Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*

Sayuri Tsukahara,^{1,2,6} Akira Kawabe,^{3,6,7} Akie Kobayashi,¹ Tasuku Ito,^{1,4} Tomoyuki Aizu,⁵ Tadasu Shin-i,⁵ Atsushi Toyoda,⁵ Asao Fujiyama,⁵ Yoshiaki Tarutani,^{1,2} and Tetsuji Kakutani^{1,2,4,8}

¹Department of Integrated Genetics, National Institute of Genetics, Shizuoka 411-8540, Japan; ²Department of Genetics, School of Life Science, The Graduate University for Advanced Studies (SOKENDAI), Shizuoka 411-8540, Japan; ³Department of Population Genetics, National Institute of Genetics, Shizuoka 411-8540, Japan; ⁴Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan; ⁵Center for Genetic Resource Information, National Institute of Genetics, Shizuoka 411-8540, Japan

The plant genome evolves with rapid proliferation of LTR-type retrotransposons, which is associated with their clustered accumulation in gene-poor regions, such as centromeres. Despite their major role for plant genome evolution, no mobile LTR element with targeted integration into gene-poor regions has been identified in plants. Here, we report such targeted integrations de novo. We and others have previously shown that an *ATCOPIA93* family retrotransposon in *Arabidopsis thaliana* is mobilized when the DNA methylation machinery is compromised. Although *ATCOPIA93* family elements are low copy number in the wild-type *A. thaliana* genome, high-copy-number related elements are found in the wild-type *Arabidopsis lyrata* genome, and they show centromere-specific localization. To understand the mechanisms for the clustered accumulation of the *A. lyrata* elements directly, we introduced one of them, named *Tal1* (*Transposon of Arabidopsis lyrata 1*), into *A. thaliana* by transformation. The introduced *Tal1* was retrotransposed in *A. thaliana*, and most of the retrotransposed copies were found in centromeric repeats of *A. thaliana*, suggesting targeted integration. The targeted integration is especially surprising because the centromeric repeat sequences differ considerably between *A. lyrata* and *A. thaliana*. Our results revealed unexpectedly dynamic controls for evolution of the transposon-rich heterochromatic regions.

[**Keywords:** centromere; tandem repeat; retrotransposon; evolution]

Supplemental material is available for this article.

Received November 22, 2011; revised version accepted February 21, 2012.

Retrotransposons are major factors causing rapid evolution of plant genomes; comparison of genome structures in closely related plant species has revealed that the changes in plant genome size and organization are mainly caused by rapid proliferation and deletion of LTR retrotransposons (Hawkins et al. 2006; Vitte and Bennetzen 2006; Hu et al. 2011). Rapid proliferation of transposons would be a potential threat to the function of the host genome because each integration may directly disrupt a protein-coding region or indirectly perturb transcription of nearby genes. However, most of the LTR retrotransposons in plant genomes are accumulated in gene-poor

regions, resulting in proliferation of these elements with less harmful effects to the host (SanMiguel et al. 1996, 1998; Rabinowicz et al. 1999). This feature causes a differentiation of gene-rich regions and transposon-rich regions within plant genomes. The latter constitute heterochromatic domains, which play a significant role in large-scale chromosome organization and its behavior (Dawe and Hiatt 2004; Grewal and Jia 2007).

How are the biased distributions of LTR retrotransposons generated? One possibility is through natural selection; if chromosomal domains with deleterious transposon insertions into genic regions are eliminated from natural populations, the transposons would be condensed in gene-poor regions. An alternative mechanism for the biased distribution is targeted integration. Targeted integrations of retrotransposons and retroviruses are known in yeast and mammals (Bushman 2003). In plants, however, despite the large proportion of the clustered LTR elements in their genomes, direct evidence is missing for targeted integration of LTR elements into gene-poor regions.

⁶These authors contributed equally to this work.

⁷Present address: Department of Bioresource and Environmental Sciences, Faculty of Life Sciences, Kyoto Sangyo University, Motoyama Kamigamo, Kyoto 606-8555, Japan.

⁸Corresponding author.

E-mail tkakutan@lab.nig.ac.jp.

Article published online ahead of print. Article and publication date are online at <http://www.genesdev.org/cgi/doi/10.1101/gad.183871.111>.

Several plant LTR retrotransposons have been shown to be mobile, such as tobacco *Tnt1* and *Tto1*, rice *Tos17*, and lotus *LORE1*, but all of them retrotranspose into genic regions, and no preferential integration into gene-poor regions has been reported for these plant retrotransposons (Hirochika et al. 1996; Okamoto and Hirochika 2000; Courtial et al. 2001; Yamazaki et al. 2001; Madsen et al. 2005; Le et al. 2007; Hou et al. 2010).

Precise transposon sequences throughout the genome of *Arabidopsis thaliana* make it an ideal model organism to analyze the behavior of transposons at the sequence level (The *Arabidopsis* Genome Initiative 2000; Le et al. 2000; Peterson-Burch et al. 2004). The genome of *A. thaliana*, as many other plant species, contains large clusters of LTR retrotransposons around centromeres. They are localized in regions flanking the cores of centromeres, which are composed of tandemly arrayed satellite repeats of 178 base pairs (bp) (Heslop-Harrison et al. 2003). The most abundant and most well-characterized subfamily of the pericentromeric LTR retrotransposons of *A. thaliana* is called *Athila* (Pelissier et al. 1995; Steimer et al. 2000; Peterson-Burch et al. 2004; Slotkin 2010). However, no mobile *Athila* has been identified so far.

Using *A. thaliana*, we and others recently reported identification of multiple families of mobile LTR retrotransposons (Mirouze et al. 2009; Tsukahara et al. 2009). These retrotransposons—*ATGP3*, *ATCOPIA13*, *ATCOPIA21*, and *ATCOPIA93* (also called *Evade*)—are silent in the wild type, but are activated when their DNA methylation is abolished in mutants of a chromatin remodeling factor gene, *DDM1* (decrease in DNA methylation 1) (Vongs et al. 1993; Jeddeloh et al. 1999). Unlike *Athila*, these mobile elements are low copy number in the genome of *A. thaliana*. In addition, examination of integration specificity for two of them, *ATCOPIA93* and *ATGP3*, revealed that they did not show targeted integration into gene-poor regions (Mirouze et al. 2009; Tsukahara et al. 2009). Interestingly, however, a closely related *Arabidopsis* species, *Arabidopsis lyrata*, has high-copy-number retrotransposons related to *ATCOPIA93*, and their sequence similarity suggests recent proliferation (Tsukahara et al. 2009). Furthermore, they are localized within centromeric repeats. Thus, we suspected that these copies in *A. lyrata*

might have the ability to retrotranspose preferentially into the centromeres.

In this study, we directly examined the mobility of one of the *ATCOPIA93*-related elements in *A. lyrata*, which we named *Tal1* (*Transposon of Arabidopsis lyrata 1*). *Tal1* was mobile in the genome of *A. thaliana* and showed targeted integration into centromeric repeats, demonstrating the first example of targeted integration of a plant retrotransposon into gene-poor regions. Direct characterization of integration specificities of these mobile elements is revealing unexpectedly dynamic controls for evolution of the heterochromatin and repeats.

Results

Tal1 was mobile in transgenic *A. thaliana* plants

We showed previously that multiple copies of *A. lyrata* elements related to *ATCOPIA93* are localized in centromeric repeats (Tsukahara et al. 2009). On the other hand, the mobile *ATCOPIA93* copy (At5g17125) is localized in the arm region of the *A. thaliana* genome; and this element frequently retrotransposes into chromosomal arm regions (Mirouze et al. 2009). The apparent difference in genomic distributions of these *COPIA93*-like elements in *A. thaliana* and *A. lyrata* could be due to differences in host environments; for example, centromeric satellite sequences are ~30% different between these two species (Kawabe and Nasuda 2005). Alternatively, the difference may reflect differences in the properties of the transposons. To test these possibilities directly, we introduced one of the *A. lyrata* copies into the *A. thaliana* genome by transformation.

We first tried to select a mobile *A. lyrata* copy. The *A. lyrata* genome contains at least 16 copies of elements similar to *ATCOPIA93*. The identities of the 5' and 3' LTRs of these elements are 99.2%–100%, suggesting that their proliferations are recent. For transformation, we chose an *A. lyrata* element with 100% identity of the 5' and 3' LTRs and with an ORF having the potential to encode functional proteins. That copy, which we named *Tal1*, was amplified by PCR from the *A. lyrata* genome, cloned into the pRI909 binary vector (Fig. 1A), and introduced into *A. thaliana* by *Agrobacterium*-mediated transformation. Because *Tal1* may be silenced by DNA

Figure 1. Activity of *Tal1* in transgenic *A. thaliana* plants. (A) The structure of T-DNA around *Tal1* in pRI909 binary vector. Black boxes indicate 5' and 3' LTRs of *Tal1*. (B) Transcription of *Tal1* examined by RT-PCR. The region amplified by PCR is shown with a black line with two arrowheads in A. T1 transgenic plants were used. GAPC was used as a positive control. Lack of *Tal1* transcript in the original host *A. lyrata* is associated with their high level of DNA methylation (not shown). (C) Southern blot analysis using undigested genomic DNA of independent lines with the *Tal1* transgene (T2 generation). The top bands represent undigested chromosomal DNA. The bottom bands correspond to the length of extrachromosomal linear DNA of *Tal1* (black arrowhead). The probe is shown in A with a gray bar. (NT) Plant not transformed; (V) plant transformed with vector.

methylation, we used both wild-type and *ddm1* mutant plants for the host. The *ddm1* mutation is known to release transcriptional silencing of diverse methylated transposons (Hirochika et al. 2000; Miura et al. 2001; Singer et al. 2001; Lippman et al. 2004; Mirouze et al. 2009; Tsukahara et al. 2009).

In transgenic plants of both wild type and *ddm1*, we could detect the transcript of *Tal1* (Fig. 1B). Generally, during retrotransposition of an LTR element, RNA for the element is reverse-transcribed to produce linear extrachromosomal DNA, which subsequently integrates into the host genome. We examined the presence of the extrachromosomal DNA in the transgenic plant lines by Southern blotting of genomic DNA not digested by restriction enzymes. In every transgenic line examined, we could detect the extrachromosomal *Tal1* DNA of the expected length (Fig. 1C).

We next examined *Tal1* retrotransposition by Southern analyses using genomic DNA from multiple transgenic lines. After digestion by EcoRV, membranes were first hybridized with a probe for the T-DNA vector region flanking *Tal1*. By hybridization to that probe, a common

signal of the expected length was detected in each transgenic line (Fig. 2B). When the same membranes were re-hybridized with a probe for *Tal1*, additional signals were detected (Fig. 2B), suggesting the presence of *Tal1* DNA in these transgenic plants outside of the original copy introduced by T-DNA.

The signal specifically seen by the *Tal1* probe had a common pattern among these transgenic lines; the common strong signals were observed around 2.0-kb and high-molecular-weight positions (>19 kb) (Fig. 2B). The signals observed around 2.0 kb match to the expected size of the linear extrachromosomal DNA digested by EcoRV (Fig. 2B, arrowheads). The other strong signals with high molecular weight imply that *Tal1* is located within DNA sequences apart from EcoRV sites. We also used another restriction enzyme, HindIII, and compared the patterns of signals (Fig. 3). In addition to signals expected for extrachromosomal DNA digested by HindIII (653 bp and 498 bp), multiple signals were detected in low-molecular-weight regions (500~800 bp), suggesting *de novo* *Tal1* insertion near HindIII sites. Taken together, these results of Southern analyses suggest that a common feature of *Tal1*

Figure 2. Retrotranspositions of *Tal1* in transgenic plants. (A) EcoRV sites (E) and probe positions (gray bars) in the T-DNA region. (B) Six independent transgenic lines (four wild-type lines and two *ddm1* lines) were characterized. Their family numbers correspond to the lane numbers in Figure 1C (the T2 plants were siblings of the same family). DNA from T1 and T2 plants were digested by EcoRV, which is insensitive to DNA methylation, and were hybridized with a probe for T-DNA (probe b) and, subsequently, with a probe for *Tal1* (probe c). The position expected for the extrachromosomal DNA fragment is indicated by a black arrowhead. The *Tal1* probe does not hybridize to endogenous *COPIA93* family retrotransposons in *A. thaliana*. The additional band seen by probe b in family 3 is likely to reflect partial or rearranged T-DNA integration. In each family, 14 individuals of T2 plants were checked by PCR for the presence of T-DNA, and eight of the plants were used in the Southern blot. For families 1, 8, and 9, we could find T2 plants without T-DNA (T-DNA-), and two of them were included in each family. Note that some of them have signals for the extrachromosomal DNA, suggesting that newly integrated copies have activity for retrotransposition. The lengths of the molecular weight markers are 19.33 kb, 3.47 kb, and 1.88 kb. (N) Plant not transformed.

Figure 3. Retrotranspositions of *Tal1* near *HindIII* sites. (A) *HindIII* sites (H) and probe positions (gray bars). (B) T2 plants, siblings of the transgenic lines used in Figure 1, were examined first with probe e for the vector and subsequently with probe d for *Tal1*. The position expected for the extrachromosomal DNA fragment is indicated by a black arrowhead.

integration sites in these lines is loci near *HindIII* sites and apart from *EcoRV* sites. We suspected that these might indicate insertions of *Tal1* in centromeric tandem arrays, which normally includes *HindIII* sites, but rarely includes *EcoRV* sites.

Tal1 preferentially integrates into 178-bp centromeric repeats

To directly examine *Tal1* insertion site specificity, we performed whole-genome resequencing of the transgenic plants (two wild-type and two *ddm1* lines in the T2 generation, one individual plant for each line) using an Illumina Genome Analyzer IIx sequencer. Eighty million to 110 million reads (read length: 110 bp) were generated for each line, covering $>50\times$ of the genomes. In order to detect flanking sequences of *Tal1*, we first screened reads containing 30 bp of sequence for the 5' or 3' end of *Tal1*. From these reads, we then subtracted hits for the T-DNA vector, which correspond to the original transgene, and internal regions of *Tal1* flanking the LTRs, which reflect duplication of the LTR sequence. Remaining reads containing >20 nucleotides (nt) were used for identifying new integration sites of *Tal1*. The numbers of such reads were 46 and 758 for two lines of wild type, and 1471 and 2039 for two lines of *ddm1* mutants.

We then examined how the flanking sequence of each read matches the consensus sequence of the centromeric satellite repeats. For each read, the proportion of the identities of the flanking 20 nt was examined along

the 178-nt centromeric satellite unit, and the maximum value was used for further analysis. Distributions of such proportions were generally bimodal (Fig. 4; Supplemental Fig. S1), reflecting sequences within the satellites and other regions. For randomly chosen genomic reads (Fig. 4B,D,F), $\sim 8\%$ corresponded to the satellites, which is consistent with previous estimates for the amount of centromeric satellites (The *Arabidopsis* Genome Initiative 2000; Hosouchi et al. 2002).

When reads flanking the new *Tal1* integrations were analyzed for the two wild-type lines, 43 out of 46 (93%) and 721 out of 758 (95%) reads belong to the population for the satellites (Fig. 4A; Supplemental Fig. S1). Thus the new *Tal1* integrations were biased strongly toward the satellites. In addition, many of the other reads (one out of three and 36 out of 37) represent nested integration into another *Tal1* (Supplemental Fig. S2). The strong bias toward the satellite was also found in the *ddm1* background. In *ddm1* lines, 1394 out of 1471 (95%) and 1794 out of 2039 (88%) reads correspond to the integrations into the satellite (Fig. 4C; Supplemental Fig. S1), and nested integration into another *Tal1* was also found (17 out of 77 and 39 out of 245) (Supplemental Fig. S2). Integration sites of *Tal1* were distributed throughout the satellite unit in both wild type and *ddm1* (Fig. 5A; Supplemental Table S1), although local hot and cold spots for integration exist (Fig. 5B; Supplemental Table S2).

We showed previously that *ATCOPIA93* of *A. thaliana*, which is related to *Tal1* of *A. lyrata*, proliferates in the *ddm1* background. We then examined the integration sites of *ATCOPIA93* in five independent *ddm1* lines. The distribution of new integration of *ATCOPIA93* was not biased toward the satellites (Fig. 4E; Supplemental Fig. S1). When compared with random reads from the same individual, the *ATCOPIA93* integration sites were even biased toward the opposite direction; their integrations into the satellites were 11 out of 634 (2%), 95 out of 2823 (3%), one out of 249 (0.4%), 182 out of 5990 (3%), and 104 out of 4644 (2%), which are significantly lower than those of random reads (Fig. 4E,F; Supplemental Fig. S1). This observation is consistent with a previous report showing frequent integrations of *ATCOPIA93* into chromosomal arm regions (Mirouze et al. 2009). Even though *ATCOPIA93* and *Tal1* are highly related, their targeting specificities differ considerably.

Sequences of *COPIA93/20* of *A. lyrata* suggest that their recent burst of integrations into centromeres occurred in independent subfamilies

The results above suggest that *Tal1* and the mobile *ATCOPIA93* (At5g17125) have different integration site specificities; the former specifically integrates into centromeric repeats, while the latter do not. In order to know which is the ancestral property, we extended the sequence analyses to similar elements in the genomes of *A. thaliana* and *A. lyrata*. Sequence comparison of *A. lyrata* copies (shown by red lines in Fig. 6) suggests that proliferation of these copies occurred in four clusters of sequences independently. On the other hand, only five

the results with additional plants: *Tal1* insertions in four transgenic lines (two *ddm1* and two wild type), and *ATCOPIA93* insertions in five independently self-pollinated *ddm1* lines. In each panel, the vertical bar represents 300 reads. The consensus sequence of the centromeric satellite used is AASCTTCTTCTTGCTTCTCAAAGCTTTSATGGTGTAGCCAAAGTCCRTATGAGTCTTTGKCT TTGTRTCTTCTAACAAGGAWACAMTACTTAGGCTTTTARGATCVGGTTGCGGTTYWAGTTCTTATACTCAATCATAACACA TGACATCWAGTCATATTYGACTCCAAAACACTAACC, with V representing A, C, or G; W representing A or T; R representing A or G; Y representing C or T; K representing G or T; M representing A or C; and S representing C or G. Integration sites other than centromeric satellites are shown in Supplemental Table S1.

copies of *A. thaliana* elements were found. We showed previously that *A. lyrata* elements related to mobile *ATCOPIA93* (cluster 1 in Fig. 6, which includes *Tal1*) are localized in centromeric repeats (Tsukahara et al. 2009). We then examined whether *A. lyrata* copies in other clusters (clusters 2, 3, and 4) are localized in centromeres.

In total, 288 unique *A. lyrata* loci of *COPIA93/20* were characterized. Among them, 272 have internal regions, and 16 are solo LTRs. In all of the four clusters, most copies are flanked by centromeric satellite sequences (72% in total) (Fig. 7). About 4% of loci are flanked by unknown sequences, but centromeric satellite sequences exist within 500 bp. In addition, most of the integrations into non-satellite regions are flanked by other transposons (Fig. 7). The majority of these transposons belong to *COPIA93/20* families or *Athila*, which are likely to be localized in pericentromeric or centromeric regions (Pelissier et al. 1995). Among the 288 loci examined, only six loci had flanking sequence with homology with genic and intergenic regions of *A. thaliana*. Almost all (95%, even if flanking sequences without detectable homology were excluded) integrations are likely to be localized within centromeric or pericentromeric nongenic regions. These trends of integration site bias were found in all four clusters of *A. lyrata* elements (Fig. 7). These results suggest that the preferential accumulation in centromeric satellites is the ancestral property of these elements before separation of the mobile *ATCOPIA93* and *Tal1*.

Discussion

Centromere-specific targeted integration of *Tal1*

In this study, we showed that *Tal1* integrated almost exclusively into centromeric repeats of *A. thaliana*. One could argue that even though the de novo *Tal1* integrations were found in centromeres, the possibility remains

that it is due to natural selection; cells or individuals with *Tal1* insertion into gene-rich regions may be immediately eliminated by natural selection. We think that possibility is unlikely for multiple reasons. First, centromeres are

Figure 5. Distribution of *Tal1* integration sites within the 178-bp centromeric repeat of *A. thaliana*. (A) Positions of *Tal1* integrations. The 178 positions are those in the sequences shown in the legend for Figure 4. The top and bottom bars indicate insertions with 5'-to-3' and 3'-to-5' directions of *Tal1*, respectively. The target site duplication of *Tal1* was normally 5 bases long, and the third nucleotide from the integration site is shown because that site would be the same for the 5' and 3' flanking sequences. The number of reads for each site is shown in Supplemental Table S2. (B) The 178 sites were divided into four classes depending on the presence or absence of *Tal1* insertion in forward (5'-to-3') and reverse (3'-to-5') orientations in two *ddm1* transgenic lines in A. The numbers in parentheses represent the values expected if insertions in forward and reverse orientations are independent. The skew ($P < 10^{-12}$) suggests positive association and the presence of local hot and cold spots for *Tal1* insertions within the satellite unit.

Figure 6. Phylogenetic relationship of *COPIA93/20* elements of *A. thaliana* and *A. lyrata*. The phylogenetic trees were made based on the nucleotide sequences of integrase (INT) and reverse transcriptase (RT) domains. *A. lyrata* copies are shown by red lines, and *A. thaliana* copies are shown by black lines. At2g07420 has been annotated as *ATCOPIA20*, and four other *A. thaliana* copies have been annotated as *ATCOPIA93*. Each of the *A. lyrata* clusters includes ~10–60 copies with an average synonymous divergence of 2%–10%, suggesting recent bursts. As >100 shotgun reads were sometimes found with identical sequences for the integrase and reverse transcriptase domains (~600 bp), some of them may represent multiple copies with identical sequences, reflecting even more recent bursts. Phylogenetic trees were constructed by the neighbor-joining method using distances corrected by the Jukes-Cantor method. Bootstrap probabilities were estimated by 500 replicates and are shown only in major branches. The number of sequences and the average number of nucleotide differences (both synonymous and replacement sites) in each cluster are shown at the right of the cluster.

not the only gene-poor regions in the genome; if exclusion from gene-rich regions is the main force to cause the biased distribution, *Tal1* would also accumulate in other gene-poor regions, such as ribosomal DNA and intergenic regions, especially in the early phase of the selection. That was not the case (Fig. 4; Supplemental Fig S1; Supplemental Table S1). Second, we could not detect any sign of elimination of cells or individuals in the *Tal1* transgenic plants. The transgenic lines were morphologically normal even after multiple rounds of self-pollinations where *Tal1* continues to be expressed (Supplemental Fig. S3), and their fertility was indistinguishable from that of the control plants. Finally, *ATCOPIA93*, a copy closely related to *Tal1*, did not show biased accumulation in

centromeres; many of the insertions were found in genic regions (Fig. 4E; Supplemental Fig. S1). It is hard to speculate that the deleterious effects of genic insertion of *ATCOPIA93* are much weaker than those of *Tal1*, because their structures are similar (Fig. 6). The accumulation of *Tal1* in centromeres is likely due to targeted integration, rather than natural selection.

What feature in centromeres does *Tal1* recognize? The centromeric satellite sequences are known to evolve fast (Mallik and Henikoff 2002). Although *A. lyrata* and *A. thaliana* are closely related species, divergence of the satellite sequences is ~30% (Kawabe and Nasuda 2005). Still, *Tal1* targeted to centromeres of *A. thaliana*. Instead of the primary sequence, the higher-order chromatin configuration, unique to centromeres, may be recognized by *Tal1*. Centromeric and pericentromeric regions tend to have a condensed structure, which is marked by DNA methylation and dimethylation of histone H3 Lys 9, and the *ddm1* mutation abolishes the condensed structure and the modifications (Gendrel et al. 2002; Soppe et al. 2002; Lippman et al. 2004). However, even in the *ddm1* mutant, most of the insertions of *Tal1* were found in centromeric repeats, suggesting that marks of centromeres

Figure 7. Genomic localization of *A. lyrata* *COPIA93/20* elements. Proportions of inserted region categories are shown for the total and each of four clusters (C1–C4). pAa, pAge1, and pAge2 are centromeric satellite sequences in *A. lyrata*. Integrations into the *COPIA93/20* family were separated from those into the other transposable element (TE) families. If centromeric satellites (either pAa, pAge1, or pAge2) or TE sequences were found within 500 bp from the insertion site, the loci were categorized as “Near” them.

recognized by *Tal1* do not depend on *DDM1* function. Centromeres are also marked by specific proteins (Kanizay and Dawe 2009), which could be more likely candidates for the target of *Tal1* integration.

Evolutionary dynamics of COPIA93/20 elements and centromeric satellites in A. lyrata and A. thaliana

A notable feature suggested from the sequence data of *COPIA93/20* elements shown in Figure 6 is that their bursts occurred in four independent lineages in *A. lyrata*, while their possible orthologous copies in *A. thaliana* remained low copy. The *lyrata*-specific bursts are also found in other subfamilies of LTR elements (Hu et al. 2011), consistent with the theoretical prediction that active transposons propagate more easily in outcrossing *A. lyrata* than in self-pollinating *A. thaliana* (Hickey 1982).

The phylogenetic analysis also suggests that targeting to centromeres is likely an ancestral property of *A. lyrata* elements, which is lost in the mobile *ATCOPIA93* (At5g17125). Interestingly, however, none of the other four copies of *A. thaliana* elements shown in Figure 6 (At1g34967, At1g43775, At4g04410, and At2g07420) localize in centromeric satellites either. Considering that, it is especially surprising that *Tal1* has an ability to target to centromeric satellite sequences of *A. thaliana* (this study), even though no copy is found to localize in satellite sequences in the natural wild-type *A. thaliana*.

It is enigmatic that no *COPIA93/20* elements are localized in centromeric satellites in *A. thaliana*. One possible explanation is that *A. thaliana* has lost the typical ancestral *COPIA93/20* copies within centromeric satellites. A transposon insertion within the satellite repeat could be lost by unequal crossing over. In addition, as the karyotype of *A. thaliana* is different from those in other related species, its centromeric sequences seem to have been replaced by the current form in the recent past (Kawabe et al. 2006). During such replacement, *COPIA93/20* localized in the centromeric satellites may have been lost, and only copies outside the satellites may have been transmitted.

In any case, the evolution of integration specificity of the retroelements plays a key role in heterochromatin dynamics. Generally, targeted integration of a retroelement depends on recognition of the target site by its integrase (Sandmeyer 1990; Bushman 2003; Gao et al. 2008). Amino acid sequences of integrases of *Tal1* and *ATCOPIA93* are >90% identical (Supplemental Fig. S4). If the difference in the targeting specificities is due to their structural difference, the responsible motifs may be clarified by examination of targeting specificity after generating retrotransposon constructs with chimeric integrases.

Perspectives

Accumulation of LTR retrotransposons around centromeric satellite repeats is a general feature found in many plant species. For example, LTR retrotransposons such as maize CRM or rice CRR are localized around centromeres and have been extensively investigated on the levels of sequence, chromosomal location, and evolution

(Cheng et al. 2002; Nagaki et al. 2005; Ma and Jackson 2006; Du et al. 2010). However, no mobile copy of the centromeric LTR element had been reported to date. *Tal1* might provide an experimental system for understanding the effects of retrotransposon insertion de novo for the evolution of centromeres.

Retroelements have been widely used for transformation vectors in human gene therapies. Especially useful vectors are those with specific integration targets (Bushman 2003). However, as far as we know, no mobile element targeting to centromeres has previously been identified in any organism. *Tal1* might provide a useful vector to engineer chromosome organization, for example, by introducing loxP sites into centromeres to shuffle chromosome arms by CRE recombinase or introducing a telomere construct (Yu et al. 2007) to make telocentric chromosomes.

Materials and methods

Plant materials

The wild-type and *ddm1* mutant of *A. thaliana* plants used were in a Col background. The mutant allele *ddm1-1* (Vongs et al. 1993) was used throughout.

Construction of the pRI909 vector containing Tal1 and transformation

A PstI-SacI fragment of the 5' half of *Tal1* and a SacI-EcoRI fragment of the 3' half of *Tal1* were amplified from genomic DNA of *A. lyrata* by nested PCR using the primers shown in Supplemental Table S3. These fragments were cloned into the PstI-EcoRI site of the pRI909 vector (Takara) by Mighty Mix (Takara). *A. thaliana* wild type and *ddm1* mutants were transformed by the standard floral dip method (Clough and Bent 1998) using a pRI909 vector containing *Tal1*. The T-DNAs transferred to *A. thaliana* carried a neomycin phosphotransferase gene (NPTII) that was used for the selection marker.

Southern analysis

Genomic DNA of *A. thaliana* was extracted from mature leaves using Nucleon PhytoPure genomic DNA extraction kits (GE Healthcare). Probes for *Tal1* were amplified using the primers described in Supplemental Table S3.

RT-PCR

For RT-PCR, total RNA was extracted from leaf tissue using the SV Total RNA Isolation system (Promega) and was treated with DNase I (Promega). From 2 µg of total RNA, cDNA was synthesized using the First Strand cDNA Synthesis kit (GE Healthcare) and a pd(N)6 primer. A one-tenth portion of the RT reaction was used as a template for PCR (total, 10 µL). PCR conditions were 3 min at 94°C; 25–33 cycles of 30 sec at 94°C, 30 sec at 55°C, and 60 sec at 72°C; and 3 min at 72°C. The PCR product was then separated by electrophoresis on a 1.5% agarose gel. Primer pairs for RT-PCR are shown in Supplemental Table S3.

Whole-genome resequencing

Tal1 integration sites were analyzed in two wild-type and two *ddm1* lines, which are siblings of plants in lanes 2, 6, 8, and 9 in

Figure 1C. The results of the four lines are included in Supplemental Figure S1, but Figure 4 includes only the results for two of them, corresponding to lanes 6 and 9. Whole-genome resequencing libraries (insert size: 200–350 bp) were prepared using a Paired End DNA Sample Prep kit (Illumina). The quality and quantity of DNA samples were measured using the Quant-iT PicoGreen dsDNA Assay kit (Life Technologies), Agilent 2100 Bioanalyzer, and Power SYBR Green PCR Master mix (Applied Biosystems). The genomic libraries were used to generate clusters on the Illumina cBot using the Paired End Cluster Generation kit (Illumina) and were sequenced on the Illumina Genome Analyzer IIx and HiSeq2000 sequencers with 111 and 101 cycles, respectively. Sequences and quality scores passing through the standard Illumina pipeline filters were retained for further analysis. Raw sequence data for the transgenic plants were deposited in the DDBJ (DNA Data Bank of Japan) Sequence Read Archive (DRA; accession nos. DRA000394, 000425–000427). Flanking sequences of transposons were identified by simply extracting reads containing terminal regions of transposons, using each of the paired end sequences independently. When 20 nt of *Tal1* flanking sequence was analyzed, a sequence with $\geq 80\%$ match to the centromeric satellite consensus was categorized as flanked by the satellite. When the 20 nt of flanking sequence matched 100% to a part of the *Tal1* sequence, it was classified as a nested integration into another *Tal1*. The integration site specificity of *Tal1* was also examined by a suppression PCR technique (Tsukahara et al. 2009). By that method, 30 out of 33 detected *Tal1* flanking sequences were found to match the centromeric satellites (data not shown), consistent with the genome resequencing results.

Analysis of *A. lyrata* genome sequence data

Transposon sequences were obtained from *A. lyrata* shotgun genome sequence data. Integrase or reverse transcriptase domain sequences were used as queries to a BLAST search at the NCBI Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi>). The obtained shotgun reads were aligned, and sequences with two or more identical reads were retained. From these conserved domain sequences, a sequence walk was done to extend alignment to the end of the transposon sequence. The LTR sequence was defined by comparing flanking sequences of different loci sequences to determine both 5' and 3' ends. The LTR sequences were used for queries to obtain other loci simultaneously. Sequences flanked to LTR regions were checked by the NCBI BLAST homepage restricted to *Arabidopsis* sequence entries.

Acknowledgments

We thank Akiko Terui for technical assistance, and Yasushi Hiromi and Tetsusi Iida for critical comments on the manuscript. This study was supported by a fellowship by the Japan Society for the promotion of Science to S.T. and grants from the Mitsubishi Foundation, the Takeda Science Foundation, and the Japanese Ministry of Education, Culture, Sports, Science, and Technology (19207002 and 19060014) to T.K.

References

- The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796–815.
- Bushman FD. 2003. Targeting survival: Integration site selection by retroviruses and LTR-retrotransposons. *Cell* **115**: 135–138.
- Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, Gu M, Blattner FR, Jiang J. 2002. Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14**: 1691–1704.
- Clough SJ, Bent AF. 1998. Floral dip: A simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* **16**: 735–743.
- Courtial B, Feuerbach F, Eberhard S, Rohmer L, Chiapello H, Camilleri C, Lucas H. 2001. Tnt1 transposition events are induced by in vitro transformation of *Arabidopsis thaliana*, and transposed copies integrate into genes. *Mol Genet Genomics* **265**: 32–42.
- Dawe RK, Hiatt EN. 2004. Plant neocentromeres: Fast, focused, and driven. *Chromosome Res* **12**: 655–669.
- Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J. 2010. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants; insights from genome-wide analysis and multi-specific comparison. *Plant J* **63**: 584–598.
- Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res* **18**: 359–369.
- Gendrel AV, Lippman Z, Yordan C, Colot V, Martienssen RA. 2002. Dependence of heterochromatic histone H3 methylation patterns on the *Arabidopsis* gene DDM1. *Science* **297**: 1871–1873.
- Grewal SI, Jia S. 2007. Heterochromatin revisited. *Nat Rev Genet* **8**: 35–46.
- Hawkins JS, Kim H, Nason JD. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res* **16**: 1252–1261.
- Heslop-Harrison JS, Brandes A, Schwarzacher T. 2003. Tandemly repeated DNA sequences and centromeric chromosomal regions of *Arabidopsis* species. *Chromosome Res* **11**: 241–253.
- Hickey DA. 1982. Selfish DNA: A sexually-transmitted nuclear parasite. *Genetics* **101**: 519–531.
- Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. 1996. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc Natl Acad Sci* **93**: 7783–7788.
- Hirochika H, Okamoto H, Kakutani T. 2000. Silencing of retrotransposons in *Arabidopsis* and reactivation by the ddm1 mutation. *Plant Cell* **12**: 357–369.
- Hosouchi T, Kumekawa N, Tsuruoka H, Kotani H. 2002. Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1, 2, and 3. *DNA Res* **9**: 117–121.
- Hou Y, Rajagopal J, Irwin PA, Voytas DF. 2010. Retrotransposon vectors for gene delivery in plants. *Mob DNA* **1**: 19. doi: 10.1186/1759-8753-1-19.
- Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* **43**: 476–481.
- Jeddeloh JA, Stokes TL, Richards EJ. 1999. Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nat Genet* **22**: 94–97.
- Kakutani T, Jeddeloh JA, Flowers S, Munakata K, Richards EJ. 1996. Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. *Proc Natl Acad Sci* **93**: 12406–12411.
- Kanizay L, Dawe RK. 2009. Centromeres: Long intergenic spaces with adaptive features. *Funct Integr Genomics* **9**: 287–292.

- Kawabe A, Nasuda S. 2005. Structure and genomic organization of centromeric repeat in *Arabidopsis* species. *Mol Genet Genomics* **272**: 593–602.
- Kawabe A, Hansson B, Hagenblad J, Forrest A, Charlesworth D. 2006. Centromere locations and associated chromosome rearrangements in *Arabidopsis lyrata* and *A. thaliana*. *Genetics* **173**: 1613–1619.
- Le QH, Wright S, Yu Z, Bureau T. 2000. Transposon diversity in *Arabidopsis thaliana*. *Proc Natl Acad Sci* **97**: 7376–7381.
- Le QH, Melayah D, Bonnivard E, Petit M, Grandbastien MA. 2007. Distribution dynamics of the Tnt1 retrotransposon in tobacco. *Mol Genet Genomics* **278**: 639–651.
- Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430**: 471–476.
- Ma J, Jackson SA. 2006. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. *Genome Res* **16**: 251–259.
- Madsen LH, Fukai E, Radutoiu S, Yost CK, Sandal N, Schauser L, Stougaard J. 2005. LOREL, an active low-copy-number TY3-gypsy retrotransposon family in the model legume *Lotus japonicus*. *Plant J* **44**: 372–381.
- Mallik HS, Henikoff S. 2002. Conflict begets complexity: The evolution of centromeres. *Curr Opin Genet Dev* **12**: 711–718.
- Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, Cao J, Weigel D, Paszkowski J, Mathieu O. 2009. Selective epigenetic control of retrotransposition in *Arabidopsis*. *Nature* **461**: 427–430.
- Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature* **411**: 212–214.
- Nagaki K, Neumann P, Zhang D, Ouyang S, Buell CR, Cheng Z, Jiang J. 2005. Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Mol Biol Evol* **22**: 845–855.
- Okamoto H, Hirochika H. 2000. Efficient insertion mutagenesis of *Arabidopsis* by tissue culture-induced activation of tobacco retrotransposon Tto1. *Plant J* **23**: 291–304.
- Pelissier T, Tutois S, Deragon JM, Tourmente S, Genestier S, Picard G. 1995. Athila, a new retroelement from *Arabidopsis thaliana*. *Plant Mol Biol* **29**: 441–452.
- Peterson-Burch BD, Nettleton D, Voytas DF. 2004. Genomic neighborhoods for *Arabidopsis* retrotransposons: A role for targeted integration in the distribution of the Metaviridae. *Genome Biol* **5**: R78. doi: 10.1186/gb-2004-5-10-r78.
- Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet* **23**: 305–308.
- Sandmeyer SB. 1990. Integration specificity of retrotransposons and retroviruses. *Annu Rev Genet* **24**: 491–518.
- SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**: 765–768.
- SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet* **20**: 43–45.
- Singer T, Yordan C, Martienssen RA. 2001. Robertson's Mutator transposon in *A. thaliana* are regulated by the chromatin-remodeling gene Decrease in DNA methylation (DDM1). *Genes Dev* **15**: 591–602.
- Slotkin RK. 2010. The epigenetic control of the Athila family retrotransposons in *Arabidopsis*. *Epigenetics* **5**: 483–490.
- Soppe WJ, Jasencakova Z, Houben A, Kakutani T, Meister A, Huang MS, Jacobsen SE, Schubert I, Frasz PF. 2002. DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in *Arabidopsis*. *EMBO J* **21**: 6549–6559.
- Steimer A, Amedo P, Afsar K, Frasz P, Mittelsten-Scheid O, Paszkowski J. 2000. Endogenous targets of transcriptional gene silencing in *Arabidopsis*. *Plant Cell* **12**: 1165–1178.
- Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. 2009. Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature* **461**: 423–426.
- Vitte C, Bennetzen JL. 2006. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proc Natl Acad Sci* **103**: 17638–17643.
- Vongs A, Kakutani T, Martienssen RA, Richards EJ. 1993. *Arabidopsis thaliana* DNA methylation mutants. *Science* **260**: 1926–1928.
- Yamazaki M, Tsugawa H, Miyao A, Yano M, Wu J, Yamamoto S, Matsumoto T, Sasaki T, Hirochika H. 2001. Rice retrotransposon Tos17 prefers low-copy-number sequences as integration targets. *Mol Genet Genomics* **125**: 336–344.
- Yu W, Han F, Gao Z, Vega JM, Birchler JA. 2007. Construction and behavior of engineered minichromosomes in maize. *Proc Natl Acad Sci* **104**: 8924–8929.

Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*

Sayuri Tsukahara, Akira Kawabe, Akie Kobayashi, et al.

Genes Dev. 2012, **26**: originally published online March 19, 2012

Access the most recent version at doi:[10.1101/gad.183871.111](https://doi.org/10.1101/gad.183871.111)

Supplemental Material <http://genesdev.cshlp.org/content/suppl/2012/03/12/gad.183871.111.DC1>

Related Content **Retrotransposon insertion targeting: a mechanism for homogenization of centromere sequences on nonhomologous chromosomes**
James A. Birchler and Gernot G. Presting
[Genes Dev.](#) April , 2012 26: 638-640

References This article cites 48 articles, 16 of which can be accessed free at:
<http://genesdev.cshlp.org/content/26/7/705.full.html#ref-list-1>
Articles cited in:
<http://genesdev.cshlp.org/content/26/7/705.full.html#related-urls>

License

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).