Potential Retroviruses in Plants: Tat*1* Is Related to a Group of *Arabidopsis thaliana* Ty3/*gypsy* Retrotransposons That Encode Envelope-Like Proteins

David A. Wright and Daniel F. Voytas

Department of Zoology and Genetics, Iowa State University, Ames, Iowa 50011

Manuscript received December 29, 1997
Accepted for publication March 9, 1998

ABSTRACT

Tat*1* was originally identified as an insertion near the *Arabidopsis thaliana SAM1* gene. We provide evidence that Tat*1* is a retrotransposon and that previously described insertions are solo long terminal repeats (LTRs) left behind after the deletion of coding regions of full-length elements. Three Tat*1* insertions were characterized that have retrotransposon features, including a primer binding site complementary to an *A. thaliana* asparagine tRNA and an open reading frame (ORF) with ~44% amino acid sequence similarity to the *gag* protein of the *Zea mays* retrotransposon *Zeon-1*. Tat*1* elements have large, polymorphic 3′ noncoding regions that may contain transduced DNA sequences; a 477-base insertion in the 3′ noncoding region of the Tat*1-3* element contains part of a related retrotransposon and sequences similar to the nontranslated leader sequence of *AT-P5C1*, a gene for pyrroline-5-carboxylate reductase. Analysis of DNA sequences generated by the *A. thaliana* genome project identified 10 families of Ty3/*gypsy* retrotransposons, which share up to 51 and 62% amino-acid similarity to the ORFs of Tat*1* and the *A. thaliana Athila* element, respectively. Phylogenetic analyses resolved the plant Ty3/*gypsy* elements into two lineages, one of which includes homologs of Tat*1* and *Athila*. Four families of *A. thaliana* elements within the Tat/*Athila* lineage encode a conserved ORF after integrase at a position occupied by the envelope gene in retroviruses and in some insect Ty3/*gypsy* retrotransposons. Like retroviral envelope genes, this ORF encodes a transmembrane domain and, in some insertions, a putative secretory signal sequence. This suggests that Tat/*Athila* retrotransposons may produce enveloped virions and may be infectious.

THE eukaryotic retrotransposons are divided into two distinct classes of elements on the basis of their structure: the long terminal repeat (LTR) retrotransposons and the LINE-like or non-LTR elements (Doolittle *et al.* 1989; Xiong and Eickbush 1990). These element classes are related by the fact that each must undergo reverse transcription of an RNA intermediate to replicate, and each generally encodes its own reverse transcriptase. The LTR retrotransposons replicate by a mechanism that resembles that of the retroviruses (Boeke and Sandmeyer 1991). They typically use a specific tRNA to prime reverse transcription, and a linear cDNA is synthesized through a series of template transfers that require redundant LTR sequences at each end of the element mRNA. This all occurs within a virus-like particle formed from proteins encoded by the retrotransposon mRNA. After reverse transcription, an integration complex directs the resulting cDNA to a new site in the genome of the host cell.

Phylogenetic analyses based on reverse transcriptase amino acid sequences resolve the LTR retrotransposons into two families: the Ty3/*gypsy* retrotransposons (*Metaviridae*) and the Ty1/*copia* elements (*Pseudoviridae*) (Boeke *et al.* 1998a; Boeke *et al.* 1998b; Xiong and Eickbush 1990). Although distinct, Ty3/*gypsy* elements are more closely related to the retroviruses than to the Ty1/*copia* elements. They also share a similar genetic organization with the retroviruses, principally in the order of integrase and reverse transcriptase in their *pol* genes. For the Ty3/*gypsy* elements, reverse transcriptase precedes integrase, and this order is reversed for the Ty1/*copia* elements. In addition, some Ty3/*gypsy* elements have an extra open reading frame (ORF) similar to retroviral envelope (*env*) proteins, which is required for viral infectivity. The *Drosophila melanogaster gypsy* retrotransposons encode an *env*-like ORF and can be transmitted between cells (Kim *et al.* 1994; Song *et al.* 1994). Thus there are two distinct lineages of infectious LTR retroelements, the retroviruses, and those Ty3/*gypsy* retrotransposons that encode envelope-like proteins. The Ty3/*gypsy* elements have been divided into two genera, the metaviruses and the errantiviruses, the latter of which include all elements with *env*-like genes (Boeke *et al.* 1998a).

In plants, retrotransposons have been extremely successful (Bennetzen 1996; Voytas 1996). The enormous size of many plant genomes demonstrates a great tolerance for repetitive DNA, a substantial proportion of which appears to be composed of retrotransposons.

Corresponding author: Daniel F. Voytas, 2208 Molecular Biology Building, Iowa State University, Ames, IA 50011.
E-mail: voytas@iastate.edu

Because of their abundance, retrotransposons have undoubtedly influenced plant gene evolution. They can cause mutations in coding sequences (Grandbastien *et al.* 1989; Hirochika *et al.* 1996; Purugganan and Wessler 1994), and the promoter regions of some plant genes contain relics of retrotransposon insertions that contribute transcriptional regulatory sequences (White *et al.* 1994). Retrotransposons also generate gene duplications: Repetitive retrotransposon sequences provide substrates for unequal crossing over, and such an event is thought to have caused a zein gene duplication in maize (White *et al.* 1994). Occasionally, cellular mRNAs are reverse transcribed and the resultant cDNA recombines into the genome giving rise to new genes, or more frequently, cDNA pseudogenes (Maestre *et al.* 1995). The transduction of gene sequences during reverse transcription, which produced the oncogenic retroviruses, has also been documented to occur for a plant retrotransposon (Bureau *et al.* 1994; Jin and Bennetzen 1994); a maize *Bs1* insertion in *Adh1* carries part of an ATPase gene and is the only known example of a retrotransposon-mediated gene transduction event.

Arabidopsis thaliana is unusual among plants in that its genome is small and its retrotransposon families are of low copy number (Konieczny *et al.* 1991; Voytas and Ausubel 1988; Voytas *et al.* 1990; Wright *et al.* 1996). Of the 28 Ty1/*copia* and non-LTR retrotransposon families identified in our laboratory, none appear transpositionally active. Each family typically has three or fewer insertions in a given ecotype, and the structure of many insertions has been compromised by mutation or deletion. Furthermore, the chromosomal locations of elements and their copy numbers often do not differ between ecotypes, suggesting that they have not transposed recently. It seems that retrotransposon activity has been suppressed in *A. thaliana* or that most repetitive DNA has been lost (Voytas 1996).

The transposable elements Tat*1* and *Athila* are the only known *A. thaliana* elements of moderate copy number. These families are represented in some ecotypes by about 10 and 30 copies, respectively (Peleman *et al.* 1991; Pelissier *et al.* 1995). Tat*1* and *Athila* were chance discoveries; each was found in a section of sequenced DNA. *Athila* is flanked by LTRs typical of retrotransposons; however, none of the insertions characterized encode *gag* or *pol* homologs (Pelissier *et al.* 1995). Tat*1* was initially discovered as a 431-base insertion in one of 11 genomic clones of the *S*-adenosylmethionine synthetase gene (*SAM1*) isolated from a λ-phage library (Peleman *et al.* 1991). Its presence in only one of the characterized clones suggested that it transposed into this site within the population of plants from which DNA was extracted for library construction. Because of its small size and lack of coding sequences, Tat*1* was thought to be a degenerate DNA transposon.

We favored an alternative hypothesis to describe Tat*1*, namely that it is a retrotransposon solo LTR. Solo LTRs arise when the two LTRs of an integrated retrotransposon recombine, deleting the internal region and leaving behind a single LTR flanked by a target site duplication. Tat*1* shares features with retrotransposon solo LTRs: It has LTR dinucleotide end-sequences (5′TG-CA3′), which are part of a 12-base inverted terminal repeat, and it created a 5-base target site duplication upon integration, typical of plant retrotransposons. In this study, we demonstrate that Tat*1* is a retrotransposon and a member of a group of related retrovirus-like Ty3/*gypsy* elements present in the genomes of monocots and dicots. Some of these elements encode a conserved *env*-like gene, suggesting that infectious LTR retroelements exist in plants.

MATERIALS AND METHODS

Plant material and Southern hybridizations: The Arabidopsis Information Service supplied the following seed stocks (Kranz and Kirchheim 1987): Col-0, La-0, Kas-1, Co-4, Sei-0, Mv-0, Ll-0, Cvi-0, Fi-3, Ba-1, Hau-0, Aa-0, Ms-0, Ag-0, Ge-0, No-0 and Mh-0. Genomic DNA was extracted using genomic tips and protocols supplied by Qiagen (Valencia, CA). For Southern hybridizations, the resulting DNA was digested with *Eco*RI, electrophoresed on 0.8% agarose, and transferred to Gene Screen Plus membranes using the manufacturer's alkaline transfer protocol (New England Nuclear, Boston, MA). All hybridizations were performed as described (Church and Gilbert 1984).

Library screening, probe preparation and PCR: Tat*1* clones were obtained by screening a Landsberg *erecta* (La-0) λ-phage library (Voytas *et al.* 1990), using a probe derived by PCR amplification of La-0 DNA. The primers for probe amplification were based on published Tat*1* sequences: (DVO158, 5′-GGGATCCGCAATTAGAATCT-3′; DVO159, 5′-CGAATTC GGTCCACTTCGGA-3′) (Peleman *et al.* 1991). Subsequent probes were restriction fragments of cloned Tat*1* elements (Figure 1), and all probes were radiolabeled by random priming (Promega, Madison, WI). Long PCR was performed using the Expand Long Template PCR System (Boehringer Mannheim, Indianapolis) with LTR-specific primers (DVO354, 5′-CCACAAGATTCTAATTGCGGATTC-3′; DVO355, 5′-CCG AAATGGACCGAACCCGACATC-3′). The protocol used was for PCR amplification of DNA up to 15 kb. The following PCR primers were used to confirm the structure of Tat*1-3*: DVO405 (5′-TTTCCAGGCTCTTGACGAGATTTG-3′) for the 3′ noncoding region, DVO385 (5′-CGACTCGAGCTCCATAGC GATG-3′) for the second ORF of Tat*1-3* (note that the seventh base was changed from an A to a G to make an *Xho*I and a *Sal*I restriction site) and DVO371 (5′-CGGATTGGGCC GAAATGGACCGAA-3′) for the 3′ LTR.

DNA sequencing: Clones were sequenced either by the DNA sequencing facility at Iowa State University or with the *fmol* sequencing kit (Promega). DNA from the λ-phage clones was initially subcloned into the vector pBluescript II KS- and transformed into the *E. coli* host strain XL1 Blue (Stratagene, La Jolla, CA) (Ausubel *et al.* 1987). Subclones in the vector pMOB were used for transposon mutagenesis with the TN 1000 sequencing kit (Gold Biotechnologies, St. Louis, MO). Transposon-specific primers were used for DNA sequencing reactions.

Sequence analysis: Sequence analysis was performed using the GCG software package (Devereux *et al.* 1984), DNA Strider 1.2 (Marck 1988), the BLAST search tool (Altschul

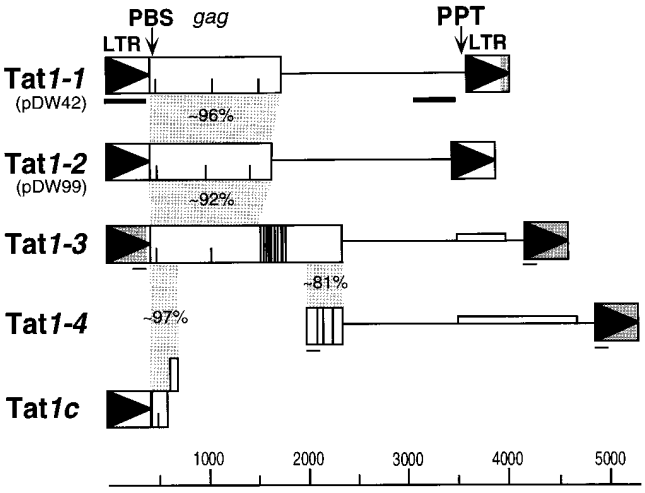

Figure 1.—Genomic organization of Tat*1* elements. Boxes with black triangles represent long terminal repeats (LTRs), and shaded portions of LTRs denote DNA sequences missing from particular clones. Open boxes indicate ORFs; the short lines are methionine codons and long lines are stop codons. Offset boxes represent a change in reading frame. Percentage nucleotide identified is indicated in the shaded regions between ORFs. The 3′ noncoding regions are depicted as thin lines; they share 98% identity between Tat*1-1* and Tat*1-2*, 96% identity between Tat*1-2* and Tat*1-3*, and 89% identity between Tat*1-3* and Tat*1-4*. The narrow boxes represent insertions unique to Tat*1-3* and Tat*1-4*. Black bars below Tat*1-1* indicate regions used for hybridization probes. Short, thin lines denote locations of primers used for PCR amplification of Tat*1-3* and Tat*1-4*. Abbreviations are as follows: PBS, primer binding site; PPT, polypurine tract. The scale is in base pairs.

et al. 1990), and the tRNAscan-SE 1.1 program (Lowe and Eddy 1997). Phylogenetic relationships were determined by the neighbor-joining distance algorithm using Phylip (Felsenstein 1993; Saitou and Nei 1987) and were based on reverse transcriptase amino-acid sequences that had been aligned with ClustalW1.7 (Thompson *et al.* 1994). Transmembrane helices were identified using the PHDhtm program (Rost *et al.* 1995). All DNA sequences have been submitted to the DDBS/EMBL/GenBank databases under the accession numbers AF056631, AF056632, AF056633, and AF056634.

RESULTS

Tat*1* is a retrotransposon: Tat*1* insertions share features with retrotransposon solo LTRs. We reasoned that if Tat*1* is a retrotransposon, then there should be full-length elements in the genome consisting of two Tat*1* sequences flanking an internal retrotransposon coding region. To test this hypothesis, additional Tat*1* elements were isolated by screening a Landsberg (La-0) genomic DNA library with a Tat*1* probe. Twenty-one λ-phage clones were isolated and Southern analysis revealed two clones (pDW42 and pDW99) each with two copies of Tat*1* (data not shown). The two Tat*1* elements in each clone were sequenced, along with the intervening DNA (Figure 1). All Tat*1* sequences shared >89% nucleotide identity with the previously characterized Tat*1a* - Tat*1c* elements (Peleman *et al.* 1991). In clone pDW99, the 5′ and 3′ Tat*1* sequences were 433 bases in length and only differed at two base positions. These Tat*1* sequences also had conserved features of LTRs, including the dinucleotide end-sequences (5′-TG-CA-3′) that were part of 12-base inverted terminal repeats. If the two Tat*1* elements in clone pDW99 were retrotransposon LTRs, then both, along with the intervening DNA, should be flanked by a target site duplication. A putative 5-base target site duplication (TATGT) was present immediately adjacent to the 5′ and 3′ Tat*1* elements, supporting the hypothesis that they and the intervening DNA inserted as a single unit. In clone pDW42, the 5′ Tat*1* was 432 bases in length and shared 98% nucleotide sequence identity to the 3′ Tat*1*. The last ~74 bases of the 3′ Tat*1* was truncated during library construction and lies adjacent to one phage arm. A target site duplication, therefore, could not be identified in this clone.

DNA sequences were analyzed for potential coding information between the 5′ and 3′ Tat*1* elements. Nearly identical ORFs of 424 and 405 amino acids were found encoded between the Tat*1* sequences in pDW42 and pDW99, respectively (Figure 1). The derived amino-acid sequences of these ORFs were used to search the DNA sequence database with the BLAST search tool, and significant similarity was found to the *Zea mays* retrotransposable element *Zeon-1* (p = 4.4e-08) (Hu *et al.* 1995). The ORFs have ~44% similarity across their entirety to the 628-amino-acid ORF encoded by *Zeon-1* (see below). The *Zeon-1* ORF includes a zinc finger motif characteristic of retrotransposon *gag* protein RNA binding domains (Hu *et al.* 1995). Although the Tat*1* ORFs do not include the zinc finger motif, the degree of similarity suggests that they are part of a related *gag* protein.

If the Tat*1* sequences in pDW42 and pDW99 defined retrotransposon insertions, a primer binding site (PBS) would be predicted to lie adjacent to the 5′ Tat*1* elements in both clones. The putative Tat*1* PBS shares similarity with PBSs of *Zeon-1* and another maize retrotransposon called *Cinful* (see below), but it is not complementary to an initiator methionine tRNA as is the case for most plant retrotransposons. Additionally, a possible polypurine tract (PPT), the primer for second-strand cDNA synthesis, was observed 1 base upstream of the 3′ Tat*1* sequence in both phage clones (5′-GAG GACTTGGGGGGCAAA-3′). We concluded from the available evidence that Tat*1* is a retrotransposon, and we have designated the 3960-base insertion in pDW42 as Tat*1-1* and the 3879-base insertion in pDW99 as Tat*1-2* (Figure 1). It is apparent that both Tat*1-1* and Tat*1-2* are nonfunctional. Their ORFs are truncated with respect to the coding information found in transposition-competent retrotransposons, and they lack obvious *pol* motifs.

In light of our findings, the previously reported Tat*1* sequences can be reinterpreted. Tat*1a* and Tat*1b*, which

are flanked by putative target site duplications, are solo LTRs. Tat*1c*, the only element without a target site duplication, is actually the 5′ LTR and part of the coding sequence for a larger Tat*1* element (Figure 1).

Copy number of Tat*1* among *A. thaliana* ecotypes: To estimate Tat*1* copy number, the 5′ LTR, *gag*, and the 3′ noncoding region were used as separate probes in Southern hybridizations (Figure 2). The Southern filters contained genomic DNA from 17 ecotypes representing wild populations of *A. thaliana* from around the world. This collection of ecotypes had previously been used to evaluate retrotransposon population dynamics (Konieczny *et al.* 1991; Voytas *et al.* 1990; Wright *et al.* 1996). Based on the hybridization with the *gag* probe, element copy number ranges from two to approximately ten copies per ecotype (Figure 2). The copy number of the LTRs is higher, likely due to the presence of two LTRs flanking full-length elements or solo LTRs scattered throughout the genome. The Tat*1* copy number contrasts with the copy numbers (typically less than three per ecotype) observed for 28 other *A. thaliana* retrotransposon families (Konieczny *et al.* 1991; Voytas *et al.* 1990; Wright *et al.* 1996). In addition, the Tat*1*-hybridizing restriction fragments are highly polymorphic among strains. This degree of polymorphism, coupled with the high copy number, suggested that Tat*1* has been active in transposition since the separation of the ecotypes.

The Tat*1* 3′ noncoding region contains DNA sequences from elsewhere in the genome: In an attempt to identify a complete and functional Tat*1* element, LTR-specific primers were used in PCR reactions optimized for amplification of large DNA fragments. Most full-length retrotransposable elements are between 5 and 6 kb in length. DNAs from all 17 ecotypes were used as templates, and each gave amplification products of ~3.2 kb, the size predicted for Tat*1-1* and Tat*1-2* (data not shown). In La-0, however, a 3.8-kb PCR product was also recovered. This PCR product was cloned, sequenced and called Tat*1-3*. This insertion is expected to be about 4.6 kb in total length if the LTR sequences are included (Figure 1).

Tat*1-3* differed from Tat*1-1* and Tat*1-2* in that it had two ORFs separated by stop codons and a 477-base insertion in the 3′ noncoding region. The first ORF (365 amino acids) was similar to but shorter than the ORFs of the other Tat*1* elements (Figure 1). The sequences constituting the second ORF (188 amino acids) were not present in the other Tat*1* insertions and were not related to other sequences in the DNA databases. Database searches with the 477-base insertion in the 3′ noncoding region, however, revealed three regions of similarity to other genomic sequences (Figure 3). A region of 113 bases matched a region of 26-bp repeats in the 5′ untranslated sequence of the *AT-P5C1* mRNA, which encodes pyrroline-5-carboxylate reductase ($p = 2.1e\text{-}19$) (Figure 3B) (Verbruggen *et al.* 1993). In addition, 50 bases appear to be a remnant of another retrotransposon related to Tat*1*. These 50 bases are 71% identical to the 3′ end of the Tat*1-3* LTR and the putative primer

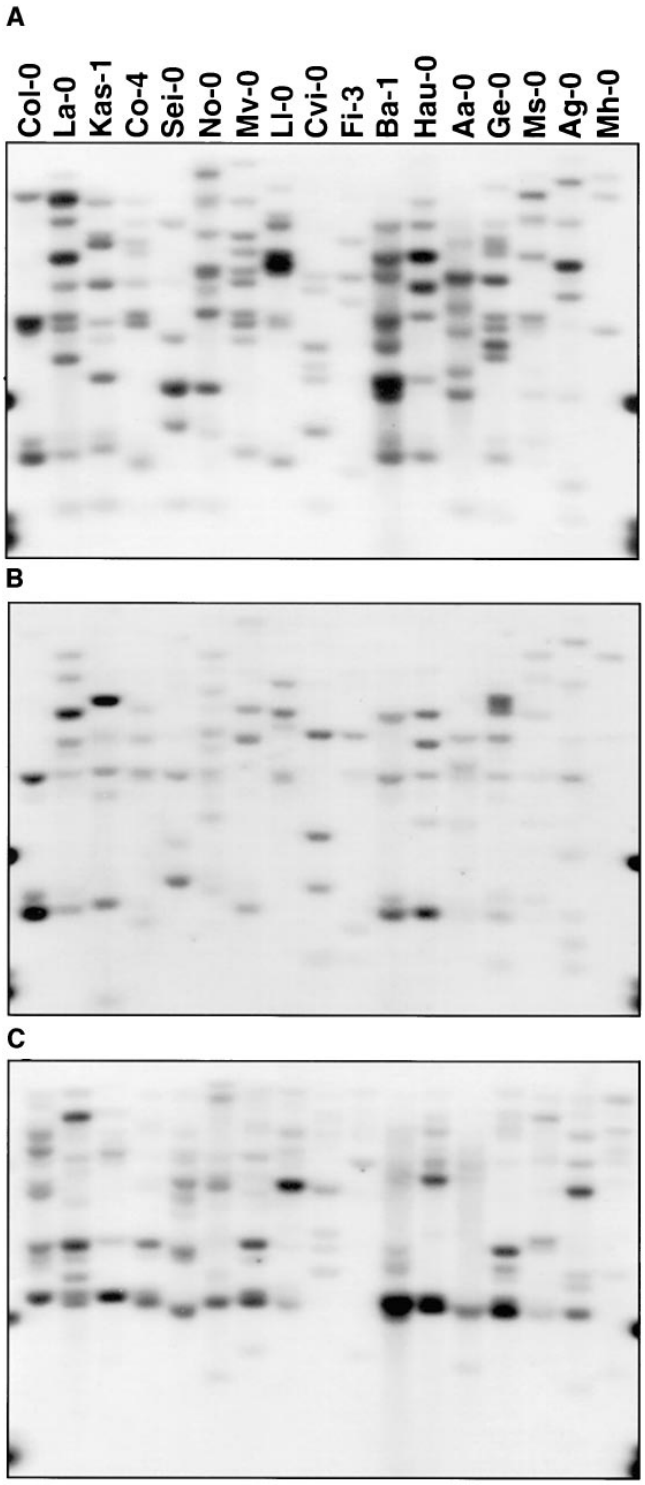

Figure 2.—Copy number of Tat*1* in 17 ecotypes of *A. thaliana*. Ecotype DNAs were digested with *Eco*RI and the Southern filter was hybridized with radiolabeled probes for the Tat*1-1* LTR (A), *gag* (B), and 3′ noncoding region (C). Lanes in A are labeled with the corresponding ecotypes; the same filter was stripped and used for all three hybridizations.

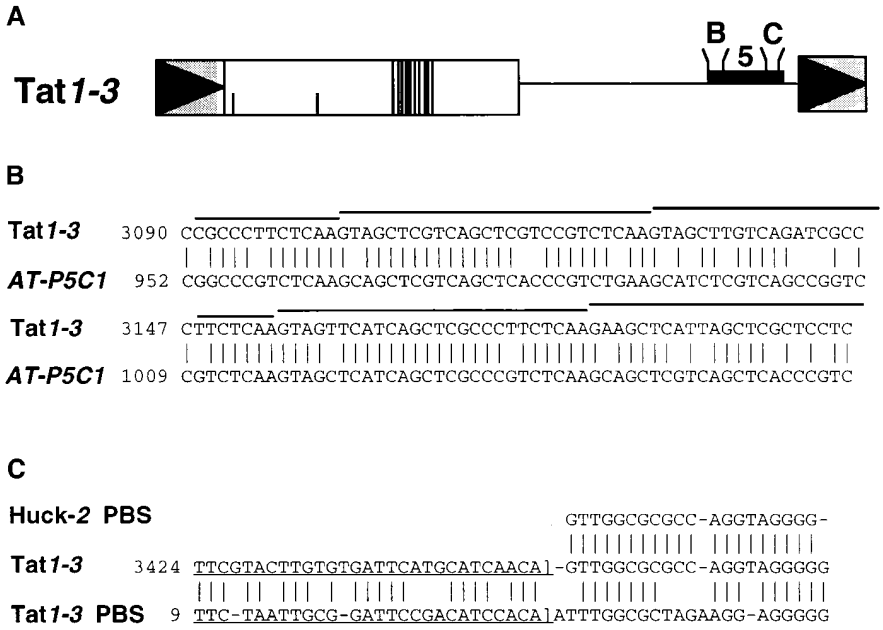

Figure 3.—Putative transduced sequences in the 3′ noncoding region of Tat*1-3*. (A) The genomic organization of Tat*1-3* as described in Figure 1. B and C designate sequences in the 477-base insertion that match sequences elsewhere in the *A. thaliana* genome and are shown in detail. The number 5 denotes sequences found on chromosome *5*. (B) The lines over the alignment denote a 26-base repeat found in the 5′ nontranslated region of *AT-P5C1*, a gene encoding pyrroline-5-carboxylate reductase (Verbruggen *et al.* 1993). Numbers indicate base positions within Ta*1-3* and *AT-P5C1* (GenBank accession number M76538). (C) A sequence that matches the primer binding site of the *Z. mays* retrotransposon Huck*1-2*, and to a lesser degree, the PBS and 5′ LTR end-sequences of Tat*1-3*. Underlined sequences represent the Tat*1-3* LTR or LTR-similar sequences in the 3′-end of Tat*1-3*.

binding site (Figure 3C). The putative primer binding site, however, is more closely related to those of other plant retrotransposons such as Huck-2 (SanMiguel *et al.* 1996) (Figure 3C). Finally, sequences in the remainder of the insertion showed significant similarity to a region on chromosome *5* (data not shown). To confirm that Tat*1-3* was not a PCR artifact, two additional primer pairs were used in separate amplifications. Both amplifications gave PCR products of the predicted sizes, which were cloned and confirmed to be Tat*1-3* by DNA sequencing.

PCR amplifications with the additional primer pairs also yielded a product 0.8 kb longer than that expected for Tat*1-3*. This product was cloned, sequenced, and found to be another Tat*1* element, designated Tat*1-4* (Figure 1). This element has sequences similar to a Tat*1* LTR, polypurine tract, and the second ORF of Tat*1-3*. In Tat*1-4*, 1182 bases of DNA are found in the 3′ noncoding region at the position corresponding to the 477-base insertion in Tat*1-3*. This region does not match any sequences in the DNA databases.

Other Tat*1*-like elements in *A. thaliana*: A BLAST search of DNA sequences generated by the *A. thaliana* genome project identified two more solo LTRs similar to Tat*1* (Figure 4). All share similarities throughout, but most strikingly, they are very well conserved at the 5′ and 3′ ends where it is expected integrase would bind (Brait erman and Boeke 1994). These conserved end-sequences suggest that the integrases encoded by full-length elements are also related, and that the LTRs have evolved under functional constraints; that is, they are not simply degenerate Tat*1* LTRs. The two new LTRs are designated as Tat*2-1* and Tat*3-1*. Tat*2-1* is 418 bases long, is flanked by a 5-base target site duplication (CTATT) and is ~63% identical to the Tat*1-2* 5′ LTR. Tat*3-1* is 463 bases long and is also flanked by a target site duplication (ATATT). Tat*3-1* is ~53% identical to the Tat*1-2* 5′ LTR.

Tat*1* and *Athila* are related to Ty3/*gypsy* retrotransposons: Further analysis of data from the *A. thaliana* genome project revealed two slightly degenerate retrotransposons with similarity to the Tat*1* ORF. These elements were identified within the sequence of the P1 phage clones MXA21 (Accession AB005247; bases 54,977–66,874) and MX110 (Accession AB005248; bases 24,125–35,848). Each has two LTRs, a putative PBS, and long ORFs between their LTRs. The genetic organization of these elements is depicted in Figures 5A and 6A. Amino-acid sequence analysis indicated the presence of an RNA binding domain that defines *gag* in both elements (data not shown). This region is followed by conserved reverse transcriptase, RNase H, and integrase amino-acid sequence domains characteristic of *pol* (data not shown). Classification of eukaryotic retrotransposons into the Ty1/*copia* elements (*Pseudoviridae*) and Ty3/*gypsy* elements (*Metaviridae*) is based on *pol* gene structure (Boeke *et al.* 1998a; Boeke *et al.* 1998b). The domain order of the *pol* genes (reverse transcriptase precedes integrase) and similarities among their encoded reverse transcriptases (see below) identifies these elements as the first full-length *A. thaliana* Ty3/*gypsy* elements.

Because the characterized Tat*1* insertions do not encode *pol* genes, this element family could not be classified. However, the amino-acid sequence of the Tat*1-2* ORF is 51% similar to the *gag* region of the MXA21 retrotransposon (Figure 5B). Since plant retrotransposons within the Ty1/*copia* or Ty3/*gypsy* families, even those with highly similar *pol* genes, share little amino-acid sequence similarity in their *gag* regions, Tat*1* is

Figure 4.—Additional *A. thaliana* Tat retrotransposons. Alignment of the 5′ LTR from Tat*1-2* with two related solo LTRs found in the available sequence from the *A. thaliana* genome project. Tat*2-1* is from the ESSA I contig fragment 4 (Accession Z97339, bases 115,028-115,445) and Tat*3-1* is from BAC F11P17 (Accession AC002294, bases 81,040-81,502). The solo LTRs are flanked by the target site duplications CTATT and ATATT, respectively (not shown).

likely a Ty3/*gypsy* element. This conclusion is further supported by the report that the Tat-like *Zeon-1* retrotransposon is very similar to a *Z. mays* Ty3/*gypsy* element called *cinful* (Bennetzen 1996); however, only the 5′ LTR and putative PBS sequences are available in the sequence database for analysis (Accession U68402). Because of the extent of similarity to Tat*1*, we have named the MXA21 insertion Tat*4-1*.

The *gag* region of the MX110 element is 62% similar (p = 1.1e-193) to the first ORF of *Athila*, which has previously been unclassified (Pelissier *et al.* 1995) (data not shown). This implies that *Athila* is also a Ty3/*gypsy* element, and we have designated the MX110 insertion as Athila*1-1* (Figure 6A). Our classification of *Athila* as a Ty3/*gypsy* element is further supported by the observation that the *Athila gag* amino-acid sequences share significant similarity to the *gag* protein encoded by the cyclops-*2* Ty3/*gypsy* retrotransposon of pea (Accession AJ000640; p = 1.1e-46; data not shown). Further analysis of the available *A. thaliana* genome sequences identified three additional *Athila* homologs. They include an additional Athila*1* element, designated Athila*1-2*, and two more distantly related *Athila*-like elements, designated Athila*2-1* and Athila*3-1* (Figure 6A).

In addition to similarities among their *gag* amino-acid sequences, the Tat elements have short LTRs (<550 bp) and long 3′ noncoding regions (>2 kb) (Table 1, Figure 5A). In contrast, the *Athila*-like elements have long LTRs (>1.2 kb) and are very large retrotransposons (>11 kb) (Table 1, Figure 6A). One additional feature to note about both the *Athila*-like and Tat-like elements is the high degree of sequence degeneracy of their internal coding regions. This contrasts with the near sequence identity of their 5′ and 3′ LTRs, which is typically greater than 95% (Table 1). Because a single template is used in the synthesis of both LTRs, LTR sequences are usually identical at the time of integration. The degree of sequence similarity between the LTRs suggests that most elements integrated relatively recently. The polymorphisms observed in the internal domains of these insertions, therefore, may have been present in their progenitors, and these elements may have been replicated *in trans.*

A novel, conserved coding region in *Athila* elements: A surprising feature of Athila*1-1* is the presence of an additional ORF after integrase (Figure 6A). Like *gag*, this ORF shares significant similarity across its entirety (p = 3.8e-08) to the second ORF of *Athila.* This ORF is also encoded by the Athila*2-1* and Athila*3-1* elements, although it is somewhat more degenerate. The presence of this coding sequence among these divergent retrotransposons suggests that it plays a functional role in the element replication cycle. However, the ORF shows no similarity to retrotransposon *gag* or *pol* genes. The retroviruses and some Ty3/*gypsy* retrotransposons encode an *env* gene after integrase. Although not well conserved in primary sequence, both viral and retrotransposon envelope proteins share some structural similarities. They are typically translated from spliced mRNAs and the primary translation product encodes a signal peptide and a transmembrane domain near the C terminus. All four families of *Athila* elements encode a domain near the center of the ORF that is strongly predicted to be a transmembrane region (70–90% confidence, depending on the element analyzed) (Rost *et al.* 1995) (Figure 6B). Two retrotransposons, *Athila* and Athila*2-1*, also have a hydrophobic transmembrane domain near the 5′ end of their *env*-like ORFs, which may serve as a secretory signal sequence (von Heijne 1986).

Two lineages of plant Ty3/*gypsy* retrotransposons:

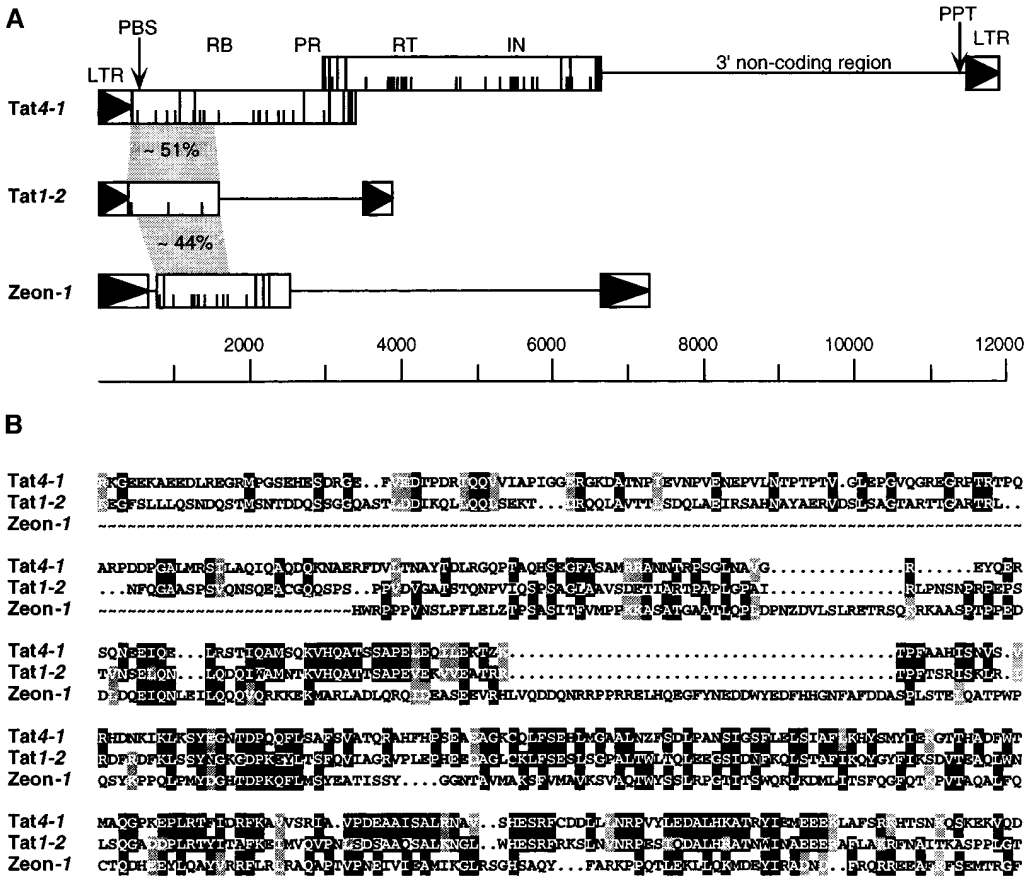

Figure 5.—Genomic organization of plant Ty3/*gypsy* retrotransposons related to Tat*1*. (A) Element features are as described in Figure 1. Tat*4-1* is from *A. thaliana* and *Zeon-1* is from *Z. mays*. The numbers in the gray areas between elements reflect percentage amino-acid similarity to Tat*1-2*. The scale is in base pairs. Abbreviations not in Figure 1 are as follows: RB, RNA binding domain; PR, protease; RT, reverse transcriptase; IN, integrase. (B) Amino-acid sequence alignment of the *gag* genes of Tat*4-1*, Tat*1-2* and *Zeon-1*. Black boxes identify identical amino-acid residues; gray boxes are similar residues (I=L=V, K=R, D=E).

Relationships among Ty3/*gypsy* retrotransposons from *A. thaliana* and other organisms were assessed by constructing a neighbor-joining tree of their reverse transcriptase amino-acid sequences (Figure 7). Included in the analysis were reverse transcriptases from two additional families of *A. thaliana* Ty3/*gypsy* elements that we identified from the unannotated genome sequence data (designated Tma elements; Tma*1-1* and Tma*3-1*); two other Tma element families were identified in the genome sequence that did not encode complete reverse transcriptases (Tma*2-1* and Tma*4-1*; Table 1). Also included in the phylogenetic analyses were reverse transcriptases from a faba bean retrotransposon and the cyclops-*2* element from pea. The plant Ty3/*gypsy* group retrotransposons resolved into two lineages: One was made up of del*1* from lily, the IFG7 retrotransposon from pine, *reina* from *Z. mays*, and Tma*1-1* and Tma*3-1*. This group of elements formed a single branch closely related to numerous fungal retrotransposons (branch 1). The second branch (branch 2) was well separated from all other known Ty3/*gypsy* group elements, and was further resolved into two lineages: Athila*1-1*, cyclops-*2* and the faba bean reverse transcriptase formed one lineage (the *Athila* branch), and Tat*4-1* and Grande*1-4* from *Zea diploperennis* formed a separate, distinct branch (the Tat branch).

Primer binding sites: Most plant Ty1/*copia* retrotransposons as well as the branch 1 Ty3/*gypsy* elements have PBSs complementary to the 3′ end of an initiator methionine tRNA. This is not the case for any of the branch 2 Ty3/*gypsy* elements. We compared the putative PBSs of Tat-branch and *Athila*-branch elements to known plant tRNA genes as well as to the 11 tRNA genes that had been identified to date in sequences generated by the *A. thaliana* genome project. In addition, we searched the unannotated *A. thaliana* genome sequences and identified 30 more *A. thaliana* tRNA genes using the program tRNAscan-SE (Lowe and Eddy 1997) (data not shown). The PBS of Tat*1* is complementary to 10 bases at the 3′ end of the asparagine tRNA for the AAC codon; these 10 bases are followed by a 2-base mismatch and 6 additional bases of perfect complementarity (Figure 8A). The Tat*4-1* PBS is complementary to 20 bases at the 3′ end of the arginine tRNA for the AGG codon with one mismatch 10 bases from the 3′ end; Huck-*2*, Grande-*zm1*, Grande*1-4*, and the retrotranspo-

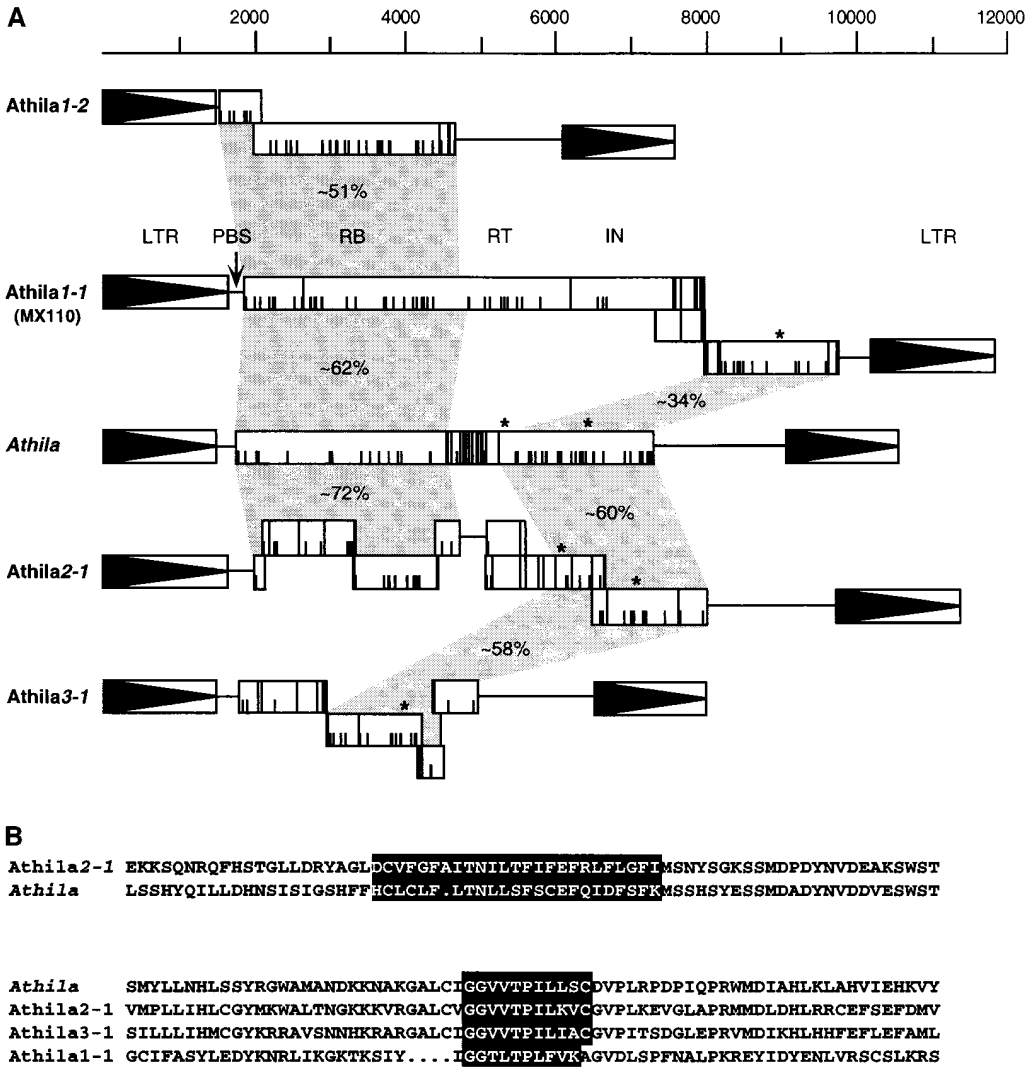

Figure 6.—Genomic organization of plant Ty3/*gypsy* retrotransposons related to *Athila*. (A) Element features are as described in Figure 1. The numbers in the gray areas between elements reflect percentage amino-acid similarity to *Athila*. Asterisks denote potential transmembrane domains. Athila*1-2* is from BAC IG009D12, Athila*2-1* is from BAC IG007N22 (bases 59,949-71,254) and Athila*3-1* is from BAC TM025C13. The scale is in base pairs. (B) Amino-acid sequence alignments of the regions encoding putative transmembrane domains. The topmost alignment is for the 5′-most transmembrane domain in *Athila* and Athila*2-1*, which may serve as a secretory signal sequence. The lower alignment is for the transmembrane domain found in the second ORF of the four different retrotransposon families. Boxed amino-acid residues were predicted to be part of the transmembrane domain by the PHDhtm program (70–90% confidence) (Rost *et al.* 1995).

son-like insertion in the 3′ noncoding region of Tat*1-3* all have 20-base perfect complementarity to this tRNA (Figure 8B). The PBS of Athila*1-1* is perfectly complementary to 15 bases at the 3′ end of the aspartic acid tRNA for the GAC codon, and *Athila* and Athila*2-1* have 13 bases of complementarity to this tRNA (Figure 8C). At this time there is no known plant tRNA complementary to the PBS of *Zeon-1*, which has the same PBS as the maize retrotransposon *cinful*. As more tRNA sequences become available, a candidate primer may be identified for these elements.

DISCUSSION

Tat*1* is related to plant Ty3/*gypsy* retrotransposons: Tat*1* was originally identified as an insertional polymorphism downstream of the *A. thaliana SAM1* gene (Peleman *et al.* 1991). Of 11 genomic clones characterized from a λ-phage library, one contained a 431-base insertion that was designated Tat*1a*. The small size of this insertion suggested that it was a DNA transposon. Because Tat*1a* was present in only one of the *SAM1* clones characterized, it was thought to have transposed to this site in one of the plants from which DNA was extracted for library construction. Tat*1*, therefore, was considered a likely candidate for an active *A. thaliana* transposon.

We considered a different interpretation of the Tat*1* data, namely that the 431-base Tat*1a* insertion was a retrotransposon solo LTR. Solo LTRs are left behind as a consequence of recombination between LTRs of full-length elements. The characterization of additional Tat*1* insertions supported our hypothesis. For example, we identified an insertion, designated Tat*1-2*, which has two 433-base LTRs (each >91% identical to Tat*1a*) and has a flanking 5-base target site duplication. Three bases after the Tat*1-2* 5′ LTR is a putative PBS with 10 bases of complementary to the 3′ end of an *A. thaliana* asparagine tRNA. One base upstream of the 3′ LTR is a polypurine tract. Tat*1* elements encode a short ORF that is highly similar (~44%) to the *gag* protein of the maize element *Zeon-1*. This ORF is even more similar (~51%) to the *gag* protein of the *A. thaliana* Ty3/*gypsy* element

TABLE 1

Features of *A. thaliana* Ty3/*gypsy* retrotransposons[a]

Insertion	Genus	Total size	LTR size (5′,3′)	% LTR Identity	TSD	PBS	PPT	*gag*/RT/IN/*env*-like ORF
Athila[b]	E	10,505	1539, 1552	99.8	TTACG	Asp	+	+/−/−/+
Athila*1-1*	E	~12,000	>1324, >1331	~99	—	Asp	−	+/+/+/+
Athila*1-2*	E	7,559	1386, 1419	98.3	CGGGT	Asp	+	+/−/−/−
Athila*2-1*	E	11,297	1744, 1752	95.6	—	Asp	+	+/−/−/+
Athila*3-1*	E	~8,100	>1200, >1200	~95	—	—	+	−/−/−/+
Tat*1-1*	E	~4,034	432, ~432	~98	ND	Asn	+	+/−/−/−
Tat*1-2*	E	3,879	433, 433	99.5	TATGT	Asn	+	+/−/−/−
Tat*4-1*	E	11,898	453, 452	96.5	GTGAA	Arg	+	+/+/+/−
Tma*1-1*[c]	M	7,801	1164, 1158	96.2	ATATC	i-Met	+	+/+/+/−
Tma*2-1*	M	8,429	1161, 1488	90.1	AAAT	i-Met	+	+/+/+/−
Tma*3-1*	M	7,768	1155, 1054	93.8	CAAAG	i-Met	+	+/+/+/−
Tma*4-1*	M	~4,550	>1200, >1200	~97	—	—	+	−/−/+/−

[a] Features that could not be identified from the DNA sequences, likely due to sequence degeneracy or deletion, are indicated by (—); ND, not determined due to lack of data; TSD, target site duplication; PBS, primer binding site; PPT, polypurine tract; M, metavirus; E, errantivirus.

[b] Features are for *Athila* accession X81801, with the exception of the TSD, which is for the insertion λH3 (Pelissier *et al.* 1995).

[c] Tma = T, transposon; m, metavirus; a, *A. thaliana*; Tma*2-1* is from BAC IG007N22 and Tma*4-1* is from ESSA contig 7 (Accession number Z97342, bases 76,621–82,535).

Tat*4-1*, which we identified from the DNA sequence of the *A. thaliana* P1 phage clone MXA21. Since the *gag* proteins of retrotransposons are generally not well conserved, this suggests that Tat*1* is a Ty3/*gypsy* retrotransposon. By this reasoning, the previously characterized *Athila* element also appears to be a Ty3/*gypsy* retrotransposon. It shares ~62% amino-acid similarity between its first ORF and the *gag* protein of Athila*1-1*, an *A. thaliana* Ty3/*gypsy* element that we identified from the sequence of the P1 phage clone MX110. Although it is possible that the *gag* sequence similarity between these elements is the consequence of xenologous recombination, we do not believe this is the case, because Tat*1* and *Athila* share a number of other features with related plant Ty3/*gypsy* elements (see below). With the exception of degenerate Ty3/*gypsy* reverse transcriptase sequences in the *A. thaliana* mitochondrial genome (Knoop *et al.* 1996), the elements described in this report are the first *A. thaliana* Ty3/*gypsy* retrotransposons and are among a handful described to date in plants.

None of the three characterized Tat*1* insertions encode reverse transcriptase or integrase motifs typically associated with functional retrotransposons. This may be the result of internal deletions, as suggested by size polymorphisms among the elements and their encoded ORFs. The related maize element *Zeon-1* also does not encode proteins necessary for transposition (Hu *et al.* 1995). Both Tat*1* and *Zeon-1*, therefore, may be replicated by one or more master elements that provide functions *in trans*. This mechanism for transposition is further supported by the observation that all of the characterized *A. thaliana* Ty3/*gypsy* insertions have LTRs that share >95% nucleotide identity. This contrasts with often highly degenerate internal coding sequences. For the *A. thaliana* Ta*1* elements, a family of Ty1/*copia* retrotransposons, LTR sequences of given insertions also share ~95% nucleotide identity, yet their coding regions are largely intact and carry only a few premature stop codons or frameshifts (Voytas *et al.* 1990). The extent of internal coding sequence degeneracy among the Ty3/*gypsy* elements relative to the near identity of their LTRs implies that transcripts from defective elements were acted upon *in trans* to generate these insertions. We were unable to identify candidate Tat*1* master elements. Nonetheless, because Tat*1* integrated recently near *SAM1*, a master element or a related retrotransposon that can act on Tat*1* mRNA is likely present in the *A. thaliana* genome.

Southern hybridization analyses also suggest that Tat*1* is transpositionally active. Up to 10 copies of Tat*1* insertions are found in the genomes of the 17 diverse *A. thaliana* ecotypes analyzed. The extensive levels of observed restriction fragment length polymorphism among ecotypes is also consistent with transposition, although polymorphisms generated by recombination cannot be excluded. Like Tat*1*, the *Athila* elements are highly polymorphic and of moderate copy number (up to 30) (Pelissier *et al.* 1995). The copy number and polymorphic nature of the Tat*1* and *Athila* elements contrasts sharply with the 28 LINE-like and Ty1/*copia* elements previously characterized (Konieczny *et al.* 1991; Voytas *et al.* 1990; Wright *et al.* 1996). These elements typically number no more than three insertions per ecotype), and they generally exhibit very uni-

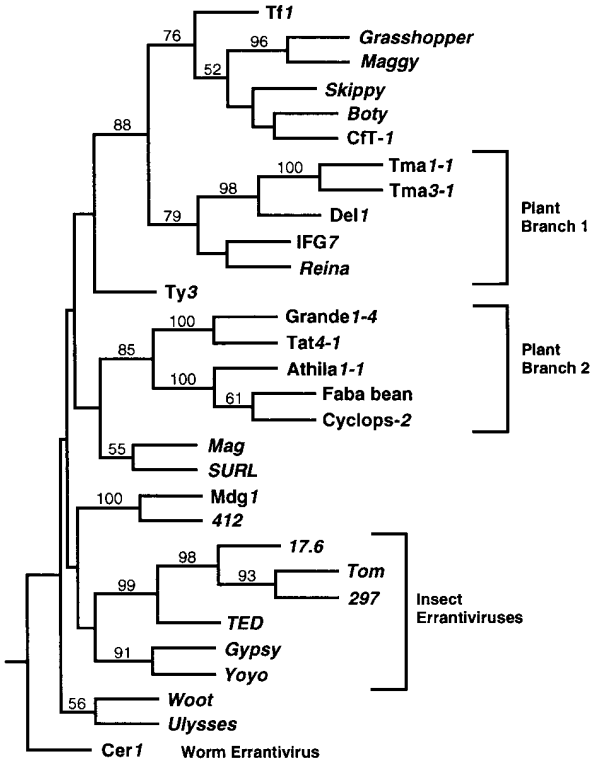

Figure 7.—Phylogenetic relationships of Ty3/*gypsy* retrotransposons. Relationships of reverse transcriptase amino-acid sequences were determined by the neighbor-joining distance algorithm using Phylip (Felsenstein 1993; Saitou and Nei 1987). The tree was rooted using retrovirus sequences. Bootstrap values of greater than 50% of 100 replicates are shown as numbers above branch nodes. Nonplant elements containing *env*-like ORFs are labeled as errantiviruses. The remaining nonplant elements do not encode *env*-like ORFs and are metaviruses. All plant elements with *env*-like ORFs are in branch 2. Elements are as follows: del*1*, *Lilium henryi* (X13886); Tma*3-1*, *A. thaliana* (BAC IG009D12); Tma*1-1*, *A. thaliana* (BAC T32N15)(AC002534); IFG7, *Pinus radiata* (Xiong and Eickbush 1990); *Reina*, *Zea mays* (U69258); Tf*1*, *Schizosaccharomyces pombe* (M38526); CfT-*1*, *Cladosporium fulvum* (Z11866); *Skippy*, *Fusarium oxysporum* (L34658); *Maggy*, *Magnaporthe grisea* (L35053); *Grasshopper*, *Magnaporthe grisea* (M77661); Faba bean element, *Vicia faba* (AB007466); Cyclops-*2*, *Pisum sativum* (AJ000640); Athila*1-1*, *A. thaliana* (P1 clone MX110) (AB005248); Grande*1-4*, *Zea diploperennis* (X97604); Tat*4-1*, *A. thaliana* (P1 clone MXA21)(AB005247); Ty3, *Saccharomyces cerevisiae* (M23367); *SURL*, *Tripneustis gratilla* (M75723); *Mag*, *Bombyx mori* (X17219); *Woot*, *Tribolium castaneum* (U09586); *Ulysses*, *Drosophila virilis* (X56645); *412*, *D. melanogaster* (X04132); *mdg1*, *D. melanogaster* (X59545); *TED*, *Trichoplusia ni* (M32662); *Tom*, *Drosophila ananassae* (Z24451); *297*, *D. melanogaster* (X03431); *17.6*, *D. melanogaster* (X01472); *Yoyo*, *Ceratitis capitata* (U60529); *gypsy*, *D. melanogaster* (M12927); Cer*1*, *Caenorhabditis elegans* (U15406).

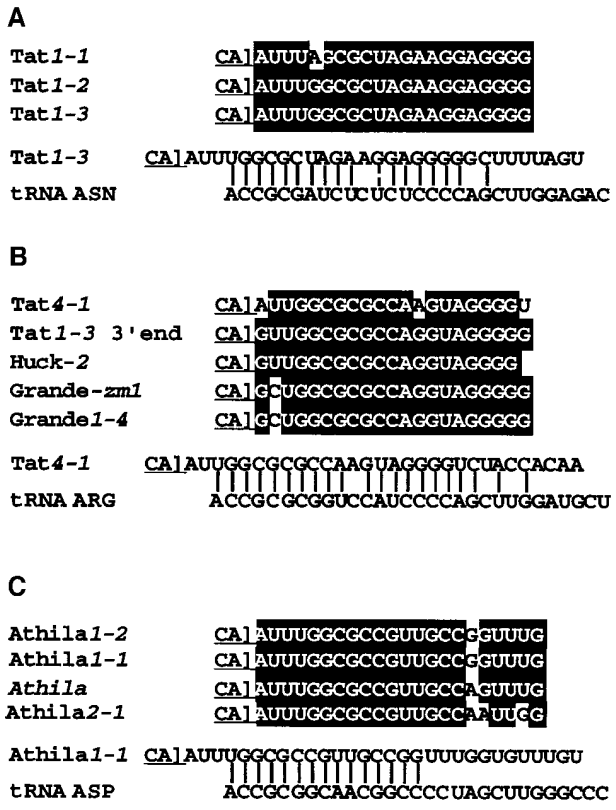

Figure 8.—Putative primer binding sites of plant Ty3/*gypsy* retrotransposons. In all panels, the shaded residues in the top figure are identical nucleotides shared among the various retrotransposon PBSs. (A) Complementarity between the 3′ end of an *A. thaliana* Asn tRNA and the Tat*1* PBS. (B) Similarities among the PBSs of Tat*4-1*, Huck-*2*, Grande-*zm1*, Grande *1-4* and the insertion in the noncoding region of Tat*1-3*. Complementarity is shown between an *A. thaliana* Arg tRNA and the PBS of Tat*4-1*. (C) Similarities between the PBS of *Athila*-like elements, and base-pairing between the 3′ end of an *A. thaliana* Asp tRNA with the Athila*1-1* PBS.

form hybridization patterns; most differences can be explained by restriction site gain or loss. Ty3/*gypsy* elements, therefore, may be more transpositionally active than other classes of *A. thaliana* retrotransposons.

Tat*1* elements may transduce genomic sequences: Tat*1*, *Zeon-1*, and Tat*4-1* have large 3′ noncoding regions (from 2 to 4.5 kb). For the Tat*1* insertions, this region is highly polymorphic and is characterized by numerous insertions/deletions. Imbedded within the 3′ noncoding region of Tat*1-3* is a 477-base insertion that contains four and a half iterations of a 26-base motif found in the leader sequence of *AT-P5C1*, an *A. thaliana* pyrroline-5-carboxylate reductase gene (Verbruggen *et al.* 1993). In addition, this 477-base insertion contains sequences that resemble part of the LTR and PBS of a putative Ty3/*gypsy* element, as well as sequences highly similar to a region of chromosome *5*. Tat*1-4* contains 1.8 kb of DNA sequence of unknown origin in place of the 477 base Tat*1-3* insertion. The large 3′ noncoding regions of *Zeon-1* and Tat*4-1* suggest that they also may carry genomic sequences. Transduction may be one mechanism by which genomic DNA sequences are incorporated into the 3′ noncoding regions of these elements. Transduction is well documented for the retroviruses,

some of which are oncogenic as a consequence of having incorporated genes for cellular growth factors. Transduction events among retrotransposons, however, are rare. The only documented retrotransposon-mediated transduction event is for the maize retroelement *Bs1*; a *Bs1* insertion within the *Adh1* gene encodes part of a cellular ATPase gene (Bureau *et al.* 1994; Jin and Bennetzen 1994). As the *A. thaliana* genome is sequenced, it will be possible to explore more definitively the origin of the sequences in the 3′ regions of these elements and the likelihood that they arose by transduction.

Plant Ty3/*gypsy* retrotransposons: The ongoing *A. thaliana* genome project has increased our understanding of plant transposable element diversity. From the available genomic DNA sequences, we have identified five *A. thaliana* Ty3/*gypsy* elements (by their characteristic reverse transcriptase sequences and *pol* gene organization) and 10 partial *A. thaliana* Ty3/*gypsy* insertions closely related to these elements. Phylogenetic analyses based on Ty3/*gypsy* reverse transcriptase amino-acid sequences resolved the plant retrotransposons into two major lineages. One is composed of del*1* from lily, *reina* from *Z. mays*, IFG7 from pine, and two *A. thaliana* Ty3/*gypsy* elements. These retrotransposons are all closely related to a group of fungal Ty3/*gypsy* retrotransposons. The second lineage includes Tat*4-1*, Athila*1-1*, and their homologs. An unusual feature of some elements in the Tat/*Athila* lineage is the presence of an additional, well-conserved ORF after the *pol* gene.

A nomenclature system proposed for the retrotransposons divides the Ty3/*gypsy* elements into two genera, the Metavirus and the Errantivirus (Boeke *et al.* 1998a). This classification is based principally on the presence of an additional ORF or *env* gene in the errantiviruses. Reverse transcriptase sequences also differ among the genera. Although little confidence can be placed in the relationships of the more basal branches of the Ty3/*gypsy* reverse transcriptase tree, the metaviruses and errantiviruses never cluster with each other, even for elements from the same species, such as those from *D. melanogaster*. Because plant Ty3/*gypsy* retrotransposons resolve into two distinct lineages, one of which contains elements with *env*-like ORFs, we propose that both genera of Metaviridae are present in plants. Specifically, we propose that del*1* and related elements are metaviruses (branch 1, Figure 7) and that Tat*4-1*, Athila*1-1* and their homologues are errantiviruses (branch 2, Figure 7). The plant errantiviruses further resolve into two major lineages, one containing Tat*4-1* (which we refer to as the Tat branch) and the other containing Athila*1-1* (which we refer to as the *Athila* branch). Not all of the insertions that we have classified as plant errantiviruses encode clear *env*-like ORFs, likely due to deletion events and sequence degeneracy. Hopefully, the ongoing genome sequencing efforts will reveal additional, more intact members of these element families to determine how well this classification system is supported.

Elements within the Tat and *Athila* branches share several other distinguishing features: Tat-branch elements have short LTRs (<550 bp) and long 3′ noncoding regions (>2 kb). Elements in the *Athila* branch have long LTRs (>1.2 kb) and are generally very large retrotransposons (>11 kb). An additional, highly polymorphic feature of the plant errantiviruses are the sequences of their putative primer binding sites. Plant Ty1/*copia* elements and all characterized plant metaviruses have PBSs complementary to an initiator methionine tRNA. This is not the case for elements in the Tat or *Athila* lineages. We identified at least three possible primer tRNAs for these retrotransposons among tRNA genes that we identified in the emerging *A. thaliana* genome sequence. Potential primers include an aspartic acid tRNA (for the *Athila* branch elements) an arginine tRNA (for Tat*4-1* and the Zea elements Huck-*2*, Grande-*zm1* and Grande*1-4*) and an asparagine tRNA (for Tat*1*).

Plant retroviruses? What is the function of the additional ORF encoded by the plant errantiviruses? Two lines of evidence suggest that it plays a role in the replication cycle of these elements: The ORF is found in multiple distinct element families, and within these elements it has evolved under functional constraints. For example, between *Athila* and Athila*1-1*, the *env*-like ORF shares ~34% similarity over >400 amino acids. Second, the ORF has a transmembrane domain, which is the most universal feature of retrovirus and animal errantivirus envelope proteins and suggests that it encodes components of a viral envelope. *Athila* and the closely related retrotransposon Athila*2-1* also encode a transmembrane domain near the N terminus of the ORF at a position typically occupied by secretory signal sequences in envelope proteins. Envelope proteins of mammalian retroviruses and animal errantiviruses share other features in common; *env* is typically encoded by a subgenomic spliced transcript, and the protein is cleaved by a cellular endopeptidase to give rise to the glycosylated surface protein and transmembrane protein of the infectious virus. Putative glycosylation sites and endopeptidase cleavage domains can also be identified in the *env*-like genes of the plant errantiviruses (data not shown). However, until a replication-competent plant errantivirus is identified, their significance remains speculative. The possibility of retroviruses in plants has been previously suggested (Bennetzen 1996). If they do exist, they have likely evolved unique mechanisms for transmission, including the ability to overcome the obstacle for infection presented by the plant cell well. At this point we cannot exclude the possibility that these elements have originated from nonplant hosts via horizontal transfer. Nonetheless, they appear to be widespread among plant genomes, as they are prevalent in the genomes of both monocots (maize) and dicots (*A. thaliana*, pea, faba bean).

Most LTR retrotransposons replicate strictly within the confines of their host cells. The finding that the *gypsy* retrotransposon of *D. melanogaster* has an infectious extracellular stage, however, has made it evident that infectious LTR retroelements are not limited to the vertebrate retroviruses (Kim *et al.* 1994; Song *et al.* 1994). Our discovery of retrotransposons with a third ORF in the plant kingdom suggests that infectious LTR retroelements are pervasive. Their presence, coupled with evidence that some Tat*1* elements and the maize *Bs1* elements transduce genomic sequences, argues that barriers for interspecies gene flow may not be very rigid in plants. Animal retroviruses are notorious causal agents of disease, and if plant retroviruses exist, they may play a role in plant disease or disease processes. Infectious retrotransposons would offer many potential applications in plant biotechnology. For example, the unusually large sizes of these elements, and in particular, their large 3′ noncoding regions, suggest that they can be modified to carry additional genes for use as vectors for plant gene transfer. It is our hope that the plant genome efforts will uncover replication-competent members of this unusual group of retrotransposons that can be used to test directly the biological significance of their envelope-like ORFs and the likelihood that these elements are transmitted extracellularly.

We thank members of the Voytas lab for helpful comments on the manuscript. This is Journal Paper No. J-17759 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 3120, and was supported by Hatch Act and State of Iowa funds.

LITERATURE CITED

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

Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman *et al.*, 1987 *Current Protocols in Molecular Biology*. Greene/Wiley Interscience, New York.

Bennetzen, J. L., 1996 The contributions of retroelements to plant genome organization, function and evolution. Trends Microbiol. **4:** 347–353.

Boeke, J. D., T. Eickbush, S. B. Sandmeyer and D. F. Voytas, 1998a Metaviridae, in *Virus Taxonomy: ICTV VIIth Report*, edited by F. A. Murphy. Springer-Verlag, New York.

Boeke, J. D., T. Eickbush, S. B. Sandmeyer and D. F. Voytas, 1998b Pseudoviridae, in *Virus Taxonomy: ICTV VIIth Report*, edited by F. A. Murphy. Springer-Verlag, New York.

Boeke, J. D., and S. B. Sandmeyer, 1991 Yeast transposable elements, pp. 193–261 in *The Molecular and Cellular Biology of the Yeast* Saccharomyces, edited by J. Broach, E. Jones and J. Pringle. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Braiterman, L. T., and J. D. Boeke, 1994 Ty*1 in vitro* integration: affects of mutations in *cis* and in *trans*. Mol. Cell. Biol. **14:** 5731–5740.

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

Church, G. M., and W. Gilbert, 1984 Genomic sequencing. Proc. Natl. Acad. Sci. USA **81:** 1991–1995.

Devereux, J., P. Haeberli and O. Smithies, 1984 A comprehensive set of sequence analysis 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.

Felsenstein, J., 1993 *PHYLIP (Phylogeny Inference Package)*. Department of Genetics, University of Washington, Seattle.

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.

Hirochika, H., K. Sugimoto, Y. Otsuki, H. Tsugawa and M. Kanda, 1996 Retrotransposons of rice involved in mutations induced by tissue culture. Proc. Natl. Acad. Sci. USA **93:** 7783–7788.

Hu, W., O. P. Das and J. Messing, 1995 Zeon-*1*, a member of a new maize retrotransposon family. Mol. Gen. Genet. **248:** 471–480.

Jin, Y.-K., and J. L. Bennetzen, 1994 Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs*1* retroelement of maize. Plant Cell **6:** 1177–1186.

Kim, A., C. Terzian, P. Santamaria, A. Pelisson, N. Purd'homme *et al.*, 1994 Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **91:** 1285–1289.

Knoop, V., M. Unseld, J. Marienfeld, P. Brandt, S. Sunkel *et al.*, 1996 *Copia*-, *gypsy*- and LINE-like retrotransposon fragments in the mitochondrial genome of *Arabidopsis thaliana*. Genetics **142:** 579–585.

Konieczny, A., D. Voytas, M. Cummings and F. Ausubel, 1991 A superfamily of retrotransposable elements in *Arabidopsis thaliana*. Genetics **127:** 801–809.

Kranz, A., and B. Kirchheim, 1987 Genetic resources in *Arabidopsis*. *Arabidopsis* Inf. Serv. **24**.

Lowe, T. M., and S. R. Eddy, 1997 tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence Nucleic Acids Res. **25:** 955–964.

Maestre, J., T. Tchenio, O. Dhellin and T. Heidmann, 1995 mRNA retroposition in human cells: processed pseudogene formation. EMBO J. **14:** 6333–6338.

Marck, C., 1988 *DNA Strider*: a C program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Res. **16:** 1829–1836.

Peleman, J., B. Cottyn, W. Van Camp, M. Van Montagu and D. Inze, 1991 Transient occurrence of extrachromsomal DNA of an *Arabidopsis thaliana* transposon-like element, Tat*1*. Proc. Natl. Acad. Sci. USA **88:** 3618–3622.

Pelissier, T., S. Tutois, J. Deragon, S. Tourmente, S. Genestier *et al.*, 1995 *Athila*, a new retroelement from *Arabidopsis thaliana*. Plant Mol. Biol. **29:** 441–452.

Purugganan, M. D., and S. R. Wessler, 1994 Molecular evolution of *magellan*, a maize Ty*3*/*gypsy*-like retrotransposon. Proc. Natl. Acad. Sci. USA **91:** 11674–11678.

Rost, B., R. Casadio, P. Fariselli and C. Sander, 1995 Prediction of helical transmembrane segments at 95% accuracy. Protein Sci. **4:** 521–533.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SanMiguel, P., A. Tikhonov, Y. Jin, N. Motchoulskaia, D. Zakharrov *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765–768.

Song, S., T. Gerasimova, M. Kurkulos, J. Boeke and V. Corces, 1994 An *env*-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus. Genes Dev. **8:** 2046–2057.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Verbruggen, N., R. Villarroel and M. Van Montagu, 1993 Osmoregulation of a pyrroline-5-carboxylate reductase gene in *Arabidopsis thaliana*. Plant Physiol. **103:** 771–781.

von Heijne, G., 1986 A new method for predicting signal sequence cleavage sites. Nucleic Acids Res. **14:** 4683–4690.

Voytas, D. F., 1996 Retroelements in genome organization. Science **274:** 737–738.

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.

White, S. E., L. F. Habera and S. R. Wessler, 1994 Retrotransposons in the flanking regions of normal plant genes: a role for *copia*-like elements in the evolution of gene structure and expression. Proc. Natl. Acad. Sci. USA **91:** 11792–11796.

Wright, D. A., N. Ke, J. Smalle, B. M. Hauge, H. M. Goodman and D. F. Voytas, 1996 Multiple non-LTR retrotransposons in the genome of *Arabidopsis thaliana.* Genetics **142:** 569–578.

Xiong, Y., and T. H. Eickbush, 1990 Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9:** 3353–3362.

Communicating editor: V. Sundaresan