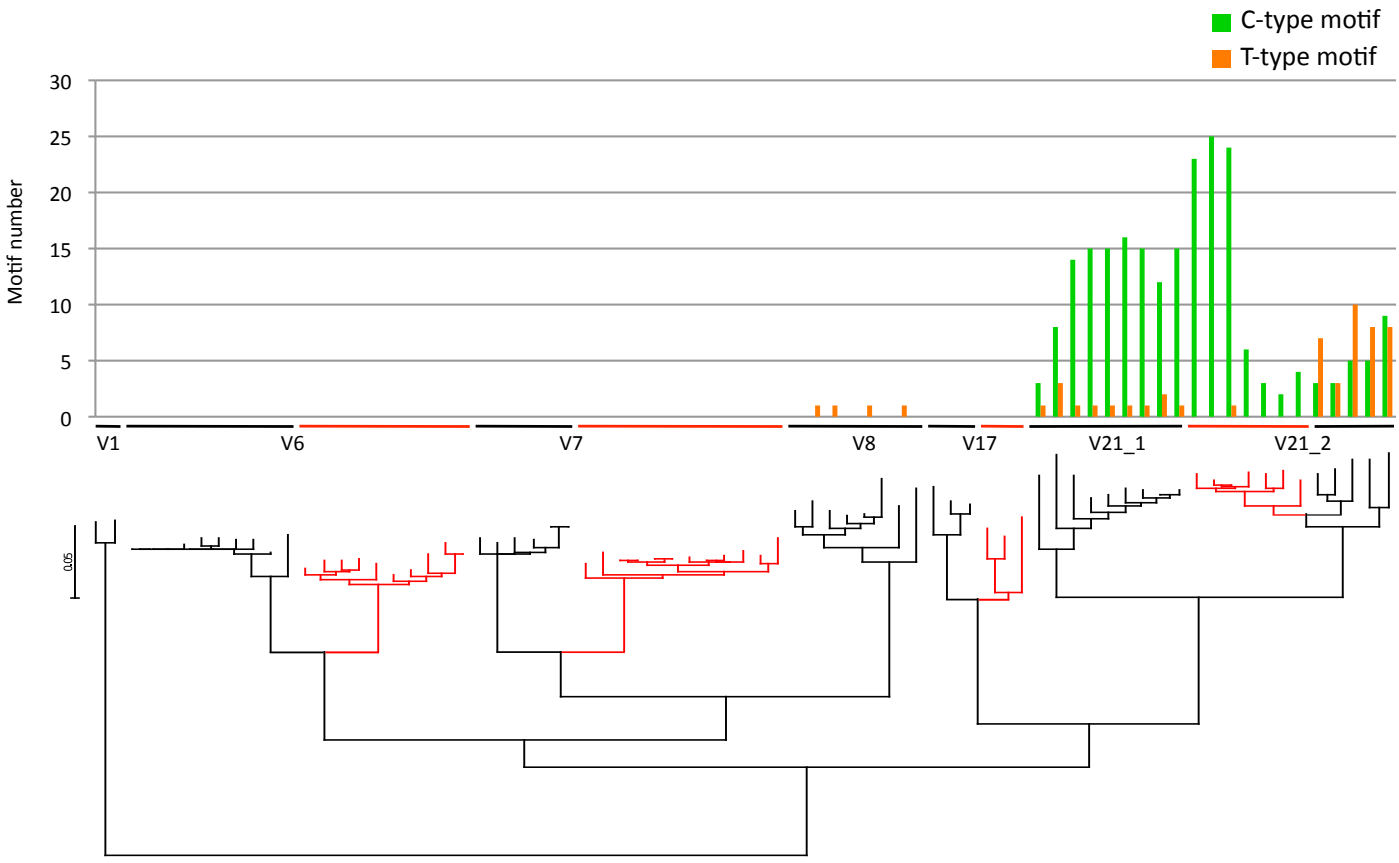


Figure 8. Rapid turnovers of VANC recognition motifs during evolution of VANDAL TEs

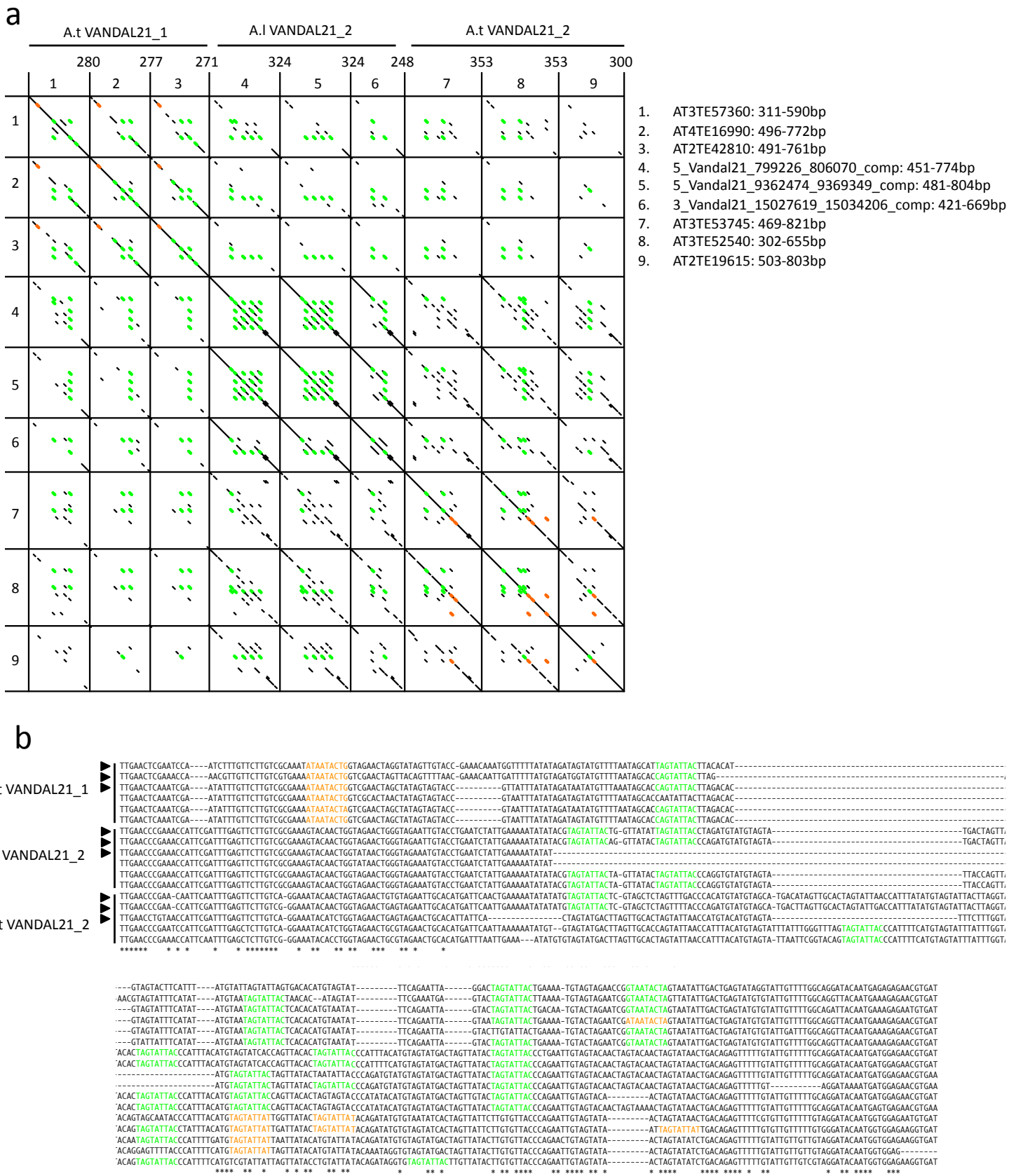
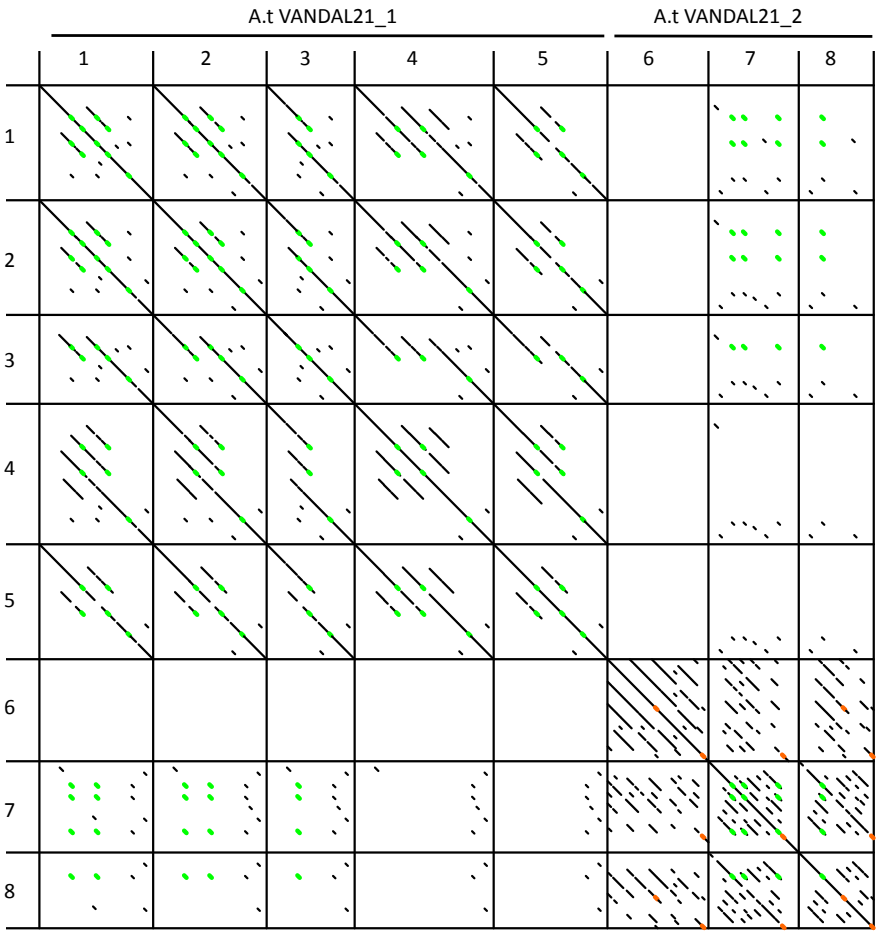


Figure 9. Formation of tandem repeats are associated with rapid accumulation of VANC recognition motifs

c



1. AT2TE05755: 5551-5988bp
2. AT3TE48070: 6535-6970bp
3. AT5TE49755: 6458-6795bp
4. AT2TE06955: 6467-7000bp
5. AT2TE42810: 6367-6803bp
6. AT4TE23345: 5913-6300bp
7. AT3TE52540: 5948-6294bp
8. AT2TE19615: 6099-6387bp

d

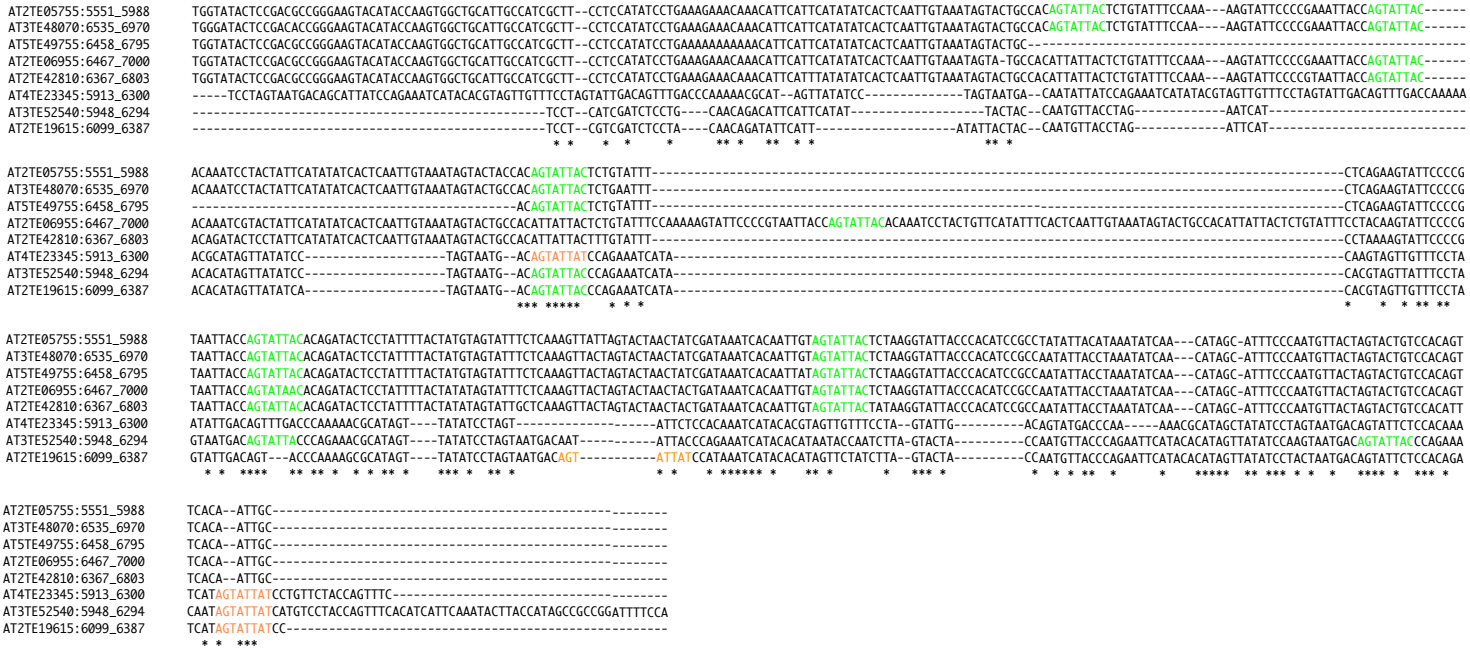
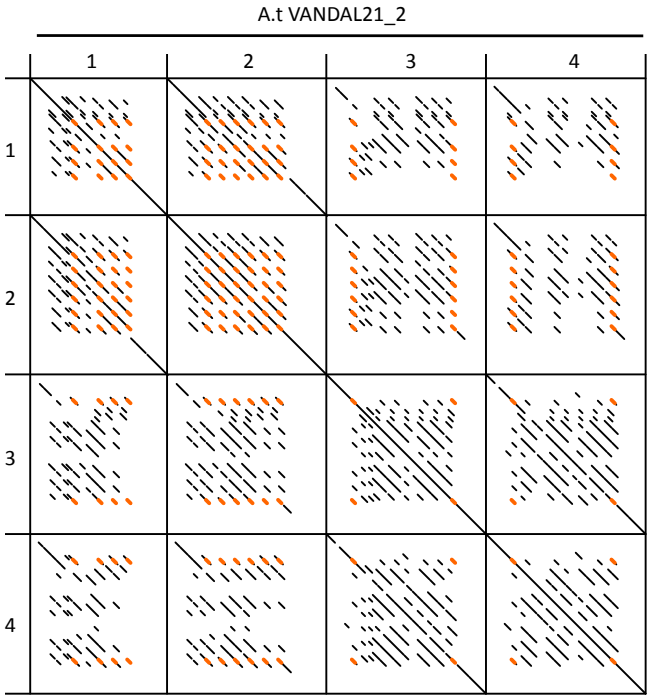


Figure 9. Formation of tandem repeats are associated with rapid accumulation of VANC recognition motifs (continued)

e



1. AT3TE52540: 4241-4600bp
2. AT3TE53745: 4400-4818bp
3. AT3TE47555: 3611-4030bp
4. AT2TE19615: 4176-4495bp

f

```
AT3TE52540:4241-4600      GGGTTGTGTGAGCTTAATTTATGCCTTTGTTTTGAAAGACTTTTATTGAGTATTCCCTTGTATTCCCAATATTATCGTTACCAAGTATTCCCTTGTA----------GGGTTGTGTGAGCTTAATTTATGCCTTTGTTTTGAAAGACTTTTATTGAGTATTCCCTTG
AT3TE53745:4400-4818      GGGTTGTGTGAGCTTCATTTATGCCTTTGTTTTGTAAGACTTTTATTGAGTATTCCCTTGTATTCCCAATATTATCGTTATCAAGTATTCCCTTGTATTCCCAGTATTATCGTTAAATCTGGGTTGTGTGAGCTTCATTTATGCCTTTGTTTTGTAAGACTTTTATTGAGTATTCCCTTG
AT3TE47555:3611-4030      GGTTTGTGTGAACTTAATTTCTGTCTTTGTTTTGTAAGACTTTTATTGAGTATTCTCTTATATTCCCAGTATTATCGTTATCAAGTTTTCCATTGTATTCTCAGTATTCCATTGTATTCTGGTTTGTGTGAACTTAATTTCTGTCTTTGTTTTGTAAGACTTTTATTGAGTATTCTCTTA
AT2TE19615:4176_4495      GGTTTGTGTGAACTTAATTTCTGCCTTTGTTTTTTAAGACTTTTATTGAGTATTCTCTTGTATTCCCAGTATTATCGTTATCAAGTATTCCATTGTATTCTCTGTATTATCGTTAAATCTGGTTTGTGTGAACTTAATTTCTGCCTTTGTTTTTTAAGACTTTTATTGAGTATTCTCTTG
                          ** ******** *** ***** ** *********  ********************** *** ******** ********** ****** ***** ******                                  ** ******** *** ***** ** *********  ********************** ***

AT3TE52540:4241-4600      TGTAATCGTTAAATCTTAGTTTTCCCTTGTATTCTCAGTATTATCGTTAAATCCTAGTATTCCATTGTATTCCCAGTATTATCGTTAAATCCTAGTATTCCATTGTATTCCCAGTATTAT----------CTCCTTGAATGCT---TCCAACGATTCAACTCTTTTAT
AT3TE53745:4400-4818      TATTATCGTTAAATCCTAGTATTCCCTTGTATTCTCAGTATTATCGTTAAATCCTAATATTCCATTGTATTCCCAGTATTATCGTTAAATCCTAGTATTCCATTGTATTCCCAGTATTATCGTTAAATCCTAACATATCAAATTCCTTGAATGCT---TTCAACGATTCAACTCTTTTAT
AT3TE47555:3611-4030      TATTATCGTTAAATCTTAGTTTTCTCTTGTATTCTCGGTATTATCCTCAAATCATAGTATTCCCTTGTATTCTCTGTATTATCGTTAAATCATAGTTTTCCCTTGTATTCTCTGTATTATCGTTAAATCATAGATT------TCCCTTGTATTCTCAGTATTATCTTTACATCCTAACAT
AT2TE19615:4176_4495      TATTATCGTTAAATCTTAGTTTTCCCTTGTATTCTCGGTATTATCTTCAAATCATAGTATTCTCTTGTATTCTCTGTATTATCTTTAAATCTTAGTTTTCCCTTGTATTCTCTGTATTATCGTTAAATCTTAGTTT------TCCCTTGTATTCTCAGTATTATCGTTAAATCCTAACAT
                          * * *********** **** *** ********** ******** * ***** ** *****  ******** * ******** ******* **** **** ******** * *******                        ***** ** **   *  *   **  *  **   **

AT3TE52540:4241-4600      -CCAAATATAAAATCATTGTTATAATTTCATTC---AGCTGGTCTACCATCTTCTTATCCAAGCTA
AT3TE53745:4400-4818      -CCAAATCTGAAATCATTGTTATAATTTCATTC---AGCTGGTCTACCATCTTCTTATCCAAGCTA
AT3TE47555:3611-4030      ATCAAATTCCAATTCATATCCAAATTGAAATCCAAAAAAAAGCATACCAAATTCAAATCCAAACAA
AT2TE19615:4176_4495      ATCAAATTCCAATTCATATCTAAATTGAAATCCAAAAAAAAGCATACCAAATTCAAATCCAAACAA
                            *****   ** ****    * * *   ** *   *    *  *****  ***  ****
```

Figure 9. Formation of tandem repeats are associated with rapid accumulation of VANC recognition motifs (continued)

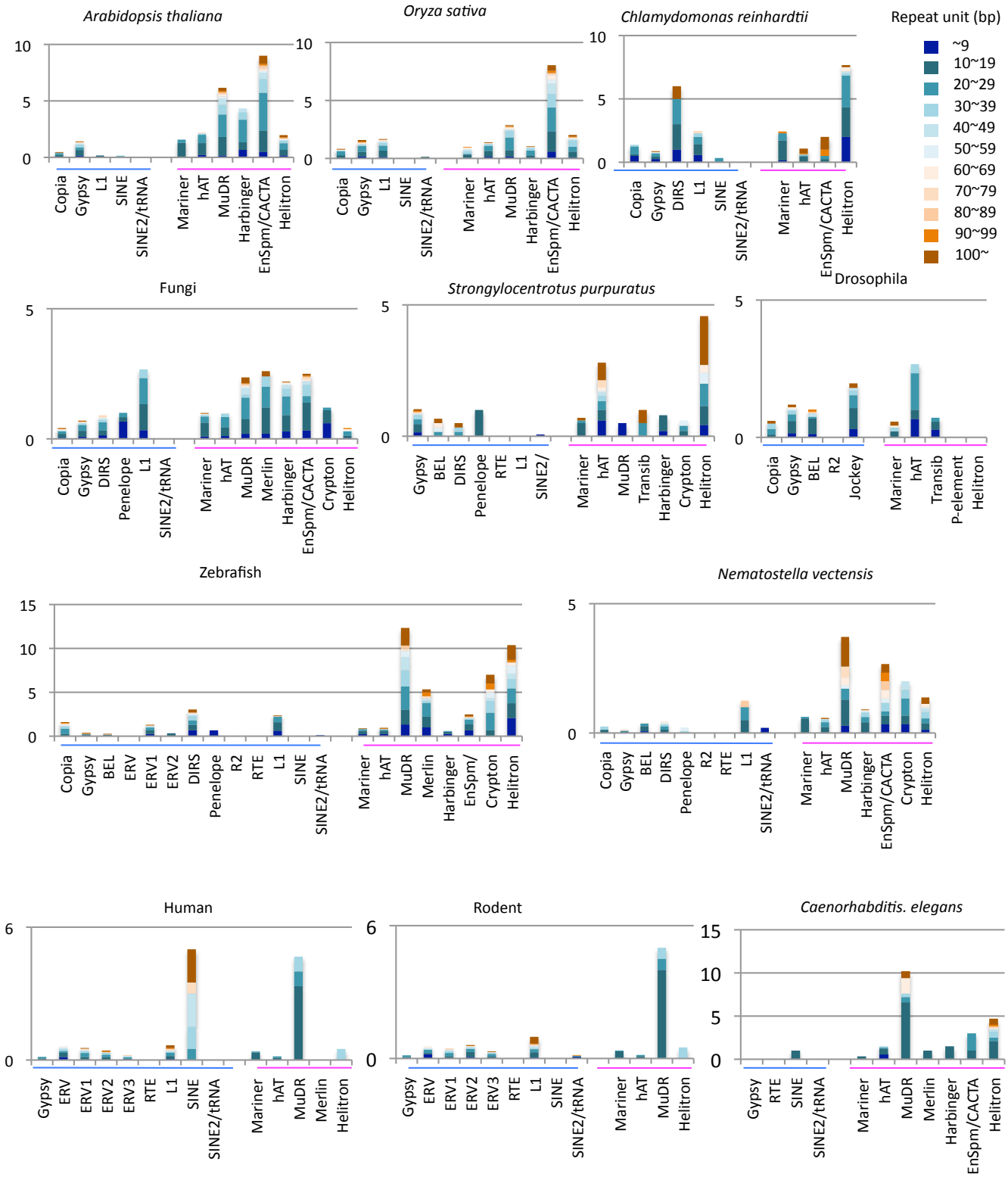


Figure 10. Tandem repeats in eukaryotic TEs

Supplemental Figure legends

Supplemental Figure 1. Sequence-specific loss of DNA methylation by VANC in multiple transgenic lines

(a, b) Scatter plots comparing effects on DNA methylation in TEs between two biological replicates of ΔAB transgenic plants on CHG sites (a) and CHH sites (b) indicated as (#1) and (#2), respectively. Significance of decrease in DNA methylation was assessed as described in Figure 2.

Supplemental Figure 2. Production of anti-VANC polyclonal antibody

(a-c) Purification of 6xHis-VANC protein from bacteria. Expression of 6xHis-VANC was examined by CBB staining (a), and western blotting with anti-6xHis antibody (b). Recombinant 6xHis-VANC was purified by nickel column(c). Black triangles indicate size of 6xHis-VANC protein.

(d, e) Validation of anti-VANC polyclonal antibody for western blot and immuniprecipitation. Schematic diagram showing the procedure of western blotting and immunoprecipitation (d). Wesern blotting of nucler extracts and immunoprecipitated samples (e). Wild type plants and ΔAB transgenic plants were used. For immunoprecipitation, either VANC immunized or unimmunized antiserum was used. VANC protein and its degradates are shown as black and white triangles, respectively.

Supplemental Figure 3. VANC preferentially binds to VANDAL21 TEs

500kb window of chromosome 2 showing the results of ChIP-seq. VANDAL21 TEs were surrounded by orange boxes.

Supplemental Figure 4. Phylogenic relationship among VANDAL21 and its closely-related families

The phylogenic tree is same with Figure 8 except IDs of transposable element (TAIR10) for A. thaliana TEs, and chromosome number and start position for *A. lyrata* TEs are noted. The bootstrap probabilities (%) with 1000 replications for major clusters are indicated beside the branches.

Supplemental Figure 5. Sequence divergence among VANDAL families in *A. thaliana*

Dot-plot (Harr-plot) among VANDAL families. Black boxes indicate ORFs of each TE. ORFs encode Transposase domain are colored in blue. Regions with 10bp exact match are shown by dots. Presence of C-type and T-type motifs are shown as green and orange dots, respectively.

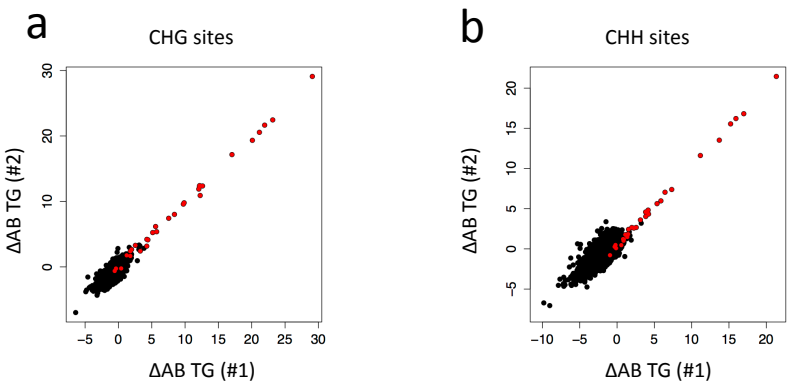
Supplemental Figure 6. integrative genome views at VANC binding regions

Integrative views showing the DNA methylation profiles and expression profiles of WT and ΔAB transgenic plants, and FLAG-VANC distribution. Bars indicate the proportion of methylated cytosines at CG sites, CHG sites and CHH sites of of WT and ΔAB transgenic plants. Expression levels of WT and ΔAB transgenic plants are shown

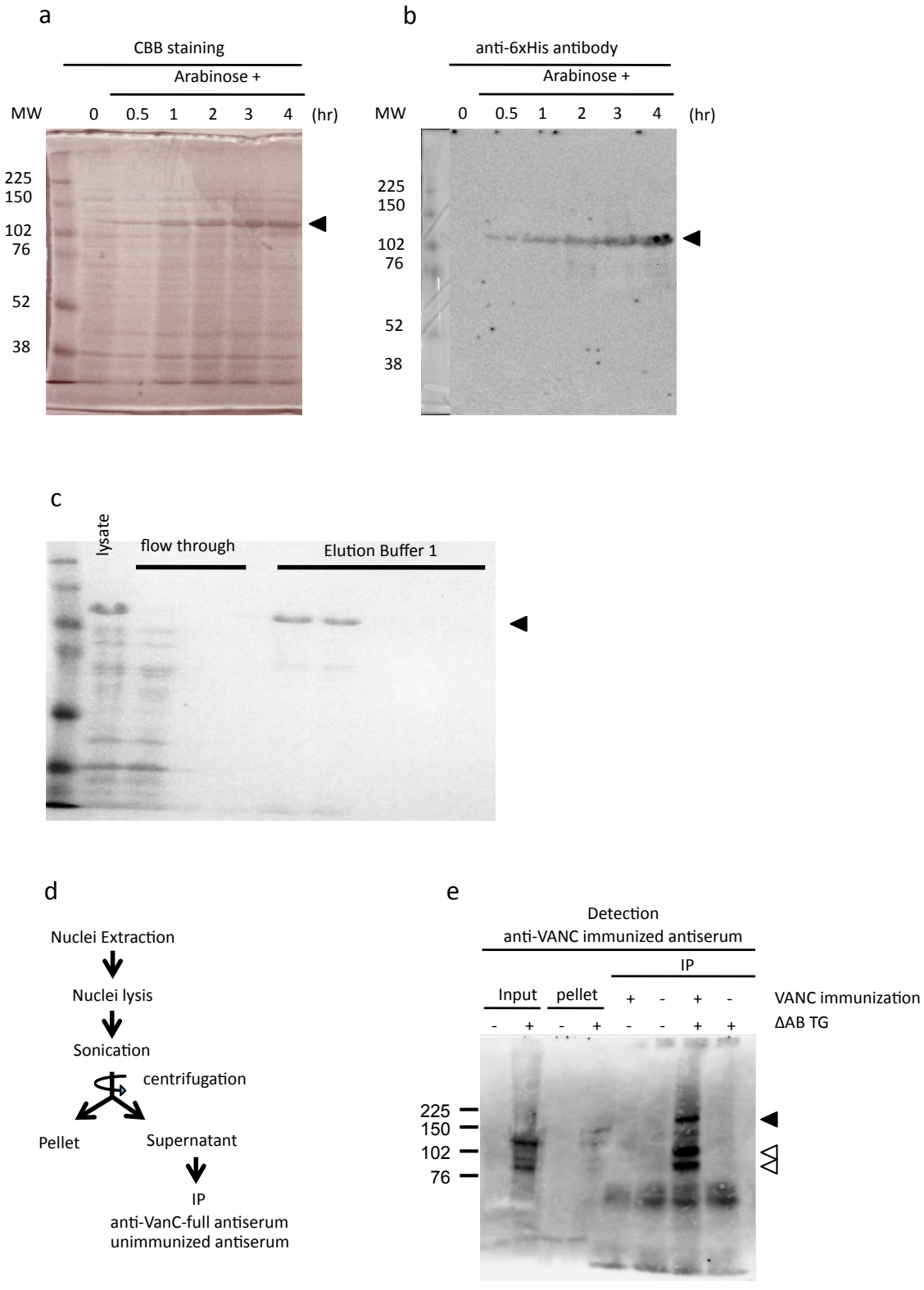
as coverage of RNA-seq signals. FLAG-VANC indicates the coverage of FLAG-VANC signals. Black boxes, grey boxes, and grey lines indicate VANDAL21 TEs, coding regions, and introns, respectively. Positions of the C-type ("YAGTATTAC") motif and the T-type ("YAGTATTAT") motif are shown as green and orange bars, respectively. Bars on top and beneath the grey line indicate that motifs exist on forward and reverse strand, respectively. Each number corresponds to the positions pointed in Figure 3

Oligonucleotide name	Sequence	Purpose
VanC C FLAG inFu F	TGACAAATAATTGGTAATATCGCGTAA	Invserse PCR for ΔAB_Hiun in pBluescript II SK (-)
VanC C FLAG inFu R	TGTAGTCTGGAACAGTATATCCGGTAT	Invserse PCR for ΔAB_Hiun in pBluescript II SK (-)
3×FLAG VanC F	CTGTTCCAGACTACAAAGACGATGACGA	Amplification of 3xFLAG sequence
3×FLAG VanC R	ACCAATTATTTGTCATCATCATCTTTAT	Amplification of 3xFLAG sequence
Van21C pBlu Sma1F	GAATTCCTGCAGCCCGAATAATCGTCTGGCCAGTCCCTT	Amplification of VANC fragment
Van21C pBlu Sma1R	ACTAGTGGATCCCCCTGTGTTATCCTATTGTTCCTTAATC	Amplification of VANC fragment
Motif-Plus F1	ACAATGAGCTTCGTATTGCTCAGTATTACCACCCTGTTCTTCTTACCGA	EMSA; The “CAGTATTAC” site is colored in red
Motif-Plus R1	CTCGGTAAGAAGAACAGGGTGGTAATACTGAGCAATACGAAGCTCATTG	EMSA; The “CAGTATTAC” site is colored in red
act2 F1	ACAATGAGCTTCGTATTGCTCCTGAAGAGCACCCTGTTCTTCTTACCGA	EMSA
act2 R1	CTCGGTAAGAAGAACAGGGTGCTCTTCAGGAGCAATACGAAGCTCATTG	EMSA
VanC seq F1	catccgaaccacctttactctt	VANC sequencing
VanC seq R1	CTCCCTCATCCTCCACAGAC	VANC sequencing
VanC seq F2	ATGCAAACTGATGAGGTGGA	VANC sequencing
VanC seq R2	CTAATACCATAGCGGATGGGA	VANC sequencing
VanC seq F3	AAGCACATCTACCACCTGCCT	VANC sequencing
VanC seq R3	GACCAAGACGCTTCATCCAACT	VANC sequencing
VanC seq F4	Ggttagtatttcctataattcc	VANC sequencing
VanC seq R4	gttttcacatgttactagac	VANC sequencing
VanC 5prime F1	CCGCCTAAGACGCGTGGAGGA	VANC sequencing
VanC 3prime R1	TGGAACAGTATATCCGGTATTAC	VANC sequencing
VanC 5prime AttB1	GGGGACAAGTTTGTACAAAAAAGCAGGCTTTCCGCCTAAGACGCGTGGAGG	Cloning VANC cDNA to pDEST vectors
VanC 3prime AttB2	GGGGACCACTTTGTACAAGAAAGCTGGGTTTATGGAACAGTATATCCGGTA	Cloning VANC cDNA to pDEST vectors

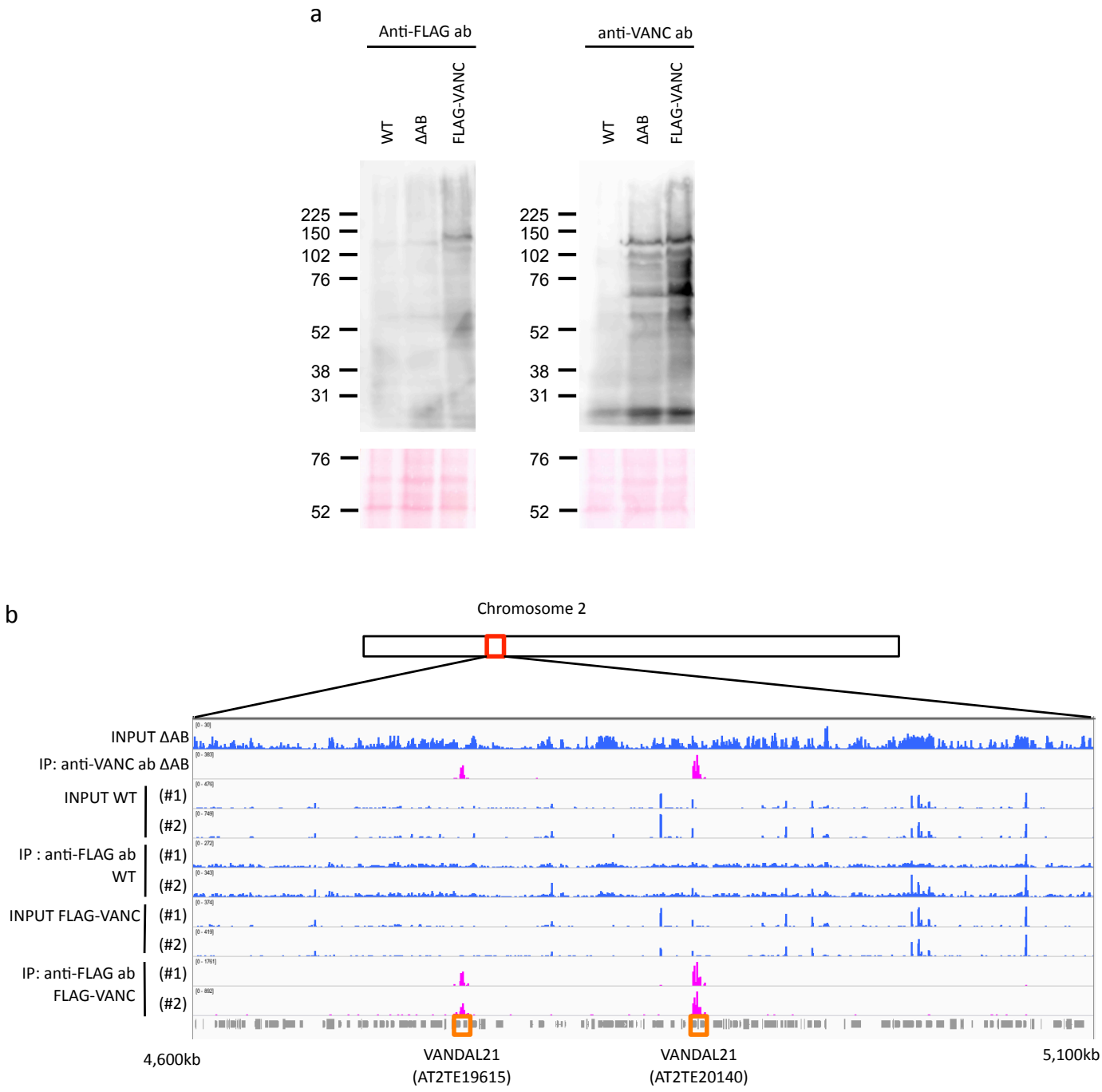
Supplemental Table 1. Oligonucleotides used in this study



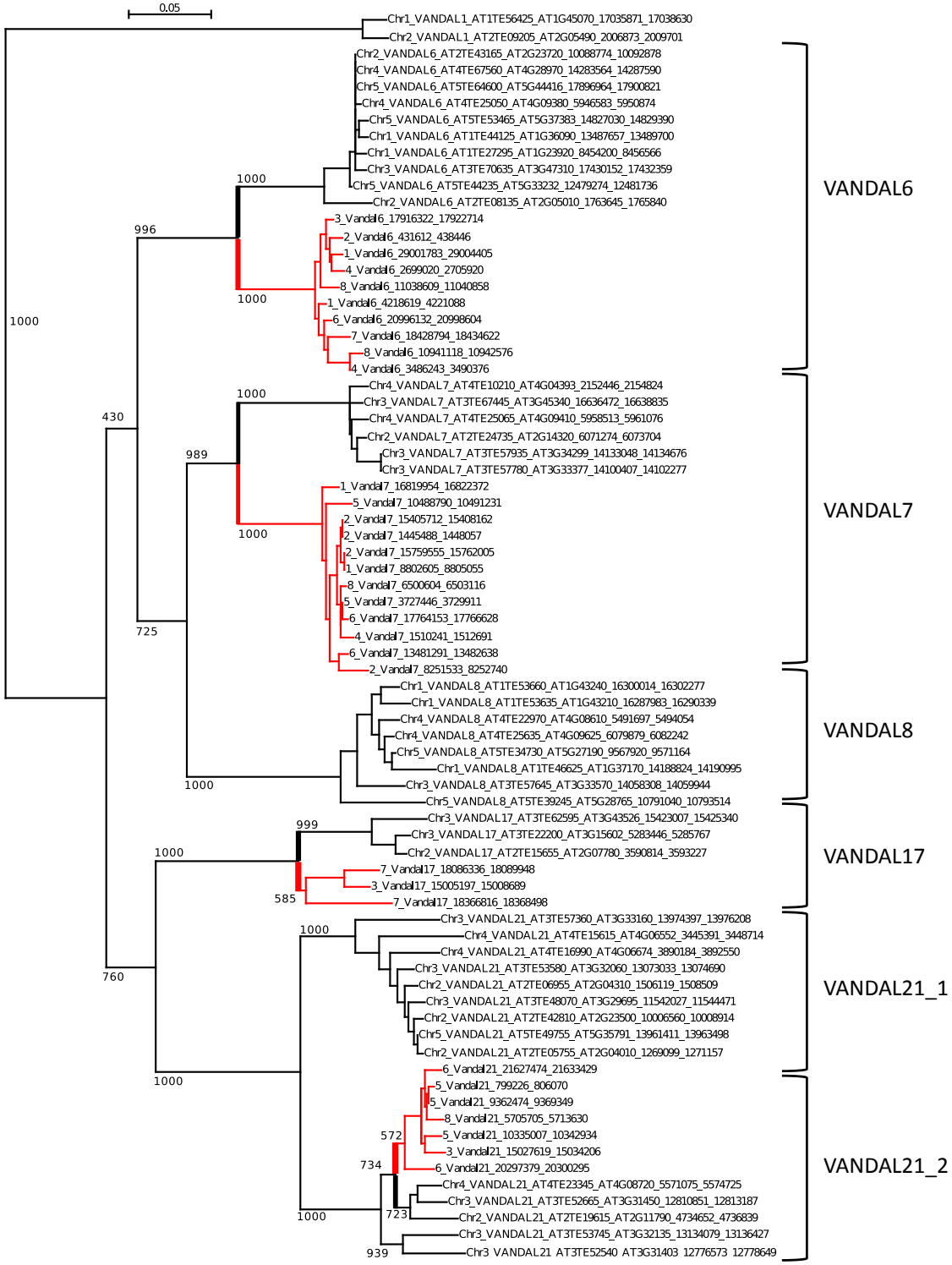
Supplemental Figure 1. Sequence-specific loss of DNA methylation by VANC in multiple transgenic lines



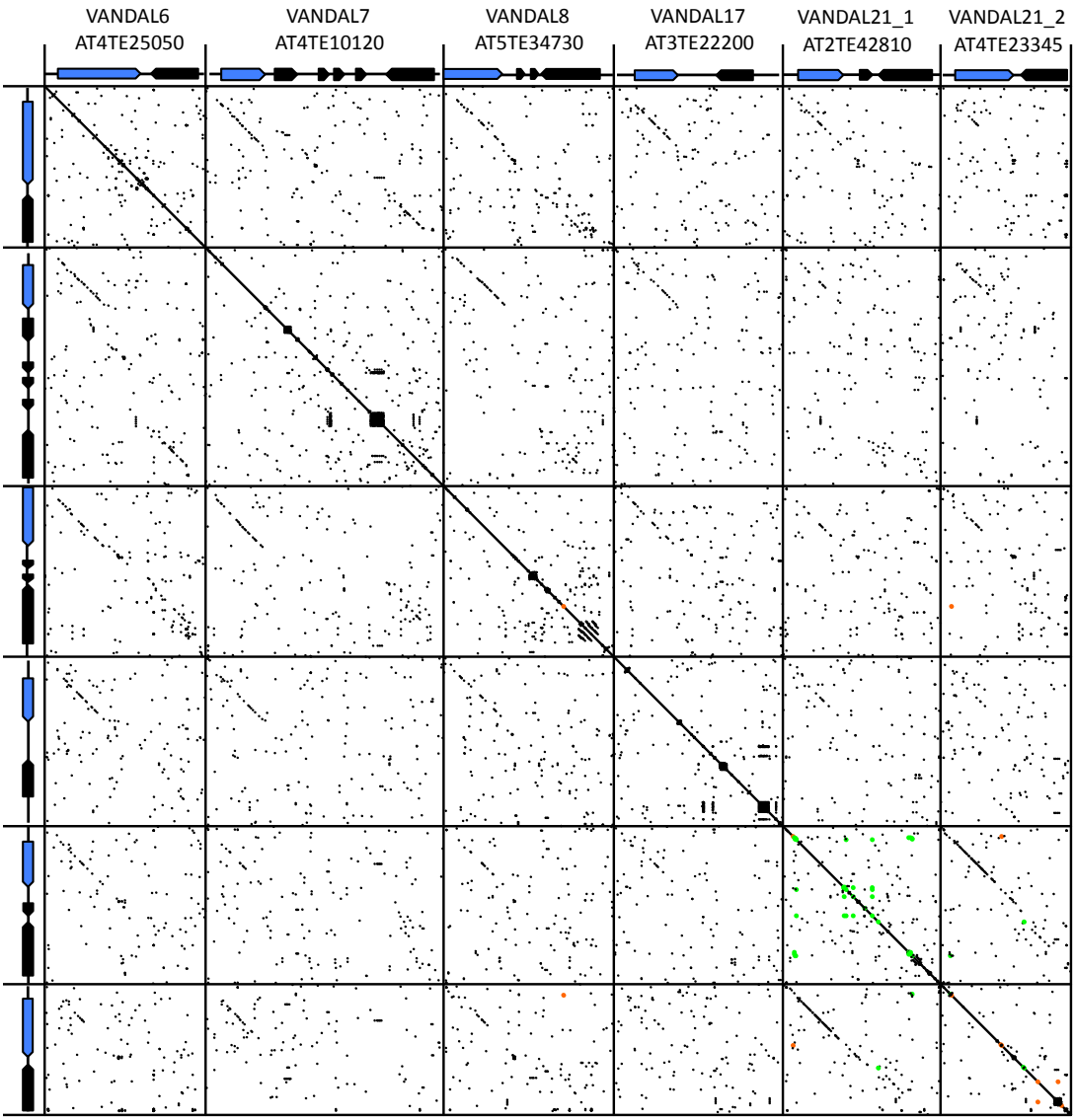
Supplemental Figure 2. Production of anti-VANC polyclonal antibody



Supplemental Figure 3. VANC preferentially binds to VANDAL21 TEs

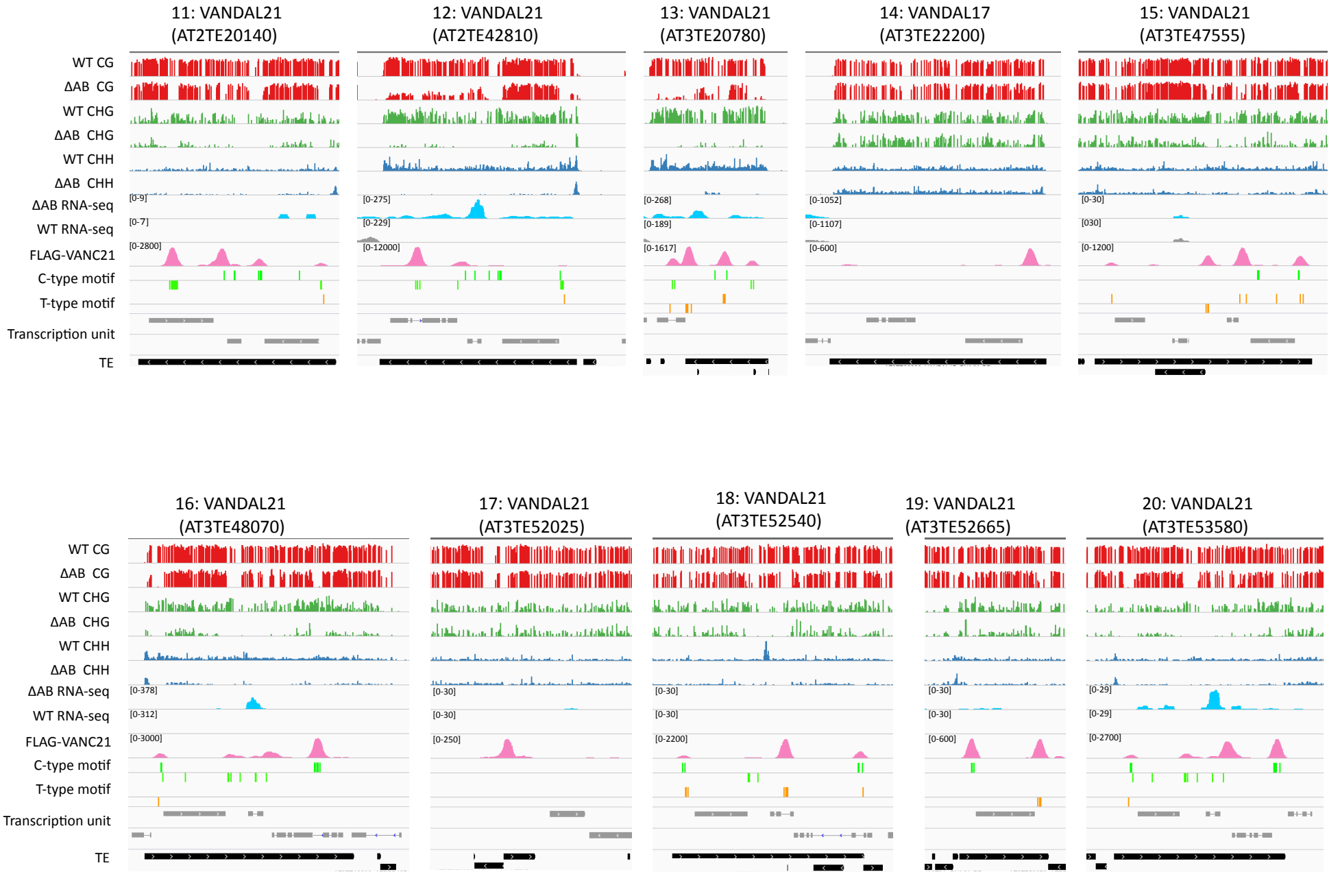


Supplemental Figure 4. Phylogenic relationship among VANDAL21 and its closely-related families

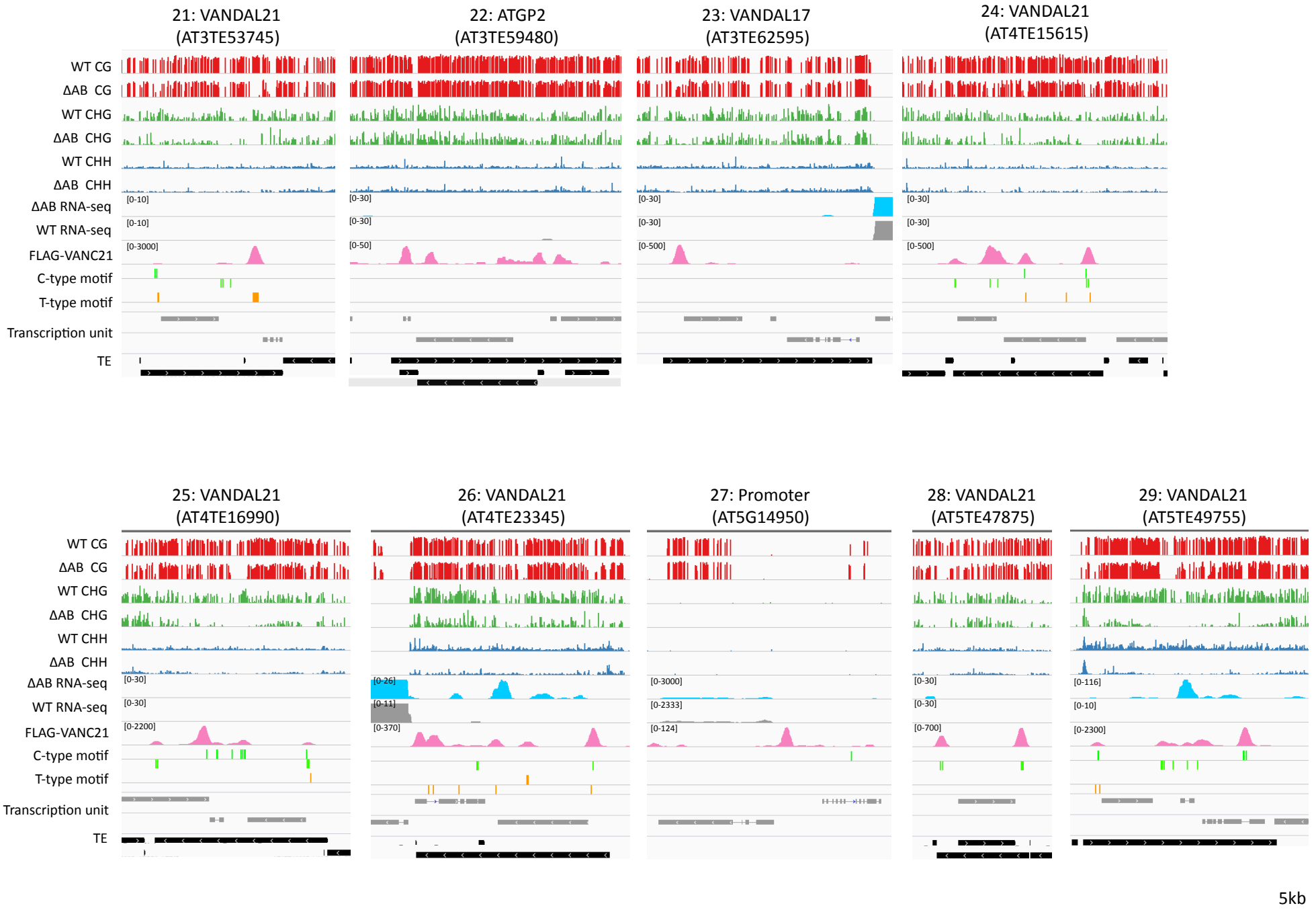


Supplemental Figure 5. Sequence divergence among VANDAL families in *A. thaliana*

Supplemental Figure 6. integrative genome views at VANC binding regions



Supplemental Figure 6. integrative genome views at VANC binding regions (continued)



Supplemental Figure 6. integrative genome views at VANC binding regions (continued)