FARE, a New Family of Foldback Transposons in Arabidopsis

Aaron J. Windsor and Candace S. Waddell

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

Manuscript received June 23, 2000
Accepted for publication August 16, 2000

ABSTRACT

A new family of transposons, *FARE*, has been identified in Arabidopsis. The structure of these elements is typical of foldback transposons, a distinct subset of mobile DNA elements found in both plants and animals. The ends of *FARE* elements are long, conserved inverted repeat sequences typically 550 bp in length. These inverted repeats are modular in organization and are predicted to confer extensive secondary structure to the elements. *FARE* elements are present in high copy number, are heterogeneous in size, and can be divided into two subgroups. *FARE*1's average 1.1 kb in length and are composed entirely of the long inverted repeats. *FARE*2's are larger, up to 16.7 kb in length, and contain a large internal region in addition to the inverted repeat ends. The internal region is predicted to encode three proteins, one of which bears homology to a known transposase. *FARE*1.1 was isolated as an insertion polymorphism between the ecotypes Columbia and Nossen. This, coupled with the presence of 9-bp target-site duplications, strongly suggests that *FARE* elements have transposed recently. The termini of *FARE* elements and other foldback transposons are imperfect palindromic sequences, a unique organization that further distinguishes these elements from other mobile DNAs.

TRANSPOSABLE elements (TEs) are ubiquitous features of prokaryotic and eukaryotic genomes and have been implicated in a host of phenomena associated with genome restructuring. TEs are generally categorized by their mode of transposition; class I elements transpose *via* an RNA intermediate while class II elements transpose through a DNA intermediate. The class II elements fall into two major subgroups: the terminal inverted repeat (TIR) elements and the long inverted repeat (IVR) elements, also known as the foldback transposons (FTs). The majority of characterized class II elements are TIR elements. These transposons are defined by their termini, which are short, perfect (or nearly perfect) inverted repeats generally 10–40 nucleotides in length. The internal sequences of the TIR elements encode one or more proteins involved in transposition, including the transposase. TIR elements can be autonomous or nonautonomous. Generally, nonautonomous elements have functional end sequences but do not encode a functional transposase.

The FTs are a group of elements with specific structural characteristics that distinguish them from the TIR elements. As their name implies, the FT elements are capable of forming extensive secondary structure as a consequence of the sequences contained in their ends (Potter *et al.* 1980). The ends of FT elements are large, modular, imperfect IVRs that range in size from ~300 bp to several kilobases. Most FTs consist entirely of their IVR ends and contain no protein coding sequences. To date, few FTs have been identified and the best characterized of these are the *FB* elements of *D. melanogaster* (Bingham and Zachar 1989; Potter *et al.* 1989) and the TU elements of *Strongylocentrotus purpuratus* (Hoffman-Liebermann *et al.* 1985, 1989).

The *D. melanogaster FB* element ends are IVRs that contain three modules, each of which is composed of multiple copies of a short repeated sequence in direct orientation. The size of the *FB* IVRs is variable, even within a single element (Truett *et al.* 1981; Potter 1982). This heterogeneity is attributed to ectopic recombination and/or DNA polymerase slippage between the short, tandemly arranged sequences in the IVRs. While most *FB* elements are composed solely of the IVRs, a limited number of elements, the *FB-NOF*s, contain a 4-kb internal region that is predicted to encode one to three proteins. *FB* transposition is dependent on the presence of an intact *FB-NOF* element (Harden and Ashburner 1990), and one of the predicted proteins, the product of open reading frame (ORF)1, has been shown to be expressed (Smyth-Templeton and Potter 1989). It is assumed that *FB* elements transpose through a DNA intermediate, but no specifics are known about the transposition mechanism.

Within the last few years, the first FTs have also been identified in the plant kingdom. The *So*FT1 element was initially isolated from tomato and displays all the features characteristic of FTs (Rebatchouk and Narita 1997). *So*FT1 has no protein coding capacity and nothing is known about the element's mechanism of transposition. It is assumed, however, that the element has transposed as it was identified as a polymorphism, and the

Corresponding author: Candace S. Waddell, Department of Biology, McGill University, 1205 Dr. Penfield Ave., Montreal, Quebec H3A 1B1, Canada. E-mail: candace_waddell@maclan.mcgill.ca

element is flanked by a target-site duplication. Related elements are also present in the genomes of several solanaceous plant species. A second family of plant FTs, *Hairpin* elements, has been reported in Arabidopsis (Adé and Belzile 1999). While predicted to form foldback secondary structure, these elements are not typical of the FTs. The IVRs of *Hairpin* elements are quite small (*ca.* 110 nucleotides), the ends lack modular organization, and the elements are very homogeneous in size. Based upon their published structure, it is possible that *Hairpin* elements may actually represent a novel class of miniature inverted-repeat transposable elements (MITEs; Bureau and Wessler 1992; Bureau *et al.* 1996) or extreme deletion derivatives of a larger, currently uncharacterized TIR element.

In this work we describe a new family of FTs in *Arabidopsis thaliana*, which we have designated *FARE* (Foldback Arabidopsis Repeat Element). *FARE* elements have the potential to form extensive secondary structure based on the presence of large, modular, imperfect IVR ends. The general organization of the *FARE* elements is most like that of the *FB* family of foldback transposons. *FARE* elements have transposed in recent evolutionary time, as evidenced by the fact that the first *FARE* was identified as a sequence polymorphism between the Columbia and Nossen ecotypes. One class of *FARE* elements, *FARE*2, is predicted to have protein coding capacity. Three proteins are encoded by these elements and their sequence suggests that at least one may possess transposase activity.

MATERIALS AND METHODS

Arabidopsis lines: Line16-10C is in the Nossen (No-0) ecotype background. Seed for all other ecotypes utilized in this study, except Rschew (RLD1), was obtained from the Arabidopsis Biological Resource Center, Ohio State University. RLD1 seed was obtained from Lehle Seeds, Round Rock, TX.

Plant material and DNA isolation: Arabidopsis genomic DNA was prepared according to Dellaporta *et al.* (1983) from between 200 and 400 mg of whole seedlings pooled from liquid cultures. Recovered genomic DNA was suspended in 200 μl of 10 mm Tris-Cl, pH 7.6. Liquid cultures were grown as follows. Seeds were surface sterilized in a solution of 1.5% sodium hypochlorite and 0.02% SDS for 5 min, rinsed three times with sterile, double-distilled water, and stratified for 4 days at 4° in the absence of light. Seeds were germinated and grown at a density of one seed per milliliter in 50 ml of liquid GM media in 250-ml flasks. GM media is composed of 1× MS basal salts, 1% sucrose, 0.5 g/liter MES (2-[*N*-morpholino] ethanesulfonic acid), 1 mg/liter thiamine, 0.5 mg/liter pyridoxin, 0.5 mg/liter nicotinic acid, 100 mg/liter myoinositol, with pH adjusted to 5.7 with 1 n KOH (Valvekens *et al.* 1988). Liquid cultures were grown at 24° for 2 wk on an orbital shaker (120 rpm) under continuous light at a photon flux of ~90 μmol/m^2/sec.

PCR and molecular analysis: PCR was performed in a Perkin-Elmer (Norwalk, CT) DNA Thermal Cycler 480 in a total volume of 100 μl using standard conditions for Pharmacia (Piscataway, NJ) *Taq* polymerase. PCR primers were obtained from BioCorp, Inc. (Montreal). A 5.0-μl aliquot from each genomic DNA preparation was used for PCR. Negative controls, lacking template DNA, were run for all reactions. PCR products were visualized on agarose gels.

Amplification of the *FARE*1.1 insertion site was carried out utilizing primers rtyBP-AW9 (5′-ATAGTTGACCCACTAGA CGC-3′) and rtyBP-AW3 (5′-TCTTTCTCAAGTAAGTATTAG GTC-3′), which correspond to positions 29914–29933 and 34004–33981 of accession AC007048, respectively. Samples were incubated at 94° for 2 min followed by 35 cycles of 94° for 10 sec, 51° for 30 sec, and 68° for 3 min. At the end of 25 cycles, an additional 2.5 units of *Taq* polymerase was added to each reaction. As a final step, the samples were incubated at 72° for 10 min. Products from three independent No-0 reactions were purified for subcloning using the QIAEX II gel extraction system (QIAGEN, Valencia, CA) according to the manufacturer's instructions. The purified No-0 products were TA-cloned into the *Eco*RV site of pBluescript II SK(−). T-tailed pBluescript II SK(−) was prepared by incubating *Eco*RV-digested vector at 72° for 20 min in the presence of 1× Pharmacia PCR buffer, 0.5 mm dTTP, and 1.0 unit Pharmacia *Taq* polymerase. *Taq* polymerase was heat-inactivated by incubation at 98° for 10 min. Ligations and transformations of *Escherichia coli* were carried out according to standard protocols (Ausubel *et al.* 1996). Six independent subcloned No-0 products, two from each starting PCR reaction, were sequenced using the SequiTherm EXCEL II DNA Sequencing Kit-LC (Epicenter Technologies, Madison, WI) with M13-forward and reverse primers according to the manufacturer's instructions. Sequencing reactions were run and read on a Li-Cor LONG READIR 4200 automated sequencing apparatus as described by the manufacturer.

The primers ArgoIR-L#1 (5′-GAAAAATTCTTTCTAAT GCC-3′) and ArgoIR-L#2 (5′-GTTAACTTAAAACAATTTCC-3′) were used to amplify the terminal 135 bp of the *FARE*1 left ends. A 294-bp fragment from the *FARE*2 internal region, corresponding to exon 1 of *CDS*3, was amplified using the primers ORF3_15867 (5′-CTCTCTCAAGGAGAAACGG-3′) and ORF3_16160 (5′-GAACAAATCTACAGAGAAGG-3′). Reactions were incubated at 94° for 2 min followed by 30 cycles of 94° for 10 sec, 45° for 15 sec, and 68° for 15 sec (*FARE*1 left end reaction) or for 30 sec (*FARE*2 *CDS*3 reaction). The *FARE*1 left end and *FARE*2 *CDS*3 products were subjected to *Dra*I or *Hin*dIII digestion, respectively, to verify their specificity.

The 294-bp *CDS*3 PCR product from Columbia (Col-0) was used in Southern analysis to probe for the presence of *FARE*2 elements in Col-0, No-0, L*er* (Landsberg *erecta*), Ws (Wassilewskija), RLD1 (Rschew), and En-2 (Enkheim). After purification with QIAEX II, the probe fragment was labeled with [α-^{32}P]dCTP (New England Nuclear Life Science Products Boston) utilizing the Multiprime DNA Labeling System (Amersham Pharmacia Biotech) according to the manufacturer's instructions. Genomic DNA samples (700 ng per sample) were digested overnight with *Eco*RI and fractionated on a 30 cm, 0.7% agarose gel for 900 Vhr. DNA was transferred to a GeneScreen Plus hybridization transfer membrane (New England Nuclear Life Science Products) according to the manufacturer's instructions and hybridized to the *CDS*3 probe at 65°. The blot was washed at 65°, twice in 0.2× SSC, 0.1% SDS and once in 0.1× SSC, 0.1% SDS.

DNA sequence analysis: All computer-assisted sequence analyses were carried out utilizing the following on-line resources:

NCBI BLAST, http://www.ncbi.nlm.nih.gov/BLAST/ (Altschul *et al.* 1997)
NCBI Entrez, http://www.ncbi.nlm.nih.gov/Entrez/nucleotide.html
GeneBee service, http://www.genebee.msu.su/
GENSCAN, http://CCR-081.mit.edu/GENSCAN.html (Burge and Karlin 1997)

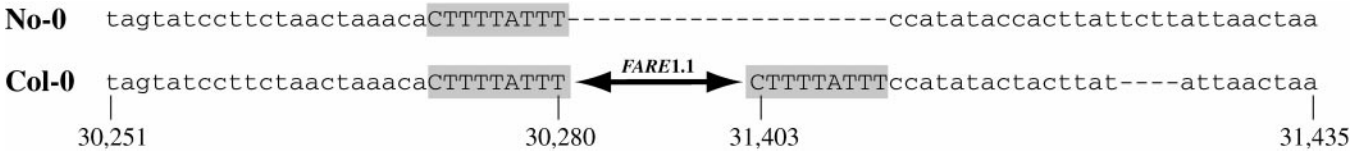


FIGURE 1.—The *FARE*1.1 integration site in Columbia (bottom sequence, Col-0). The top sequence (No-0) is the homologous region from Nossen. Numbers indicate nucleotide positions in accession no. AC007048 (chromosome II, section 118). Shaded boxes indicate the *FARE*1.1 9-bp target-site duplication. The double arrow, labeled *FARE*1.1, represents the 1122-bp *FARE*1.1 insertion in Col-0.

ProfileScan Server, http://www.isrec.isb-sib.ch/software/PFSCAN_form.html
mfold, http://mfold2.wustl.edu/~mfold/dna/form1.cgi.

Database searches were performed with low complexity filters turned off. All other settings were the default values for the Advanced BLAST utility. At the time of writing (June 6, 2000), 622,873 sequences, or 1,987,066,614 total letters, were available on the database. DNA folding predictions by mfold make use of the free energies determined by SANTALUCIA (1998) and the salt correction established by J. SANTALUCIA JR., M. ZUKER, A. BOMMARITO and R. J. IRANI (unpublished results).

RESULTS

Identification of *FARE*1.1: The first member of the *FARE* family of transposable elements, *FARE*1.1, was initially identified during the characterization of a genomic rearrangement in *A. thaliana* (A. J. WINDSOR and C. S. WADDELL, unpublished results). While sharing no causal relationship with the rearrangement, *FARE*1.1 represents one component of a sequence polymorphism between the ecotypes Columbia (Col-0) and Nossen (No-0) in section 118 of chromosome II (AC007048). We sequenced the No-0 region and compared it to the homologous Col-0 sequence in GenBank. This comparison revealed the presence of *FARE*1.1, a 1122-bp insertion, in Col-0 whose structure is very similar to that of transposons. The element displays an inverted repeat organization that is composed of a complex pattern of tandemly repeated sequences. In addition, it is flanked by a 9-bp direct repeat that is present in No-0 in an unduplicated form (Figure 1). The production of target-site duplications is a hallmark of TE integration. Analysis of the *FARE*1.1 sequence demonstrates that the element is A-T rich (73%) and does not contain any ORFs, suggesting that the element does not represent an autonomous TE.

***FARE*1.1 is composed of modular inverted repeats:** The ends of *FARE*1.1 are primarily composed of three distinct domains that are defined by the presence of different repeating units. The arrangement of these repeats results in the formation of an element with two symmetrical halves, the single-stranded version of which is predicted to have striking secondary structure (Figure 2). Thus, the ends of *FARE*1.1 appear to be composed of two large, imperfect, inverted repeats. The alignment of the first 561 nucleotides of the element, or the left end (L-end), with the reverse complement of the final 561 nucleotides, or the right end (R-end), demonstrates the modular nature of the *FARE*1.1 ends (Figure 3). Note that the terminal 16 nucleotides of each end do not contribute to this organization and are not convincing inverted repeats (Figures 2 and 3). Domain I, which spans positions 17–113 in the L-end and positions 1106–1009 in the R-end, is made up of six directly oriented repeats of the consensus sequence, $TAC_{(3)}T_{(4)}$. While the spacing of the repeats is conserved between the L- and R-ends, the intervening sequences separating the repeat units are more variable in base composition (Figure 3).

A variable region of 31 nucleotides on the L-end and 30 nucleotides on the R-end separates domain I of *FARE*1.1 from domain II (Figure 3). Domain II extends from position 145 to position 312 and from position 978 to position 825 in the L- and R-ends, respectively. It is composed of seven or eight direct repeats of the consensus sequence, $T_{(3–5)}C_{(3)}GCCA_{(3–5)}$, arranged as arrays in each end of the element. The L-end contains one additional copy of this repeat (positions 299–312) that is in inverted orientation relative to all other L-end domain II repeats (Figure 3).

The most internal region of *FARE*1.1, or domain III, extends from position 315 to position 561 in the L-end and from positions 823 to 572 in the R-end of the element (Figure 3). Domain III displays the highest

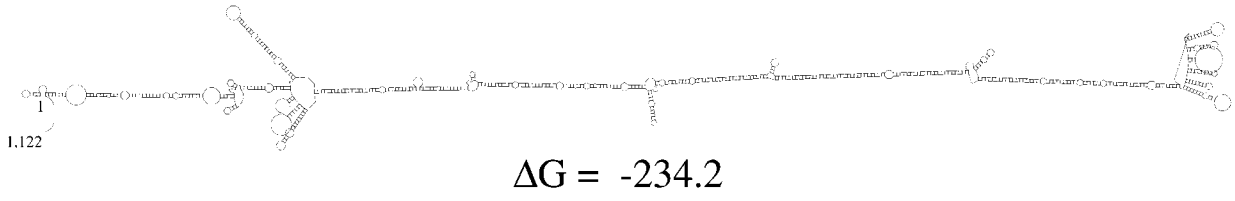
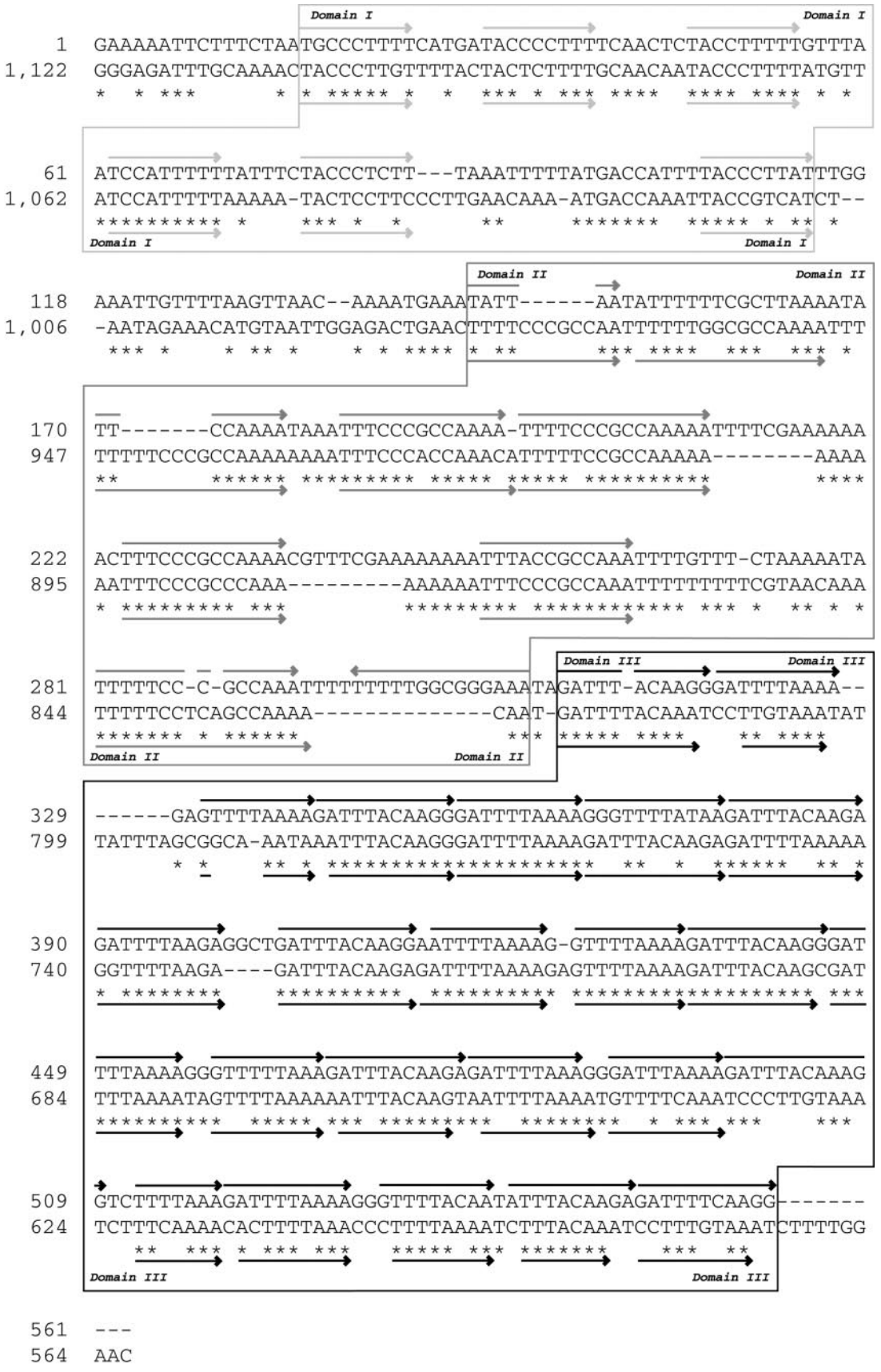


FIGURE 2.—The predicted single-stranded DNA secondary structure of a *FARE*1 element. The terminal 120 nucleotides from each end of *FARE*1.1 have been replaced with the *FARE*1 consensus sequences. The first (1) and terminal (1122) nucleotides are indicated. The ΔG value is the free energy of the system and a measurement of the structure's stability.


Domain I
Domain I
1 GAAAAATTCTTTCTAATGCCCTTTTCATGATACCCCTTTTCAACTCTACCTTTTTGTTTA
1,122 GGGAGATTTGCAAAACTACCCTTGTTTTACTACTCTTTTGCAACAATACCCTTTTATGTT
61 ATCCATTTTTTATTTCTACCCTCTT---TAAATTTTTATGACCATTTTACCCTTATTTGG
1,062 ATCCATTTTTAAAAA-TACTCCTTCCCTTGAACAAA-ATGACCAAATTACCGTCATCT--
Domain I
Domain I
Domain II
Domain II
118 AAATTGTTTTAAGTTAAC--AAAATGAAATATT------AATATTTTTTCGCTTAAAATA
1,006 -AATAGAAACATGTAATTGGAGACTGAACTTTTCCCGCCAATTTTTTGGCGCCAAAATTT
170 TT-------CCAAAATAAATTTCCCGCCAAAA-TTTTCCCGCCAAAAATTTTCGAAAAAA
947 TTTTTCCCGCCAAAAAAAATTTCCCACCAAACATTTTTCCGCCAAAAA--------AAAA
222 ACTTTCCCGCCAAAACGTTTCGAAAAAAAATTTACCGCCAAATTTTGTTT-CTAAAAATA
895 AATTTCCCGCCCAAA---------AAAAAATTTCCCGCCAAATTTTTTTTTCGTAACAAA
Domain III
Domain III
281 TTTTTCC-C-GCCAAATTTTTTTTTGGCGGGAAATAGATTT-ACAAGGGATTTTAAAA--
844 TTTTTCCTCAGCCAAAA--------------CAAT-GATTTTACAAATCCTTGTAAATAT
Domain II
Domain II
329 ------GAGTTTTAAAAGATTTACAAGGGATTTTAAAAGGGTTTTATAAGATTTACAAGA
799 TATTTAGCGGCA-AATAAATTTACAAGGGATTTTAAAAGATTTACAAGAGATTTTAAAAA
390 GATTTTAAGAGGCTGATTTACAAGGAATTTTAAAAG-GTTTTAAAAGATTTACAAGGGAT
740 GGTTTTAAGA----GATTTACAAGAGATTTTAAAAGAGTTTTAAAAGATTTACAAGCGAT
449 TTTAAAAGGGTTTTTAAAGATTTACAAGAGATTTTAAAGGGATTTAAAAGATTTACAAAG
684 TTTAAAATAGTTTTAAAAAATTTACAAGTAATTTTAAAATGTTTTCAAATCCCTTGTAAA
509 GTCTTTTAAAGATTTTAAAAGGGTTTTACAATATTTACAAGAGATTTTCAAGG-------
624 TCTTTCAAAACACTTTTAAACCCTTTTAAAATCTTTACAAATCCTTTGTAAATCTTTTGG
Domain III
Domain III
561 ---
564 AAC

complexity of the *FARE*1.1 domains, being composed of two related repeat units. The consensus sequence of the first repeat is $GAT_{(3)}ACAAG^{G}/_{A}$ and that of the second is $G^{A}/_{-}T_{(4)}A_{(4)}$. The repeats are arranged in direct orientation with few intervening sequences. The organization of domain III is highly conserved between the ends of *FARE*1.1.

***FARE*1.1 is a member of a multiple copy family of elements in the Col-0 genome:** To determine if *FARE*1.1 is a unique element in the genome or if it is a member of a multicopy family, the terminal 50 nucleotides from each end of the element were queried against the available Arabidopsis sequence. No elements identical to *FARE*1.1 were identified, but many highly related sequences reside in the Arabidopsis genome. At the time of writing, 83 intact or nearly intact *FARE* elements have been identified whose terminal 50 nucleotides share >85% identity with those of *FARE*1.1. This result demonstrates that *FARE*1.1 belongs to a multicopy family of elements with conserved ends in Arabidopsis (Table 1). We have designated this group *FARE*1. While *FARE*1 ends are highly conserved, the elements themselves are heterogeneous in size.

Twenty-seven elements with completely intact termini were randomly selected for further investigation. The first 120 nucleotides of each end, which includes domain I and the 16 terminal nucleotides, are highly conserved. The elements display an average identity of 91% for the L-end and 93% for the R-end when compared to the *FARE*1 consensus ends (Table 1). *FARE*1's are small, ranging in size from 0.5 to 2.6 kb (Table 1), and all have an A-T content of 75% (±2%). While no two *FARE*1's are identical, the organization of their IVRs is well conserved and none of the elements manifest inherent protein coding capacity. The observed variability in size is correlated with expansions and contractions of the highly repetitive domains II and III (data not shown).

Notable exceptions to the generalities of *FARE*1 structure are *FARE*1.32, *FARE*1.33, and *FARE*1.34 (Table 1), each of which contains a large insertion representing different TEs. *FARE*1.33 and *FARE*1.34 harbor remnant retroelements and *FARE*1.32 is disrupted by a putative *Ac*-like element (data not shown). The TE insertions are present at different positions within the host *FARE*1's. Further, the insertions break the symmetry of the host elements, thereby disrupting the predicted secondary structure of these *FARE*1's. These observations argue that *FARE*1.32, *FARE*1.33, and *FARE*1.34 were targets for TE integration in the past and that any coding capacity conferred by the inserted elements is unrelated to *FARE*1 function. Apart from the insertions, *FARE*1.32, *FARE*1.33, and *FARE*1.34 share all of the features of *FARE*1 elements.

***FARE*2's are a second related group of elements:** Our database queries also identified a second distinct group of elements that are related to *FARE*1, which we have designated *FARE*2. At the time of writing, 13 *FARE*2's with intact or mostly intact ends have been identified and 8 of these elements have been characterized in detail (Table 1). These elements are much larger than the *FARE*1's, ranging in size from 8.5 to 16.7 kb (Table 1) with an A-T content of 67% (±1%). While sharing features with *FARE*1's, *FARE*2's are distinct and the ends of these elements differ from those of *FARE*1's at both the structural and nucleotide levels. Unlike *FARE*1's, *FARE*2's are not composed solely of IVR sequences; rather, these elements have clearly definable ends separated by a large internal region that harbors three hypothetical coding sequences (*CDS*'s). The *FARE*2's are highly conserved, and alignment of the elements demonstrates that all members of the class are deletion derivatives of a large, ancestral element (data not shown).

The ends of individual *FARE*2's display an average of 94% identity with the consensus ends (Table 1). The *FARE*2 consensus sequences for both the L- and R-end regions share 74% identity with their *FARE*1 counterparts (Figure 4A). In total, 31 conserved substitutions are observed in the terminal 120 nucleotides of each consensus *FARE*2 end relative to the *FARE*1 consensus ends; however, the structure of the region, including the domain I repeats, is conserved between the two groups of elements (Figures 4A and 5A). These results are summarized as a phylogenetic tree in Figure 4B and indicate that *FARE*2's are more related to one another than to *FARE*1's.

Internal to the terminal 120 nucleotides of the *FARE*2 ends, the elements contain large, imperfect inverted repeats (Figure 5A). These 0.4-kb repeats display homology with domain II of the *FARE*1's but lack the well-defined subrepeats (data not shown). Functionally, the *FARE*2 inverted repeats are predicted to contribute to extensive secondary structure similar to that predicted for *FARE*1's (data not shown).

The bulk of a given *FARE*2 is composed of an internal region several kilobases in length (Figure 5A) and deletions within this region account for the variability observed in the sizes of the elements. This region is composed of predicted coding sequence as well as repetitive and A-T-rich sequences. The domain III repeats identi-

FIGURE 3.—Alignment of the L-end (top strand) and the R-end (bottom strand) of *FARE*1.1. The R-end is represented in the reverse complement. Numbering corresponds to the base position in the full-length *FARE*1.1 sequence; the terminal nucleotide of the L-end is designated position 1, the terminal nucleotide of the R-end is designated 1122. Gaps introduced into the alignment are indicated with dashes. Asterisks (*) indicate identity. Boxes denote domains identified by the presence of specific repeated units: light gray corresponds to domain I; gray, domain II; and black, domain III. Arrows represent repeat units.

TABLE 1

Characteristics of *FARE*1 and *FARE*2 elements

Element	Size (kb)	Identity (%)[a]		Direct repeat[b]	Chromosome location			
		L-end	R-end		No.	Accession no.	L-end[c]	R-end[c]
*FARE*1.1	1.1	94	96	CTTTTATTT/CTTTTATTT	2	AC007048	30,281	31,402
*FARE*1.2[d]	0.5	90	92	CATGCAATA/CATGCAATA	2	AC005967	10,963	11,403
*FARE*1.3	0.6	85	92	aaccAATAA/tttaAATAA	5	AB009054	68,281	67,659
*FARE*1.4	0.7	96	95	TAACTTATT/TAACTTATT	2	AC006053	50,676	51,420
*FARE*1.5	0.9	95	96	AAACTAAAa/AAACTAAAt	1	AC004473	5,878	5,011
*FARE*1.6	0.9	95	95	ATATTAAAA/ATATTAAAA	2	AC006217	7,726	8,638
*FARE*1.7	1.0	93	94	aAAAgGTAT/tAAAaGTAT	4	AF076274	10,716	9,692
*FARE*1.8	1.0	95	96	TTACAATTA/TTACAATTA	3	AP000736	59,852	58,827
*FARE*1.9	1.0	96	94	TcccTTTAA/TtatTTTAA	4	AC002330	69,167	68,130
*FARE*1.11	1.1	96	93	AAACAAAAA/AAACAAAAA	5	AB006707	41,234	42,309
*FARE*1.12	1.1	95	92	TTTATTAAA/TTTATTAAA	2	AC007063	2,459	1,379
*FARE*1.13	1.1	95	89	TATAAATAA/TATAAATAA	4	AL035526	29,363	30,475
*FARE*1.14	1.1	92	92	GATTATATa/GATTATATt	4	AL117386	29,555	28,426
*FARE*1.15	1.1	92	92	GATTATATA/GATTATATA	4	AL117386	12,251	11,119
*FARE*1.18	1.2	88	93	AAAATTCTa/AAAATTCTg	4	AF072897	33,005	31,833
*FARE*1.19	1.2	91	94	aAAAAtAAT/tAAAAaAAT	5	AB023037	43,374	44,529
*FARE*1.20	1.2	91	95	Not detected	1	AC002311	10,858	12,025
*FARE*1.21	1.2	94	95	AATAATCAA/AATAATCAA	5	AC006259	12,778	11,591
*FARE*1.24	1.5	88	92	AATACAATT/AATACAATT	5	AB013393	25,578	24,110
*FARE*1.25	1.8	87	91	TTGAGAATT/TTGAGAATT	5	AB025638	20,593	22,345
*FARE*1.26	1.9	86	94	ATAAACAAA/ATAAACAAA	2	AC007267	42,028	43,882
*FARE*1.27	1.9	87	91	ATATTTTTG/ATATTTTTG	2	AC006436	47,732	49,627
*FARE*1.28	1.9	86	93	CATTTTAAA/CATTTTAAA	3	AB018114	70,033	68,106
*FARE*1.30	2.0	84	93	AATAATATA/AATAATATA	2	AC007267	41,770	39,735
*FARE*1.31	2.6	87	94	Not detected	2	AC006429	47,339	44,692
*FARE*1.32	5.9	85	92	ACCcATTTT/ACCaATTTT	2	AC006298	21,816	15,962
*FARE*1.33	6.9	92	91	AATTATAAA/AATTATAAA	5	AB016877	34,164	41,014
*FARE*1.34	13.0	95	95	ggTTTTAAA/tcTTTTAAA	2	AC006920	61,145	48,170
*FARE*2.1	8.5	99	97	TATTATTAT/TATTATTAT	2	AC006217	8,782	17,240
*FARE*2.3	12.5	84	87	TTTgTTTTt/TTTtTTTTg	4	AC006266	43,026	30,488
*FARE*2.6	15.4	97	94	TTGTTTTTT/TTGTTTTTT	3	AL096860	17,385	1,990
*FARE*2.7	15.5	96	96	AAAGAATTA/AAAGAATTA	2	AC006298	45,029	29,539
*FARE*2.8	15.5	94	92	Not detected	2	AC007197	51,808	36,269
*FARE*2.9	15.6	94	97	TTAATTTTT/TTAATTTTT	2	AC007211	24,130	39,775
*FARE*2.10	15.8	96	98	TTAAGACAA/TTAAGACAA	2	AC005936	64,318	80,132
*FARE*2.11	16.7	93	98	ATaAAAATA/ATgAAAATA	2	[e]	1,244	5,671

With the exception of *FARE*1.1, elements are arranged from smallest to largest according to class (*FARE*1 or *FARE*2).

[a] The percentage identity of the first 120 nucleotides of the indicated end as compared to the consensus sequence. *FARE*1 ends are compared to the *FARE*1 consensus for a given end; *FARE*2 ends are compared to the *FARE*2 consensus for a given end.

[b] Target-site duplications. Identities between repeats are indicated with underlined uppercase lettering; mismatches are indicated with lowercase lettering.

[c] The position of the terminal nucleotide for each end within the accession is indicated.

[d] *FARE*1.2 completely lacks the domain III region observed in other *FARE*1 elements and is likely a deletion derivative.

[e] *FARE*2.11 spans two BACs. The terminal nucleotide of the L-end is in accession no. AC006420; the terminal position of the R-end is in accession no. AC007235.

fied in *FARE*1's are observed in this internal region and are interspersed in direct orientation throughout the noncoding sequences and the introns of all three *CDS*'s (Figure 5A).

The *FARE*2 internal region is predicted to encode up to three proteins in the largest elements. The predicted *CDS*1 product (CDS1) is a soluble, globular protein of 739 amino acids that contains a putative nuclear localization signal (NLS) as well as a single zinc finger CCHC motif (Figure 5B). The CCHC motif is generally associated with RNA binding proteins encoded by retroviruses (Katz and Jentoft 1989); however, it is also found in a number of eukaryotic proteins involved in ssDNA and dsDNA binding (Xu *et al.* 1992; Webb and McMaster 1993). The CDS1 shares limited identity with MURA, which is one of two proteins encoded by the autonomous maize TIR element, *MuDR*. The region of identity (27%) extends from residue 174 to the C terminus of

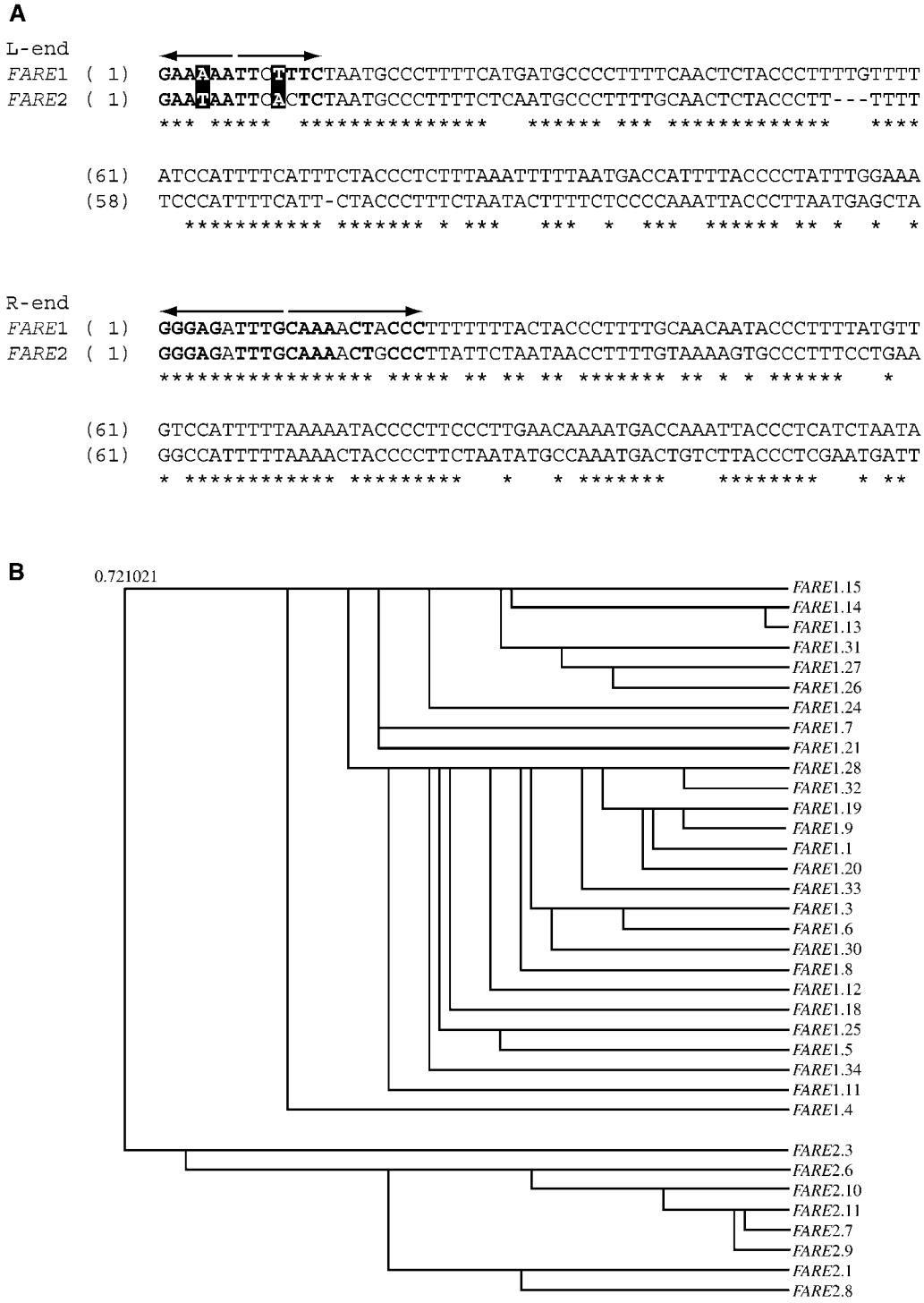


FIGURE 4.—*FARE*1 and *FARE*2 elements are highly related. (A) Alignment of the terminal 120 nucleotides of the *FARE*1 and *FARE*2 consensus L- and R-ends. R-end sequences are in the reverse complement. Nucleotide positions are indicated in parentheses. Identities are marked with asterisks (*). Arrows indicate terminal palindromic sequences and bold letters denote bases contributing to the palindrome. Residues highlighted in black indicate reciprocal substitutions in the *FARE*2 L-end terminus that preserve the palindromic nature of the sequence. (B) An unrooted phylogenetic tree of the *FARE*1 and *FARE*2 R-ends. The numerical value of the first branch, which separates *FARE*1 and *FARE*2 R-ends, is a weighted homology score demonstrating that the ends are highly related but distinguishable. The tree was constructed using the GeneBee service.

CDS1 (Figure 5B). No similarities are observed in the N-terminal regions of the two proteins. Experimental results are consistent with MURA being a transposase; it binds to specific sequences in the TIRs of the *Mu* elements (BENITO and WALBOT 1997) and shares homology with the transposases of several bacterial insertion sequence elements (EISEN *et al.* 1994).

*CDS*2 and *CDS*3 (Figure 5A) are predicted to encode soluble, globular proteins of 783 and 866 amino acids, respectively. These hypothetical proteins have strong, negative charges and share no functional homologies with any known protein sequences. The predicted *CDS*2 product (CDS2) is characterized by a large, glutamic acid-rich region in its C terminus that accounts for the negative character of the protein. Out of the 372 C-terminal amino acids of the predicted protein, 94 are glutamic acid residues (data not shown). The predicted *CDS*3 product (CDS3) contains a single NLS in its C

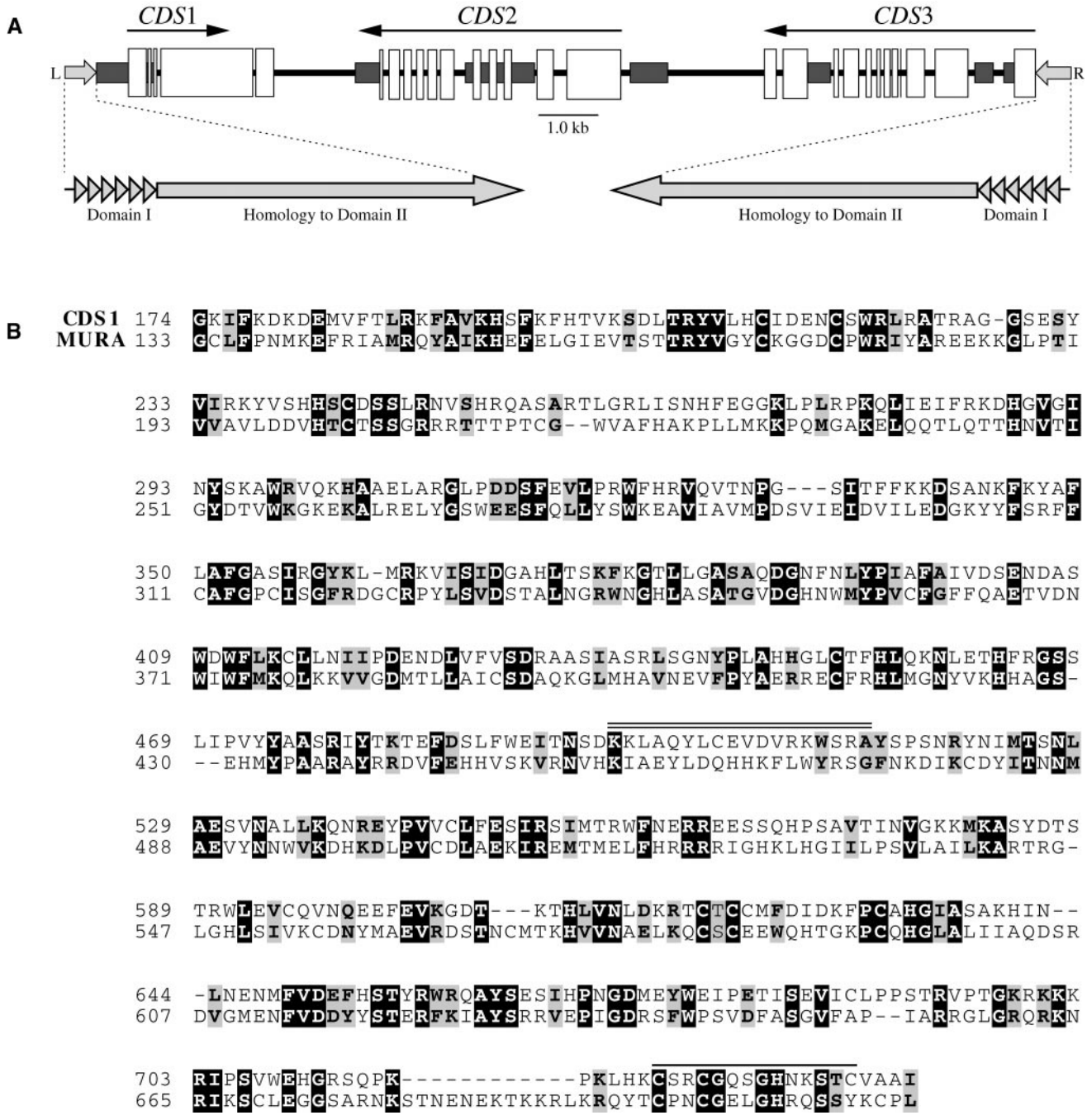


FIGURE 5.—Characteristics of *FARE*2 elements. (A) Composite cartoon depicting *FARE*2 elements. The *FARE*2 ends are represented as gray arrows. The ends have been enlarged to show details: thin lines represent the 16 terminal nucleotides of the ends, triangles signify the domain I repeats, and the large gray arrows represent the inverted repeat structure that bears homology to domain II of the *FARE*1 elements. Hypothetical *CDS*'s are labeled and open rectangles represent predicted exons. Arrows above each *CDS* indicate the direction of transcription. Solid boxes represent regions containing the domain III repeats. Each *CDS* represents the most complete version among the *FARE*2 elements: *CDS*1 is taken from *FARE*2.6, *CDS*2 is taken from *FARE*2.8, and *CDS*3 is derived from *FARE*2.6 and *FARE*2.11. (B) Alignment of CDS1 (top) and MURA (bottom) proteins. Residue positions in the full-length proteins are shown. The native MURA protein is 823 amino acids. Identities are indicated by black boxes; conserved changes are indicated by shaded boxes. Gaps introduced into the alignment are indicated with dashes. The predicted NLS and zinc finger CCHC motifs of the CDS1 product are indicated by a double line and a single line, respectively. The consensus amino acid sequence for the zinc finger CCHC motif is $CX_{(2)}CX_{(4)}HX_{(4)}C$.

terminus, but displays no other functional homologies. Database searches using the nucleotide sequences of *CDS*1, *CDS*2, and *CDS*3 reveal that these sequences are always found associated with *FARE*2's in the Arabidopsis genome. We note that there are other sequences in addition to *CDS*1 that are predicted to encode proteins

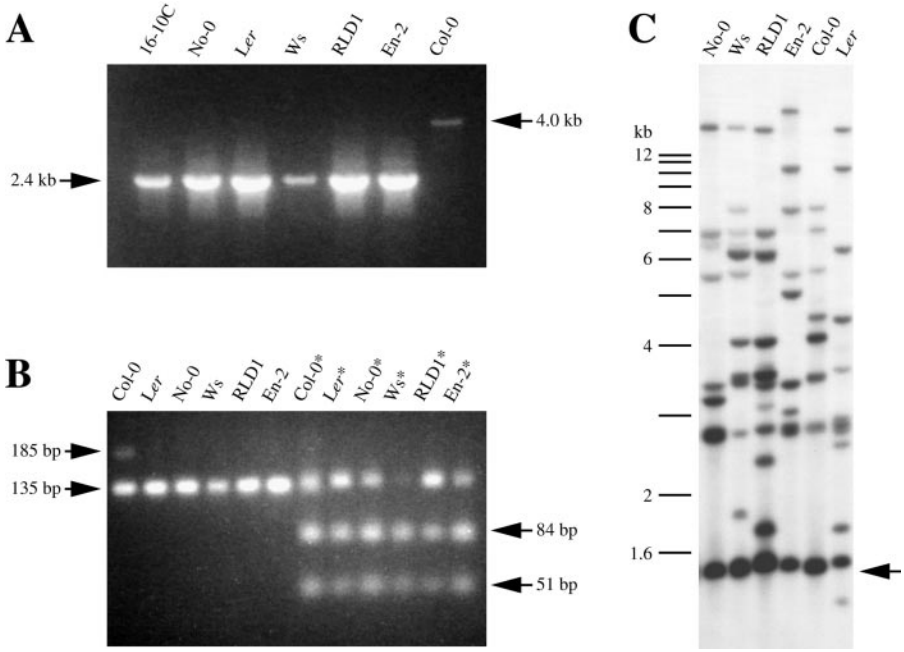


FIGURE 6.—PCR and Southern analysis of *FARE* elements in Nossen (No-0), Landsberg *erecta* (L*er*), Wassilewskija (Ws), Rschew (RLD1), Enkheim (En-2), and Columbia (Col-0). Negative control lanes were blank for all PCR reactions (data not shown). (A) PCR amplification of the *FARE*1.1 insertion site in six Arabidopsis ecotypes. Fragment sizes are indicated. 16-10C harbors a large chromosome II inversion and is the No-0 derivative in which the *FARE*1.1 polymorphism was first identified. The 4.0-kb PCR product obtained from Col-0 represents the polymorphism and the associated *FARE*1.1 insertion. The 2.4-kb fragment obtained from the other ecotypes demonstrates the absence of the *FARE*1.1 insertion in these ecotypes. (B) PCR amplification of the *FARE*1 L-end from six ecotypes. This specific product is 135 bp in length and is cleaved by *Dra*I to yield an 84- and a 51-bp fragment. Digested samples are indicated by an asterisk (*). (C) Hybridization of *FARE*2 coding sequence to genomic DNA of six ecotypes. Genomic DNA was digested with *Eco*RI and probed with exon 1 of *CDS*3. All *FARE*2 elements identified in the database have an *Eco*RI site in the first intron of *CDS*3, positioned ~1.6 kb from the R terminus. Five of the identified *FARE*2 elements also have an *Eco*RI site ~260 bp from the R terminus; in these elements, the *CDS*3 probe will recognize a fragment of ~1.4 kb (indicated with an arrow). DNA length markers are indicated on the left.

with amino acid homology to MURA (data not shown). However, the *CDS*1 nucleotide sequence is distinct and exhibits no significant identity to these other Arabidopsis coding regions.

The termini of *FARE* elements are imperfect palindromes: Examination of the consensus *FARE*1 L-end terminus reveals the existence of a nearly perfect palindromic motif. The palindrome extends from the terminal guanine to the cytosine at position 13 (Figure 4A). The two half-sites of the palindrome are not identical due to the presence of an additional cytosine in the inner half-site at position 9. Similarly, the consensus R-end terminus also contains a nearly perfect palindromic motif. The sequence of this motif is distinct from that observed in the L-end and the motif is larger, extending from the terminal guanine to the cytosine at position 21 (Figure 4A). As observed in the L-end terminus, the inner half-site of this palindrome contains one additional nucleotide, an adenine at position 18.

Similar palindromic sequences are found in the *FARE*2 termini. The consensus *FARE*2 L-end terminus contains three substitutions relative to the *FARE*1 terminus. Two of these substitutions, at positions 4 (A to T) and 10 (T to A), are reciprocal substitutions in the half-sites that preserve the palindromic nature of the sequence (Figure 4A). The consensus *FARE*2 R-end terminus is virtually identical to that of *FARE*1 with only a single nucleotide change at position 18 (A to G; Figure 4A).

Nine-base pair target-site duplications are characteristic of the *FARE* family of TEs: Twenty-two of the *FARE* insertions characterized are flanked by perfect 9-bp target-site duplications, 11 display partial duplications, and 3 lack the presence of a duplication altogether (Table 1). Thus, a 9-bp target-site duplication represents a general characteristic of *FARE*1 and *FARE*2 element integration. Analysis of the recovered target-site duplications indicates a strong bias for A-T-rich sequences (Table 1).

***FARE* elements are not restricted to Col-0:** *FARE*1.1 was initially identified by PCR as one component of a polymorphism existing between Col-0 and No-0. PCR amplification of the region yields a 4.0-kb product in Col-0 *vs.* a 2.4-kb product in No-0 (Figure 6A). The analysis of four additional ecotypes, L*er*, Ws, RLD1, and En-2, demonstrates that they, like No-0, lack *FARE*1.1 at this location.

To determine if *FARE*1's are present in the genomes of these other ecotypes, primers were designed to specifically amplify a fragment from the L-end of *FARE*1's. Amplification confirmed the presence of the elements in all six ecotypes (Figure 6B) and the specificity of the products was verified by digesting with *Dra*I. A single *Dra*I site is present in the amplified region of more than half of the *FARE*1 L-ends identified in Col-0. A second, faint 185-bp fragment was recovered from Col-0, but not from the other ecotypes (Figure 6B). Given the similarity of *FARE*1 and *FARE*2 elements, we postulate that a misprimed reaction involving a *FARE*2 end was the source of the 185-bp fragment.

To identify *FARE*2's in ecotypes other than Col-0, primers were designed to specifically amplify a 294-bp fragment from exon1 of *CDS*3. The successful amplification of this product indicated that *FARE*2's are present

in all six ecotypes (data not shown). The Col-0 *CDS*3 PCR product was used as a probe against genomic DNA. The results shown in Figure 6C indicate that multiple *FARE*2 hybridizing bands exist in all ecotypes. On the basis of the number and intensity of observed bands, the copy number estimates range from 15 to 25 *FARE*2 elements depending on the ecotype. Common hybridizing bands exist between ecotypes; however, unique bands are also observed (Figure 6C). These unique bands may represent the transposition of *FARE*2 elements after the divergence of the ecotypes.

DISCUSSION

We have identified a new family of foldback TEs in *A. thaliana* that we have designated the *FARE* elements. The first member of this family, *FARE*1.1, was initially identified as a 1.1-kb insertion found on chromosome II of Col-0 but absent from No-0. *FARE*1.1 is composed entirely of two large, modular, imperfect IVRs with the potential to form striking secondary structure. All of the characteristics of *FARE*1.1 are typical of FTs, a group of transposons with long IVRs of modular organization and the demonstrated capacity to form secondary structure (Potter *et al.* 1980, 1989). Database searches utilizing the 50 terminal nucleotides of *FARE*1.1 demonstrate that the Arabidopsis genome contains many *FARE* elements; we have characterized 36 of these in detail. The *FARE* elements fall into two highly related groups, the *FARE*1's and the *FARE*2's, that can be distinguished both structurally and at the nucleotide level. The IVRs of *FARE*1 and *FARE*2 elements are very similar and the 16 terminal nucleotides from each end fail to contribute to the predicted foldback secondary structure. *FARE*2's differ from *FARE*1's in that they contain a large internal region that is predicted to encode one to three proteins. Both groups of elements are heterogeneous in size.

*FARE*1's are small (0.5–2.6 kb), A-T rich, and are composed entirely of two large, imperfect IVRs. These IVRs are organized into three modules: domain I, domain II, and domain III. Each domain is composed of distinct repeating units in direct orientation. The elements do not display any inherent coding capacity, suggesting that *FARE*1's are not autonomous. In terms of their gross structure, *FARE*1's most resemble the *FB* elements of *D. melanogaster*, which have a high A-T content (Potter 1982), do not display protein coding capacity, and likewise have long IVRs composed of three domains (Potter *et al.* 1989).

*FARE*2's are large (8.0–16.7 kb), A-T rich, and closely related to the *FARE*1's. The *FARE*2 ends are modular and display homology to domains I and II of the *FARE*1's. A large, internal region physically separates the *FARE*2 ends and is predicted to encode up to three proteins. The domain III repeats, initially identified in the *FARE*1's, are interspersed throughout this internal region and do not contribute to the secondary structure of the ends. The predicted *CDS*'s are unique to *FARE*2's and no similar sequences are found elsewhere in the genome. This suggests that the *CDS*'s may encode transposition functions. Indeed, CDS1 shares limited identity with the MURA transposase of maize. Further, CDS1 and CDS3 contain NLSs, a feature of proteins capable of interacting with nuclear DNA. Currently, no specific homologues have been identified for the CDS2 or CDS3 proteins. No expressed sequence tags have been identified that correspond to the predicted transcription products of any of the *FARE*2 *CDS*'s. This is not surprising, as proteins encoded by TEs are often expressed at low levels. In addition, point mutations and deletions have rendered most, if not all, of the *FARE*2 coding sequences defective (data not shown), and it is likely that functional versions of the *CDS*'s, if they exist at all, are present at low or single copy number. Structurally, *FARE*2's are very similar to the *FB-NOF* elements of *D. melanogaster*, a unique class of *FB* elements that have protein coding capacity (Smyth-Templeton and Potter 1989; Harden and Ashburner 1990).

We do not know if any of the *FARE*1's or *FARE*2's are still capable of transposition; however, there are two compelling lines of evidence that suggest that *FARE* elements have transposed in the genome of Arabidopsis. The first is the presence of *FARE*1.1 in Col-0 and its absence from No-0, L*er*, Ws, RLD1, and En-2. This demonstrates that *FARE*1.1 has transposed in Col-0, albeit sometime after the divergence of this ecotype from the others. The second line of evidence is the observation that 92% of *FARE*1 and *FARE*2 elements are flanked by recognizable target-site duplications. A majority of these, 67%, are perfect 9-bp duplications and another 27% are nearly perfect duplications with only one or two nucleotide changes. Target-site duplications are the direct result of transposition events. The presence of imperfect target-site duplications at some *FARE* insertion sites may indicate the relative age of these insertions; older duplications are predicted to be subject to base substitution and to accumulate mutations. Interestingly, *FB* elements also produce 9-bp target site duplications (Harden and Ashburner 1990), and this similarity implies a conserved mechanism for transposition between these foldback transposons. We note that *Mu* also produces 9-bp target-site duplications upon integration (Chandler and Hardeman 1992), supporting the idea that the limited homology of the *FARE*2 CDS1 to MURA is based on function.

The analysis of the target-site duplications produced by *FARE* elements demonstrates a propensity for insertion into A-T-rich sequences. Generally, such regions are noncoding and none of the *FARE* elements have disrupted known genes or predicted coding sequences (data not shown). Two *FARE*1's, however, were found in close proximity to genes. The R-end of *FARE*1.25 is situated 284 bp 5′ to the translational start of *PHT*3, an inorganic phosphate transporter (Mitsukawa *et al.*

1997). Likewise, the L-end of *FARE*1.2 is located 548 bp from the translational start of *rpoPT*, a DNA-dependent RNA polymerase (Hedtke *et al.* 1997). Given the structure of *FARE* elements, it is possible that insertions 5′ to coding sequences could influence the expression of these sequences; however, no experiments have been performed to address this possibility with respect to either *PHT*3 or *rpoPT*.

Chromosomes II and IV are the only fully sequenced chromosomes of Arabidopsis and account for an estimated 32% of the total genome (Lin *et al.* 1999; Mayer *et al.* 1999). We have analyzed the distribution of the *FARE* elements on these two chromosomes and were unable to identify insertions in the nucleolar organizers or in the centromeres, two regions that have been shown to accumulate other TEs and repetitive sequences (Lin *et al.* 1999; Mayer *et al.* 1999; Parinov *et al.* 1999). By sequence analysis, we determined that chromosome II contains 38 *FARE* insertions and chromosome IV, which is slightly larger, contains 43 (data not shown). Approximately half of these insertions are partial elements or unpaired ends and are, therefore, clearly defective. By extrapolation, the total number of *FARE* sequences in the Arabidopsis genome may approach 250, making *FARE* the highest copy number class II element family identified in this species. *FB* elements, which are present in 30–60 copies in the *D. melanogaster* genome (Truett *et al.* 1981; Silber *et al.* 1989), have been implicated in genomic restructuring events such as rearrangements, inversions, and translocations (Bingham and Zachar 1989; Lovering *et al.* 1991). Given the copy number, the highly repetitive nature of the *FARE* elements, and the observation of partial elements, it is reasonable to postulate that *FARE* elements may be associated with similar processes in Arabidopsis.

It is evident that *FARE*1's and *FARE*2's are derived from a common ancestor, but the nature of their relationship is debatable. We have considered two models. The first is that *FARE*1's are deletion derivatives of an ancestral *FARE*2 element. The fact that *FARE*2's manifest all of the features associated with the *FARE*1's (domains I, II, III, and the palindromic termini) as well as additional characteristics, such as the *CDS*'s, argues in favor of this model. Alternatively, *FARE*2's may have arisen through the insertion of a second TE into an ancestral *FARE*1. Under this premise, domain II of the *FARE*2 ends has evolved from the inverted repeats of an ancient TIR element. The size of the inverted repeats (*ca.* 450 bp), as well as the limited identity of CDS1 with MURA, suggests that this second TE is distantly related to the maize *Mu* element. One prediction of this model is that additional copies of this second TE should be present in the Arabidopsis genome as independent mobile elements. Currently, no such elements have been identified. While we favor the model that *FARE*1's are deletion derivatives of an ancestral *FARE*2 element, we cannot exclude the alternative possibility.

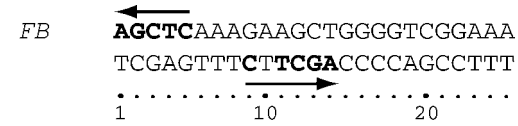


Figure 7.—Sequence of *D. melanogaster FB* termini. The left and right ends of *FB*4 (Potter 1982) are identical; therefore, only one sequence is shown. Numbers indicate nucleotide positions. Arrows indicate regions of dyad symmetry and boldface letters denote bases contributing to the symmetry.

Similar models have been proposed for *FB* and *FB-NOF* elements in *D. melanogaster*. *FB-NOF* elements contain an internal region, the "loop," which is absent from the smaller *FB* elements. It has been argued that (1) *FB* elements represent deletion derivatives of *FB-NOF* (Bingham and Zachar 1989) or (2) the loop of *FB-NOF* elements, which contains protein coding capacity, represents a second TE (Harden and Ashburner 1990). Like *FARE*2's, the internal region of *FB-NOF* is unique to this group of elements and is predicted to encode up to three proteins. No homologies have been identified for any of these proteins.

While there is strong evidence that *FB* transposition is dependent on *FB-NOF*, we can only speculate as to whether or not such a relationship exists between *FARE*1 and *FARE*2 elements. The ends of *FARE*1's and *FARE*2's are highly related but not identical and there are many conserved changes between the two groups of elements. Even single nucleotide changes in the ends of many TIR elements abolish their transposition competency. However, *FARE* elements are FTs, and little is known about the mechanism(s) of transposition of this group of elements. Therefore, *FARE*1's and *FARE*2's may rely on transposition functions that are distinct from those characterized in TIR elements. These functions may recognize structural features of the ends instead of, or in addition to, specific sequences.

The sequences of the *FARE* termini distinguish these elements from most other class II transposable elements, including other FTs, whose termini are characterized as perfect or nearly perfect inverted repeats. While not representing complements of each other, the termini of the *FARE* elements are highly conserved and display nearly perfect palindromic sequences. As a structural motif, palindromic sequences, as well as other sequences with dyad symmetry, have been implicated in the activities of a wide range of DNA binding proteins including transcriptional regulators, site-specific recombinases, and type II restriction endonucleases. We have examined the sequences of the *D. melanogaster FB* termini and note that these also contain dyad symmetry (Figure 7). As observed in *FARE* elements, the inner and outer half-sites differ in length by a single nucleotide. Sequence analysis reveals that dyad symmetry is also present in the termini of *TU* and *So*FT elements (data not shown). Taken together, these observations suggest

a functional requirement for this organization in the activity of *FARE* elements and other FTs.

To facilitate transposition, the transposition machinery must be able to delimit the ends of a given element and distinguish transposon sequences from those of the host. In the case of TIR elements, this is accomplished by the presence of sequences composing the TIRs and the close association of the transposase with these sequences to form an active complex (Beall and Rio 1997; Gorbunova and Levy 1997). Similarly, one might postulate that the palindromic regions of the *FARE* termini provide a basis for delimiting the ends of the elements during transposition. Type II restriction endonucleases recognize palindromic sites in order to carry out their functions. These enzymes initially bind nonspecifically to DNA and recognize the specific target sequence as a function of groove geometry (Pingoud and Jeltsch 1997). Termini recognition based on groove geometry could allow *FARE*2-encoded transposition functions to interact with both *FARE*2 and *FARE*1 elements. This situation is analogous to "star activity," a phenomenon in which some type II restriction enzymes act upon sequences that differ from the canonical recognition site by a single nucleotide (Pingoud and Jeltsch 1997). In addition, the differences between the L- and R-end palindromes of *FARE* elements suggest that at least two protein functions are necessary for recognition and cleavage at the *FARE* ends. The fact that no functional homologies exist for CDS2 and CDS3 or the products encoded by the *FB-NOF* ORFs is also noteworthy within the context of this model. The genes encoding type II restriction endonucleases share little sequence homology and the enzymes themselves are known to utilize novel and disparate structures to accomplish the tasks of DNA recognition and cleavage (Pingoud and Jeltsch 1997).

We thank J. Poupart, P. Lasko, V. Gorbunova, and M. Deyholos for comments on the manuscript. Seed stocks, unless otherwise noted, were obtained from the Arabidopsis Biological Resource Center at Ohio State University. This work was supported by a grant to C.S.W. from the Natural Sciences and Engineering Research Council of Canada.

Note added in proof: Our nucleotide sequence data from No-0 have been deposited in GenBank, accession no. AF311319. *FARE*1.2 has recently appeared in the literature as MULE12 (Q. H. Le, S. Wright, Z. Yu and T. Bureau, 2000, Transposon diversity in *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **97:** 7376–7381).

LITERATURE CITED

Adé, J., and F. J. Belzile, 1999 *Hairpin* elements, the first family of foldback transposons (FTs) in *Arabidopsis thaliana*. Plant J. **19:** 591–597.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.* 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman *et al.* (Editors), 1996 *Current Protocols in Molecular Biology*. John Wiley & Sons, New York.

Beall, E. L., and D. C. Rio, 1997 Drosophila P-element transposase is a novel site-specific endonuclease. Genes Dev. **11:** 2137–2151.

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:** 5165–5175.

Bingham, P. M., and Z. Zachar, 1989 Retrotransposons and the FB transposon from *Drosophila melanogaster*, pp. 485–502 in *Mobile DNA*, edited by D. E. Berg and M. M. Howe. American Society for Microbiology, Washington, DC.

Bureau, T. E., and S. R. Wessler, 1992 Tourist: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell **4:** 1283–1294.

Bureau, T. E., P. C. Ronald and S. R. Wessler, 1996 A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. Proc. Natl. Acad. Sci. USA **93:** 8524–8529.

Burge, C., and S. Karlin, 1997 Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268:** 78–94.

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

Dellaporta, S. L., J. Wood and J. B. Hicks, 1983 A plant DNA minipreparation: version II. Plant Mol. Biol. Rep. **1:** 19–21.

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

Gorbunova, V., and A. A. Levy, 1997 Circularized *Ac/Ds* transposons: formation, structure and fate. Genetics **145:** 1161–1169.

Harden, N., and M. Ashburner, 1990 Characterization of the FB-NOF transposable element of *Drosophila melanogaster*. Genetics **126:** 387–400.

Hedtke, B., T. Borner and A. Weihe, 1997 Mitochondrial and chloroplast phage-type RNA polymerases in Arabidopsis. Science **277:** 809–811.

Hoffman-Liebermann, B., D. Liebermann, L. H. Kedes and S. N. Cohen, 1985 TU elements: a heterogeneous family of modularly structured eucaryotic transposons. Mol. Cell. Biol. **5:** 991–1001.

Hoffman-Liebermann, B., D. Liebermann and S. N. Cohen, 1989 TU elements and puppy sequences, pp. 575–592 in *Mobile DNA*, edited by D. E. Berg and M. M. Howe. American Society for Microbiology, Washington, DC.

Katz, R. A., and J. E. Jentoft, 1989 What is the role of the cys-his motif in retroviral nucleocapsid (NC) proteins? Bioessays **11:** 176–181.

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

Lovering, R., N. Harden and M. Ashburner, 1991 The molecular structure of *TE146* and its derivatives in *Drosophila melanogaster*. Genetics **128:** 357–372.

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.

Mitsukawa, N., S. Okumura and D. Shibata, 1997 High-affinity phosphate transporter genes of Arabidopsis thaliana. Soil Sci. Plant Nutr. **43:** 971–974.

Parinov, S., D.-Y. Mayalagu, W. C. Yang, M. Kumaran and V. Sundaresan, 1999 Analysis of flanking sequences from *Dissociation* insertion lines: a database for reverse genetics in Arabidopsis. Plant Cell **11:** 2263–2270.

Pingoud, A., and A. Jeltsch, 1997 Recognition and cleavage of DNA by type-II restriction endonucleases. Eur. J. Biochem. **246:** 1–22.

Potter, S. S., 1982 DNA sequence of a foldback transposable element in Drosophila. Nature **297:** 201–204.

Potter, S. S., M. Truett, M. Phillips and A. Maher, 1980 Eucaryotic transposable elements with inverted terminal repeats. Cell **17:** 429–439.

Potter, S. S., B. Heineke, S. Kaur, G. Jones, J. Lloyd *et al.*, 1989 Drosophila foldback elements, primate L1 elements, and transgenic mice. Prog. Nucleic Acid Res. Mol. Biol. **36:** 3–23.

Rebatchouk, D., and J. O. Narita, 1997 Foldback transposable elements in plants. Plant Mol. Biol. **34:** 831–835.

Santalucia, J., Jr., 1998 A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. USA **95:** 1460–1465.

Silber, J., C. Bazin, F. Lemeunier, S. Aulard and M. Volovitch, 1989 Distribution and conservation of the Foldback transposable element in Drosophila. J. Mol. Evol. **28:** 220–224.

Smyth-Templeton, N., and S. S. Potter, 1989 Complete foldback transposon elements encode a novel protein found in *Drosophila melanogaster*. EMBO J. **8:** 1887–1894.

Truett, M. A., R. S. Jones and S. S. Potter, 1981 Unusual structure of the FB family of transposable elements in Drosophila. Cell **24:** 753–763.

Valvekens, D., M. Van Montagu and M. Van Lijsebettens, 1988 *Agrobacterium tumefaciens*-mediated transformation of *Arabidopsis thaliana* root explants by using kanamycin selection. Proc. Natl. Acad. Sci. USA **85:** 5536–5540.

Webb, J. R., and W. R. McMaster, 1993 Molecular cloning and expression of a *Leishmania* major gene encoding a single-stranded DNA-binding protein containing nine 'CCHC' zinc finger motifs. J. Biol. Chem. **268:** 13994–14002.

Xu, H.-P., T. Rajavashisth, N. Grewal, V. Jung, M. Riggs *et al.*, 1992 A gene encoding a protein with seven zinc finger domains acts on the sexual differentiation pathways of *Schizosaccharomyces pombe*. Mol. Biol. Cell **3:** 721–734.

Communicating editor: J. A. Birchler