A Superfamily of *Arabidopsis thaliana* Retrotransposons

Andrzej Konieczny,* Daniel F. Voytas,*[,1] 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 October 2, 1990
Accepted for publication December 17, 1990

ABSTRACT

We describe a superfamily of *Arabidopsis thaliana* retrotransposable elements that consists of at least ten related families designated Ta*1*–Ta*10*. The Ta*1* family has been described previously. Two genomic clones representing the Ta*2* and Ta*3* elements were isolated from an *A. thaliana* (race Landsberg *erecta*) λ library using sequences derived from the reverse transcriptase region of Ta*1* as hybridization probes. Nucleotide sequence analysis showed that the Ta*1*, Ta*2* and Ta*3* families share >75% amino acid identity in pairwise comparisons of their reverse transcriptase and RNase H genes. In addition to Ta*1*, Ta*2* and Ta*3*, we identified seven other related retrotransposon families in Landsberg *erecta*, Ta*4*–Ta*10*, using degenerate primers and the polymerase chain reaction to amplify a highly conserved region of retrotransposon-encoded reverse transcriptase. One to two copies of elements Ta*2*–Ta*10* are present in the genomes of the *A. thaliana* races Landsberg *erecta* and Columbia indicating that the superfamily comprises at least 0.1% of the *A. thaliana* genome. The nucleotide sequences of the reverse transcriptase regions of the ten element families place them in the category of copia-like retrotransposons and phylogenetic analysis of the amino acid sequences suggests that horizontal transfer may have played a role in their evolution.

MOBILE genetic elements that proliferate by reverse transcription comprise a substantial fraction of eukaryotic genomic DNA. For mice, humans and *Drosophila melanogaster*, it has been estimated that as much as 10% of the genome consists of reverse transcribing elements (*i.e.*, retrotransposons and endogenous retroviruses, TEMIN 1985; BINGHAM and ZACHAR 1989). In plants, retrotransposable elements have only recently been described, but the rapidly growing catalog of plant retrotransposons suggests that these elements may be as commonplace in plant genomes as they are in the genomes of other higher eukaryotes (VOYTAS and AUSUBEL 1988; GRANDBASTIEN, SPIELMANN and CABOCHE 1989; JIN and BENNETZEN 1989; JOHNS *et al.* 1989; LUCAS, MOORE and FLAVELL 1989; SMYTH *et al.* 1989).

Most eukaryotic reverse transcribing elements (retroelements) identified thus far can be divided in two major groups with respect to their structural similarities (TEMIN 1985). Retroviruses and some transposable elements containing long terminal direct repeats (LTRs) make up the first group. A second group consisting of fungal mitochondrial introns and a variety of retrotransposons that lack LTRs has been referred to as non-LTR retrotransposons (XIONG and EICKBUSH 1988).

Like retroviruses, LTR retrotransposons consist of a large internal domain (3–5 kbp) flanked by LTRs (300–500 bp). Transcription initiation and termination signals are carried within the LTRs and direct the synthesis of RNA transcripts that encode the protein products of the *pol* gene which complete the transposition process. The *pol* gene encodes several enzymatic activities including protease, RNase H, reverse transcriptase (RT) and integrase. The reverse transcriptase region of the *pol* gene is the most highly conserved sequence of the retroelements (DOOLITTLE *et al.* 1989).

Within the class of LTR retrotransposons, two major lineages can be distinguished which differ in the linear arrangement of the putative enzymatic functions encoded by the *pol* gene. In one lineage, the order is: protease, reverse transcriptase, RNase H; integrase. Elements in this lineage, which are more closely related to the retroviruses than to other classes of retrotransposons, include the *gypsy, 17.6, 297* and *412* elements from *D. melanogaster* (MARLOR, PARKHURST and CORCES 1986; SAIGO *et al.* 1984; INOUYE, YUKI and SAIGO 1986; YUKI, ISHIMARU and SAIGO 1986), the Ty*3* elements of yeast (HANSEN, CHALKER and SANDMAYER 1988), and the *del* elements of lily (SMYTH *et al.* 1989). In the second lineage the order of enzymatic activities encoded by *pol* is: protease; integrase; reverse transcriptase; RNase H. The second lineage includes the *copia* and *1731* elements from *D. melanogaster* (MOUNT and RUBIN 1985; FOURCADE-PERRONET *et al.* 1988); the Ty*1* and Ty*2* elements from yeast (CLARE and FARABAUGH 1985; WARMING-

[1] Current address: Department of Botany, Iowa State University, Ames, Iowa 50011.

TON *et al.* 1985); the Tnt1 element from *Nicotiana tabacum* (GRANDBASTIEN, SPIELMANN and CABOCHE 1989) as well as by Ta*1* elements from *Arabidopsis thaliana* (VOYTAS and AUSUBEL 1988; VOYTAS *et al.* 1990). We refer to this latter lineage as "copia-like" retrotransposons.

The present study was motivated by our finding that probes derived from the reverse transcriptase region of Ta*1* hybridize weakly to non-Ta*1* *A. thaliana* sequences. We report here the identification of ten families of copia-like elements which represent most, if not all, of the members of a superfamily of copia-like retrotransposable elements in *A. thaliana*. This study is the first comprehensive analysis of an entire superfamily of transposable elements within the genome of a single species.

MATERIALS AND METHODS

Plant material: Seeds of *A. thaliana* races Landsberg, Columbia and Kashmir were obtained from the *Arabidopsis* Information Service (KRANZ and KIRCHHEIM 1987). Landsberg carries the recessive mutation *erecta*.

DNA manipulations: *A. thaliana* DNA was isolated from whole plants using a standard procedure (AUSUBEL *et al.* 1990). The clones, λ31-3 and λ31-5, were isolated from a genomic library of Landsberg *erecta* DNA constructed in lambda FIX (Stratagene) according to manufacturer's instructions. Recombinant phage were plated on the *Escherichia coli mcrA mcrB* strain ER1458 (RALEIGH and WILSON 1986). Plaque hybridizations were conducted as previously described using a Ta*1* element reverse transcriptase probe (INT, Figure 1, VOYTAS *et al.* 1990).

For Southern blot analyses, 5 μg of *A. thaliana* genomic DNA were digested with restriction endonucleases listed in the legend to Figure 4, subjected to electrophoresis in 0.8% agarose gels, and transferred to Gene Screen Plus nylon membranes (New England Nuclear). DNA probes (Figure 1) were labeled by random priming (AUSUBEL *et al.* 1990) and hybridized to the filters using conditions recommended by the manufacturer. Filters were washed at 65° in 0.2 × SSC.

DNA sequences were obtained using the dideoxy method (AUSUBEL *et al.* 1990) with Sequenase (U.S. Biochemical Corp.) and both single- and double-stranded DNA templates. Nested deletions were generated with Bal31 nuclease and cloned into M13 phage vectors (AUSUBEL *et al.* 1990). DNA sequences were assembled on a VAX computer (Digital Equipment Corporation) using the Multiple Sequence Editor (W. GILBERT, unpublished). Amino acid alignments were performed with ALIGN (DAYHOFF, BARKER and HUNT 1983). Other DNA and protein sequence data analyses were performed with the programs of the University of Wisconsin Genetics Computer Group (DEVEREUX, HAEBERLI and SMITHIES 1984). The DNA sequences of Ta*2* and Ta*3* and the partial sequences of Ta*4*-Ta*10* have been submitted to Genbank.

The polymerase chain reaction (PCR) was used to clone putative reverse transcriptase regions from *A. thaliana* Landsberg *erecta* DNA using 1μg of DNA as the template for the reactions. The oligonucleotide primers were synthesized based on the consensus sequence for the reverse transcriptase region of the *pol* gene shown in Figure 3: Primer 1: 5′AYRTCRTCNACRTANAGNAG-3′; primer 2: 5′AARACNGCNTTYTTRMAYGG-3′, where M = A + C, N = A + C + G + T, R = A + G and Y = T + C. The primers are oriented to amplify a 268-bp fragment of reverse transcriptase. PCR was performed using reagents provided in the GenAmp DNA Amplification Kit (Perkin Elmer-Cetus). Conditions for the reaction were: denaturation for 30 sec at 93°; annealing for 30 sec at 50°; polymerization for 3 min at 72°. The cycle was repeated 30 times. The amplification products were gel purified and cloned in the vectors M13mp18 (YANISCH-PERRON, VIEIRA and MESSING 1985) or pUC13 (NORRANDER, KEMPE and MESSING 1983) digested with *Sma*I. DNA sequence data were obtained for 34 independent clones.

Phylogenetic analysis: DNA sequences were assembled and translated using the programs of the University of Wisconsin Genetics Computer Group (DEVEREUX, HAEBERLI and SMITHIES 1984). Derived amino acid sequences were chosen from 11 clones thought to represent distinct elements in the Landsberg *erecta* race based on Southern blot hybridizations and preliminary sequence comparisons. These sequences, along with the corresponding sequences from other available related retrotransposons, were aligned using the computer program TreeAlign (HEIN, 1989a,b). The gap penalty used was $g_k = 10 + (3 \times k)$.

The TreeAlign program simultaneously aligns a set of sequences and reconstructs a phylogenetic tree using nearest neighbor interchanges and the parsimony criterion (HEIN, 1989a,b). However, it is known that nearest neighbor interchanges is not the most effective method of finding the most parsimonious phylogeny (SWOFFORD 1990). Therefore the aligned sequences were entered into a prerelease version of the computer program MacClade (MADDISON and MADDISON 1991), and the TreeAlign tree was reconstructed using the tree manipulation features of the MacClade program. This tree was then used as the initial tree for further phylogenetic analysis using the computer program PAUP, version 3.0g (SWOFFORD 1990). Each amino acid position was scored as an unordered character and regions representing the amino acids coded for by the PCR primers were eliminated from the analysis. Both subtree pruning-regrafting and tree bisection-reconnection branch-swapping algorithms were used. The phylogenetic tree was rooted using the outgroup method with the Ty*1* retrotransposon of yeast as the outgroup.

RESULTS

Identification of two *A. thaliana* retrotransposable elements related to Ta*1*: Reverse transcriptase is the most highly conserved protein encoded by retroviruses and retrotransposons (DOOLITTLE *et al.* 1989). In Southern blot analyses used to characterize the *A. thaliana* Ta*1* retrotransposable element family, we typically observed a number of sequences which hybridized weakly to Ta*1* reverse transcriptase probes (VOYTAS *et al.* 1990). To clone these cross-hybridizing sequences, we screened an *A. thaliana* (Landsberg *erecta*) genomic library constructed in λFIX using the Ta*1* reverse transcriptase probes INT and INT3 shown in Figure 1. As shown in Figure 1, two clones, λ31-3 and λ31-5, were identified and then characterized by restriction endonuclease and Southern blot hybridization analysis to delimit the regions that hybridized to the Ta*1* reverse transcriptase probes (data not shown).

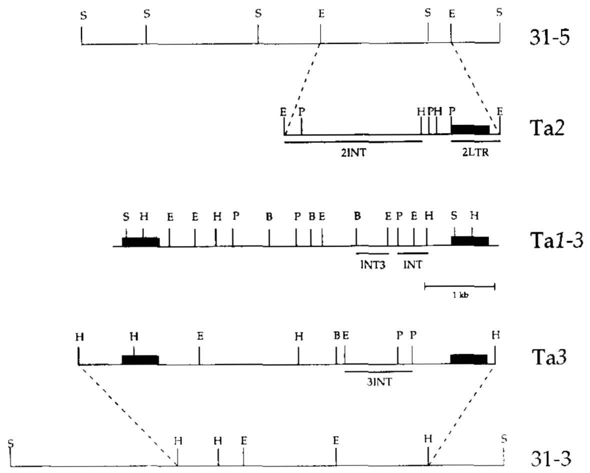

FIGURE 1.—Restriction maps of the Ta*1–3*, Ta2 and Ta*3* retrotransposable elements. Black boxes represent the LTRs for each of the elements. INT and INT3 are restriction fragments used as hybridization probes for identification and initial characterization of Ta2 and Ta*3*. 2LTR is the hybridization probe used to look for a second Ta2 LTR. 2INT is the hybridization probe used for determination of copy number of Ta2 and 3INT is the probe used for determination of Ta*3* copy number. Restriction endonuclease sites are as follows: B = *Bgl*II; E = *Eco*RI; H = *Hin*dIII; P = *Pst*I; S = *Sal*I. A description of Ta*1–3* has been previously reported (VOYTAS and AUSUBEL 1988).

Ta2: The DNA sequence of a 3-kbp *Eco*RI fragment from λ31–5 that hybridized to the Ta*1* probes was determined and shown to contain three short reading frames (Figure 2) which share a high degree of amino acid sequence similarity with the terminal portion of the Ta*1* open reading frame: 84.5% similarity with part of the Ta*1* INT protein (104 of 123 residues) and 86.6% similarity with the Ta*1* reverse transcriptase and RNase H (252 of 299 residues, Table 1). In contrast, members of the Ta*1* element family all share >96% nucleotide identity at both the nucleotide and amino acid level (VOYTAS *et al.* 1990). We therefore concluded that λ31–5 carried a retrotransposon that we named Ta2 that was related to but distinct from Ta*1*.

Because copia-like retrotransposons are flanked by LTRs, we expected that the sequences downstream of the Ta2 open reading frames would be repeated upstream of the element. However, when sequences encompassing the putative Ta2 3′ LTR were used as hybridization probes (2LTR, Figure 1) to localize a λ31–5 restriction fragment containing the 5′ LTR, no hybridizing sequences were found. Because λ31–5 contains at least 6.5 kbp of DNA upstream of the putative Ta2 3′ LTR which is sufficient to encode a complete retrotransposon, we concluded that the 5′ LTR of Ta2 had been deleted. A member of Ta*1* element family, Ta*1–2*, has undergone a similar deletion of its 5′ LTR (VOYTAS *et al.* 1990).

The following observations suggest that the 3′ LTR of Ta2 is intact. First, a polypurine tract (PPT), the putative primer for second strand DNA synthesis during reverse transcription, is invariantly present at the junction of the internal sequence domain and the 3′ LTR of retroviruses and LTR-retrotransposons (VARMUS and BROWN 1989). Sixty-six base pairs downstream of the third Ta2 ORF (Ta2c, Figure 2) is a 17-bp purine-rich sequence (Figure 3). This sequence is identical to the PPT of Ta*1*, and 16 of 17 nucleotides are identical to the PPT of Ta*3* (see below). Second, immediately downstream of the Ta2 polypurine tract is the dinucleotide TG and 520 nucleotides downstream of this is the dinucleotide CA (the invariant LTR end-sequences of retroviruses and retrotransposons; VARMUS and BROWN 1989) (Figure 3). Ta*1* and Ta*3* LTRs also begin and end with these same dinucleotides (see below). Third, the putative Ta2 LTR shares 60.4% nucleotide identity with Ta*1* LTRs and 55.6% nucleotide identity with the LTRs of the Ta*3* element. In addition, a stretch of 70 nucleotides within the putative Ta2 LTR shares 91.4% identity (64 out of 70 residues) with the Ta*1* LTRs (data not shown).

Ta3: A 6.0-kbp region of λ31–3, the second clone that hybridized to the Ta*1* reverse transcriptase probes, was sequenced and revealed a single 4139 bp open reading frame that shares ~75% nucleotide and

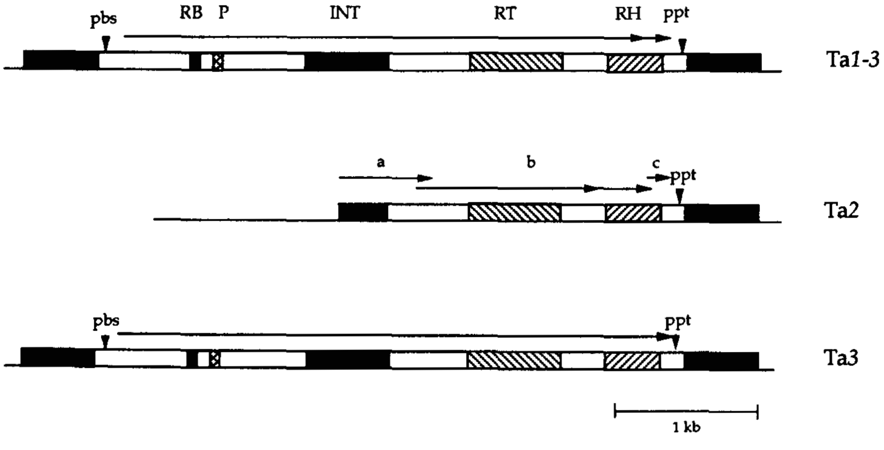

FIGURE 2.—Structural features of Ta*1*-*3*, Ta2 and Ta*3*. Black boxes represent LTRs. Shaded boxes represent conserved amino acid sequence domains common among retrotransposons and retroviruses: RB = RNA binding domain, P = protease, INT = integrase, RT = reverse transcriptase, RH = RNase H, PBS = primer binding site, and PPT = polypurine tract. Arrows over the elements represent open reading frames and arrowheads represent stop codons.

TABLE 1

Amino acid similarities within internal domains of *A. thaliana* retrotransposons

Elements compared	RNA binding domain (%)	Protease (%)	Integrase (%)	Tether (%)	Reverse transcriptase (%)	RNase H (%)
Ta*1-3 vs.* Ta2	N/A[a]	N/A	84.5 (123)[b]	70.8 (182)	86.6 (299)	78.2 (138)
Ta*1-3 vs.* Ta*3*	66.0 (30)	83.0 (30)	82.1 (241)	51.3 (195)	81.4 (216)	75.5 (92)
Ta*2 vs.* Ta*3*	N/A	N/A	78.6 (122)	52.8 (195)	82.8 (216)	81.5 (92)
Ta*3 vs.* Tnt*1*	45.4 (33)	33.0 (12)	43.9 (241)	30.7 (195)	55.0 (216)	56.0 (92)

The alignment was obtained using PIRALIGN program. Tether = a nonconserved region between integrase and reverse transcriptase domains.

[a] N/A = not applicable.

[b] Numbers in parentheses represent the number of amino acids compared.

```
PRIMER BINDING SITE
                                5' LTR ─────┐
Ta1-3  AAGGTTTAAGGTTCGTTTGGTAACA AG TGGTATCAGAGCCATTGGTTCTTGCGAGCTATG
Ta3    TCGTATTGGGATCTGTTTTACAACA AG TGGTATCAGAGCGAGGCTTACTCGTTTCTTGAT
Tnt1   TTTGGTAAGGGGTTTATTCCCAACA AC TGGTATCAGAGCACAGGTTCTGCTCGTTCACTG

POLYPURINE TRACT
                                      ┌──── 3' LTR
Ta1-3  AGTAATTCACGGTTGGAATAGGATCAAGGTGGAGAT TGTTGGAGTTATGATCCAATTCCTA
Ta2    GGTATGAAGAAGAGGAATGAGGATCAAGGTGGAGAT TGTTAAGAAGTGATCCTATTCGGTT
Ta3    ATGATATTGAGATGGGAATGGGATCAAGGTGGAGAT TGTTATGATTATGATCCAATTCGGG
Tnt1   TACCTCCTCTGGATGAATGAGACTGGAGGGGGAGAT TGATGATGTCCATCTCATTGAAGAA
```

FIGURE 3.—Nucleotide sequence comparisons of the primer binding sites and polypurine tracts of the *A. thaliana* retrotransposons and the Tnt*1* elements of *N. tabacum*. Consensus sequences are in plain type. The 12 bp of the primer binding site which are identical to plant tRNA are underlined. The differences from the consensus are in bold.

~72% amino acid identity to both the Ta*1* and Ta*2* open reading frames. We designated this element Ta*3* since it is equally distinct from both the Ta*1* and Ta*2* elements. Unlike Ta*2*, Ta*3* appears to be a structurally complete element, with an internal domain consisting of a single ORF flanked by two LTRs. The 3′ LTR (499 bp) and the 5′ LTR (485 bp) share 96.1% identity and differ by a short 14 bp insertion/deletion and several nucleotide substitutions (data not shown). The percent identity between the Ta*3* LTRs (96.1%) is similar to that observed between the LTRs of a given Ta*1* element copy (e.g. 98.4%, for Ta*1–3;* VOYTAS *et al.* 1990).

The Ta*3* internal domain contains two sites that most likely serve to prime DNA synthesis by reverse transcription. Adjacent to the Ta*3* 5′ LTR is a 12 bp sequence identical to plant $tRNA^{i}_{met}$ (GAUSS and SPRINZL 1983) (Figure 3); analogous sequences are found in most retrotransposons (including Ta*1*) and retroviruses and prime first strand DNA synthesis. Like Ta*1* and Ta*2*, Ta*3* also has a polypurine tract adjacent to the 3′ LTR (Figure 3) for priming second strand DNA synthesis. Immediately flanking Ta*3* are two 5-bp direct repeats (ATCTC), most probably target site duplications generated upon integration of the element into the genome as was shown for Ta*1* (VOYTAS and AUSUBEL 1988).

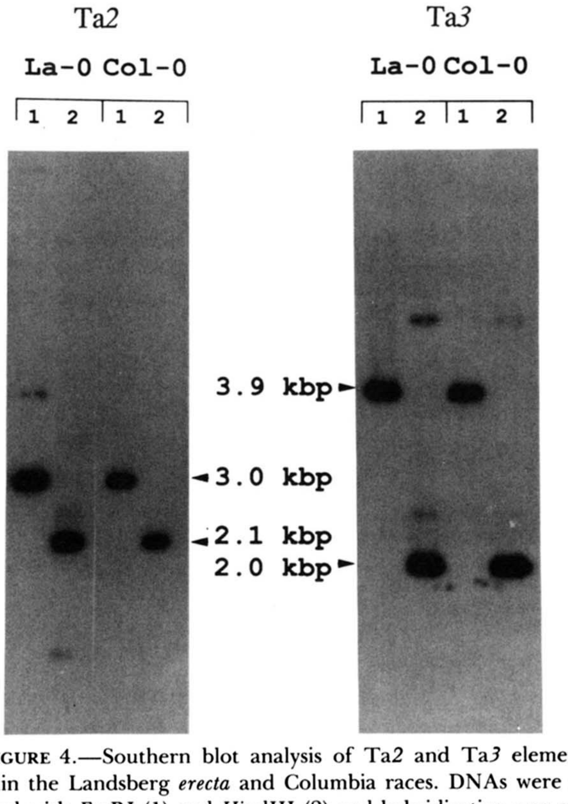

FIGURE 4.—Southern blot analysis of Ta*2* and Ta*3* elements within the Landsberg *erecta* and Columbia races. DNAs were digested with *Eco*RI (1) and *Hin*dIII (2) and hybridization was performed as described in MATERIALS AND METHODS using 2INT and 3INT probes specific for internal domains of Ta2 and Ta*3* respectively (see Figure 1 for probes).

Copy numbers of the Ta2 and Ta3 element families: The copy number of the Ta2 and Ta*3* element families was determined by Southern blot analysis. Southern filters were prepared with DNA isolated from three *A. thaliana* races digested separately with *Eco*RI and *Hin*dIII. Based on the restriction maps of the cloned elements, each element copy should be visualized as a uniquely sized restriction fragment when hybridized with appropriate internal domain probes (2INT and 3INT, Figure 1). Both the Ta*2* and Ta*3* elements are present as a single copy within the Columbia, Landsberg *erecta* (Figure 4) and Kashmir (data

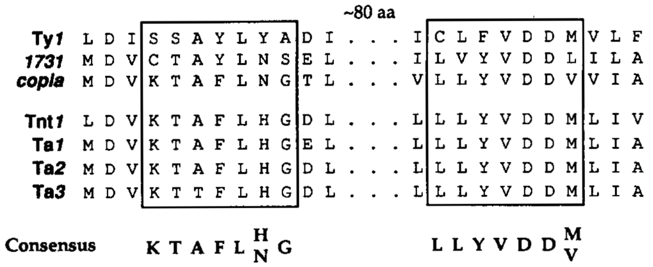
~80 aa
Ty1 L D I S S A Y L Y A D I . . . I C L F V D D M V L F
1731 M D V C T A Y L N S E L . . . I L V Y V D D L I L A
copia M D V K T A F L N G T L . . . V L L Y V D D V V I A
Tnt1 L D V K T A F L H G D L . . . L L L Y V D D M L I V
Ta1 M D V K T A F L H G E L . . . L L L Y V D D M L I A
Ta2 M D V K T A F L H G D L . . . L L L Y V D D M L I A
Ta3 M D V K T T F L H G D L . . . L L L Y V D D M L I A
Consensus K T A F L H/N G L L Y V D D M/V

FIGURE 5.—Schematic alignment of reverse transcriptase sequences of seven members of the "*copia*-like" class of retrotransposons. The primers for PCR were synthesized based upon the consensus amino acid sequence (see MATERIALS AND METHODS).

not shown) geographical races. Although these elements exist as single insertions, we refer to Ta*2* and Ta*3* as element families.

Identification and sequence analysis of other *A. thaliana copia*-like elements: We used the PCR to amplify the reverse transcriptase region of presumptive retroelements from *A. thaliana* Landsberg *erecta* genomic DNA by utilizing a pair of degenerate oligonucleotide primers that correspond to highly conserved regions of the reverse transcriptase region of the *pol* gene (Figure 5). The products of amplification migrated on acrylamide gels as a major band of 268 bp which was the expected size (data not shown). The fragments representing the 268-bp band were cloned and a total of 34 independent clones were sequenced as described in MATERIALS AND METHODS.

Sequence analysis: When the putative amino acid sequences encoded by the 34 clones were compared with those of other transposable elements, we found that the elements could be divided into ten distinct families (Table 2). The criterion for assignment to a family was >90% amino acid identity in pairwise comparisons. The alignment of amino acid sequence of one member from each of the ten families (including Ta*1*, Ta*2* and Ta*3*) is presented in Figure 6. Figure 6 also shows the corresponding regions of the *D. melanogaster* elements *copia* (MOUNT and RUBIN 1985) and *1731* (FOURCADE-PERRONET *et al.* 1988), the tobacco Tnt*1* element (GRANDBASTIEN, SPIELMANN and CABOCHE 1989), and the yeast Ty*1* element (CLARE and FARABAUGH 1985). The alignment includes three regions (boxed and numbered 1–3) that are highly conserved among all of the sequences shown in Figure 6; a similar pattern of conserved sequences is observed when all known reverse transcriptases are compared (XIONG and EICKBUSH 1988). The *A. thaliana* reverse transcriptase sequences obtained by PCR are similar to those of other retrotransposons; the level of amino acid similarity between the families extends from 37% (between Ta*8* and Ta*1–3*) to 85% (between Ta*8* and Ta*9*). This level of amino acid similarity for elements of the same class is considered high (*e.g.*, Visna and MuLV show 25% similarity; XIONG and EICKBUSH 1988). These data indicate that the products of amplification are not likely to be artifactual but represent true *A. thaliana* retrotransposons.

TABLE 2

Products of PCR arranged in families based on amino acid similarity

Element family	No. of clones recovered from PRC products	Average nucleotide divergence between clones based on pairwise comparisons (%)	No. of copies
Ta*1*	11	<1.0	1-3[a]
Ta*2*	2	1.7	1[b]
Ta*3*	0	N/A[c]	1[b]
Ta*4*	3	4.2	1
Ta*5*	2	0	1
Ta*6*	1	N/A	2
Ta*7*	4	<1.0	1
Ta*8*	2	0	1
Ta*9*	1	N/A	1
Ta*10*	9	1.3	1

[a] From VOYTAS *et al.* (1990).
[b] Determined as described in the section on copy number of Ta*2* and Ta*3*.
[c] N/A = not applicable.

Among the 34 clones sequenced that were obtained by PCR, we identified 11 clones that correspond to Ta*1* and 2 clones that correspond to Ta*2* (Table 2), but did not find any sequences representing Ta*3*. This may be due to the fact that the conserved amino acid domain of Ta*3* that was used as the basis for the lefthand primer differs by one amino acid from the consensus sequence (KTTFL *vs.* KTAFL; Figure 5). Another possibility for not finding sequences corresponding to Ta*3* is that we found that some reverse transcriptase sequences were preferentially amplified compared to others; 11 Ta*1* clones were obtained but only one each for Ta*6* and Ta*9* (Table 2).

DNA sequences of clones within a given family showed a variation ranging from no differences (Ta*5* and Ta*8*) to 4.2% (Ta*4* family) (Table 2). Some variation in the sequences may be due to replication errors which occurred during amplification. To quantify the level of PCR-generated errors, the reverse transcriptase nucleotide sequences of Ta*1* and Ta*2* obtained from PCR clones were compared with the sequences obtained from lambda clones (see above and VOYTAS *et al.* 1990). A mean variation of 3.7% (standard deviation = 1.1%) was observed indicating that the sequence divergence among clones of a given family is most likely due to amplification errors, whereas divergence between families most likely reflects evolutionary change. Another piece of evidence suggesting that amplification errors were responsible for most of the variation within a given element family is that each family, with the exception of Ta*1*, has only one or two members in both the Landsberg *erecta* and Columbia ecotypes as determined by Southern blot hybridization (Table 2; Figure 4; data not shown).

```
Plurality   KTAFLHGDLEEELYMEQPEGF VDN..PDK VCLLKKSLYGLKQSPRQWNK.F N.FM-.EQGFIRSEHD.CVYVK-QVS.G. FI-YLLLYVDDM

Ta10        .......E...TV..YP.A.. ..:RN.GY ..K.N.A......A..A..AR. VAYV.-KM..VQ.KC.AS-LFV.YKHGQD KA..........
Ta9         .........T.TVF.T..A.. ..PAKR.Y ..H.H...........A.FDK. STYL.L.F...C.MS.PS-LFI.YIKGKD I..S........
Ta8         .........S.TV..S..A.. ..PIK..Y ..HFD...........A.FSK. STYF.L.F..FC.MS.PS-LFV.YIKGKD I.ML........
Ta7         .......Y.D.TI..D...Y  IHEKY... .....R.............NK. .E...QRI.YE..KY.S...Y..ILLS.D Y...........
Ta6         .......E.DQV......... EV.PEK.Q .............A......K. HA...LSLQ.A.....S......E.NP.E .V..........
Ta5         .......Q.DQP.F....K.. EV.PEL.Q ......A......A......R. .A...MD.K.T..VS.S......E..N.E CV..........
Ta4         .......E............V .SVGNE.. .............A......R. .K...TNEK.Q..D..Q......TMNN.E LV..........
Ta3         ..T.................. .SDSSEN. ..R.................R. DR...SS.Q.......A......H..EHD ............
Ta2         ..................... ITDKTDK. ....................R. GR...N..K....A..A.........EQ.- .V..........
Ta1-3       .......E............C ISEDGEN. ....................R. .R...ID.N.......A.........EQE HL..........
Tnt1        ............I........ EVAGKKHM ..K.N.........A....YMK. DS...KS.TYLKTYS.P...F..RF.ENN ...I........
1731        C..Y.NSE.KDTV..K..Q.. T.AAN..Q .L..R.AI......G.E..SKL DGVL.KDL..KACN.EP.-LYQ..SGQ.N LM.LI.V....L
Ty1         SS.Y.YA.IK....IRP.PHL GM.-DK-- LIR......E....GAN.YETI KSYLIQQC.MEEVRGWS..FKNS..TI-- --.--C.F....
copia       .....N.T.K..I..RL.Q.I SC.SDN-- ..K.N.AI.....AA.C.FEV. EQAL.K.CE.VN.SV.R.I.ILDKGNINE N...V......V

                      1                           2                                               3
```

FIGURE 6.—Amino acid sequence alignment of the reverse transcriptase regions of the *A. thaliana* Ta*1*-Ta*10* superfamily of retrotransposons. Each family is represented by one sequence. Dots represent amino acids identical to the consensus sequence (top line). Related elements (*copia* and 1731 of *D. melanogaster*, Tnt*1* of *N. tabacum* and Ty*1* of yeast) are included in the alignment.

TABLE 3

Nucleotide and amino acid sequence comparisons of the Ta*1-3*, Ta*2* and Ta*3* reverse transcriptase and RNase H gene

	Nucleotides compared	Changes observed	Percent nucleotide identity	Percent amino acid identity	Percent silent changes	Percent replacement changes
Ta*1-3* *vs.* Ta*2*[a]	1265	323	74.5	78.1	60.0	40.0
Ta*1-3* *vs.* Ta*3*	1269	328	74.2	78.2	59.8	40.2
Ta*2* *vs.* Ta*3*	1265	328	74.1	77.2	58.4	41.6

[a] One gap of three nucleotides and one gap of one nucleotide was inserted for alignment.

Phylogenetic analysis: The alignment of the derived amino acid sequences for the copia-like retrotransposons shown in Figure 5 formed the basis of the character state matrix used in the phylogenetic analysis. The phylogenetic analysis used 57 informative characters and resulted in a single most parsimonious tree of length 306. The consistency index excluding autapomorphies was 0.79. An interesting feature of the tree is that the Ta elements do not form a monophyletic assemblage, but rather fall into two general groups. One group appears to have shared a more recent common ancestor with the Tnt*1* retrotransposon of tobacco; a relationship supported by six unambiguous character state changes. The other group appears to have shared a more recent common ancestor with the retrotransposons from *D. melanogaster.* This group is represented by Ta*8*, Ta*9* and Ta*10*. These elements are joined with *1731* of *D. melanogaster* by four unambiguous character state changes, and together with *1731* share two character state changes with *copia*.

DISCUSSION

Ta*2* and Ta*3* are copia-like *A. thaliana* retrotransposons: The partial nucleotide sequence of Ta*2* and the complete sequence of Ta*3* indicate that these elements have an overall structure typical of retrotransposable elements (Figure 2). The 5′ LTR and part of the internal domain of Ta*2* have been deleted. Ta*3* is bounded by two LTRs that are terminated by the consensus sequence 5′TG . . . CA 3′ (Figure 3) found in other retrotransposons (TEMIN 1985). The internal sequence of Ta*3* includes most of the *cis*- and *trans*-acting elements found in retrotransposons and retroviruses. A single large Ta*3* ORF encodes 1355 amino acids of putative Gag and Pol proteins (Figure 2). *Cis* sequences necessary to prime first (primer binding site, PBS) and second (polypurine tract, PPT) strand DNA synthesis were found at the 5′ and 3′ ends of the internal domain of Ta*3*, respectively (Figures 2 and 3).

The 3-kbp *Eco*RI fragment of Ta*2*, and the entire sequence of Ta*3*, encode open reading frames which share significant amino acid sequence identity to each other and to the polyprotein of another *A. thaliana* retrotransposon family, Ta*1*. The amino acid sequence similarity to Ta*1* occurs across a single reading frame for Ta*3*, and across three short overlapping reading frames for Ta2, one of which carries a stop codon (Figure 2). The amino acid sequences of the Ta*1*, Ta2 and Ta*3* families are colinear. Only a single gap must be inserted into the Ta2 ORF and two gaps into Ta*3* ORF in order to optimize alignment.

For most retrotransposons and retroviruses, the reverse transcriptase gene precedes the integrase gene (DOOLITTLE *et al.* 1989). A distinct lineage of retrotransposon is comprised of "copia-like" elements for which the order of these coding regions is reversed (XIONG and EICKBUSH 1988; DOOLITTLE *et al.* 1989). Like Ta*1*, the integrase region of both Ta2 and Ta*3* lies upstream of the reverse transcriptase region, clearly placing them among the copia-like group of retrotransposons.

Activity of Ta2 and Ta*3*: We have previously demonstrated that the Ta*1* elements are likely incapable of transposition due to deletions and nucleotide changes which have occurred among the various element copies (VOYTAS *et al.* 1990). The deletion of the 5′ LTR suffered by Ta2 and the organization of the Ta2 ORF indicate that this element is also nonfunctional. The Ta*3* element appears to be structurally intact and the encoded protein product does not carry any frameshifts or premature termination codons. However, because the mechanism of transposition results in identical LTRs upon integration (VARMUS and BROWN 1989), the nucleotide differences between the 3′ and 5′ LTRs of Ta*3* indicate that this element has accumulated mutations subsequent to its insertion in the genome. The fact that Ta*3* is also present in the genome as a single copy (Table 2) makes it likely that Ta*3*, like Ta*1* and Ta2, is nonfunctional.

The majority of the nucleotide differences between Ta*1*, Ta2 and Ta*3* are not due to random mutational events incurred since the loss of function. Several lines of evidence indicate that these elements have evolved independently under functional constraints for significant periods of time before insertion into their present sites within the *A. thaliana* genome. First, the protein coding regions and the *cis* sequences required for retrotransposition of the three element families have evolved much slower than the non-protein-coding regions of the internal element domain and the LTRs. The LTRs only share 50–60% nucleotide identity whereas the internal domains are approximately 74% similar (over the entire length). The noncoding regions of the internal domain (excluding the priming sites) also show only between 50% and 60% nucleotide identity (data not shown). Blocks of similarity between the Ta*1* and Ta2 LTRs suggest that these may be *cis* sequences which correspond to transcription initiation and termination regions (VARMUS and BROWN 1989; BOEKE 1989). The sequences which precede and follow the open reading frame that most likely serve as priming sites for reverse transcription are also highly conserved.

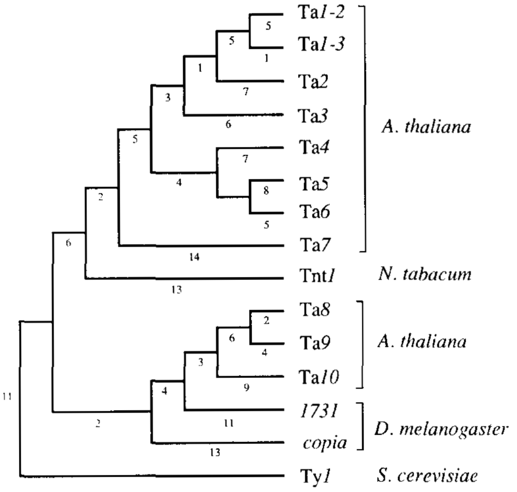

FIGURE 7.—Most parsimonious phylogenetic tree of the *A. thaliana copia*-like retrotransposons. Ty*1* was used as the outgroup for rooting. Tree length = 306; consistency index (excluding autapomorphies) = 0.79. Numerals adjacent to branches represent the number of unambiguous character state changes that can be assigned to that branch.

A second line of evidence that Ta*1*, Ta2 and Ta*3* have evolved independently in *A. thaliana* is that the regions which encode putative enzymatic functions have evolved much slower than the amino acids which separate these regions. This is particularly evident for the "tether" region between the integrase and reverse transcriptase of Ta*1* and Ta2 (Table 1). The amino acid similarity between these two elements falls off dramatically in this region. This holds true for comparisons of these elements with the related retrotransposon Tnt*1* (Table 1).

Third, nucleotide substitutions which occur between element copies show a bias for silent amino acid changes. Random mutations in a coding sequence would be expected to result in ~¾ amino acid replacements and ~¼ silent substitutions (LEWONTIN 1989). Comparisons between the element families indicate that this trend is almost reversed. The three elements show approximately 60% silent changes and 40% amino acid replacements (Table 3). While the proteins encoded by these elements may no longer be functional, they appear to have been significantly constrained during at least part of their evolutionary history.

Origins of the Ta retrotransposons: Two contrasting, but not mutually exclusive mechanisms are thought to be responsible for the distribution of retrotransposons: vertical inheritance and horizontal transfer. If retrotransposons are ancestral and predate the origin of lineages that contain them, then the pattern of their inheritance would be expected to be vertical. This is the case for many genes, such as those coding for rRNA and histones. The observation that the retrotransposon families Ta*1*, Ta*2* and Ta*3* are present in widely dispersed geographical races suggests that the transposition events associated with these elements may predate the speciation of *A. thaliana*. These data are consistent with vertical inheritance. Vertical inheritance combined with the processes of replicative transposition and random sequence loss should lead to a pattern in which element families are more closely related to each other within an organism than between widely divergent organisms. The most parsimonious phylogeny of the *A. thaliana* copia-like elements (Figure 7) suggests that Ta*1*-Ta*7* share a common ancestor and, together with Tnt*1*, form a monophyletic clade that is composed exclusively of plant retrotransposons, a pattern consistent with vertical inheritance.

Structural similarities among retrotransposons from widely diverged species have led to speculation that these elements have also been transferred horizontally (*e.g.*, DOOLITTLE *et al.* 1989; SMYTH *et al.* 1989). Among the closely related elements found in different species are the gypsy group of elements (*gypsy, 17.6, 297* and *412* of *D. melanogaster*; the Ty*3* element of yeast; and the *del* elements of lily) and the copia group of elements (*copia* and *1731* of *D. melanogaster*; Ty*1* of yeast; Tnt*1* of *N. tabacum*; and the Ta*1*-Ta*10* superfamily of *A. thaliana*). Two additional lines of evidence have been cited for horizontal transfer of the *D. melanogaster copia* elements in particular: First, the codon usage of these elements is distinctly different from *D. melanogaster* cellular genes (MOUNT and RUBIN 1985) and, second, the apparent absence of *copia* from some species of Drosophila (RUBIN 1983; STACEY *et al.* 1986).

Phylogenetic analysis of retrotransposon sequence data has previously been interpreted as evidence for horizontal transfer of the *copia* group of elements between Drosophila and yeast (XIONG and EICKBUSH 1988; DOOLITTLE *et al.* 1989). However, one limitation of these previous studies is that it was only possible to examine single transposable elements from widely diverged organisms (Drosophila and yeast). A more comprehensive study of horizontal transfer would involve sequence analysis of several elements from a single organism in comparison to those of other organisms.

Our present study provides evidence for horizontal transfer based on a comparison of a superfamily of *copia*-like elements composed of ten different families. The most parsimonious tree for the entire *copia*-like group of retrotransposons in *A. thaliana* (Figure 7) shows a clade that is composed of both *A. thaliana* (Ta*8*–Ta*10*) and *D. melanogaster* (*copia* and *1731*) retrotransposons, a pattern consistent with horizontal transfer. This implies that Ta*8*–Ta*10* share a more recent common ancestor with *D. melanogaster* retrotransposons than they do with other *A. thaliana* retrotransposons. Additional evidence that supports the notion of horizontal transfer is the observation that the different retrotransposon families are represented in most cases by a single distinct element. This suggests that much of the evolution of these elements occurred independently and outside the genome of *A. thaliana*. Since resident in the *A. thaliana* genome, it appears that most of these elements have failed to proliferate.

How common are retrotransposable elements within the *A. thaliana* genome? Because we initially identified the Ta*1* retrotransposable element family by analyzing restriction fragment length polymorphisms across a small fraction (~0.14%) of the *A. thaliana* genome, we wanted to determine if retrotransposable elements were commonplace in this species. Using PCR and a set of degenerate primers (Figure 5), we amplified and cloned highly conserved reverse transcriptase sequences of *A. thaliana* copia-like elements. Among 34 clones sequenced, we identified nine distinct families of elements. A tenth family (Ta*3*) was identified on the basis of cross-hybridization to reverse transcriptase probes derived from Ta*1*. Most of these ten element families that we studied exist as single copy insertions (Table 2). Based on these data and the size of the *A. thaliana* genome (100 Mb; HAUGE *et al.* 1991) we estimate that the Ta*1*–Ta*10* superfamily consists of ~0.1% of the *A. thaliana* genome.

It is possible that additional copia-like retrotransposons could be detected by the polymerase chain reaction using different sets of primers. This seems likely due to the fact that we did not clone any PCR amplified sequences corresponding to Ta*3*. The most likely reason is that the Ta*3* reverse transcriptase contains an amino acid substitution in the highly conserved sequence KTAFLHG used as the basis for synthesizing one of the PCR primers. If this is true, the superfamily of *copia*-like elements in *A. thaliana* may be larger than suggested by our experiments. On the other hand, it is also possible that Ta*3* was not identified simply because only 34 clones were sequenced.

M.P.C. was supported by National Institutes of Health Genetics Training Grant GM07620. This work was supported by a grant from Hoechst AG to Massachusetts General Hospital.

LITERATURE CITED

AUSUBEL, F. M., R. BRENT, R. E. KINGSTON, D. D. MOORE, J. G. SEIDMAN, J. A. SMITH and K. STRUHL, 1990 *Current Protocols in Molecular Biology*. Greene Publishing Associates/Wiley Interscience, New York.

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 cerev-*

isae, pp. 335–374 in *Mobile DNA*, edited by D. E. BERG, and M. M. HOWE. American Society for Microbiology, Washington, D.C.

CLARE, J., and P. FARABAUGH, 1985 Nucleotide sequence of yeast Ty element: evidence for an unusual mechanism of gene expression. Proc. Natl. Acad. Sci. USA **82:** 2829–2833.

DAYHOFF, M. O., W. C. BARKER and L. T. HUNT, 1983 Establishing homologies in protein sequences. Methods Enzymol. **91:** 534–545.

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.

FOURCADE-PERONNET, F., L. D'AURIOL, J. BECKER, F. GALIBERT and M. BEST-BELPOMME, 1988 Primary structure and functional organization of *Drosophila* 1731 retrotransposon. Nucleic Acids Res. **16:** 6113–6125.

GAUSS, D. H., and M. SPRINZL, 1983 Compilation of transfer RNA sequences and modified nucleosides in transfer RNA, pp. 129–226 in *A Laboratory Manual of Genetic Analysis, Identification and Sequence Determination*, edited by P. F. AGRIS and R. A. KOPPER. Alan R. Liss, New York.

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

HANSEN, L. J., D. L. CHALKER and S. B. SANDMEYER, 1988 Ty3, a yeast retrotransposon associated with tRNA genes, has homology to animal retroviruses. Mol. Cell. Biol. **8:** 5245–5256.

HAUGE, B. M., J. GIRAUDAT, S. HANLEY, I. HWANG, T. KOCHI and H. M. GOODMAN, 1991 Physical mapping of the *Arabidopsis* genome and its applications, in *Plant Molecular Biology*, edited by R. G. HERMAN. Plenum, New York (in press).

HEIN, J., 1989a A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Mol. Biol. Evol. **6:** 649–668.

HEIN, J., 1989b A tree reconstruction method that is economical in the number of pairwise comparisons used. Mol. Biol. Evol. **6:** 669–684.

INOUYE, S., S. YUKI and K. SAIGO, 1986 Complete nucleotide sequence and genome organization of a *Drosophila* transposable genetic element, 297. Eur. J. Biochem. **154:** 417–425.

JIN, Y.-K., and J. L. BENNETZEN, 1989 Structure and coding properties of Bs*1*, a maize retrovirus-like transposon. Proc. Natl. Acad. Sci. USA **86:** 6235–6239.

JOHNS, M. A., M. S. BABCOCK, S. M. FUERSTENBERG, S. I. FUERSTENBERG, M. FREELING and R. B. SIMPSON, 1989 An unusually compact retrotransposon in maize. Plant Mol. Biol. **12:** 633–642.

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

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

LUCAS, H., G. MOORE and R. B. FLAVELL, 1989 Characterization of retrotransposon-like element in wheat. UCLA Symp. Mol. Cell. Biol. **18:** 282.

MADDISON, W. P., and D. R. MADDISON, 1991 MacClade: Interactive Analysis of Phylogeny and Character Evolution, version 2.99B6 (test version of MacClade 3.0). Sinauer Associates, Sunderland, Mass.

MARLOR, R., S. PARKHURST and V. CORCES, 1986 The *Drosophila melanogaster* gypsy transposable element encodes putative gene products homologous to retroviral proteins. Mol. Cell. Biol. **6:** 1129–1134.

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.

NORRANDER, J., T. KEMPE and J. MESSING, 1983 Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. Gene **26:** 101–106.

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

RUBIN, G. M., 1983 Dispersed repetitive DNAs in *Drosophila*, pp. 329–361 in *Mobile Genetic Elements*, edited by J. A. SHAPIRO. Academic Press, New York.

SAIGO, K., W. KUGIMIYA, Y. MATSUO, S. INOUYE, K. YOSHIOKA and S. YUKI, 1984 Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. Nature **312:** 659–661.

SMYTH, D. R., P. KALITSIS, J. L. JOSEPH and J. W. SENTRY, 1989 Plant retrotransposon from *Lilium henryi* is related to Ty3 of yeast and the gypsy group of *Drosophila*. Proc. Natl. Acad. Sci. USA **86:** 5015–5019.

STACEY, S. N., R. A. LANSMAN, H. W. BROCK and T. A. GRIGLIATTI, 1986 Distribution and conservation of mobile elements in the genus *Drosophila*. Mol. Biol. Evol. **3:** 522–534.

SWOFFORD, D. L., 1990 Phylogenetic analysis using parsimony, PAUP version 3.0g. Ill. Nat. Hist. Surv.

TEMIN, H. M., 1985 Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. Mol. Biol. Evol. **2:** 455–468.

VARMUS, H., and P. BROWN, 1989 Retroviruses, pp. 53–108 in *Mobile DNA*, edited by D. E. BERG, and M. M. HOWE. American Society for Microbiology, Washington, D.C.

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

VOYTAS, D. F., A. KONIECZNY, M. P. CUMMINGS and F. M. AUSUBEL, 1990 The structure, distribution and evolution of the Ta*1* retrotransposable element family of *Arabidopsis thaliana*. Genetics **126:** 713–721.

WARMINGTON, J. R., R. B. WARING, C. S. NEWLON, K. J. INDGE and S. G. OLIVER, 1985 Nucleotide sequence characterization of Ty *1–17*, a class II transposons from yeast. Nucleic Acids Res. **13:** 6679–6693.

XIONG, Y., and T. H. EICKBUSH, 1988 Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol. Biol. Evol. **5:** 675–690.

YANISCH-PERRON, C., J. VIEIRA and J. MESSING, 1985 Improved M13 phage cloning vectors and host strains. Nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33:** 103–119.

YUKI, S., S. ISHIMARU and K. SAIGO, 1986 Identification of genes for reverse transcriptase-like enzymes in two *Drosophila* retrotransposons, 412 and gypsy: a rapid detection method of reverse transcriptase genes using YXDD box probes. Nucleic Acids Res. **14:** 3017–3030.

Communicating editor: B. BURR