The Structure, Distribution and Evolution of the Ta*1* Retrotransposable Element Family of *Arabidopsis thaliana*

Daniel F. Voytas,* Andrzej Konieczny,* Michael P. Cummings† and Frederick M. Ausubel*

**Department of Genetics, Harvard Medical School and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, and †Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138*

Manuscript received May 7, 1990
Accepted for publication July 21, 1990

ABSTRACT

The Ta*1* elements are a low copy number, *copia*-like retrotransposable element family of *Arabidopsis thaliana*. Six Ta*1* insertions comprise all of the Ta*1* element copies found in three geographically diverse *A. thaliana* races. These six elements occupy three distinct target sites: Ta*1-1* is located on chromosome *5* and is common to all three races (Col-0, Kas-1 and La-0). Ta*1-2* is present in two races on chromosome *4* (Kas-1 and La-0), and Ta*1-3*, also located on chromosome *4*, is present only in one race (La-0). The six Ta*1* insertions share >96% nucleotide identity, yet are likely to be incapable of further transposition due to deletions or nucleotide changes that alter either the coding capacity of the elements or conserved protein domains required for retrotransposition. Nucleotide sequence comparisons of these elements and the distribution of Ta*1* among 12 additional *A. thaliana* geographical races suggest that Ta*1-1* predated the global dispersal of *A. thaliana*. As the species spread throughout the world, two additional transposition events occurred which gave rise first to Ta*1-2* and finally to Ta*1-3*.

TRANSPOSABLE genetic elements are an apparently ubiquitous component of eukaryotic genomes (BERG and HOWE 1989). They have been identified in virtually every organism which has been subjected to molecular scrutiny and typically comprise a substantial fraction of eukaryotic genomic DNA (*e.g.*, estimated at 10% of the *Drosophila melanogaster* genome; FINNEGAN and FAWCETT 1986). Transposable elements are responsible for a wide variety of genetic effects, including mutations, chromosomal deletions and rearrangements. Transposition, therefore, is believed to play a major role in genome evolution despite the fact that the long term consequences of transposition on genes, chromosomes, genomes, populations and species are largely unknown.

Although reasons for the persistence of transposable elements in eukaryotic genomes are not understood, two factors are important in considering transposable element evolution and population dynamics (reviewed in AJIOKA and HARTL 1989): (1) the genetic variability which results from transposition must not compromise host fitness; and (2) the transposable elements must attain sufficient copy numbers to offset deleterious mutations which they incur through transposition or while residing in the genome.

To obtain greater insight into transposable element population dynamics and evolution, we undertook a complete analysis of the structure and distribution of the Ta*1* retrotransposon family which we recently discovered in the crucifer *Arabidopsis thaliana* (L.) Heynh (VOYTAS and AUSUBEL 1988). Retrotransposable elements have been studied extensively in yeast (BOEKE 1989) and Drosophila (BINGHAM and ZACHAR 1989) and have been found in numerous other organisms (DOOLITTLE *et al.* 1989). Among plants, the *A. thaliana* transposable element insertion, Ta*1-3*, was the first transposable element shown to carry all of the structural and coding features characteristic of retroviruses and eukaryotic virus-like retrotransposons (VOYTAS and AUSUBEL 1988).

Like retrotransposons and integrated retroviral proviruses, Ta*1-3* consists of a large central domain (4.2 kbp) bounded by long terminal direct repeats (LTRs, 0.5 kbp; VOYTAS and AUSUBEL 1988). The Ta*1-3* LTRs terminate in short inverted repeats with LTR end-sequences identical to those of other virus-like retro-elements (5′-TG. . .CA-3′). The central domain sequences adjacent to the 5′ LTR of Ta*1-3* are identical to the 3′ terminus of plant $tRNA^{i}_{met}$, and hybridization between these sequences and a plant tRNA likely serves to prime first strand DNA synthesis by reverse transcription. Within the central domain adjacent to the 3′ LTR is a short oligo-purine stretch of DNA which may prime second strand DNA synthesis.

The central domain of Ta*1-3* encodes a single open reading frame, the derived amino acid sequence of which shares strong similarity to conserved protein coding regions of retroviruses and retrotransposons (VOYTAS and AUSUBEL 1988). The Ta*1* elements share a particularly high degree of structural and coding similarities with the *D. melanogaster copia* ele-

ments and a recently described family of retrotransposable elements from *Nicotiana tabacum*, Tnt*1* (MOUNT and RUBIN 1985; VOYTAS and AUSUBEL 1988; GRANDBASTIEN, SPIELMANN and CABOCHE 1989). Unlike other well-characterized retrotransposons, Ta*1*, *copia* and Tnt*1* encode all of their genetic information in a single open reading frame and have a reversed order for their integrase and reverse transcriptase genes. Pairwise comparisons of 250 amino acids which characterize each of the reverse transcriptase and integrase domains of these elements show between 37% and 54% amino acid identity (VOYTAS and AUSUBEL 1988; GRANDBASTIEN, SPIELMANN and CABOCHE 1989).

Unlike most retrotransposable element families, relatively few Ta*1* insertions are found within the *A. thaliana* genome. The low copy-number of these elements has made it possible to undertake a detailed study of the structure and distribution of the Ta*1* elements among different *A. thaliana* races. A complete structural analysis of the full complement of transposable element copies among wild populations has yet to be undertaken in any species. In this paper we describe experiments which assessed the structural integrity of the Ta*1* elements, the relationships among the Ta*1* element copies, and the likely manner in which these elements spread throughout the *A. thaliana* genome over the course of the global dispersal of the species.

MATERIALS AND METHODS

***Arabidopsis thaliana* races:** The *A. thaliana* geographical races used in this study were obtained from the *Arabidopsis* Information Service, Frankfurt, West Germany, with the exception of Mv-0, which was isolated from a naturalized population growing on Martha's Vineyard, Massachusetts. The races represent *A. thaliana* populations from the following locations (KRANZ and KIRCHHEIM 1987): La-0, West Germany; Col-0, West Germany; Kas-1, India; Co-4, Portugal; Sei-0, Italy; Mv-0, United States; Ll-0 Spain; Cvi-0, Cape Verde Islands; Fi-3, Finland; Ba-1, Great Britain; Hau-0, Denmark; Aa-0, West Germany; Ms-0, Soviet Union; Ag-0, France; Mh-0, Poland. The La-0 race carries the recessive mutation *erecta*, which confers a short, upright growth habit (REDEI 1962). Both La-0 and Col-0 are standard laboratory strains, widely used for both classical and molecular genetic analyses (MEYEROWITZ 1987).

DNA manipulations: *A. thaliana* DNA isolations were performed by methods previously described (AUSUBEL *et al.* 1987). For Southern blot analyses, 1 μg of *A. thaliana* genomic DNA was subjected to electrophoresis on 0.8% agarose gels and transfered to Gene Screen Plus nylon membranes (New England Nuclear). DNA probes were labeled by random priming (Boehringer Mannheim) and hybridized to filters using conditions recommended by the manufacturer (New England Nuclear). Filters were washed at 65° in 0.2× SSC (T_m = 75°).

The cloning of the Ta*1-1* element from the Kas-1 (Kashmir) race has been previously reported (VOYTAS and AUSUBEL 1988). The Ta*1-1* element from the Col-0 (Columbia) race was isolated from a library constructed in pUC12 using size-selected DNAs digested to completion with *Xba*I. Recombinant clones were identified by colony hybridizations (AUSUBEL *et al.* 1987) using a DNA probe which flanks the *A. thaliana* chalcone synthase structural gene (FEINBAUM and AUSUBEL 1988). The La-0 (Landsberg) copy of Ta*1-1* was cloned from a total genomic DNA library constructed in lambda FIX (Stratagene, see below). Plaque lifts were performed with Colony/Plaque Screen membranes (New England Nuclear) according to instructions provided by the manufacturer. Recombinants were identified using hybridization probes specific to the Ta*1* LTR (Figure 3) and sequences which flank the chalcone synthase gene (FEINBAUM and AUSUBEL 1988).

The Ta*1-2* and Ta*1-3* elements were isolated from Landsberg and Kashmir genomic DNA libraries constructed in the vector lambda FIX (Stratagene) using *Mbo*I partial digests of these DNAs that had been size-fractionated on low-melting temperature agarose gels. The vector DNA was digested with *Xho*I, and both vector and insert DNAs were partially filled-in with the appropriate nucleotides to prevent vector/vector ligation and multiple inserts. Ligation reactions were packaged with Gigapack Gold packaging extracts (Stratagene) and plaque lifts were performed as described above.

Initial attempts to clone the Ta*1-2* and Ta*1-3* elements from either the Landsberg or Kashmir libraries were unsuccessful. Plaque hybridizations using a central element probe from Ta*1-1* (INT, Figure 3) failed to identify Ta*1* clones within the Landsberg library. Of 41 independent clones isolated from the Kashmir library, all contained the deleted element copy Ta*1-1* (see RESULTS) and not the sought after element, Ta*1-2*.

To prevent repeated cloning of the Ta*1-1* elements, a probe was isolated which hybridized specifically to the Ta*1-2* and Ta*1-3* elements. This probe (INT1, Figure 3), was part of a 3 kbp *Hin*dIII fragment which was determined to be unique to Ta*1-2* and Ta*1-3* by genomic Southern blot analyses (data not shown). The 3-kpb *Hin*dIII fragment was isolated from a library constructed in lambda ZAPII (Stratagene) using size selected La-0 DNAs digested to completion with *Hin*dIII. The INT probe was used for phage isolation. To account for the possibility that the Ta*1-2* and Ta*1-3* elements may contain methylated cytosines causing recombinant phage carrying them to be degraded by restriction systems present in many common *Escherichia coli* laboratory strains, the lambda FIX libraries were plated on the *mcrA*$^-$, *mcrB*$^-$ strain, ER1458 (RALEIGH and WILSON 1986). The Ta*1-2* and Ta*1-3* elements were only isolated from phage plated on ER1458 using the INT1 probe (Figure 3).

DNA sequencing: The Ta*1* elements were sequenced with Klenow, Sequenase (US Biochemical Corp.) or Taq polymerase (Stratagene) using both single- and double-stranded DNA templates (AUSUBEL *et al.* 1987). Nested deletions were created with either Bal 31 nuclease or exonuclease III (AUSUBEL *et al.* 1987). Oligonucleotides were synthesized to prime sequencing reactions using a Biosearch DNA Synthesizer (New Brunswick Scientific). The DNA sequence was obtained on both strands for each of the Ta*1* element copies.

DNA sequence and phylogenetic analysis: DNA sequences were assembled on a VAX computer (Digital Equipment Corporation) using the Multiple Sequencing Editor (W. GILBERT, unpublished data). Amino acid sequence alignments were performed with the program ALIGN (NEEDLEMAN and WUNSCH 1970), and all subsequent analyses were performed with the programs of the University of Wisconsin Genetics Computer Group (DEVEREUX, HAEBERLI and SMI-

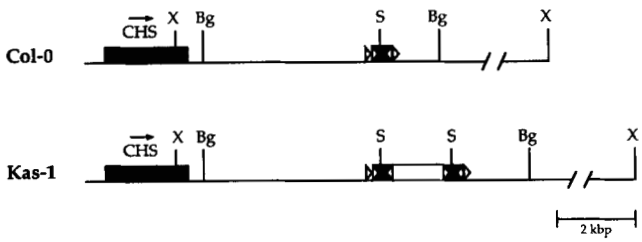

FIGURE 1.—The Ta*1-1* insertion in the Col-0 (Columbia) and Kas-1 (Kashmir) races. The chalcone synthase gene (CHS) is represented by a black box with the arrow indicating the direction of transcription (FEINBAUM and AUSUBEL 1988). Arrowheads flanking the Ta*1-1* insertions depict target site duplications. Black boxes represent LTRs with arrowheads signifying the LTR inverted repeat end-sequences. Restriction sites are as follows: Bg, *Bgl*II; S, *Sal*I; X, *Xba*I.

THIES 1984). The DNA sequence of each of the Ta*1* element copies has been submitted to Genbank.

The aligned nucleotide sequences for the reverse transcriptase/RNase H genes, the polypurine tracts, and the 3′ LTRs were used for phylogenetic analysis. The nucleotide sequence for the polypurine tract and the 3′ LTR of the tobacco retrotransposon, Tnt*1* (GRANDBASTIEN, SPIELMANN and CABOCHE 1989), could not be unambiguously aligned with those of the Ta*1* elements and therefore was treated as missing data. Each nucleotide position was treated as an individual, four state (A, C, G, T), unordered character. Each continguous block of inserted/deleted nucleotides was treated as an individual, two state (present, absent), unordered character. A prerelease version of the computer program MacClade (MADDISON and MADDISON 1990) was used for all data and tree manipulations except for determining the topology, which was obtained by an exhaustive search using the computer program PAUP, version 3.0g (SWOFFORD 1990). The phylogenetic tree was rooted using the outgroup method, with Tnt*1* as the outgroup. All numerical results from the test version of MacClade were checked for veracity with PAUP.

RESULTS

Structural organization and distribution of Ta*1* elements from the Columbia, Kashmir and Landsberg races: The Ta*1* elements were initially identified through restriction fragment length polymorphism (RFLP) analyses of genomic DNA isolated from 16 *A. thaliana* races (VOYTAS and AUSUBEL 1988). These experiments were designed specifically to identify RFLPs which may have arisen by DNA transposition. Among the hybridization probes used to detect polymorphisms was a genomic lambda phage clone which contained a 15 kbp insert that included the structural gene for chalcone synthase. In 14 of the races analyzed, this clone detected a 6 kbp *Bgl*II fragment downstream of the chalcone synthase gene (see Figure 1, and Col-0 and La-0 in Figure 2). In contrast, an 8.3 kbp *Bgl*II fragment was observed in the Kas-1 race (Kashmir; Figures 1 and 2) and the race Ll-0 (data not shown). Additional Southern blot analyses revealed that whereas both of these polymorphic restriction fragments were the same size, the Ll-0 polymorphism was due to a *Bgl*II site loss, while the Kashmir polymorphism was due to a 2.3-kbp DNA insertion (data not shown). An approximately 20-kbp *Xba*I fragment which contained the polymorphic *Bgl*II fragment was cloned from the Kashmir and Col-0 (Columbia) races (see Figure 1). Comparisons of these two genomic clones revealed that the Kashmir insertion was completely encompassed by a 2.3 kbp *Sal*I fragment (Figure 1, data not shown). The nucleotide sequence was obtained for the Kashmir insertion and flanking DNAs.

Ta1-1: The Kashmir insertion, designated Ta*1-1*, was flanked by two ~500 bp long terminal direct repeats (LTRs) and found to encode a single open reading frame (Figures 3 and 4). This open reading frame showed significant amino acid sequence identity to the *D. melanogaster copia* element reverse transcriptase (data not shown; see also VOYTAS and AUSUBEL 1988). The Ta*1-1* open reading frame, however, only encompassed the carboxyl-terminal half of the *copia* element protein. Missing from Ta*1-1* was the coding region corresponding to the *copia gag* gene (data not shown; MOUNT and RUBIN 1985). This suggested that Ta*1-1* had suffered a deletion of the central domain. Subsequent characterization of addition of Ta*1* element copies confirmed this observation, and demonstrated that the Ta*1-1* deletion begins immediately within the 5′ LTR and extends through 2.4 kbp of the central domain (Figure 3).

A solo Ta*1* LTR and no additional Ta*1* hybridizing sequences were found in the Columbia DNA downstream of chalcone synthase (Figure 1, data not shown). Because the Columbia LTR is contained within the 6-kbp *Bgl*II fragment in which the Kashmir insertion was initially identified, and because this 6 kbp *Bgl*II restriction fragment is not polymorphic for most of the races, it seemed likely that all of the *A. thaliana* races examined carry a solo Ta*1* LTR flanking the chalcone synthase gene. This prediction was confirmed by Southern blot analysis of the 16 race DNAs using an LTR-specific probe (*e.g.*, Figure 2; see also subsequent sections and Figure 7), and further supported by the cloning and sequencing of a Ta*1* LTR at this site in a third race, La-0 (Landsberg; data not shown).

The nucleotide sequence of the Columbia and Landsberg LTRs and the Ta*1-1* element from Kashmir demonstrated that all three elements are located at precisely the same chromosomal position. These elements are immediately flanked by identical 5 bp direct repeats (5′-CTTTC-3′), the presumptive target size duplication created upon element integration. The sequences remain nearly identical (>95%) for up to 200 bp either side of the LTRs among the three races (data not shown). The central domain sequences were apparently lost in Columbia and Landsberg by homologous recombination through the direct repeat

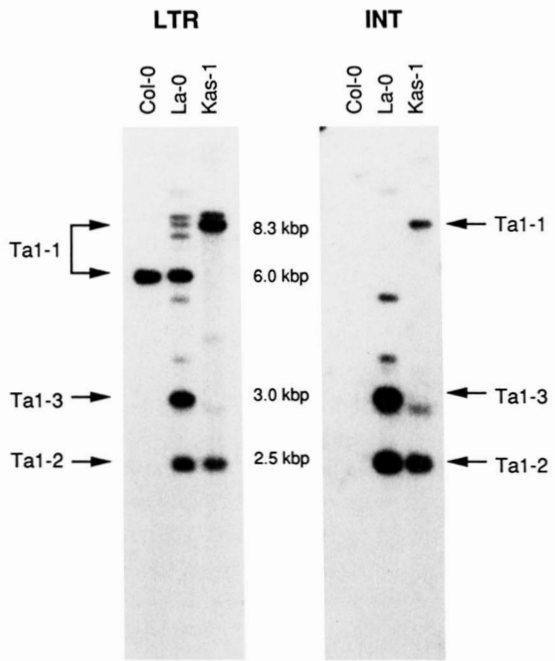

FIGURE 2.—Southern blot analysis of the Ta*1* elements within the Col-0 (Columbia), La-0 (Landsberg) and Kas-1 (Kashmir) races. DNAs were digested with *Bgl*II and hybridization experiments were conducted as described in MATERIALS AND METHODS using LTR-specific and central domain probes (INT, see Figure 3 for probes used). The Ta*1* hybridizing bands representing each of the different element copies are indicated by arrows, with sizes referring to their molecular lengths.

sequence of the LTRs, leaving behind a solo LTR. The Columbia and Landsberg LTRs have also been designated Ta*1-1*, as they are undoubtedly the remnants of a single element integration event which occurred at this site before the dispersal of the three races. The chalcone synthase gene (and thereby Ta*1-1*) has been mapped by restriction fragment length polymorphism analysis to *A. thaliana* chromosome *5* (CHANG *et al.* 1988).

Ta1 element copy number: The Ta*1* element copy number was determined by Southern blot analysis for the races Columbia, Kashmir and Landsberg (Figure 2). Based on the restriction endonuclease map of the Kashmir element, Ta*1-1*, DNAs from these races were digested with enzymes that cut within the central domain and 3′ flanking DNA. This enabled each Ta*1* copy to be visualized as a uniquely-sized restriction fragment on Southern filters hybridized with either LTR-specific or appropriate central domain probes (*e.g.*, Figure 2). The results of several such experiments demonstrated that Columbia, Kashmir and Landsberg carry one, two and three Ta*1* element copies, respectively (*e.g.*, Figure 2). For Columbia, the single Ta*1-1* LTR is the only Ta*1* hybridizing sequence in its genome (Figure 2).

A number of weakly hybridizing bands typically appear when filters are hybridized with either probes to the LTR or central domain (Figure 2). This suggested that sequences similar to Ta*1* are present in the *A. thaliana* genome. Characterization of these sequences has led to the identification of several additional *A. thaliana* retrotransposable element families which are structurally similar to Ta*1* (A. KONIECZNY, D. F. VOYTAS, M. P. CUMMINGS and F. M. AUSUBEL, in preparation).

Ta1-2: A second Ta*1* element is present in the genome of the Kashmir and Landsberg races. A single 2.5 kbp *Bgl*II fragment hybridizes to both central domain and LTR probes in both of these races (Figure 2). This element insertion has been designated Ta*1-2*. Genomic lambda phage libraries constructed from Landsberg and Kashmir DNA were used to clone the Ta*1-2* elements. As described in MATERIALS AND METHODS, these elements are probably methylated in the *A. thaliana* genome since *mcrA*$^-$, *mcrB*$^-$ bacterial hosts were required to isolate the recombinant phage carrying these elements.

The complete nucleotide sequence was obtained for the Ta*1-2* elements from Landsberg and Kashmir (data not shown). Genomic DNA flanking the 3′ LTRs of these insertions are identical, and both elements share an identical 5 bp target site (5′-TTTAT-3′). These two insertions, therefore, represent a single integration event which occurred before the dispersal of the Landsberg and Kashmir races. The empty Ta*1-2* target site was not characterized from Columbia due to the repetitive nature of the sequences which flank this insertion (data not shown). The Kashmir element has suffered a deletion of its 5′ LTR which extends ~60 bp into the central domain and includes the tRNA primer binding site and the beginning of the Ta*1* open reading frame (Figures 3 and 4). As the restriction maps and nucleotide sequence of the genomic DNA upstream of these elements show little similarity (data not shown), it appears that a relatively large deletion event occurred in the DNA flanking the Kashmir element which encompassed the 5′ LTR. The Ta*1-2* elements have been mapped by restriction fragment length polymorphism analysis to *A. thaliana* chromosome *4* (H.-G. NAM, W. LOOS and H. GOODMAN, unpublished results).

Ta1-3: In addition to Ta*1-1* and Ta*1-2*, the Landsberg race carries a third Ta*1* element copy, Ta*1-3* as demonstrated by the 3.0 kbp *Bgl*II fragment which hybridizes to both the INT and LTR probes (Figure 2). Like the Ta*1-2* elements, the Ta*1-3* element is likely methylated in the *A. thaliana* genome (see MATERIALS AND METHODS). Ta*1*-3 does not appear to have suffered any significant deletions since it carries all of the structural and coding features typical of eukaryotic retrotransposons (VOYTAS and AUSUBEL 1988). Ta*1-3* is linked to Ta*1-2* on chromosome *4*, although the precise map position of these elements relative to other markers has not yet been determined

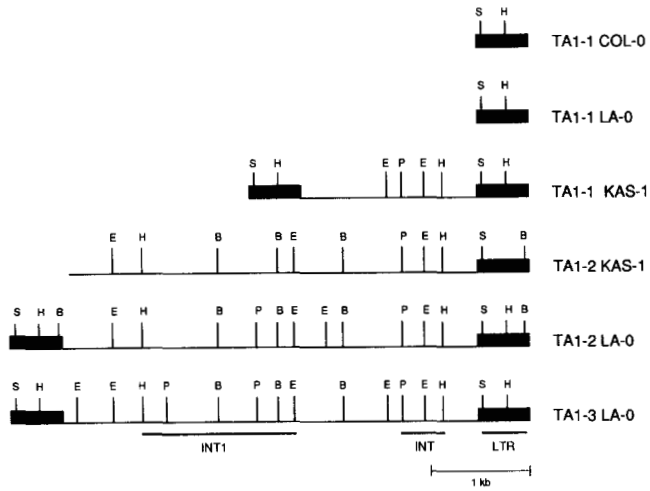

FIGURE 3.—Restriction endonuclease maps of the Ta*1* element copies present in the Col-0 (Columbia), La-0 (Landsberg) and Kas-1 (Kashmir) races. Black boxes represent LTRs and elements are aligned with respect to their 3′ LTR. Underlined sequences of Ta*1-3*, La-0 indicate restriction fragments used as probes in Southern blot (Figure 2) and cloning experiments (see MATERIALS AND METHODS). Restriction enzyme sites are as follows: S, *Sal*I; H, *Hin*dIII; E, *Eco*RI; P, *Pst*I; B, *Bgl*II.

(H.-G. NAM, W. LOOS and H. GOODMAN, unpublished results).

The nucleotide sequence of Ta*1-3* from Landsberg has been previously reported, as has analysis of the empty Ta*1-3* target site (5′-ATCAA-3′) from Columbia (VOYTAS and AUSUBEL 1988). These results suggested that Columbia has never carried a copy of Ta*1-3*, and that the transposition of Ta*1-3* to this site took place after the separation of these races.

Nucleotide sequence comparisons of the Ta*1* central domain among the Ta*1* element family members: The central domain sequences of Ta*1* are strikingly similar among the four elements analyzed. Pairwise comparisons reveal that all elements share >96% nucleotide sequence identity (Table 1). This is only slightly less similarity than that observed between functional copies of Ty*1* elements from yeast (98.9%; BOEKE *et al.* 1988). The greatest degree of similarity between Ta*1* element copies exists between the two Ta*1-2* elements from Kashmir and Landsberg (98.8%). The fact that the Landsberg copy of Ta*1-2* is more similar to its cognate from Kashmir (98.8%) than to the Ta*1-3* element in Landsberg (96.7%) suggests that high levels of concerted evolution are not occurring between Ta*1* elements in the genome of a given race.

The only Ta*1* element copies which do not appear to have suffered appreciable deletions and therefore may have the potential to transpose are the Ta*1-2* and Ta*1-3* elements from the Landsberg race (Figure 3). Several lines of evidence based upon nucleotide sequence comparisons among the Ta*1* elements suggest that these elements are not functional.

1. Because the Ta*1* element copies share such high levels of nucleotide sequence identity, a consensus Ta*1* sequence was determined as well as a consensus for the derived amino acid sequence of the major Ta*1* open reading frame within the central domain (data not shown). Several mutations among the Ta*1* element copies affect the size of the Ta*1* consensus protein. The Ta*1-2* element copies from Kashmir and Landsberg each carry two single base pair insertions/deletions (one of which they share in common) which result in a frameshift of the Ta*1* consensus reading frame (Figure 4). In addition, two nucleotide changes in the Kashmir element result in stop codons, while the Landsberg copy of Ta*1-2* has a single stop codon, all of which truncate the protein product of the consensus open reading frame (Figure 4).

We have previously reported the size of the Ta*1-3* open reading frame as 1291 amino acids (VOYTAS and AUSUBEL 1988). It is apparent from a consensus of the derived translation products of the other Ta*1* element copies, that this open reading frame should extend for an additional 71 amino acids (Figure 5). The premature termination of the Ta*1-3* open reading frame is due to a single base change that results in a stop codon. The terminal 71 amino acids of the Ta*1* consensus protein encompass a conserved amino acid domain which shares homology to the RNase H proteins of various retroviruses and retrotransposons (Figure 5; DOOLITTLE *et al.* 1989).

2. Many of the nucleotide substitutions which occur between the various Ta*1* element copies result in non-conservative replacements of amino acids which are nearly invariant among related retrotransposable elements and retroviruses. For example, a conserved cysteine which constitutes part of the zinc finger of the RNA binding domain in numerous retrotransposons and retroviruses (COVEY 1986) is replaced by a tyrosine in Ta*1-3* (Figure 6). This cysteine is invariant in the RNA binding domain, and the nonconservative substitution of this cysteine for a tyrosine (FRENCH and ROBSON 1983) would probably compromise the function of this protein domain and likewise the ability of this element to engage in active transposition.

3. Protein coding sequences which are not subject to selective evolutionary pressures would be expected to accumulate nucleotide changes which result in approximately 3/4 amino acid replacements and 1/4 silent substitutions (LEWONTIN 1989). Conversely, highly constrained protein coding sequences show a strong bias for silent nucleotide substitutions. The number of silent and replacement changes that have occurred between the Ta*1* elements are roughly equally divided between these two classes of mutations (Table 1). This indicates that the Ta*1* sequences are not highly constrained and the Ta*1* elements have

TABLE 1

Nucleotide comparisons of the Ta*1* coding region

	Ta*1-2* Kas-1					Ta*1-2* La-0					Ta*1-3* La-0				
	NC[a]	CO[b]	%I[c]	%S[d]	%R[e]	NC	CO	%I	%S	%R	NC	CO	%I	%S	%R
Ta*1-1* Kas-1	1763[f]	59	96.7	54.2	45.8	1760[g]	65	96.3	52.3	47.7	1764[h]	48	97.3	50.0	50.0
Ta*1-2* Kas-1						4050[i]	49	98.8	32.7	67.3	4054[j]	130	96.8	46.2	53.8
Ta*1-2* La-0											4087[k]	133	96.7	41.9	57.1

[a] NC = nucleotides compared.
[b] CO = nucleotide changes observed.
[c] %I = percent nucleotide identity.
[d] %S = percent silent amino acid changes.
[e] %R = percent replacement amino acid changes.
[f] A 1-base gap added for alignment.
[g] A gap of 3 bases and a gap of 1 base added for alignment.
[h] No gaps added.
[i] A gap of one nucleotide in common between the two elements; a gap of 3 bases and two gaps each of 1 base added for alignment.
[j] Two gaps each of 1 base added for alignment.
[k] A gap of 3 bases and two gaps each of 1 base added for alignment.

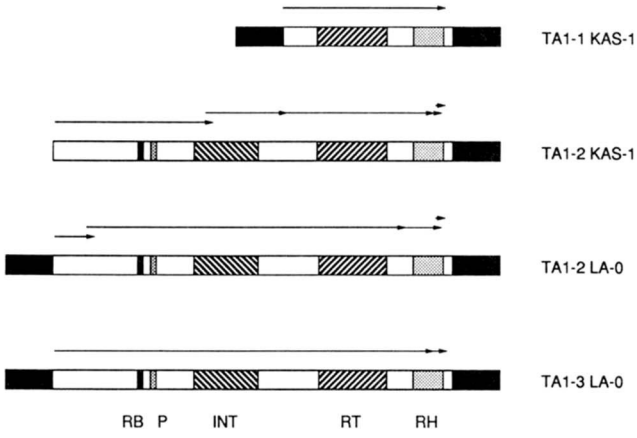

FIGURE 4.—Organization of open reading frames and conserved amino acid sequence domains among the Ta*1* element copies. LTRs are represented by black boxes and elements are aligned with respect to their 3′ LTR. Boxes within the internal portion of the element represent conserved amino acid domains: RB, RNA binding domain; P, protease domain; INT, integrase domain; RT, reverse transcriptase domain; RH, RNase H domain. Arrows depict open reading frames and arrowheads depict stop codons. Offset arrows within the central domain of the Ta*1-2* element copies represent breaks in the open reading frame due to single nucleotide insertions/deletions.

likely been subject to the random accumulation of nucleotide changes.

The distribution of the Ta*1* elements among 15 *A. thaliana* geographical races: We undertook a survey of the Ta*1* elements present in 12 additional races to address two questions: (1) do any of the races have additional copies of Ta*1*; and (2) how did the Ta*1* element family spread over the course of global dispersal of *A. thaliana*?

To assess the number of elements present in each of the races and to determine if they represent one of the already characterized insertion sites, race DNAs were digested with restriction endonucleases that cut within the central domain and flanking DNA to generate restriction fragments characteristic of each of the known element insertions. Southern filters prepared from these DNAs were hybridized with radiolabeled probes to the central domain or LTRs (*e.g.*, INT; Figure 3; data not shown).

The *A. thaliana* races examined contain one, two or three element copies (Figure 7), indicating that the Ta*1* family has not transposed appreciably over the course of the dispersal of the species. All of the races carry a copy of Ta*1-1*, and with the exception of Kashmir, this insertion is a single LTR. For 4/15 races including Columbia, Ta*1-1* is the only Ta*1* element within the genome. Like Kasmir, 5/15 races carry copies of both Ta*1-1* and Ta*1-2*, and like Landsberg, 4/15 carry all three of the characterized element copies. There are two exceptions to this pattern, namely the races Ba-1 from England and Co-4 from Portugal (Figure 7). Neither of these races appear to carry a copy of Ta*1-2*. By analogy to the Ta*1-2* copy in Kashmir, it is possible that a copy of this element was present in these races and subsequently lost from the genome due to a similar, yet more encompassing deletion event. In the case of Co-4, it is uncertain if the additional copy of Ta*1* in this race (Ta*1–4*, Figure 7) represents a unique transposition event, of if this element is actually Ta*1-2*, and sufficient restriction site polymorphisms have occurred making it appear as a unique element insertion.

Phylogenetic comparisons of the Ta*1* sequences were conducted to assess relatedness among the Ta*1* element copies and the tobacco retrotransposon, Tnt*1* (GRANDBASTIEN, SPIELMANN and CABOCHE 1989). There were 50 phylogenetically informative characters used in the analysis, which resulted in a single most parsimonious tree of length 72 (Figure 8). The consistency index was 0.83, excluding autapomorphies. Among the features of the tree are 13 unambiguous character state changes supporting the sepa-

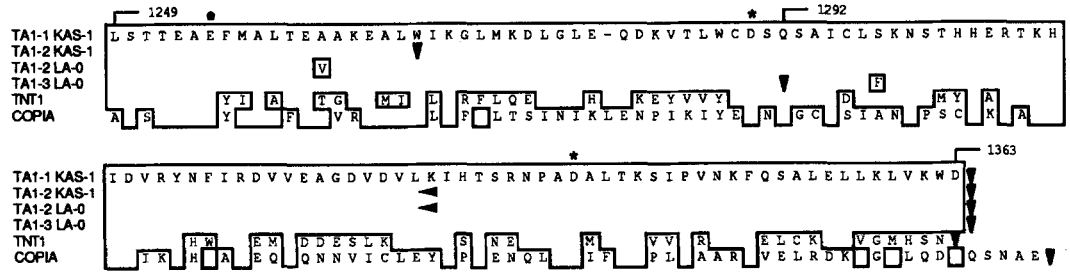

FIGURE 5.—Amino acid alignment of the terminal portion of the open reading frames among *copia*, Tnt*1* and the Ta*1* element copies. Dashes represent breaks in the sequence introduced to optimize alignment. The sequence of the Kas-1 (Kashmir) element, Ta*1-1*, represents the consensus amino acid sequence among the Ta*1* element copies. Identical amino acids among the remaining elements are boxed. Vertical arrowheads indicate stop codons, and arrowheads pointing to the right indicate a shift in the reading frame due to a single nucleotide deletion. The numbers of the amino acids refer to the open reading frame from the La-0 (Landsberg) element, Ta*1-3*. Starred amino acids are invariant among 26 RNase H proteins encoded by various retroviruses, retrotransposons and *E. coli* (DOOLITTLE *et al.* 1989).

HIV-1	C	F	N	C	G	K	E	G	H	T	A	R	N	C
RSV	C	Y	T	C	G	S	P	G	H	Y	Q	A	Q	C
MMULV	C	A	Y	C	K	E	K	G	H	W	A	K	D	C
COPIA	C	H	H	C	G	R	E	G	H	I	K	K	D	C
TNT1	C	Y	N	C	N	Q	P	G	H	F	K	A	D	C
TA1-2 KAS-1	C	W	Y	C	K	K	E	G	H	V	K	K	D	C
TA1-2 LA-0	C	W	Y	C	K	K	E	G	H	V	K	K	D	C
TA1-3 LA-0	C	W	Y	C	K	K	E	G	H	V	K	K	D	**Y**

FIGURE 6.—Nonconservative amino acid substitution in the Ta*1-3* RNA binding domain. The amino acid sequences of RNA binding domains are aligned (GRANDBASTIEN, SPIELMANN and CABOCHE 1989) for human immunodeficiency virus, type 1 (HIV-1), Rous sarcoma virus (RSV), murine Moloney leukemia virus (MMULV), the *D. melanogaster copia* element, the *N. tabacum* Tnt*1* element and the Ta*1* element copies. The invariant cysteines, glycine and histidines are boxed and the cysteine to tyrosine replacement in Ta*1-3* is bold-faced.

ration of the Ta*1-1* clade from the Ta*1-2*/Ta*1-3* clade, which suggests that Ta*1-2* and Ta*1-3* shared a more recent common ancestor. The two copies of Ta*1-2* examined from the Kashmir and Landsberg races are separated from Ta*1-3* by 25 unambiguous character state changes.

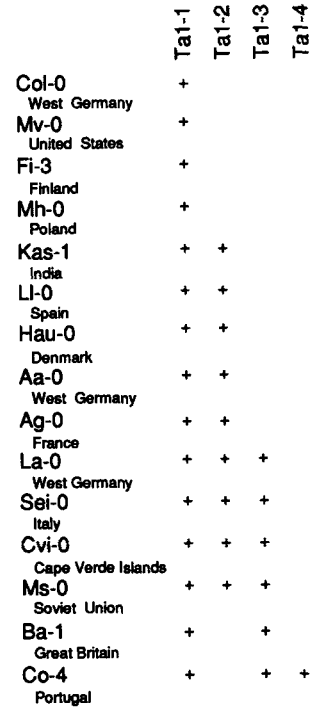

FIGURE 7.—Distribution of the Ta*1* element copies among 15 diverse geographical races of *A. thaliana*. Race origins are indicated (KRANZ and KIRSCHHEIM 1987), and Col-0, Kas-1 and La-0 represent the races Columbia, Kashmir and Landsberg, respectively. +'s indicate the presence of a particular Ta*1* insertion.

DISCUSSION

Ta*1* is no longer capable of transposition: It appears likely that the Ta*1* elements are no longer active within the *A. thaliana* genome. These elements occupy at most only three distinct chromosomal positions within the races analyzed. In addition, the nucleotide sequences of the Ta*1* elements from Columbia, Kashmir and Landsberg indicate that all of these element copies have suffered either crippling deletions or nucleotide changes.

The persistence of a transposable element family within the genome of an organism depends on two factors. First, if the transposable element family is to remain active, it must propagate itself to a copy number sufficient to offset deleterious mutations that occur either through the transposition process or while residing in the genome. Second, the transposition activity required to establish the element family must not compromise the fitness of the host. Because these factors are necessarily interrelated, it is likely they both play a role in ultimately dictating whether or not a transposable element family remains active. Indeed, these considerations have been used in mathematical models to predict either the spread or extinction of transposable element families (CHARLESWORTH and CHARLESWORTH 1983; LANGLEY, BROOKFIELD and KAPLAN 1983; KAPLAN, DARDEN and LANGLEY 1985; CHARLESWORTH and LANGLEY 1986; MONTGOMERY, CHARLESWORTH and LANGLEY 1987; LANGLEY *et al.* 1988).

The *A. thaliana* genome (70 Mb) is the smallest known genome among higher plants (LEUTWILER,

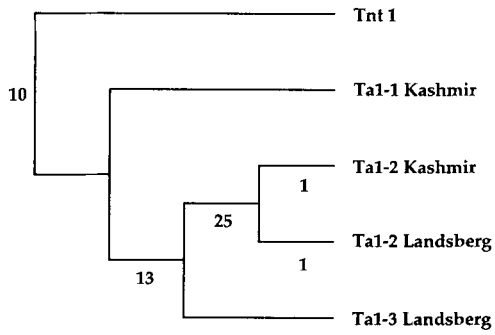

FIGURE 8.—Most parsimonious phylogenetic tree showing the relationships among the Ta*1* elements and Tnt*1*. Numbers below the branch show the number of unambiguous character state changes supporting that branch. Tree length = 72; Consistency index (excluding autapomorphies) = 0.83.

HOUGH-EVANS and MEYEROWITZ 1984) and presumably offers fewer target sites for transposition which would not have a deleterious effect on host fitness. While this may be true, active transposable element families are found in both *Drosophila melanogaster* and *Caenorhabitis elegans* (reviewed in BERG and HOWE 1989) which have similarly small genomes (MANNING, SCHMID and DAVIDSON 1975; SULSTON and BRENNER 1974). Indeed, we have recently identified several additional retrotransposable element families in *A. thaliana* that are related to Ta*1* (A. KONIECZNY, D. F. VOYTAS, M. P. CUMMINGS and F. M. AUSUBEL, in preparation), suggesting that *A. thaliana* is capable of withstanding a relatively high level of transposable element activity.

The failure of the Ta*1* family to spread is more likely due to the inability of these elements to propagate themselves to a sufficient copy number to ensure that at least some functional element copies persist in the face mutation. The source of these mutations can be the reverse transcriptase mediated transposition process, which is known to be error-prone (HOLLAND *et al.* 1982), and DNA replication that occurs between transposition events while the element is resident in the host genome. It is possible that in some *A. thaliana* races, the Ta*1* elements did achieve high copy number, but the distribution of the Ta*1* elements among the 15 geographically diverse races analyzed in this study suggests that these elements accumulated mutations before becoming successfully established in this species.

The evolution of the Ta*1* family: Retrotransposition is a replicative process which results in the accumulation of element copies. While retrotranposon insertions can be lost from the genome through deletion, there are no documented reports of retrotransposon excision. Based on these features of retrotransposition, the present-day distribution of the Ta*1* elements among the 15 races suggests the manner in which these elements entered and spread through the *A. thaliana* genome over the course of the species global dispersal. All of the races carry a copy of Ta*1-1*, indicating that this element predated the other insertions. For 4/15 of the races, Ta*1-1* is the only Ta*1* copy present in the genome and exists as a solo LTR. Since Ta*1-2* is present in the majority of the races (9/15) it is likely that this insertion was the second Ta*1* transposition event in *A. thaliana*. Because 4/16 races which have Ta*1-3* also have Ta*1-2*, Ta*1-3* is likely the most recent Ta*1* transposon insertion. This ordering of transposition events is fully supported by the phylogenetic analysis of the Ta*1* sequences.

There are at least three models by which the spread of the Ta*1* elements could have occurred in *A. thaliana*, none of which can be excluded by the distributional data and phylogenetic analysis (1) the original Ta*1* element, Ta*1-1*, could have given rise to Ta*1-2* which subsequently gave rise to Ta1-*3*; (2) the Ta*1-1* element could have given rise directly to both Ta*1-2* and Ta*1-3*; and (3) each of the elements could have entered the genome independently by horizontal transfer without ever having been derived from an ancestral *A. thaliana* Ta*1* insertion.

It is unlikely that the pattern of Ta*1* element insertions among the races represents a successive loss due the deletion of element copies rather than a successive spread of these elements. The strongest evidence comes from the characterization of an empty Ta*1-3* target site in Columbia (Columbia only carries a copy of Ta*1-1*). The nucleotide sequence of this region is identical to the sequence which flanks the Ta*1-3* insertion in Landsberg (VOYTAS and AUSUBEL 1988). In addition, the 5-bp target site which was duplicated upon the insertion of Ta*1-3* exists as a single copy in Columbia.

Where did the Ta*1* elements originate? While the Ta*1-1* insertion obviously predated the species dispersal, it is still >96% identical at the nucleotide level to the more recent insertions, (*e.g.*, Ta*1-2*, Ta*1-3*). This suggests that both the entrance of this transposable element family into the species and the species dispersal were relatively recent events. We are currently testing the presence and distribution of Ta*1* elements in other species of *Arabidopsis* to determine how widespread these elements are within the genus and if the Ta*1* elements predate the divergence of the *Arabidopsis* species. These experiments should also address the question of whether the Ta*1* family entered *A. thaliana* by some mechanism of horizontal transfer or was inherited vertically over the course of the evolution of the genus *Arabidopsis*.

We would like to thank H.-G. NAM, W. LOOS and H. GOODMAN for performing the RFLP analyses, and J. DOYLE and S. RODERMEL for critical reading of the manuscript. M.P.C. was supported by National Institutes of Health Genetics Training Grant GM07620. This work was funded by a grant from Hoechst AG to Massachusetts General Hospital.

LITERATURE CITED

AJIOKA, J. W., and D. L. HARTL, 1989 Population dynamics of transposable elements, pp. 939–958 in *Mobile DNA*, edited by D. E. BERG and M. M. HOWE. American Society for Microbiology, Washington, D.C.

AUSUBEL, F. M., R. BRENT, R. E. KINGSTON, D. D. MOORE, J. G. SEIDMAN, J. A. SMITH and K. STRUHL, 1987 *Current Protocols in Molecular Biology*. Greene Publishing Associates/Wiley Interscience, New York.

BERG, D. E., and M. M. HOWE, 1989 *Mobile DNA*. American Society for Microbiology, Washington, D.C.

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, D.C.

BOEKE, J. D., 1989, Transposable elements in *Saccharomyces cerevisiae*, pp. 335–374 in *Mobile DNA* edited by D. E. BERG and M. M. HOWE. American Society for Microbiology, Washington, D.C.

BOEKE, J. D., D. EICHINGER, C. CASTRLLON and G. R. FINK, 1988 The *Saccharomyces cerevisiae* genome contains functional and nonfunctional copies of transposon Ty*1*. Mol. Cell. Biol. **8:** 1432–1442.

CHANG, C., J. L. BOWMAN, A. W. DEJOHN, E. S. LANDER and E. M. MEYEROWITZ, 1988 Restriction fragment length polymorphism map for *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **85:** 6856–6860.

CHARLESWORTH, B., and D, CHARLESWORTH, 1983 The population dynamics of transposable elements. Genet. Res. **42:** 1–27.

CHARLESWORTH, B., and C. H. LANGLEY, 1986 The evolution of self-regulated transposition of transposable elements. Genetics **112:** 359–383.

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.

DEVEREUX, J., P. HAEBERLI and O. SMITHIES, 1984 A comprehensive set of sequence programs for the VAX. Nucleic Acids Res. **12:** 387–395.

DOOLITTLE, R. F., D.-F. FENG, M. S. JOHNSON and M. A. MCCLURE, 1989 Origins and evolutionary relationships of retroviruses. Q. Rev. Biol. **64:** 1–30.

FEINBAUM, R. L., and F. M. AUSUBEL, 1988 Transcriptional regulation of the *Arabidopsis thaliana* chalcone synthase gene. Mol. Cell. Biol. **8:** 1985–1992.

FINNEGAN, D. J., and D. H. FAWCETT, 1986 Transposable elements in *Drosophila melanogaster*, pp. 1–62 in *Oxford Surveys of Eukaryotic Genes*, Vol. 3, edited by N. MACLEAN. Oxford University Press, Oxford, England.

FRENCH, S., and B. ROBSON, 1983 What is a conservative substitution? J. Mol. Evol. **19:** 171–175.

GRANDBASTIEN, M.-A., A. SPIELMANN and M. CABOCHE, 1989 Tnt*1*, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature **337:** 376–380.

HOLLAND, J., K. SPINDLER, F. HORODYSKI, E. GRABAU, S. NICHOL and S. VANDEPOL, 1982 Rapid evolution of RNA genomes. Science **215:** 1577–1585.

KAPLAN, N., DARDEN, T. and C. H. LANGLEY, 1985 Evolution and extinction of transposable elements in Mendelian populations. Genetics **109:** 459–480.

KRANZ, A. R., and B. KIRCHHEIM, 1987 Genetic resources in Arabidopsis. Arabidopsis Inform. Serv. **24.**

LANGLEY, C. H., J. F. Y. BROOKFIELD and N. L. KAPLAN, 1983 Transposable elements in Mendelian populations. I. A theory. Genetics **104:** 457–472.

LANGLEY, C. H., E. MONTGOMERY, R. HUDSON, N. KAPLAN and B. CHARLESWORTH, 1988 On the role of unequal exchange in the containment of transposable element copy number. Genet. Res. **52:** 223–235.

LEUTWILER, L. S., B. R. HOUGH-EVANS and E. M. MEYEROWITZ, 1984 The DNA of *Arabidopsis thaliana*. Mol. Gen. Genet. **194:** 15–23.

LEWONTIN, R. C., 1989 Inferring the number of evolutionary events from DNA coding sequence differences. Mol. Biol. Evol. **6:** 15–32.

MADDISON, W. P., and D. R. MADDISON, 1990 MacClade: interactive analysis of phylogeny and character evolution, version 2.99β2. Test version of MacClade 3.0, Sinauer Associates, Sunderland Mass.

MANNING, J. E., C. W. SCHMID and N. DAVIDSON, 1975 Interspersion of repetitive and nonrepetitive DNA sequence in the *Drosophila melanogaster* genome. Cell **4:** 141–155.

MEYEROWITZ, E. M., 1987 *Arabidopsis thaliana*. Annu. Rev. Genet. **21:** 93–111.

MONTGOMERY, E., B. CHARLESWORTH and C. H. LANGLEY, 1987 A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. Genet. Res. **49:** 31–41.

MOUNT, S. M., and G. M. RUBIN, 1985 Complete nucleotide sequence of the *Drosophila* transposable element *copia*: homology between *copia* and retroviral proteins. Mol. Cell. Biol. **5:** 1630–1638.

NEEDLEMAN, S. B., and C. D. WUNSCH, 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48:** 443–453.

RALEIGH, E. A., and G. WILSON, 1986 *Escherichia coli* K-12 restricts DNA containing 5-methylcytosine. Proc. Natl. Acad. Sci. USA **83:** 9070–9074.

REDEI, G. P., 1962 Single locus heterosis. Z. Verebungsl. **93:** 164–170.

SULSTON, J. E., and S. BRENNER, 1974 The DNA of *Caenorhabditis elegans*. Genetics **77:** 795–104.

SWOFFORD, D. L., 1990 Phylogenetic analysis using parsimony, PAUP version 3.0g. Illinois Natural History Survey, Champaign, Ill.

VOYTAS, D. F., and F. M. AUSUBEL, 1988 A *copia*-like transposable element family in *Arabidopsis thaliana*. Nature **336:** 242–244.

Communicating editor: M. R. HANSON