Mutator-like Elements in *Arabidopsis thaliana*: Structure, Diversity and Evolution

Zhihui Yu, Stephen I. Wright and Thomas E. Bureau

Department of Biology, McGill University, Montreal, Quebec, H3A 1B1 Canada

Manuscript received May 18, 2000
Accepted for publication September 11, 2000

ABSTRACT

While genome-wide surveys of abundance and diversity of mobile elements have been conducted for some class I transposable element families, little is known about the nature of class II transposable elements on this scale. In this report, we present the results from analysis of the sequence and structural diversity of **Mu**tator-**l**ike **e**lements (MULEs) in the genome of *Arabidopsis thaliana* (Columbia). Sequence similarity searches and subsequent characterization suggest that MULEs exhibit extreme structure, sequence, and size heterogeneity. Multiple alignments at the nucleotide and amino acid levels reveal conserved, potentially transposition-related sequence motifs. While many MULEs share common structural features to *Mu* elements in maize, some groups lack characteristic long terminal inverted repeats. High sequence similarity and phylogenetic analyses based on nucleotide sequence alignments indicate that many of these elements with diverse structural features may remain transpositionally competent and that multiple MULE lineages may have been evolving independently over long time scales. Finally, there is evidence that MULEs are capable of the acquisition of host DNA segments, which may have implications for adaptive evolution, both at the element and host levels.

THE *Mutator* (*Mu*) system is a diverse family of class II transposable elements (TEs) found in maize. Robertson (1978) first identified *Mu* elements through a heritable high forward mutation rate exhibited by lines derived from a single maize stock. To date, at least six different classes have been identified in maize *Mutator* lines (Bennetzen 1996). *Mu* elements have long (≈200 bp) and highly conserved terminal inverted repeats (TIRs). However, the internal sequences are often heterogeneous (Chandler and Hardeman 1992). Upon insertion, *Mu* elements typically generate a 9-bp target site duplication (TSD) of flanking DNA (Bennetzen 1996). Transposition of *Mu* elements is primarily regulated by a member of the *MuDR* class of the elements, which contain both *mudrA* and *mudrB* genes (Hershberger *et al.* 1991, 1995; Lisch *et al.* 1995). *mudrA* encodes the transposase, MURA (Benito and Walbot 1997; Lisch *et al.* 1999), which may be related to the transposases of some insertion sequences (IS) in bacteria (Eisen *et al.* 1994), whereas *mudrB* is nonfunctional for all aspects of *Mutator* activity (Lisch *et al.* 1999). As with other mobile elements, some *Mu* elements lacking a functional *mudrA* are capable of transposition if MURA is supplied *in trans* (Chandler and Hardeman 1992; Bennetzen 1996). The *Mutator* system in maize has been demonstrated to be an active agent in creating mutation and has been developed as a highly efficient transposon-tagging tool for maize gene isolation (Walbot 1992). In addition to maize, *mudrA*-related genes are apparently expressed in *Oryza sativa* (Eisen *et al.* 1994), *Gossypium hirsutum* (GI:5046879), and *Glycine max* (GI:7640129). However, no systematic study on the distribution, diversity, and evolution of **Mu**tator-**l**ike **e**lements (MULEs) has been conducted in any higher plant species other than maize.

Arabidopsis thaliana has become a model organism for genetic analysis of many aspects of plant biology and is the first plant species to be targeted for complete genome sequencing (Meinke *et al.* 1998). This sequence information provides an exceptional opportunity to identify mobile elements and to characterize their patterns of diversity at the whole-genome level. The Arabidopsis genome has recently been shown to harbor numerous TEs, including MULEs (Lin *et al.* 1999; Mayer *et al.* 1999; Le *et al.* 2000). In this report, we analyze the sequence, structural diversity, and phylogenetic relationship of the MULE groups that contain member(s) encoding a putative MURA-related protein in *A. thaliana*.

MATERIALS AND METHODS

Data mining: Sequences surveyed in this study correspond to 243 randomly selected large-insert DNA clones (∼17.2 Mb) from the Arabidopsis Genome Initiative (AGI), as described by Le *et al.* (2000). Specifically, sequenced clones released before December 1998 were chosen for systematic screening and classifying MULEs. Additional members were then periodically mined up to December 1999. Two computer-based approaches were employed to identify MULEs. The first method involved using Arabidopsis genomic sequences as queries in

Corresponding author: Thomas E. Bureau, Department of Biology, McGill University, 1205 Dr. Penfield Ave., Montreal, Quebec, H3A 1B1 Canada. E-mail: thomas_bureau@maclan.mcgill.ca

BLAST (version 2.0; Altschul *et al.* 1990; http://www.ncbi.nlm.nih.gov/blast/) searches, as described by Le *et al.* (2000). In addition, each DNA segment (typically the sequence from one large-insert clone) was compared against its reverse complement using the program BLAST 2 Sequences (Tatusova and Madden 1999) to identify long TIR structures. The elements were classified into groups on the basis of shared nucleotide sequence similarity (BLAST score > 80). Long TIRs were defined as terminal-most regions sharing >80% sequence identity over ≥100 contiguous base pairs. A detailed description of the mined MULEs presented in this report can be accessed on our World Wide Web site at http://soave.biol.mcgill.ca/clonebase/.

Sequence analysis and molecular cloning: Both PCR- and computer-based approaches were employed to document past transposition events and to confirm the position of termini for some elements by identifying RESites (*i.e.*, sequences that are *r*elated to *e*mpty *site*s; Le *et al.* 2000). In the PCR-based protocol, genomic DNA was isolated from 10 ecotypes of *A. thaliana*: No-0, Sn-1, Ws, Nd-1, Tsu-1, Rld-1, Di-G, Tol-0, S96, and Be-0 (Arabidopsis Biological Resource Center; http://aims.cps.msu.edu/aims). PCR primers were designed corresponding to the regions flanking putative MULEs. A primer name was composed of three parts, namely, (i) ATC (**A***rabidopsis* **t***haliana* **c**lone), (ii) the GI (Geninfo Identifier) number of the clone harboring the MULE, and (iii) the corresponding position in the clone where the primer sequence was derived. The primer pair used to amplify RESites for MULE-1:GI2182289 was ATCGI2182289-38427 (5′-GTGAGGCAACACGTCATCATCTC-3′) and ATCGI2182289-40214 (5′-CTGGTCTTGAACCTCGTTCATCC-3′); for MULE-23:GI3063438, it was ATCGI3063438-86192 (5′-CCACCTTTAATCCGGGAGAATTC-3′) and ATCGI3063438-99055 (5′-CACGATGGAACTCCAGTCAG-3′); and for MULE-24:GI2760316, it was ATCGI2760316-88054 (5′-CATGTAACCCTTCATGGGTGG-3′) and ATCGI2760316-93177 (5′-TGGGATTCCAATTTGTCAGCCTG-3′). PCR amplifications were carried out using annealing temperatures ranging from 50–65° as previously described (Bureau and Wessler 1994). Amplified fragments were cloned into a pCR2.1 vector (Invitrogen, Carlsbad, CA) and subsequently sequenced using a SequiTherm EXCEL II kit (Epicentre, Madison, WI). The resulting DNA sequences were compared with the corresponding sequences at element insertion sites to confirm the position of element termini and TSDs. Alternatively, the regions flanking putative MULEs were used as BLAST queries to identify related sequences that lacked MULEs (Le *et al.* 2000).

Information concerning the position, sequence, and structure of putative open reading frames (ORFs) within mined MULEs was inferred from the annotation of surveyed clones (AGI, http://www.arabidopsis.org/AGI/AGI_sum_table.html). Multiple sequence alignments of the members within individual MULE groups were performed using DIALIGN 2.1 (http://bibiserv.techfak.uni-bielefeld.de/dialign; Morgenstern 1999) and graphically displayed with PlotSimilarity as part of the GCG program suit (version 10.0; Genetic Computer Group, University of Wisconsin, Madison). The terminal-most consensus sequences (100 nucleotides in length) of individual MULE groups were derived from the corresponding alignments. In addition, transposon insertions within the MULEs were identified using these alignments. Information concerning the potential expression of the putative ORFs was inferred from searches against GenBank expressed sequence tag (EST) databases. ProfileScan (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html; Gribskow *et al.* 1987) and Pfam HMM Search (http://pfam.wustl.edu/hmmsearch.shtml; Bateman *et al.* 2000) were used to determine the location of conserved motif(s) within analyzed protein sequences. Analysis of substitution patterns and determination of significant deviation from neutral expectations (*i.e.*, $K_a/K_s = 1$) were generated using the program K-Estimator (version 5.3; Comeron 1995; Comeron *et al.* 1999). Sliding window analysis of sequence diversity (calculated as π, the average parities difference) across aligned sequences was conducted using the program DnaSP (version 3.14; Rozas and Rozas 1999).

Phylogenetic analysis: Maize *mudrA* and Arabidopsis *mudrA*-related ORFs were compared by pairwise alignment using BLASTX and multiple alignment using MULTALIN (http://pbil.ibcp.fr / cgi-bin / npsa_automat.pl?page= / NPSA / npsa_server.html; Corpet 1988) to identify the most conserved region for use in phylogenetic analysis. Using maize *mudrA* as an outgroup, unrooted phylogenetic trees were derived from both distance-based (neighbor-joining) and character-based (parsimony) approaches using programs in the PHYLIP package (version 3.75c; Felsenstein 1993). Nucleotide distances were computed using the Kimura option of DNADIST. SEQBOOT was used to generate 100 bootstrap replicates, each of which was then analyzed by NEIGHBOR and DNAPARS. The final majority-rule consensus trees were derived using CONSENSE.

RESULTS

As reported previously (Le *et al.* 2000), 28 MULE groups, representing a total of 108 elements, were identified with systematic survey of 17.2 Mb of sequenced

TABLE 1

Summary of mined MULE groups in *A. thaliana*

Group	No. of elements	Size range (bp)	TIR size range (bp)	No. of elements with *mudA*-related ORF[a]	TSD size (bp)
MULE-1	20	492–3,952	103–408	1	9–12
MULE-2	9	444–4,809	101–222	1	7–11
MULE-3	2	1,213–3,771	107–158	1	10
MULE-16	1	3,646	292	1	6–7
MULE-24	7	1,075–4,445	100–319	2	9–10
MULE-27	7	552–4,703	141–307	1	9–11
MULE-9	16	2,338–17,078	NA[b]	7	9
MULE-19	4	7,119–8,188	NA	4	8–9
MULE-23	6	12,267–19,397	NA	6	8–9

[a] Only one putative *mudrA*-related ORF was identified per element.
[b] Not applicable.

GI2182289	38872	GTAAAATGAT<u>TTTAAGAAGA</u>	MULE-1	<u>TTTAAGAAGA</u>TAATATTATA	39930
No-0	237	GTAAAATGACTTTAAGAAGA		TAATATTATA	267
GI4544405	76472	TTTAATTGTA<u>AAATCTAAAC</u>	MULE-2	<u>AAATCTAAAC</u>ACTAACTACT	76955
GI6598390	65962	TTTAATTGTAAAATGTAAAT		CTTAACTACT	65933
GI3299824	86839	TAAAAATAAT<u>GTATGTACCT</u>	MULE-3	<u>GTATGTACCT</u>ATTTTAAACA	87945
No-0	35	TAAAAACAATGTATGTACCT		ATTTTAAACA	5
GI2443899	20128	CAACGAGTGATATC<u>TTAAAA</u>	MULE-16	<u>TAAATAA</u>TTAACAATTATAA	23812
GI3241925	67660	CTACGAGTGTAATCTTAGAA		TTAACAATTTTAA	67628
GI2760316	88370	GGGATTCTAAA<u>GATTCTAAA</u>	MULE-24	<u>GATTCTAAA</u>GAATTGAATTG	92845
No-0	162	GGGATTCTAAAGATTCTAAA		GAATTGAATTG	193
GI4309747	50697	AGCTTAGTCGG<u>TAAAGGAAT</u>	MULE-27	<u>TAAAGGAAT</u>GTTGTTTTATC	51324
GI6449475	70995	AGCTAAGTCGGTAAAGGAAA		GTTGTTATATC	71025
GI4325365	37313	AGCGGCTTTG<u>GATATGAATA</u>	MULE-9	<u>ATATGAATA</u>AGGTACTCAAC	51299
GI4589444	9492	AGCGGCTTTAGATATGACTA		AGGTTCTCAAC	9462
GI6598686	80879	CCTTCCACCCT<u>CTTATAATC</u>	MULE-19	<u>CAAATAATC</u>CCAGATTTTGA	73721
GI3299824	120504	CCTTCCTCCCTCTTCTAATC		CCAGATTTTGA	120534
GI3063438	86330	TGTTCATGACT<u>TATTCTTTC</u>	MULE-23	<u>TATTCTTTC</u>TTCCATT-GAG	98636
No-0	196	TGTTCATGACTTATTCTTTC		TTCCATTAGAG	227

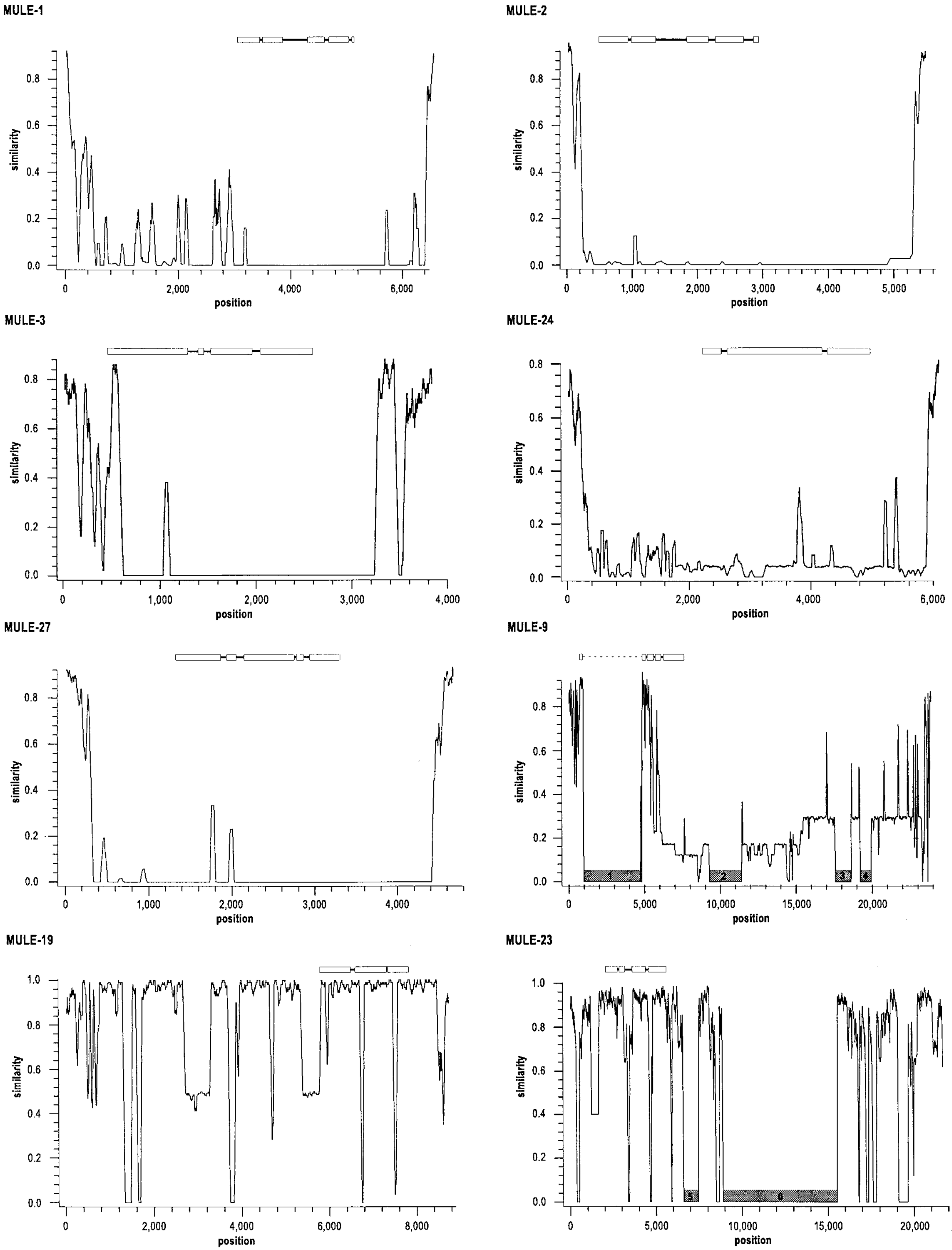
FIGURE 1.—RESites of some mined MULE group members. The MULE-associated TSDs are underlined. GI (geninfo) numbers and nucleotide positions in corresponding clones or amplified DNA fragments from *A. thaliana* ecotype No-0 are indicated. RESite analysis could not resolve the precise termini or TSD of MULE-16.

Arabidopsis genome. Nine of the reported MULE groups (72 elements in total) were found to contain the element(s) encoding a putative protein sharing ~25% similarity to MURA in maize. However, none of the elements was found to harbor a *mudrB*-related ORF. Table 1 summarizes the primary features and diversity of these groups. Detailed information of the mined elements described in this report as well as newly identified members are available on our web site at http://soave.biol.mcgill.ca/clonebase/. By analyzing flanking DNA sequences between an insertion and its corresponding RESite, the positions of both MULE termini and TSDs were confirmed for representative members from all nine MULE groups (Figure 1). Moreover, this analysis provides convincing evidence that the mined MULEs are indeed TEs.

Diversity of MULEs: Among the nine MULE groups, six contain elements with long TIRs (TIR-MULEs, Table 1). In general, the TIR-MULEs are structurally similar to *Mu* elements in maize (BENNETZEN 1996), with long TIRs (100 to 408 bp) and typically 9-bp TSDs (among the surveyed elements 49% have 9-bp TSDs, 39% have 10-bp TSDs, 5% have TSDs larger than 10 bp, and 7% have TSDs shorter than 9 bp). Fifteen percent of the TIR-MULEs contain a *mudrA*-related ORF and none contains a second ORF. Within a group, the element(s) harboring a *mudrA*-related ORF share(s) high sequence similarity (>80%) with other members only at the TIRs (Figure 2). Significant variation in element abundance is also observed among MULE groups. For example, only 1 member was identified for the MULE-16 group in our survey, compared to 20 members in the MULE-1 group. Within the latter group, 12 members share >90% sequence identity across their entire sequence. They share similarity only with the TIR sequences of the other 8 members in the same group.

The three other MULE groups (in total 26 elements were analyzed) also contain elements encoding MURA-related proteins, and 92% of their members also have a 9-bp TSD (Table 1 and Figure 1). However, MULEs in these groups display the following characteristics that have not been reported for *Mu* elements in maize or the TIR-MULEs described previously. First, the 5′ terminus and inverse complement of the 3′ terminus of these individual elements share much lower (<60%) sequence similarity compared to the TIR-MULEs in Arabidopsis and *Mu* elements in maize, which typically display >80% sequence similarity between a given element's long TIRs (CHANDLER and HARDEMAN 1992; BENNETZEN 1996; Figure 3). Second, members within a group share relatively high sequence similarity (up to 95%) across their entire length (Figure 2). Third, the majority of the elements (69%) are very large in size, ranging from ~7.1 kb to 19.4 kb. Eight out of 16 members of the MULE-9 group are relatively smaller in size (~2 to 3 kb). Multiple alignment analysis revealed that the smaller MULE-9 members were most likely derived from larger members (data not shown). Fourth, many of the large elements contain one or two ORFs in addition to the ORF related to maize MURA; the others encode hypothetical or unknown proteins. No EST information for any of the contained ORFs was available in our survey of EST databases. Given consistently low sequence similarity at their termini compared to the long TIRs of maize *Mu* elements and the Arabidopsis TIR-MULEs, we designated these elements as non-TIR-MULEs.

MULE diversity was also reflected in variation within *mudrA*-related ORFs. Of 22 sequences analyzed, the size of the putative ORFs varied from 2249 bp to 4356 bp. In addition, the *mudrA*-related ORFs were often composed of different numbers of exons (*i.e.*, 1–7) and introns (*i.e.*, 0–6). Pairwise comparison between maize


MULE-1
MULE-2
MULE-3
MULE-24
MULE-27
MULE-9
MULE-19
MULE-23
similarity
position

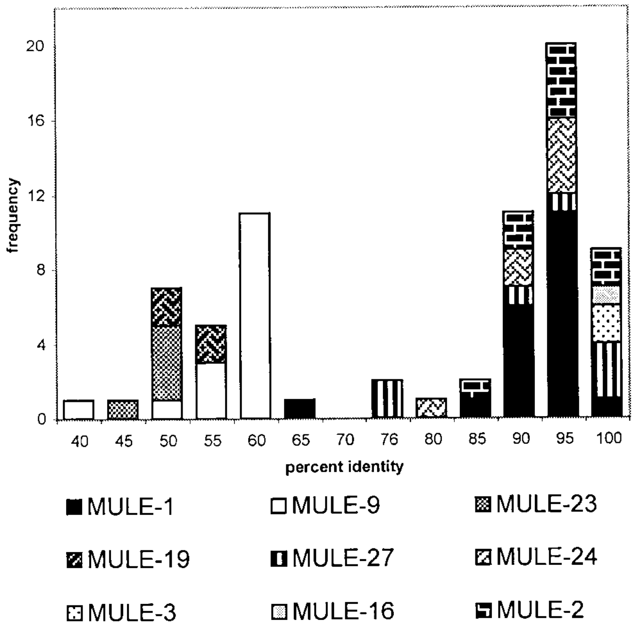


FIGURE 3.—Frequency distribution of sequence similarity at the termini of each individual MULE element. The first 100 bp of each element were aligned to the reverse complement of the last 100 bp, and the percentage similarity calculated. MULE-9, -19, and -23 are non-TIR MULEs, while MULE-1, -2, -3, -16, -24, and -27 are TIR-MULEs.

mudrA and each of the *mudrA*-related ORFs (data not shown) revealed that nucleotide substitutions, insertions, and deletions all contributed in generating this diversity.

In addition to sequence, structural, size, and element-abundance variation, we also found evidence for acquisition of host DNA segments into the internal regions of 5 of the 64 TIR-MULEs analyzed (Table 2). The size of the acquired DNA fragments range from 94 to 570 bp and make up the major portion of the internal regions of the corresponding elements. The acquired DNA sequences are 85–88% identical to the original host DNA segments. Strikingly, all of the acquired DNA segments correspond to the 5′ region (including 5′ untranslated region, 5′ flanking region, and the first one or two exons/introns) of transcription factors or developmentally regulated genes.

With one exception, MULE-1:GI2182289 (chromosome 1), the acquired gene sequences do not form ORFs. This element shows significant sequence similarity (Le *et al.* 2000) with a region spanning the first two exons and the first intron of the Arabidopsis homeobox-leucine zipper gene, *Athb-1* [Ruberti *et al.* (1991); also referred to as *HAT5* (Schena and Davis 1994); Figure 4A]. The acquisition of the *Athb-1* gene segment results in the formation of a novel putative ORF (Figure 4B) encoding a 71-amino-acid polypeptide. This putative protein shares 88% amino acid sequence similarity (Figure 4C) with the N-terminal sequence of the Athb-1 that includes an acidic domain (Figure 4B). Analysis of sequence diversity across the region of similarity between the putative gene from MULE-1:GI2182289 and the *Athb-1* gene indicates that noncoding regions have diverged more extensively than exons (Figure 4D). Calculation of substitution patterns between these two ORFs using the method of Comeron (1995) provides an estimated ratio of nonsynonymous to synonymous substitutions (K_a/K_s) of 0.6733, which is not significantly different from 1 ($P > 0.05$). Subsequent analysis has also revealed a second MULE-1 (GI613649; chromosome 4) with high nucleotide similarity to the same region of *Athb-1* (Figure 4A). The *Athb-1*-related region of MULE-1:GI613649 has numerous frameshifts and stop codons relative to *Athb-1* (Figure 4C), but the reconstructed amino acid sequence shares 80% similarity to the same region of Athb-1. As with the initially identified segment, a region corresponding to the location of the first intron of *Athb-1* is also present. No expression information of the putative gene in MULE-1:GI2182289 was identified through a survey of EST databases.

In a previous report (Le *et al.* 2000) we provided evidence demonstrating that recombination between different non-TIR-MULEs may generate MULE diversity. Furthermore, we found that nested transposon insertions also contribute to the MULE diversity. As described in Table 3, nested insertions of both class I and II TEs have been identified within six non-TIR MULEs (23% of the total non-TIR-MULEs identified). These insertions have variable sizes (ranging from ~0.73 kb to 6.67 kb) and have either TIR or long terminal repeat (LTR) structures. Two of the TE insertions also contain putative transposon-related ORFs. In addition, one TE insertion in MULE-23:GI6007863 may belong to a novel type of transposon. This TE has a 325-bp long TIR structure and is flanked by a 5-bp direct repeat (Table 3).

FIGURE 2.—Similarity plot of multiple sequence alignments of members from different MULE groups. Sequence similarity was determined using DIALIGN 2.1 (Morgenstern 1999) and displayed using PlotSimilarity (UWGCG) with a sliding window 50 bp in size. Both nucleotide and indel variation lead to a reduction in similarity estimates. The approximate positions of the *mudrA*-related ORFs and annotated exons (open boxes) and introns (solid bars) are indicated. The dashed line within the diagram of the *mudrA*-related ORF in MULE-9 represents a region corresponding to a TE insertion. The *mudrA*-related sequences in non-TIR-MULE groups are >85% identical to each other. The shaded regions in MULE-9 and -23 represent the sites where other TE insertions (see Table 3) were identified (1, insertion of an *En/Spm*-like element; 2, insertion of an *Athila*-like solo LTR element; 3, insertion of a MULE-3 element; 4, insertion of a *Tag-1*-like element; 5, insertion of a *Tat1*-like solo LTR; 6, insertion of an unclassifiable element that contains a truncated *Ty3/gypsy*-like integrase domain). As only one member was identified for MULE-16, a multiple alignment was not performed.

TABLE 2

MULE acquisition of host gene segments

MULE	Position of acquired sequence within the element	Size of acquired segment (bp)	Description of host gene[a]
MULE-1: GI3702730	25,192–25,683	501	*cdc*-related gene (GI6714475:68568-68068)
MULE-1: GI2815519	31,163–31,259	94	*fimbrin 2* (GI2811231:168-261)
MULE-1: GI4678340	25,325–25,608	292	*SCR*-related gene (GI7329669:13739-13448)
MULE-1: GI2182289	39,116–39,607	500	*Athb-1* (GI6016704:5584-6083)
MULE-24: GI3193305	15,717–16,278	570	*AtHSP101* (GI6715467:744-1313)

[a] GI number and position of corresponding clone.

The internal sequence has coding capacity for a putative protein that is 75 and 42% identical to the integrase domains of *Ty3/gypsy* retrotransposons in *A. thaliana* (Lin *et al.* 1999) and *Ananas comosus* (Thomson *et al.* 1998), respectively. This putative insertion element may reflect a novel class II element that has sustained an insertion of a truncated *Ty3/gyspy*-related retrotransposon. Alternatively, this sequence may represent a novel type of terminal inverted-repeat-containing retrotransposon (Zuker *et al.* 1984; Garrett *et al.* 1989).

Conserved sequence motifs: Figure 5 shows the consensus of the first 100 terminal-most sequences for each of the nine MULE groups. No sequence identical to the maize MURA binding site (Benito and Walbot 1997) was observed within any of the consensus sequences. Comparison of the consensus sequences revealed different levels of sequence conservation. First, the sequences are highly conserved within a MULE group. However, the overall sequence similarity is low between the terminal sequences of members from different MULE groups. Second, subterminal sequence motifs (12 bp to 24 bp in length) were shared between the terminal regions of individual non-TIR-MULE groups. Third, the terminal regions were typically A + T-rich (>60%). Nucleotide distribution within individual MULE groups (data not shown) revealed a general mosaic distribution pattern between A + T-rich and G + C-rich regions. Fourth, a general motif, 5′-$R_{(1\text{-}4)}$-3′ (R = G or A) followed by a short A + T-rich cluster, was identified at the distant ends of all the consensus sequences except MULE-16. This motif could also be found within the subterminal regions of many MULEs (data not shown).

The MURA-related proteins encoded by the mined MULEs were also analyzed for DNA-binding motif(s). Using ProfileScan and Pfam HMM, we identified a motif, **C**X2**C**X4**H**X4**C** (X represents any amino acid), at the C-terminal region of 16 Arabidopsis MURA-related proteins (67% of the total analyzed proteins; Figure 6). This motif also exists in a rice MURA-related protein, a number of known nuclear binding proteins, and other transposases (Figure 6). The C-terminal region of maize MURA has a similar motif, **C**X2**C**X4**H**X6**C.** Analyses of the N-terminal regions do not reveal any known motif.

Phylogeny of TIR and non-TIR-MULEs: A conserved

TABLE 3

Insertions of other TEs into the MULEs

Inserted MULE	TE type	Position	Size (bp)	Coding capacity	TIR size (bp)	TSD
MULE-9: GI3299824	MULE-3	86,859–87,925	1,066	None	158	gtatgtacct
MULE-9: GI3299824	*En/Spm*-like	91,459–95,242	3,783	*En/Spm*-like transposase	13	ggt
MULE-9: GI6136349	solo-LTR (*Athila*-like)	12,128–14,273	2,146	None	5	ccatt
MULE-9: GI3128140	*Tag-1*-like	50,787–51,517	731	None	21	cttatgag
MULE-23: GI6007863	Unknown	119,225–125,890	6,666	gag-pol polyprotein	325	atttg
MULE-23: GI6007863	solo-LTR (*Tat1*-like)	117,197–118,083	983	None	5	ataag

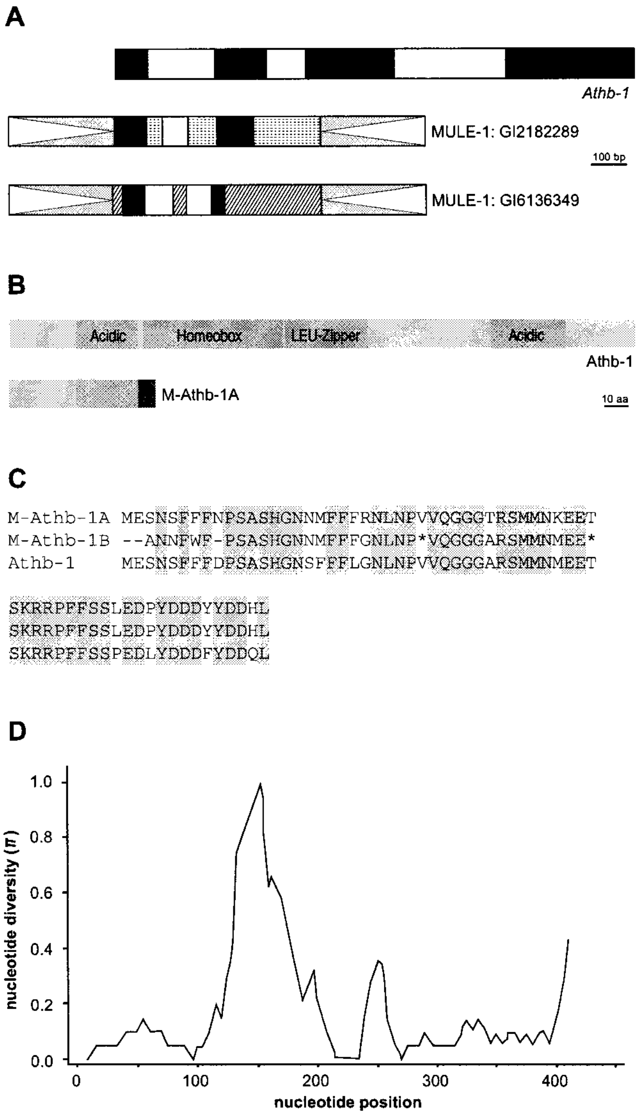


FIGURE 4.—Acquisition of the *Athb-1* gene by MULE-1:GI2182289 and MULE-1:GI6136349. (A) Illustration of the *Athb-1* gene and the element structures. Solid boxes represent exons; open boxes represent introns; shaded boxes with arrows represent TIRs; slash-lined boxes represent the internal region of MULE-1:GI6136349; and dash-lined boxes represent the internal region of MULE-1:GI2182289. The corresponding DNA sequences present in both dashed and slashed boxes have sequence similarity <50%; the corresponding sequences present in shaded boxes have sequence similarity >80%; and the DNA sequences present in both solid and open boxes of the elements have >86% sequence similarity with the corresponding DNA sequence in the *Athb-1* gene. (B) Structural relationship between the Athb-1 and the putative protein, M-Athb-1A. (C) Multiple alignment of the amino acid sequence shared between the putative protein encoded by MULE-1:GI2182289 (M-Athb-1A), the derived polypeptide from MULE-1:GI6136439 (M-Athb-1B) and the N-terminal region of the Athb-1. Identical amino acids are shaded. Asterisks represent positions where a frameshift was introduced to achieve an optimal alignment. (D) Sliding window of nucleotide sequence diversity (π) across the region of similarity between MULE-1:GI2182289 and the *Athb-1*. Sequences corresponding to an intron are located between positions 88 and 267 while the remaining regions correspond to exons.

region (270 nucleotides in total) was identified within the maize *mudrA* and the Arabidopsis *mudrA*-related ORFs (Figure 7) and used for phylogenetic analysis of the nine MULE groups. We utilized two methods, neighbor-joining and parsimony, to establish evolutionary relationships. Using maize *mudrA* as an outgroup sequence, both methods generated unrooted majority-rule trees with similar topologies. The consensus tree derived by the neighbor-joining method is shown in Figure 8. These phylogenetic relationships are consistent with our classification of MULE groups based on BLAST search results, since elements from one group are monophyletic, with high bootstrap support (>93%), and are separated by much shorter branch lengths than the elements between groups. The phylogeny also indicates that the non-TIR-MULE groups are more closely related to each other than they are to any of the TIR-MULE groups and that the non-TIR-MULEs that encode a MURA-related protein may have undergone recent amplification.

DISCUSSION

Genome sequencing projects allow for detailed analysis of the patterns and extent of transposon diversity in the genomes of model organisms. Our data suggest that the MULEs in *A. thaliana* exhibit both extreme structural and sequence heterogeneity. In fact, the observed variation indicates that the MULE superfamily may be one of the most diverse mobile element superfamilies in the plant kingdom. The presence of element insertions of varying ages may partly account for MULE diversity. The existence of numerous truncated MULEs (Lin *et al.* 1999; Mayer *et al.* 1999; Le *et al.* 2000) and the high level of divergence between MULE groups indicates that these elements might be an ancient mobile element system in the Arabidopsis genome and that many elements may no longer be transpositionally active. However, the existence of MULEs with significant sequence identity (>90%) and the identification of RESites from the closely related ecotypes suggest that many MULEs may have been recently mobile. The high level of diversity may also reflect the potential ability of MULEs to remain transpositionally competent with the presence of few conserved sequence motifs.

Non-TIR-MULEs are a novel type of plant class II TE. In contrast to the TIR-MULEs, as well as *Mu* elements in maize, these elements are characterized by low sequence similarity between termini of individual elements and the absence of long TIR structures. One might expect that these non-TIR-MULEs represent truncated, and presently inactive, elements. However, these elements are also characterized by their abundance in the genome, high level of homogeneity between members of individual groups, and a relatively high frequency of elements encoding a putative MURA-related protein. These features, combined with phylogenetic analysis, indicate that these elements are able to transpose in

```
MULE-1
A:GGGGAAAAATGTCAATTAATCCCCSAACTTTCAAAAAATGGYCATTTTATACRTCAACTTCGTATGYGGCCGTTTAAAACATGAMCWAAACGTTAACTAA
B:GGGGAAAAATGTCATTTAATRCCCCAACTTTRAAAAAATGGYCATTTTATACATCAACTTCGTATGTGGTCGTTTGAAACATGAAGTAAACGTTGACTAA

MULE-2
A:GGGGAAAAATGTTATTTAATACCTSAACTAACAAAAARTGGCCAAATTAACCGTGAAYTYYTGAAATGGCCGTTTTAAAYCTCAACAAAAAGTTGACTTC
B:GGGGAAAAATGTTATTTAATACCTGAACTTTCAAAAAATGGCCAAATTAACCGTGAACTCTTGAAATGGTCGTTTTATACCTCAACAAAAAATTGACTTC

MULE-3
A:GGG-AAACGGTYTATTTCCCCATGRGAACTATCRYATCWCGCYAGATRGAGCCCGATCTTTRRCCTCTGTCTRTATACCCACRARAYTGAAGTTAAATGYC
B:G-RWAAACGGTCTATTTYCCCATGRGAACTATCRYATCTSGCTAGATRAAGCCCGATCTTTRACCTCTTGCTRTTTACCCAYGAAATGAAAGTTAAAYGYC

MULE-16
A:GGGCTGCCTATAGGAGCCCCCCTCAAGTATATTGTTTCAACTGGGTTCCACCCAAACTCTTAGCCTCTTTCTATTTCCCCCTAATCAACTTAAATTCTTC------------
B:GG------------AGGCCCCCTCAAATATATTGTTTCAACTGGGTTCCACCCAAACTCTTAGCCTCTTTCTATTTCCCCCTAATCAACTTAAATTCTTCTCAATTCCACCT

MULE-24
A:GG-GAAAAAASYYKAAAAAAYCNTCATTTATTTTTAATTTGGCCGTTTAATACCTGTABTWTTWAAWTTGKMMGWAMAATACMTAASTTWATTTYKATTK
B:GGGGAAAAAASCCWAAAAAAYCYYSATTTATTTTCAATTTGGCCGTTTAATATCTSYAYTATTTAATTTGGMMGAAWAATACYTAAGTTAATTTTKATTN

MULE-27
A:GGGAAAAATGTCAAAACAAATCACGAACTTTCAAATTTGGGACGATTTAATCTTGAACTTCACRGAAGACAAATAAATCGTAAAGTTTTTGTTGACATTC-
B:GGGAAAAATGTCAAAA-AAATYACGAACTTTCAAATTTGGGACGATTTAATCTTGAACTTCACGAAAGACAAATAAATCGTAAAGTTTTTGTTGACTTTCG

MULE-9
A:GGGGTAATTTGCAGGGCACACCCKTTGACC-ATGTTTTATTTYCAGGGATTGKCAAAGTCAAAT------GTGGTTTTCACSWTTTGTCTTTTCR--SCTGMAATTTGA
B:-AAGAAATTTTCACARAAGT------GGCCGAAAAAAGAATTGCTCTCATGGTTTAT--CAAAGTGAGATCTGGTTTTCACCATTTGTCTTTTCAATSRTTTTCAGGAC

MULE-19
A:GAATGATCATTTCTTG--TGTCCTTTTTGTGGGACAACTAATCTGCAAAACCACTTTAGAGTCTTCATATCCCCTTTGAGTTAATGATTTGTGTTAAGACGA-------
B:GAATAATCGT--CTGGCCAGTCCCTTTTTTGTAAAGAGDAAGC--CAAAAAATCCTTGTAAAAAACAGAAATAATTTTTGTT-----TTTTTTTAATTCGAAATATAT

MULE-23
A:GGGTTAATTTACT-GAATGACCAAATTTGACATCCTTATTAAGAGAATGACCTTCTCCCAGA--AAAGT-TGGGTATATAACGTTTTTTACCATGCGAAATAT-
B:GGGTATATGCACTAGA-TGACCA----TGAAATCAATAATAATATAGAAACTAACTATAAGTCTATAGTGTCTGTCGGAGACGTATCTGAAATTACCGCAACCCC
```

FIGURE 5.—Consensus sequences (100 terminal-most nucleotides) of the nine MULE groups. A conserved nucleotide was assigned when the aligned nucleotides exceeded 60% similarity. As only one element was identified for MULE-16, the terminal sequences of the original element are shown. The "A" sequences represent the terminal end upstream of the start codon of the *mudrA*-related ORFs and the "B" sequences represent the other terminal end. The double-underlined sequences represent the motifs found at the terminal-most ends while the underlined sequences represent the motifs found in individual non-TIR-MULE groups. Other shorter shared subterminal repeats may be present between the terminal regions but were not indicated.

Tan1 transposase[a]	535	RCSNCFNIGHRRTQ--CS	551
retrotransposon RT1 protein[b]	203	RCYRCLEHGHNARD--CR	219
gag-pol fusion polyprotein[c]	392	KCFNCGKEGHIARN--CR	408
germline RNA helicase-4[d]	639	PCRNCGQEGHFAKD--CQ	655
"DEAD" box helicases[e]	655	PCRNCGQEGHFAKD--CQ	671
gag-env fusion protein[f]	470	PCFKCGQLGHIRAQ--CR	486
zinc finger protein 9[g]	156	NCYRCGESGHLARE--CT	172
zinc finger protein[h]	188	TCHYCGELGHKANS--CK	204
splicing factor[i]	90	KCYECGETGHFARE--CR	106
SLU7 splicing factor[j]	20	FCRNCGEAGHKEKD--CM	361
MULE:GI5441872[k]	18805	RCSRCKGYGHNKAT--CK	18852
MULE-24:GI3319339	96586	TCSNCKQIGHNKGS--CK	96633
MULE-24:GI2760316	90038	TCSNCKEIGHNKGT--CK	89991
MULE-16:GI2443899	22930	HCKSCGEAGHNALR--CK	22977
MULE-9:GI3252804	42930	QCSRCRQAGHNKKT--CK	42883
MULE-9:GI4589411	35847	QCSRCRQAGHNKKT--CK	35894
MULE-9:GI3128140	59456	QCSRCRQAGHNKKT--CK	59409
MULE-9:GI6136349	10518	QCSRCRQAGHNKKM--CK	10565
MULE-9:GI4325365	47660	QCSRCRQAGHNKKT--CK	47613
MULE-23:GI3063438	91617	RCSRCTGAGHNRAT--CK	91664
MULE-23:GI3980374	43428	RCSRCTGAGHNRAT--CK	43475
MULE-23:GI2828187	21240	RCSRCTGAGHNRAT--CK	21287
MULE-23:GI5041964	21745	RCSRCTGAGHNRAT--CK	21792
MULE-23:GI6007863	116329	RCSRCTGSDHNRAT--CK	116376
MULE-23:GI4519197	68216	RCSRCTGA*HNRAT--CK	68169
MULE-2:GI5103850	8614	TCSNCLQEGHNKKS--CK	8567
MULE-1:GI3510344	40354	HCGVCGAADHNSRH--HK	40307
MULE-3:GI2832639	33301	HCGVCGAADHNSRH--HK	33348
MULE-27:GI4388816	30968	TCLNC*GEGHNKAG--CK	31015
MURA:GI2130141[l]	696	TCPNCGELGHRQSSYKCP	712

FIGURE 6.—Multiple alignment of **C**X2**C**X4**H**X4**C** motif in putative MURA-related transposases (derived using BLASTX) and representatives of known proteins. The amino acid sequences corresponding to MURA-related transposases were derived from virtual translations of MULE nucleotide sequences (position indicated). For the remaining proteins, amino acid positions are given. Asterisks represent positions where a frameshift was introduced to achieve optimum alignment. (a) *Aspergillus niger var. awamori* (GI1805251, NYYSSONEN *et al.* 1996); (b) African malaria mosquito (GI477117, BESANSKY *et al.* 1992); (c) human immunodeficiency virus (GI4107489, GAO 1998); (d–e) *Caenorhabditis elegans* (GI3386540, direct submission to GenBank; GI2773235, direct submission to GenBank); (f) Avian endogenous retrovirus (GI6048192, SACCO *et al.* 2000); (g) *Homo sapiens* (GI105602, RAJAVASHISTH *et al.* 1989); (h) *Drosophila melanogaster* (GI847869, direct submission to GenBank); (i) *A. thaliana* (GI2582645, LOPATO *et al.* 1999); (j) *Saccharomyces cerevisiae* (GI6320293, JACQ *et al.* 1997); (k) *O. sativa* (GI5441872, direct submission to GenBank); (l) *Zea mays* (GI2130141, HERSHBERGER *et al.* 1995).

the absence of long TIR structures and that they might be evolving as an independent lineage. Similar patterns of structural diversity have been observed in a family of unusual IS elements (such as IS901, IS116, and IS902; OHTSUBO and SEKINE 1996). These elements share a group of related transposases. However, they have variable terminal structures (with/without TIRs) and share little sequence similarity within terminal regions (MAHILLON and CHANDLER 1998).

Although the non-TIR-MULEs do not have long TIRs, members of individual groups do contain degenerate sequence motifs within their subterminal regions (Figure 5). Whether these motifs have any biological significance remains unknown. For some class II elements, transposition has been shown to involve transposase binding at sequence-specific recognition sites and the assembly of a transposase dimer (HAREN *et al.* 1999; DAVIES *et al.* 2000). The non-TIR-MULE subterminal motifs may correspond to transposase recognition sequences. Alternatively, the terminal regions may harbor different *cis*-factors for transposase binding. In this scenario, the mobilization of non-TIR-MULEs would require the assembly of heterodimeric transposase complexes.

Overall, we observed low sequence similarity between the terminal regions of members from different MULE groups. Except for the few nucleotides at the distant ends, no obvious sequence motif was identified to be highly conserved among all the consensus sequences. This sequence heterogeneity indicates that the binding sites for MURA-related transposases is most likely group specific in Arabidopsis. A similar case has been observed for members of the *Tc*1/*Mariner* family of transposons (PLASTERK 1996; VAN POUDEROYEN *et al.* 1997): each group shows high sequence similarity between members, but there is low sequence similarity between members of different groups.

We have identified a general motif, 5′-$R_{(1-4)}$-3′ followed by a short A + T-rich cluster, at both the terminal-most ends and the internal regions of most of the MULEs. This motif is similar to part of the sequence (5′-CGGGAAC<u>GGTAAA</u>-3′) located in the maize *Mu1* TIR that is recognized by host factors (ZHAO and SUNDARESAN 1991) and may be necessary for cleavage and strand transfer during transposition. In addition, this motif is reminiscent of a sequence (5′-G*D*TAAA-3′; *D* = G, T, or A) found in the subterminal regions of the maize *Ac* element, which were demonstrated to be the recognition sites for the binding of nuclear proteins in maize (BECKER and KUNZE 1996) and tobacco (LEVY *et al.* 1996). In fact, similar motifs have been recognized in a variety of class II plant TEs (LEVY *et al.* 1996). It is tempting to speculate that the motif identified in our study may function as a *cis*-acting sequence in the regulation of MULE activity.

We have also identified a **C**X2**C**X4**H**X4**C** motif at the C-terminal region of the majority of MURA-related proteins in Arabidopsis. This motif also exists in all known retroviruses (with the exception of spumaretroviruses; COVEY 1986; SCHWARTZ *et al.* 1997), many nucleic binding proteins (BERG 1986), and some retrotransposons, such as *copia*-like retrotransposons from tobacco (GRANDBASTIEN *et al.* 1989), and *Ty* elements in yeast (JORDAN and MCDONALD 1999). The **C**X2**C**X4**H**X4**C** motif has been demonstrated to interact with viral RNA (COVEY 1986; DARLIX *et al.* 1995), eukaryotic pre-mRNAs (FU 1993; HEIRICHS and BAKER 1995; LOPATO *et al.* 1999), and single-stranded DNA (RAJAVASHISTH *et al.* 1989; REMACLE *et al.* 1999). Given its RNA- and DNA-binding characteristics, the **C**X2**C**X4**H**X4**C** motif at the C-terminal region of the putative MURA-related proteins might interact with the MULE DNA or RNA, possibly playing a role in MULE transposition or the regulation of MULE mobility in *A. thaliana*.

It seems that acquisition of host DNA sequences to

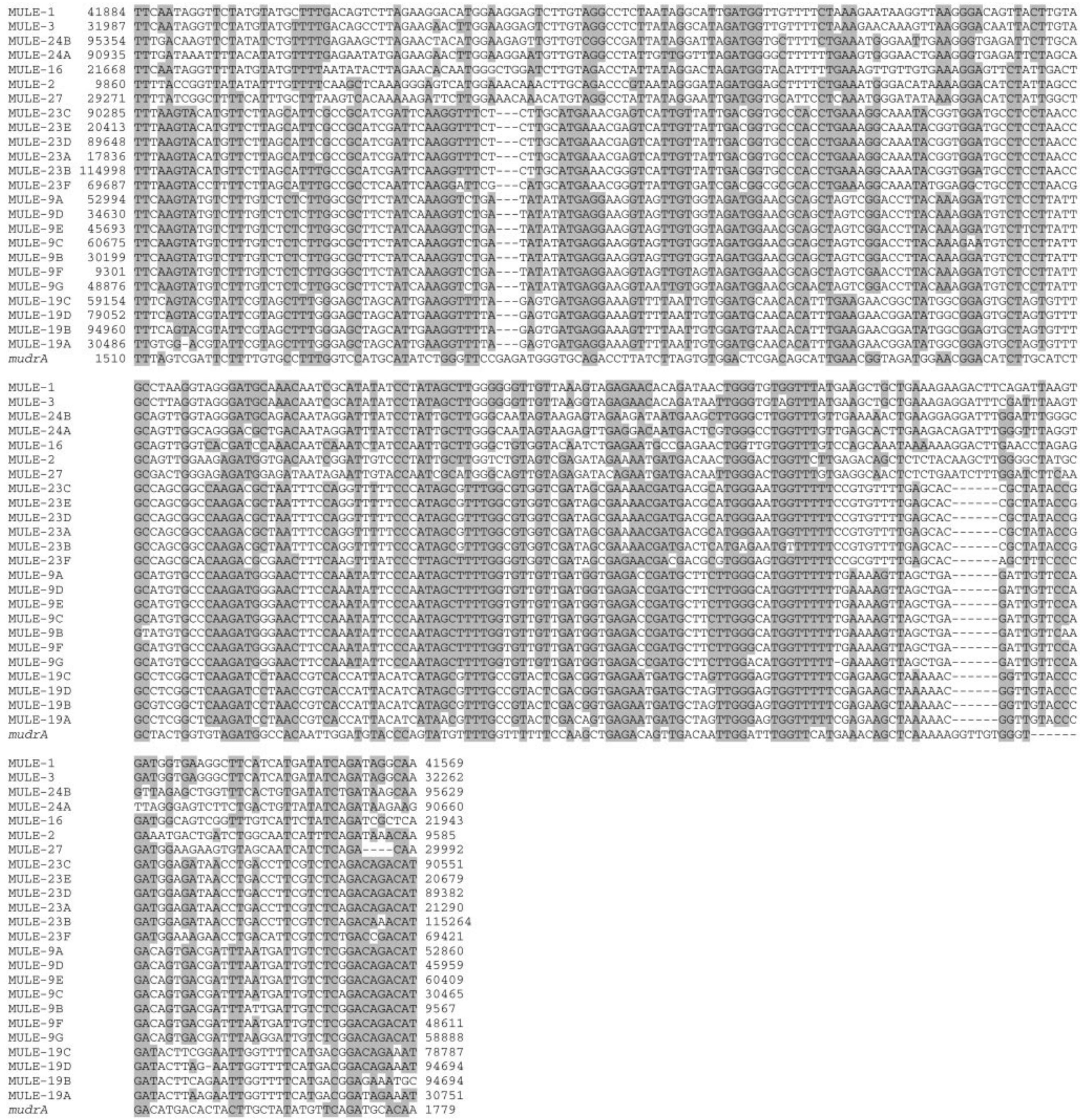


Figure 7.—Multiple alignment of the most conserved region between the Arabidopsis *mudrA*-related ORFs and the maize *mudrA* gene. Nucleotides sharing >60% similarity are shaded. The similarity was determined by the conservation mode of the program GeneDoc (Nicholas *et al.* 1997). The corresponding GI numbers for each MULE are as follows: MULE-1, 3510344; MULE-2, 5103850; MULE-3, 2832639; MULE-16, 2443899; MULE-24A, 2760316; MULE-24B, 3319339; MULE-27, 4388816; MULE-9A, 5672513; MULE-9B, 4185120; MULE-9C, 3128140; MULE-9D, 4589411; MULE-9E, 3252804; MULE-9F, 6136349; MULE-9G, 4325365; MULE-19A, 5041971; MULE-19B, 4585891; MULE-19C, 3242700; MULE-19D, 4914383; MULE-23A, 2828187; MULE-23B, 6007863; MULE-23C, 3063438; MULE-23D, 3980374; MULE-23E, 5041964; MULE-23F, 4519197. The beginning and end nucleotide positions in the corresponding clones are indicated for each sequence used in the alignment.

assemble new elements is a frequent event for TIR-MULEs. In addition to our documentation of five acquisition events in Arabidopsis, the maize *Mu2* has also been reported to have acquired a host MRS-A DNA segment (*Mu*-related sequence; Talbert and Chandler 1988; Talbert *et al.* 1989). These examples might suggest a common pathway in generation of the heterogeneity of MULE internal sequences. While the acquisition events by Arabidopsis TIR-MULEs involved the 5′ ends of cellular genes, the significance of this bias is

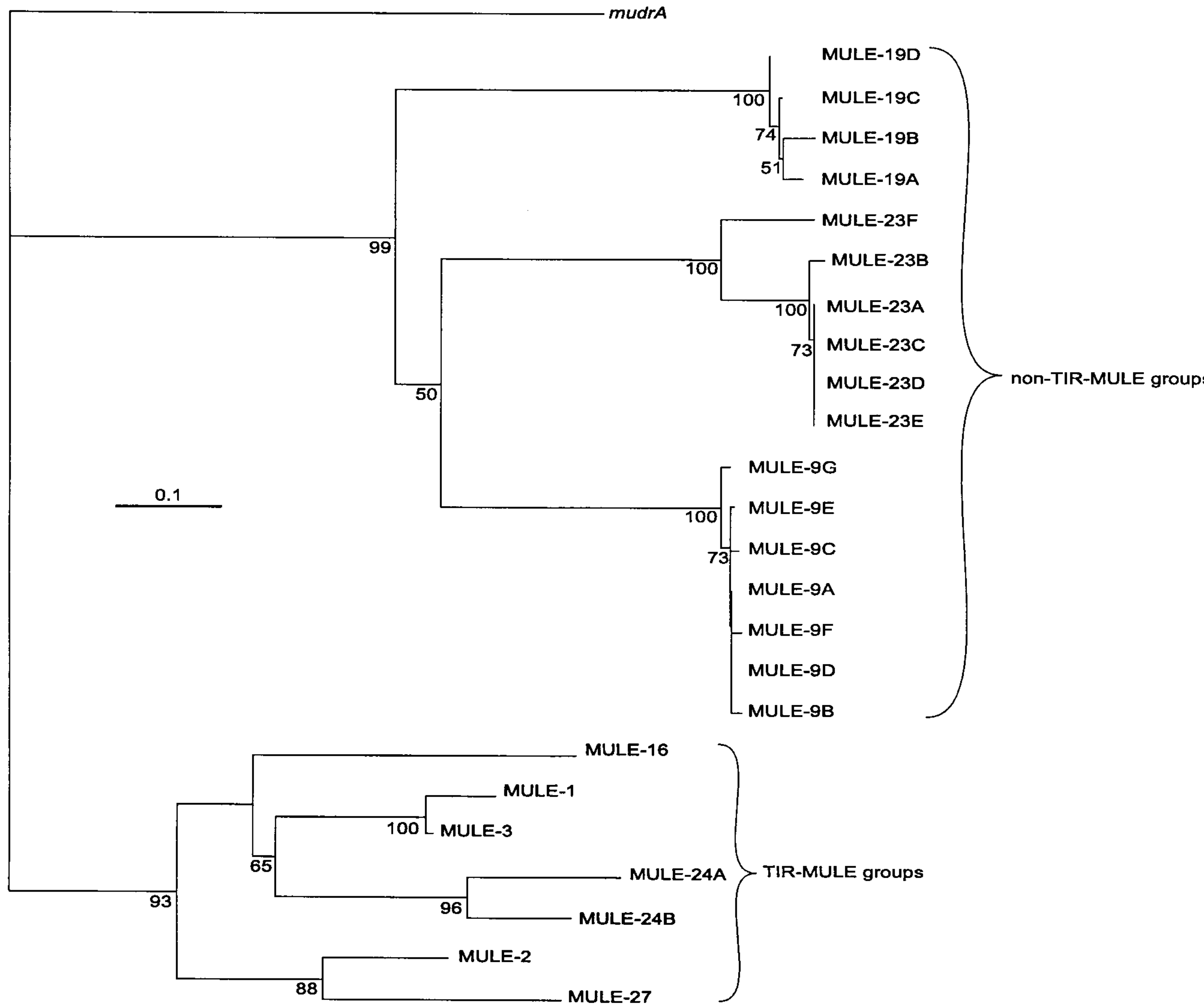


Figure 8.—A majority-rule and strict consensus tree of *mudrA*-containing MULE elements derived by the neighbor-joining method. The frequencies (>50%) of corresponding branches among 100 derived neighbor-joining trees are indicated. The corresponding GI numbers for each MULE are as indicated in the Figure 7 legend.

currently unknown. Acquisition of cellular genes does not appear to necessarily prevent transposition since two MULE-1 elements harboring *Athb-1* on different chromosomes have been identified. Class I elements have also been documented to acquire or transduce cellular genes (Bureau *et al.* 1994; Boeke and Stoye 1997). These genes can be expressed by means of an LTR promoter and in many cases lead to disease phenotypes (Vogt 1997). Likewise, acquired and modified host DNA within MULEs could be expressed from either a TIR-promoter, an acquired promoter, or a promoter in the flanking region. However, there is currently no evidence that the putative ORFs are actually expressed *in vivo* or whether these polypeptides have any function. While there is evidence for a lower level of divergence between the putative ORF and *Athb-1* in coding regions, it is unclear whether this pattern reflects selective constraint only on *Athb-1* or whether there are in fact functional constraints on the coding region of the MULE-1-related gene. The K_a/K_s ratio does not provide a strong indication of departure from neutral patterns, suggesting that the acquired exons may be nonfunctional. In addition to generating element diversity, the ability to capture sequences from hosts might be important in creating adaptive changes for MULE evolution. On the other hand, considering that genomic DNA segments captured by *Mu* elements and MULEs can transpose, likely be duplicated by means of replicative transposition, and recombine with sequences encoding functional domains, these elements might also play important roles in host gene organization and evolution (Henikoff *et al.* 1997).

The discovery of the *Mu* element family in maize involved the isolation and characterization of various members. In this study, we have characterized the sequence and structural diversity of MULEs in *A. thaliana*, thereby extending the range of the MULE superfamily. The apparent success of MULEs in the Arabidopsis genome provides an excellent opportunity for learning about the mechanisms driving the diversity and evolution of a class II TE system in eukaryotic genomes. The *Mu* element family in maize is a highly effective agent for the creation of *de novo* mutations. In fact, *Mu* element-tagging approaches have been extremely effective in the isolation and functional analysis of numerous maize genes (Walbot 1992; Maes *et al.* 1999). Introduction of active *Mu* elements into heterologous plant species, however, has not been successful (Walbot 1992). The identification and characterization of MULEs in species other than maize may therefore facilitate the development of novel element-tagging approaches.

The authors thank Julie Pourpart, Daniel J. Schoen, Anne Bruneau, Ken Hastings, and Ruying Chang for comments on our manuscript. We are also grateful to Quang Hien Le, Chris Olive, Newton Agrawal, and Boris-Antoine Legault for computer-related support. This work was funded by National Science and Engineering Research Council of Canada grants to T.E.B.

LITERATURE CITED

Altschul, S. W., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe *et al.*, 2000 The Pfam protein families database. Nucleic Acids Res. **28:** 263–266.

Becker, H., and R. Kunze, 1996 Binding site for maize nuclear proteins in the subterminal regions of the transposable element *Activator*. Mol. Gen. Genet. **251:** 428–435.

Benito, M.-I., and V. Walbot, 1997 Characterization of the maize *Mutator* transposable element MURA transposase as a DNA-binding protein. Mol. Cell. Biol. **17:** 5161–5175.

Bennetzen, J. L., 1996 The *Mutator* transposable element system of maize, pp. 195–229 in *Transposable Elements*, edited by H. Saedler and A. Gierl. Springer-Verlag, Berlin.

Berg, J. M., 1986 Potential metal-binding domains in nucleic acid binding proteins. Science **232:** 485–487.

Besansky, N. J., S. M. Paskewitz, D. M. Hamm and F. H. Collins, 1992 Distinct families of site-specific retrotransposons occupy identical positions in the rRNA genes of *Anopheles gambiae*. Mol. Cell. Biol. **12:** 5102–5110.

Boeke, J. D., and J. P. Stoye, 1997 Retrotransposons, endogenous retroviruses and the evolution of the retroelements, pp. 343–436 in *Retroviruses*, edited by J. M. Coffin, S. H. Hughes and H. E. Varmus. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bureau, T. E., and S. R. Wessler, 1994 Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. USA **91:** 1411–1415.

Bureau, T. E., S. E. White and S. R. Wessler, 1994 Transduction of a cellular gene by a plant retroelement. Cell **77:** 479–480.

Chandler, V. L., and K. J. Hardeman, 1992 The *Mu* elements of *Zea mays*. Adv. Genet. **30:** 77–122.

Comeron, J., 1995 A method for estimating the numbers of synonymous and nonsynonmous substitution per site. J. Mol. Evol. **41:** 1152–1159.

Comeron, J. M., M. Kreitman and M. Aguade, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

Corpet, F., 1988 Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. **16:** 10881–10890.

Covey, S. N., 1986 Amino acid sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. Nucleic Acids Res. **14:** 623–633.

Darlix, J. L., M. Lapadat, H. Tapolsky, H. De Rocquigny and B. P. Roques, 1995 First glimpses at structure-function relationships of nucleocapsid protein of retroviruses. J. Mol. Biol. **254:** 523–537.

Davies, D. R., I. Y. Goryshin, W. S. Reznikoff and I. Rayment, 2000 Three-dimensional structure of the *Tn5* synaptic complex transposition intermediate. Science **289:** 77–85.

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

Felsenstein, J., 1993 PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Fu, X.-D., 1993 Specific commitment of different pre-mRNAs to splicing by single SR proteins. Nature **365:** 82–85.

Gao, F., 1998 A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J. Virol. **72:** 5680–5698.

Garrett, J. E., D. S. Knutzon and D. Carroll, 1989 Composite transposable elements in the *Xenopus laevis* genome. Mol. Cell. Biol. **9:** 3018–3027.

Grandbastien, M.-A., A. Spielmann and M. Caboche, 1989 Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature **337:** 376–380.

Gribskow, M., A. D. Mclachlan and D. Eisenberg, 1987 Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. USA **84:** 4355–4358.

Haren, L., B. Ton-Hoang and M. Chandler, 1999 Integrating DNA: transposases and retroviral integrases. Annu. Rev. Microbiol. **53:** 245–281.

Heirichs, V., and B. S. Baker, 1995 The *Drosophila* SR protein RBP1 contributes to the regulation of *Doublesex* alternative splicing by recognizing RBP1 RNA target sequences. EMBO J. **14:** 3987–4000.

Henikoff, S., E. A. Greene, S. Pietrokovski, P. Bork, T. K. Attwood *et al.*, 1997 Gene families: the taxonomy of protein paralogs and chimeras. Science **278:** 609–614.

Hershberger, R. J., C. A. Warren and V. Walbot, 1991 *Mutator* activity in maize correlates with the presence and expression of the *Mu* transposable element *Mu*9. Proc. Natl. Acad. Sci. USA **88:** 10198–10202.

Hershberger, R. J., M.-I. Benito, K. Hardeman, C. Warren, V. L. Chandler *et al.*, 1995 Characterization of the major transcripts encoded by the regulatory MuDR transposable element of maize. Genetics **140:** 1087–1098.

Jacq, C., J. Alt-Morbe, B. Andre, W. Arnold, A. Bahr *et al.*, 1997 The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV. Nature **387** (6632 Suppl.): 75–78.

Jordan, I. K., and J. F. McDonald, 1999 Tempo and mode of *Ty* element evolution in *Saccharomyces cerevisiae*. Genetics **151:** 1341–1351.

Le, Q.-H., S. I. Wright, Z.-H. Yu and T. E. Bureau, 2000 Transposon discovery in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **97:** 7376–7381.

Levy, A. A., M. Fridlender, U. H. E. Rubin and Y. Sitrit, 1996 Binding of Nicotiana nuclear proteins to the subterminal regions of the *Ac* transposable element. Mol. Gen. Genet. **251:** 436–441.

Lin, X., S. Kaul, S. Roundsley, T. P. Shea, M.-I. Benito *et al.*, 1999 Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature **402:** 761–768.

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

Lopato, S., R. Gattoni, G. Fabini, J. Stevenin and A. Barta, 1999 A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. Plant Mol. Biol. **39:** 761–773.

Maes, T., P. De Keukeleire and T. Gerats, 1999 Plant tagnology. Trends Plant Sci. **4:** 90–96.

Mahillon, J., and M. Chandler, 1998 Insertion sequences. Microbiol. Mol. Biol. Rev. **62:** 725–774.

Mayer, K., C. Schuller, R. Wambutt, G. Murphy, G. Volckaert *et al.*, 1999 Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature **402:** 769–777.

Meinke, D. W., J. M. Cherry, C. Dean, S. D. Rounsley and M. Koornneef, 1998 *Arabidopsis thaliana*: a model plant for genome analysis. Science **282:** 662–682.

Morgenstern, B., 1999 DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15:** 211–218.

Nicholas, K. B., H. B. Nicholas, Jr. and D. W. Deerfield, 1997 GeneDOC: analysis and visualization of genetic variation. EMBNEW.NEWS **4:** 14.

Nyyssonen, E., M. Amutan, L. Enfield, J. Stubbs and N. S. Dunn-Coleman, 1996 The transposable element *Tan1* of *Aspergillus niger* var. awamori, a new member of the *Fot1* family. Mol. Gen. Genet. **253:** 50–56.

Ohtsubo, E., and Y. Sekine, 1996 Bacterial insertion sequences, pp. 1–26 in *Transposable Elements*, edited by H. Saedler and A. Gierl. Springer-Verlag, Berlin.

Plasterk, R. H. A., 1996 The *Tc*1/*Mariner* transposon family, pp. 125-144 in *Transposable Elements*, edited by H. Saedler and A. Gierl. Springer-Verlag, Berlin.

Rajavashisth, T. B., A. K. Taylor, A. Analibi, K. L. Svenson and L. J. Lusis, 1989 Identification of a zinc finger protein that binds to the sterol regulatory elements. Science **245:** 640–643.

Remacle, J. E., H. Kraft, W. Lerchner, G. Wuytens, C. Collart *et al.*, 1999 New mode of DNA binding of multi-zinc finger transcription factors: δEF1 family members bind with two hands to two target sites. EMBO J. **18:** 5073–5084.

Robertson, D. S., 1978 Characterization of a *Mutator* system in maize. Mutat. Res. **51:** 21–28.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Ruberti, I., G. Sessa, S. Lucchetti and G. Morelli, 1991 A novel class of plant proteins containing a homeodomain with a closely linked leucine zipper motif. EMBO J. **10:** 1787–1791.

Sacco, M. A., D. M. Flannery, K. Howes and K. Venugopal, 2000 Avian endogenous retrovirus EAV-HP shares regions of identity with avian leukosis virus subgroup J and the avian retrotransposon ART-CH. J. Virol. **74:** 1296–1306.

Schena, M., and R. W. Davis, 1994 Structure of homeobox-leucine zipper genes suggests a model for the evolution of gene families. Proc. Natl. Acad. Sci. USA **91:** 8393–8397.

Schwartz, M., D. Fiore and A. T. Panganiban, 1997 Distinct functions and requirements for the Cys-His boxes of the human immunodeficiency virus type 1 nucleocapsid protein during RNA encapsidation and replication. J. Virol. **71:** 9295–9305.

Talbert, L. E., and V. L. Chandler, 1988 Characterization of a highly conserved sequence related to *Mutator* elements in maize. Mol. Biol. Evol. **5:** 519–529.

Talbert, L. E., G. I. Patterson and V. L. Chandler, 1989 *Mu* transposable elements are structurally diverse and distributed throughout the genus Zea. J. Mol. Evol. **29:** 28–39.

Tatusova, T. A., and T. L. Madden, 1999 BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. **174:** 247–250.

Thomson, K. G., J. E. Thomas and R. G. Dietzgen, 1998 Retrotransposon-like sequences integrated into the genome of pineapple, *Ananas comosus*. Plant Mol. Biol. **38:** 461–465.

Van Pouderoyen, G. V., R. F. Ketting, A. Perrakis, R. H. A. Plasterk and T. K. Sixma, 1997 Crystal structure of the specific DNA-binding domain of *Tc3* transposase of *C. elegans* in complex with transposon DNA. EMBO J. **16:** 6044–6054.

Vogt, V. M., 1997 Retroviral virions and genomes, pp. 27–70 in *Retroviruses*, edited by J. M. Coffin, S. H. Hughes and H. E. Varmus. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Walbot, V., 1992 Strategies for mutagenesis and gene cloning using transposon tagging and T-DNA insertional mutagenesis. Annu. Rev. Plant Physiol. Plant Mol. Biol. **43:** 49–82.

Zhao, Z-Y., and V. Sundaresan, 1991 Binding sites for maize nuclear proteins in the terminal inverted repeats of the *Mu*1 transposable element. Mol. Gen. Genet. **229:** 17–26.

Zuker, C., J. Cappello, H. F. Lodish, P. George and S. Chung, 1984 Dictyostelium transposable element DIRS-1 has 350-base-pair inverted terminal repeats that contain a heat shock promoter. Proc. Natl. Acad. Sci. USA **81:** 2660–2664.

Communicating editor: J. A. Birchler