The Structure, Distribution and Evolution of the TaI Retrotransposable Element Family of Arabidopsis thaliana

Daniel F. Voytas,*, Andrzej Konieczny,† Michael P. Cummingst 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 TaI elements are a low copy number, copia-like retrotransposable element family of Arabidopsis thaliana. Six TaI insertions comprise all of the TaI element copies found in three geographically diverse A. thaliana races. These six elements occupy three distinct target sites: TaI-1 is located on chromosome 5 and is common to all three races (Col-0, Kas-1 and La-0). TaI-2 is present in two races on chromosome 4 (Kas-1 and La-0), and TaI-3, located on chromosome 4, is present only in one race (La-0). The six TaI 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 the conserved protein domain required for retrotransposition. Nucleotide sequence comparisons of these elements and the distribution of TaI among 12 additional A. thaliana geographical races suggest that TaI-1 pre-dated the global dispersal of A. thaliana. As the species spread throughout the world, two additional transposition events occurred which gave rise first to TaI-2 and finally to TaI-3.

TRANSGRESSABLE 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 TaI 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, TaI-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 proviruses, TaI-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 TaI-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 TaI-3 are identical to the 3' terminus of plant tRNA<sub>met</sub>, 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 TaI-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 TaI 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*, Tnt1 (Mount and Rubin 1985; Voytas and Ausubel 1988; Grandbastien, Spielmann and Caboche 1989). Unlike other well-characterized retrotransposons, Ta1, copia and Tnt1 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 Ta1 insertions are found within the *Arabidopsis 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 Ta1 elements among different *Arabidopsis 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 Ta1 elements, the relationships among the Ta1 element copies, and the likely manner in which these elements spread throughout the *Arabidopsis thaliana* genome over the course of the global dispersal of the species.

**MATERIALS AND METHODS**

*Arabidopsis thaliana* races: The *Arabidopsis 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 *Arabidopsis 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; Li-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: *Arabidopsis thaliana* DNA isolations were performed by methods previously described (Ausubel et al. 1987). For Southern blot analyses, 1μg of *Arabidopsis thaliana* genomic DNA was subjected to electrophoresis on 0.8% agarose gels and transferred 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°C in 0.2× SSC (Ta = 75°C).

The cloning of the Ta1-1 element from the Kas-1 (Kashmir) race has been previously reported (Voytas and Ausubel 1988). The Ta1-1 element from the Col-0 (Columbia) race was isolated from a library constructed in pUC12 using size-selected DNAs digested to completion with XhoI. Recombinant clones were identified by colony hybridizations (Ausubel et al. 1987) using a DNA probe which flanks the *Arabidopsis thaliana* chalcone synthase structural gene (Feinbaum and Ausubel 1988). The La-0 (Landsberg) copy of Ta1-1 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 Ta1 LTR (Figure 3) and sequences which flank the chalcone synthase gene (Feinbaum and Ausubel 1988).

The Ta1-2 and Ta1-3 elements were isolated from Landsberg and Kashmir genomic DNA libraries constructed in the vector lambda FIX (Stratagene) using MboI partial digests of these DNAs that had been size-fractionated on low-melting temperature agarose gels. The vector DNA was digested with XhoI, 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 Ta1-2 and Ta1-3 elements from either the Landsberg or Kashmir libraries were unsuccessful. Plaque hybridizations using a central element probe from Ta1-1 (INT, Figure 3) failed to identify Ta1 clones within the Landsberg library. Of 41 independent clones isolated from the Kashmir library, all contained the deleted element copy Ta1-1 (see RESULTS) and not the sought after element, Ta1-2.

To prevent repeated cloning of the Ta1-1 elements, a probe was isolated which hybridized specifically to the Ta1-2 and Ta1-3 elements. This probe (INT1, Figure 3), was part of a 3 kbp HindIII fragment which was determined to be unique to Ta1-2 and Ta1-3 by genomic Southern blot analyses (data not shown). The 3-kbp HindIII fragment was isolated from a library constructed in lambda ZAPII (Stratagene) using size selected La-0 DNAs digested to completion with HindIII. The INT probe was used for phage isolation. To account for the possibility that the Ta1-2 and Ta1-3 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 1989). The Ta1-2 and Ta1-3 elements were only isolated from phage plated on ER1458 using the INT1 probe (Figure 3).

**DNA sequencing:** The Ta1 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 Bal31 nuclease or exonuclease III (Ausubel et al. 1987). Oligonucleotides were synthesized to prime sequencing reactions using a Biossearch DNA Synthesizer (New Brunswick Scientific). The DNA sequence was obtained on both strands for each of the Ta1 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 Smith).
RESULTS

Structural organization and distribution of Tal elements from the Columbia, Kashmir and Landsberg races: The Tal 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 BglII 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 BglII 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 LI-0 polymorphism was due to a 2.3-kbp DNA insertion (data not shown). An approximately 20-kbp XbaI fragment which contained the polymorphic BglII 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 SalI fragment (Figure 1, data not shown). The nucleotide sequence was obtained for the Kashmir insertion and flanking DNAs.

Tal-l: The Kashmir insertion, designated Tal-l, 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 Tal-l open reading frame, however, only encompassed the carboxy-terminal half of the copia element protein. Mising from Tal-l was the coding region corresponding to the copia gag gene (data not shown; MOUNT and RUBIN 1985). This suggested that Tal-l had suffered a deletion of the central domain. Subsequent characterization of addition of Tal-l element copies confirmed this observation, and demonstrated that the Tal-l deletion begins immediately within the 5’ LTR and extends through 2.4 kbp of the central domain (Figure 3).

A solo Tal LTR and no additional Tal 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 BglII fragment in which the Kashmir insertion was initially identified, and because this 6 kbp BglII restriction fragment is not polymorphic for most of the races, it seemed likely that all of the A. thaliana races examined carry a solo Tal 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 Tal 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 Tal-l 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
sequence of the LTRs, leaving behind a solo LTR. The Columbia and Landsberg LTRs have also been designated Tal-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 Tal-1) has been mapped by restriction fragment length polymorphism analysis to A. thaliana chromosome 5 (CHANG et al. 1988).

**Tal element copy number**: The Tal 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, Tal-1, DNAs from these races were digested with enzymes that cut within the central domain and 3' flanking DNA. This enabled each Tal 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 Tal element copies, respectively (e.g., Figure 2). For Columbia, the single Tal-1 LTR is the only Tal 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 Tal 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 Tal (A. KONIEZNY, D. F. VOYTAS, M. P. CUMMINGS and F. M. AUSUBEL, in preparation).

**Tal-2**: A second Tal element is present in the genome of the Kashmir and Landsberg races. A single 2.5 kbp BglII fragment hybridizes to both central domain and LTR probes in both of these races (Figure 2). This element insertion has been designated Tal-2. Genomic lambda phage libraries constructed from Landsberg and Kashmir DNA were used to clone the Tal-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 Tal-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 Tal-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 Tal 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 Tal-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).

**Tal-3**: In addition to Tal-1 and Tal-2, the Landsberg race carries a third Tal element copy, Tal-3 as demonstrated by the 3.0 kbp BglII fragment which hybridizes to both the INT and LTR probes (Figure 2). Like the Tal-2 elements, the Tal-3 element is likely methylated in the A. thaliana genome (see MATERIALS AND METHODS). Tal-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). Tal-3 is linked to Tal-2 on chromosome 4, although the precise map position of these elements relative to other markers has not yet been determined.

---

**Figure 2.**—Southern blot analysis of the Tal elements within the Col-0 (Columbia), La-0 (Landsberg) and Kas-1 (Kashmir) races. DNAs were digested with BglII 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 Tal hybridizing bands representing each of the different element copies are indicated by arrows, with sizes referring to their molecular lengths.
sequence comparisons among the Tal elements suggest that these elements are not functional.

HindIII; Tal-3 elements from the Landsberg race (Figure 3).

Several lines of evidence based upon nucleotide sequence identity, a consensus Tal sequence was determined as well as a consensus for the derived amino acid sequence of the major Tal open reading frame within the central domain (data not shown). Several mutations among the Tal element copies affect the size of the Tal consensus protein. The Tal-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 Tal consensus reading frame (Figure 4). In addition, two nucleotide changes in the Kashmir element result in stop codons, while the Landsberg copy of Tal-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 Tal-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 Tal element copies, that this open reading frame should extend for an additional 71 amino acids (Figure 5). The premature termination of the Tal-3 open reading frame is due to a single base change that results in a stop codon. The terminal 71 amino acids of the Tal 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 Tal 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 Tal-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 Tal elements are roughly equally divided between these two classes of mutations (Table 1). This indicates that the Tal sequences are not highly constrained and the Tal elements have

**FIGURE 3.—Restriction endonuclease maps of the Tal 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 Tal-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, SauI; H, HindIII; E, EcoRI; P, PstI; B, BglII.**


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

**Nucleotide sequence comparisons of the Tal central domain among the Tal element family members:** The central domain sequences of Tal 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 elements from yeast (98.9%; BOEKEL et al. 1988). The greatest degree of similarity between Tal element copies exists between the two Tal-2 elements from Kashmir and Landsberg (98.8%). The fact that the Landsberg copy of Tal-2 is more similar to its cognate from Kashmir (98.8%) than to the Tal-3 element in Landsberg (96.7%) suggests that high levels of concerted evolution are not occurring between Tal elements in the genome of a given race.

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

Nucleotide comparisons of the Tal coding region

<table>
<thead>
<tr>
<th></th>
<th>Tal-1 Kas-1</th>
<th>Tal-2 Kas-1</th>
<th>Tal-1 La-0</th>
<th>Tal-2 La-0</th>
<th>Tal-3 La-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>1763</td>
<td>1760</td>
<td>1764</td>
<td>1764</td>
<td>1764</td>
</tr>
<tr>
<td>CO</td>
<td>59</td>
<td>65</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>%I</td>
<td>96.7</td>
<td>96.3</td>
<td>97.3</td>
<td>97.3</td>
<td>97.3</td>
</tr>
<tr>
<td>%S</td>
<td>54.2</td>
<td>52.3</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>%R</td>
<td>45.8</td>
<td>47.7</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
</tbody>
</table>

* NC = nucleotides compared.
* CO = nucleotide changes observed.
* %I = percent nucleotide identity.
* %S = percent silent amino acid changes.
* %R = percent replacement amino acid changes.

* A gap of 3 bases and a gap of 1 base added for alignment.
* 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.
* Two gaps each of 1 base added for alignment.
* 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 Tal 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 Tal-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 Tal elements among 15 A. thaliana geographical races:** We undertook a survey of the Tal elements present in 12 additional races to address two questions: (1) do any of the races have additional copies of Tal; and (2) how did the Tal 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 Tal family has not transposed appreciably over the course of the dispersal of the species. All of the races carry a copy of Tal-1, and with the exception of Kashmir, this insertion is a single LTR. For 4/15 races including Columbia, Tal-1 is the only Tal element within the genome. Like Kashmir, 5/15 races carry copies of both Tal-1 and Tal-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 Tal-2. By analogy to the Tal-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 Tal in this race (Tal-4, Figure 7) represents a unique transposition event, or if this element is actually Tal-2, and sufficient restriction site polymorphisms have occurred making it appear as a unique element insertion.

Phylogenetic comparisons of the Tal sequences were conducted to assess relatedness among the Tal element copies and the tobacco retrotransposon, Tnt1 (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-
Tal Retrotransposons of Arabidopsis

TA1-1 KAS1
TA1-2 KAS1
TA1-3 LA4

LSTSEQALEHKGWVATELGVTLMQCEGKCAVICKSNTHRTK

TA1-2 KAS1
TA1-3 LA4

LSTSEQALEHKGWVATELGVTLMQCEGKCAVICKSNTHRTK

TA1-3 LA4

LSTSEQALEHKGWVATELGVTLMQCEGKCAVICKSNTHRTK

FIGURE 5.—Amino acid alignment of the terminal portion of the open reading frames among copia, Tntl and the Tal element copies. Dashes represent breaks in the sequence introduced to optimize alignment. The sequence of the Kas-1 (Kashmir) element, Tal-1, represents the consensus amino acid sequence among the Tal 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, Tal-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
RSV
MMULV

C Y T C G S P G H Y Q A Q C

C A Y C K E G H W A K D C

C H K E G R E G H I K K D C

C Y N C N Q P G H F K A D C

C N Y C K E G H V K K D C

C W Y C K E G H V K K D C

C W Y C K E G H V K K D C

FIGURE 6.—Nonconservative amino acid substitution in the Tal-3 RNA binding domain. The amino acid sequences of RNA binding domains are aligned (Grandbastien, Spielmann and Cabocche 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 Tntl element and the Tal element copies. The invariant cysteines, glycine and histidines are boxed and the cysteine to tyrosine replacement in Tal-3 is bold-faced.

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

DISCUSSION

Tal is no longer capable of transposition: It appears likely that the Tal 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 Tal 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,
The source of these mutations can be element activity. Transposition events while the element is resident in the host genome. It is possible that in some process, which gate themselves to a sufficient copy number to ensure that at least some functional element copies persist in the genome over the course of element copies. While retrotransposition is a replicative process which results in the accumulation of element copies, it is still >96% identical at the nucleotide level to the more recent insertions, (e.g., Tal-2, Tal-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 Tal elements in other species of Arabidopsis to determine how widespread these elements are within the genus and if the Tal elements predate the divergence of the Arabidopsis species. These experiments should also address the question of whether the Tal 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


Communicating editor: M. R. HANSON