Short RNAs Can Identify New Candidate Transposable Element Families in Arabidopsis

M. Florian Mette, Johannes van der Winden, Marjori Matzke*, and Antonius J.M. Matzke

Institute of Molecular Biology, Austrian Academy of Sciences, Billrothstrasse 11, A–5020 Salzburg, Austria

Mining the Arabidopsis genome for transposable elements (TEs) by DNA sequence similarity searches and analysis protocols is revealing previously unidentified families of TEs and providing insights into TE structure, mobility, distribution, and diversity (Le et al., 2000). We suggest here that new putative TE families and partially diverged TE-like sequences can be identified by an alternate approach involving cloning and analyzing short RNAs, which are a hallmark of RNA silencing mechanisms.

RNA silencing is triggered by dsRNA that is cleaved to short RNAs 21 to 24 nucleotides (nts) in length by an RNase III-like enzyme termed Dicer (Matzke et al., 2001; Hutvágner et al., 2002). The short RNAs are thought to guide enzyme complexes that either degrade complementary RNAs in the cytoplasm (a process termed posttranscriptional gene silencing in plants and RNA interference [RNAi] in animals), or modify homologous DNA sequences in the nucleus (RNA-directed DNA methylation [RdDM]). In plants, RdDM can lead to transcriptional gene silencing if dsRNAs contain promoter sequences (Matzke et al., 2001). A major function of posttranscriptional gene silencing/RNAi and DNA methylation, which may result from RdDM in many cases, is to limit the proliferation of TEs (Matzke et al., 2000). The host defense role of RNA silencing is evidenced by the mobilization of some TEs in *Caenorhabditis elegans* mutants defective in RNAi (Ketting et al., 1999; Tabara et al., 1999) and in Arabidopsis mutants deficient in some aspect of DNA methylation or chromatin structure (Miura et al., 2001; Okamoto and Hirochika, 2001; Singer et al., 2001; Tomba et al., 2002). A role for RNA silencing in TE control is also supported by findings of sequences homologous to various TEs in collections of short RNAs cloned from different sources (Djikeng et al., 2001; Lagos-Quintana et al., 2001). The enrichment of known TE sequences in populations of short RNAs, which are presumably cleavage products of a Dicer-like enzyme, suggests that unidentified TEs might be detected through their presence in short RNA libraries. In an ongoing project to clone and sequence short RNAs approximately 17 to 27 nts in length from Arabidopsis leaves, we have isolated short RNAs that appear to be derived from previously unknown TE families and from TE-like sequences.

"40" FAMILY

One short RNA sequence has been isolated repeatedly and represents the most frequent nonstructural short RNA recovered in our study (11 independent clones comprising approximately 8% of total nonstructural RNAs). This group of short RNAs, designated the "40" family, ranges in size from 17 to 21 nts, with a fixed 5′ end and ragged 3′ ends. BLASTN searches revealed DNA sequence homology in three, unannotated intergenic regions of the Arabidopsis genome (Fig. 1). The only other highly similar sequences (identity in 20/21 nt) in the database are present in the *Oryza sativa* genome. Because the "40" short RNA family was exceptionally well represented in the population of cloned short RNAs, we investigated it further. An RNA folding program was used to examine whether the DNA sequences surrounding the short RNA "genes" could give rise to an RNA with a stable secondary structure. In all three cases, an approximately 200-bp imperfect RNA duplex, in which the short RNA is located in a semiconserved TIR, was generated (Fig. 1). Alignments of the three corresponding DNA sequences demonstrated that spacers internal to the TIRs of copies A and B display 79% DNA sequence identity, whereas the internal spacer of copy C, which is somewhat longer, shows no significant homology to A and B.

Although the potential RNA duplexes are quite long, only short RNAs derived from the TIR were cloned, indicating that this region is preferentially cleaved by a Dicer-like enzyme. Moreover, all 11 short RNAs originated exclusively from one side of the dsRNA (the 3′ end), which is the most conserved half of the TIR among the three copies. The presence of these short RNAs and their polarity were confirmed on northern blots: Only the antisense RNA probe produced a signal (Fig. 1). The "40" short RNA family appears to be relatively uniformly sized on the northern blot, displaying the same mobility as a 23-nt DNA oligonucleotide. The range of sizes that were cloned (17–21 nts) may indicate differences in the migration of short RNAs compared with DNA oligonucleotides or some degradation from the 3′ end during cloning procedures.

* Corresponding author; e-mail mmatzke@imb.oeaw.ac.at; fax 43–662–63961–29.

www.plantphysiol.org/cgi/doi/10.1104/pp.007047.

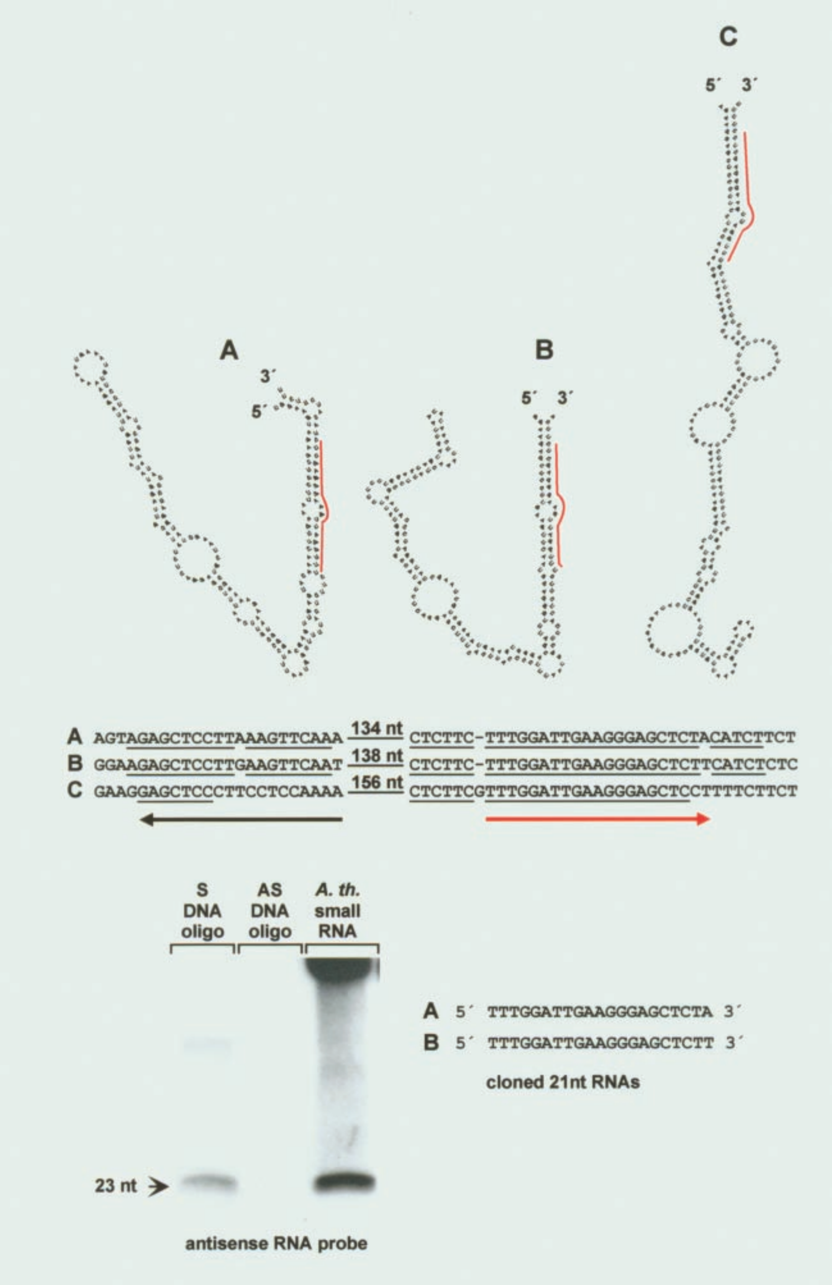
C
5´ 3´
A
3´
5´
B
5´ 3´
A AGTAGAGCTCCTTAAAGTTCAAA 134 nt CTCTTC-TTTGGATTGAAGGGAGCTCTACATCTTCT
B GGAAGAGCTCCTTGAAGTTCAAT 138 nt CTCTTC-TTTGGATTGAAGGGAGCTCTTCATCTCTC
C GAAGGAGCTCCCTTCCTCCAAAA 156 nt CTCTTCGTTTGGATTGAAGGGAGCTCCTTTTCTTCT
S DNA oligo
AS DNA oligo
A. th. small RNA
23 nt
antisense RNA probe
A 5´ TTTGGATTGAAGGGAGCTCTA 3´
B 5´ TTTGGATTGAAGGGAGCTCTT 3´
cloned 21nt RNAs

Figure 1. "40" family of short RNAs. Using the "40" short RNA sequences as queries in a BLASTN search, hits in intergenic regions in three BACs were obtained. The names of these BACs, the regions used to generate the predicted RNA secondary structures at the top, and the chromosome assignment are: A, F25P22, bases 41,361 through 41,170, chromosome 1; B, T10F20 (overlapping with T10O22), bases 26,724 through 26,919, chromosome 1; and C, T3F17, bases 28,928 through 28,714, chromosome 2. Middle, The DNA sequences conserved among copies A through C are underlined in black and the length of the spacer indicated. The long arrows below copy C denote the semiconserved terminal inverted repeats (TIRs), the right half of which gives rise to the "40" short RNA family (red arrow, red line in secondary structures at top). Bottom right, The sequences of the longest short RNA clones (21 nt) derived from copies A and B, which differ in the final nt, are shown. Bottom left, The northern blot confirms the presence of "40" short RNAs that hybridize to an antisense (AS) RNA probe in the Arabidopsis small RNA population; no hybridization signal was observed with a sense (S) RNA probe (not shown). Short RNAs were cloned according to a published procedure (Elbashir et al., 2001).

To generate the dsRNA structures that could be processed by a Dicer-like enzyme to yield the "40" family of short RNAs, the entire approximately 200-bp unit must be transcribed 5′ to 3′ from a promoter that has not yet been identified. At least copies A and B are transcribed, as exemplified by the sequences of two 21-mers, which differed in the 3′-most nt exactly according to the A and B DNA sequences (Fig. 1). Transcription of A and B must initiate either from adjacent intergenic promoters or by read-through transcription from the upstream host genes. Judging from the hybridization intensity on the northern blot, which approximates that observed with short RNAs derived from 35S promoter-driven transcripts (Mette et al., 2000), the precursor RNA for the "40" family of short RNAs is strongly transcribed.

The presence of multiple, dispersed copies of an approximately 135- to 155-bp DNA sequence flanked by relatively well conserved 20- to 30-bp TIRs in the Arabidopsis genome suggests that this small repeat family is possibly derived from a TE. Although sequence similarity generally falls off beyond the regions we have designated as TIRs, suggesting that they delimit a putative TE, we did not detect target site duplications, which would be expected from a class 2 (DNA) element. It is not yet clear whether and how members of the "40" family transpose, but their sequence heterogeneity suggests they are degenerate relics of a previously active TE family. Whether the short RNAs derived from this putative TE family play a role in controlling transposition is not clear, but the striking conservation of both the “40” short RNA sequences and the potential secondary structures from which they are presumably derived suggests an important function.

When compared with known types of TEs, the “40” family appears similar to “neisseria miniature insertion sequences” (nemis). These are abundant, small DNA insertion sequences in the chromosome of the pathogenic bacterium *Neisseriae gonorrhoeae* (Mazzone et al., 2001). Unit length nemis (approximately 160 bp) feature TIRs (26–27 bp) and can potentially fold into a robust stem-loop structure. More than 66% of nemis are found close to cellular genes. In an intriguing parallel to the “40” family, the nemi RNAs appear to be cotranscribed with cellular genes and subsequently processed, at either one or both TIRs, by RNase III (14).

“175” FAMILY

One short RNA clone 24 nt in length was found to be homologous to regions of five BACs in the BLASTN search. In each bacteria artificial chromosome (BAC), there are two hits in inverted orientation that are separated by varying lengths of spacer DNA. For the longest copy (Fig. 2, BAC F10C21), there is one mismatch to the short RNA sequence (identity in 23/24 nts); for the other four copies, sequence identity is perfect (24/24 nts). The longest sequence is annotated as a putative MudrA transposase, suggesting a MULE family. Alignments of all five sequences demonstrated that they are related by common TIRs approximately 330 bp in length that flank internal deletions of varying sizes (Fig. 2). Each element copy is flanked by a 6- to 9-bp target site duplication.

The “175” family is distinct from MULE families described so far in Arabidopsis (Yu et al., 2000), supporting the claim that short RNA sequences can draw attention to previously unidentified TE families. The longest copy, which contains the coding region of MudrA transposase, is possibly an autonomous element that has degenerated rapidly to produce a heterogeneous group of internally deleted, nonautonomous derivatives (Fig. 2). The short RNA could originate from either the left (F14F8, T4B21) or right (F14J22, F25O24) half of the TIR of an internally deleted copy (Fig. 2). The existence of short RNAs derived from the TIR region suggests transcription through the entire element and intramolecular pairing to form a dsRNA, which would probably be produced most readily with transcripts issuing from one of the more extensively deleted copies (Fig. 2).

The “175” short RNAs are less abundant than the “40” family of short RNAs, as indicated by the recovery of fewer independent clones and a fainter signal on northern blots (Fig. 2). In addition, the “175” short RNAs are longer than those in the “40” family. The “175” short RNAs migrate as a doublet slightly above the 23-nt DNA oligonucleotide standard, which is consistent with the 24-nt length of the cloned RNA. Whether the size difference between the two short RNA families reflects the activity of different dicer-like enzymes, and/or the subcellular localization of dsRNA processing (nucleus or cytoplasm), is not yet known.

Our data suggest that investigating short RNA populations can help to identify new candidate TE families and partially diverged TE-like sequences that might be missed in conventional DNA sequence analyses. In contrast to DNA sequence similarity

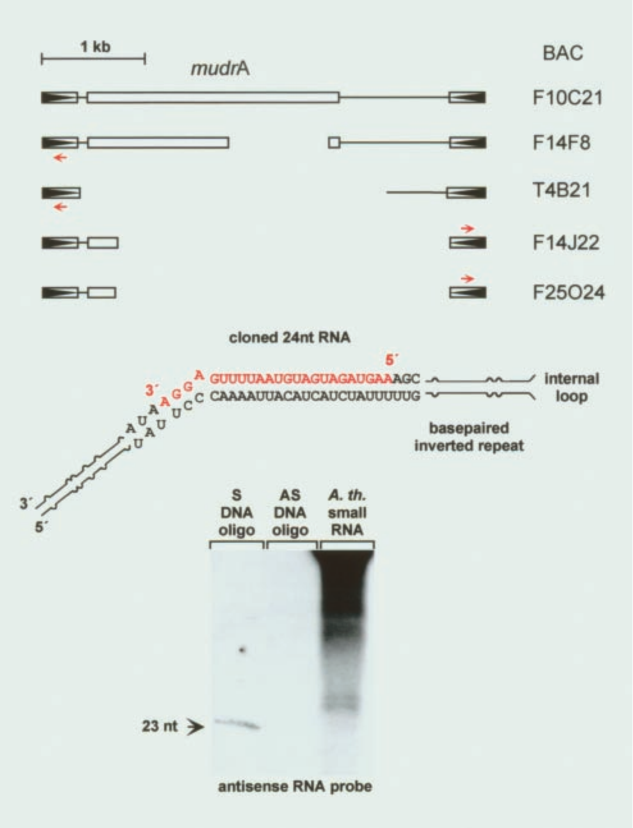

Figure 2. “175” family of short RNAs. Middle, Sequence of the 24-nt short RNA of the “175” family (red) and its position in the stem of a predicted RNA duplex generated by folding a transcript homologous to the shortest putative *Mutator*-like TE (MULE) derivative (BAC F25O24). The small red arrows indicate possible origins of the short RNA in the TIRs of the “175” MULE family. The family members are in the following regions of Arabidopsis genome: F10C21, bases 66,969 through 62,619, chromosome 1; F14F8, bases 26,965 through 30,372, chromosome 5; T4B21, bases 45,822 through 44,489, chromosome 4; F14J22, bases 50,446 through 49,368, chromosome 1; and F25O24, bases 3,075 through 4,118, chromosome 4. The northern blot confirms the presence of “175” short RNAs that hybridize to an antisense (AS) RNA probe in the Arabidopsis small RNA population. No signal was visible using a sense (S) RNA probe (not shown).

searches, this approach focuses on putative TEs that are transcribed to produce dsRNA and that might be targets of transcriptional and posttranscriptional RNA silencing mechanisms. Certain short RNAs, such as those from the "40" family, are extraordinarily well represented in the short RNA population, whereas the degree of representation of standard cellular RNA genes appears considerably less (A. Matzke and M.F. Mette, unpublished results). It is striking that the short RNAs we have described in this report always originate in regions of the genome that can potentially give rise to dsRNA, indicating that they are not random products of single stranded RNA degradation. As we found with the "175" short RNA, which revealed a putative MULE family, tiny RNAs can pinpoint a widely spaced inverted repeat comprising two halves that show high DNA sequence similarity. Moreover, as shown by the "40" family, a short RNA can also reveal imperfect inverted repeats that might be undetectable from the DNA sequence alone. Because G-U pairing is allowed in RNA secondary structures, however, an RNA duplex can form from a transcript of the region.

Much remains to be learned about areas of the genome that are transcribed to produce dsRNA precursors of short RNAs, and the identity of the RNA polymerase(s) involved. In addition to their possible role in controlling transposition, it will be interesting to determine whether short RNAs derived from TEs and TE-like sequences are involved in host gene regulation. TEs flanked by TIRs, including MITEs, *Mutator* elements, and nemis, frequently integrate next to host genes, thus potentially furnishing these genes with target sites for complementary short RNAs arising from members of the TE family that produce dsRNA. Conceivably, such TEs or their derivatives might be sources of micro-RNAs (miRNAs), at least some of which are involved in developmental timing of gene expression in *C. elegans* and possibly other animals (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001; Lai, 2002). The miRNAs are derived via Dicer cleavage of an approximately 70-nt precursor that can form an imperfect RNA duplex. Similar to the "40" family described here, miRNAs usually accumulate from only one arm of the fold-back precursor. The reason for the asymmetry in short RNA accumulation is unclear, but it might indicate preferential stabilization of the copy that can base pair with the target RNA. It is also not known why short RNAs originate only from the TIR region of the putative "40" TE family, even though the predicted RNA duplexes comprise spacer sequences. Further studies on short RNAs and the intergenic regions that encode them should help answer these questions.

Received April 11, 2002; accepted May 16, 2002.

LITERATURE CITED

Djikeng A, Shi H, Tschudi C, Ullu E (2001) RNA **7:** 1522–1530

Elbashir S, Lendeckel W, Tuschl T (2001) Genes Dev **15:** 188–200

Hutvágner G, Zamore P (2002) Curr Opin Genet Dev **12:** 225–232

Ketting R, Haverkamp T, van Luenen H, Plasterk R (1999) Cell **99:** 133–141

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Science **294:** 853–858

Lai E (2002) Nat Genet **30:** 363–364

Lau N, Lim L, Weinstein E, Bartel D (2001) Science **294:** 858–862

Le QH, Wright S, Yu Z, Bureau T (2000) Proc Natl Acad Sci USA **97:** 7376–7381

Lee R, Ambros V (2001) Science **294:** 862–864

Matzke MA, Matzke AJM, Kooter J (2001) Science **293:** 1080–1083

Matzke M, Mette MF, Matzke AJM (2000) Plant Mol Biol **43:** 401–415

Mazzone M, De Gregorio E, Lavitola A, Pagliarulo C, Alifano P, Nocera P (2001) Gene **278:** 211–222

Mette MF, Aufsatz W, van der Winden J, Matzke M, Matzke AJM (2000) EMBO J **19:** 5194–5201

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T (2001) Nature **411:** 212–214

Okamoto H, Hirochika H (2001) Trends Plant Sci **6:** 527–534

Singer T, Yordan C, Martienssen R (2001) Genes Dev **15:** 591–602

Tabara H, Sarkissian M, Kelly W, Fleenor J, Grishok A, Timmons K, Fire A, Mello C (1999) Cell **99:** 123–132

Tomba R, McCallum C, Delrow J, Henikoff J, van Steensel B, Henikoff S (2002) Curr Biol **12:** 65–68

Yu Z, Wright S, Bureau TE (2000) Genetics **156:** 2019–2031