Multiple Non-LTR Retrotransposons in the Genome of *Arabidopsis thaliana*

David A. Wright,* Ning Ke,* Jan Smalle,^{†,1} Brian M. Hauge,^{†,2} Howard M. Goodman[†]
and Daniel F. Voytas*

*Department of Zoology and Genetics, Iowa State University, Ames, Iowa 50011 and [†]Department of Genetics, Harvard Medical School and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114

Manuscript received June 13, 1995

Accepted for publication October 14, 1995

ABSTRACT

DNA sequence analysis near the *Arabidopsis thaliana* *ABI3* gene revealed the presence of a non-LTR retrotransposon insertion that we have designated *Ta11-1*. This insertion is 6.2 kb in length and encodes two overlapping reading frames with similarity to non-LTR retrotransposon proteins, including reverse transcriptase. A polymerase chain reaction assay was developed based on conserved amino acid sequences shared between the *Ta11-1* reverse transcriptase and those of non-LTR retrotransposons from other species. Seventeen additional *A. thaliana* reverse transcriptases were identified that range in nucleotide similarity from 48–88% (*Ta12*–*Ta28*). Phylogenetic analyses indicated that the *A. thaliana* sequences are more closely related to each other than to elements from other organisms, consistent with the vertical evolution of these sequences over most of their evolutionary history. One sequence, *Ta17*, is located in the mitochondrial genome. The remaining are nuclear and of low copy number among 17 diverse *A. thaliana* ecotypes tested, suggesting that they are not highly active in transposition. The paucity of retrotransposons and the small genome size of *A. thaliana* support the hypothesis that most repetitive sequences have been lost from the genome and that mechanisms may exist to prevent amplification of extant element families.

ARABIDOPSIS *thaliana* has the smallest known genome among higher plants (LEUTWILER *et al.* 1984). With only 10⁸ bp, it lies at one end of the spectrum of plant genome size that extends as high as 10¹¹ bp for some monocot species (BENNETT and SMITH 1976). An unusual feature of the *A. thaliana* genome is the scarcity of interspersed repetitive DNA (LEUTWILER *et al.* 1984; PRUITT and MEYEROWITZ 1986). Although highly repeated and moderately repeated sequences together make up ~20% of the genome, the interspersed repeat fraction constitutes only ~2% of the total nuclear DNA (MEYEROWITZ 1992). For most eucaryotes, interspersed repeat sequences are typically mobile genetic elements, the most abundant of which are the retrotransposons, mobile elements that replicate by reverse transcription of an mRNA intermediate.

Two major classes of retrotransposons have been identified that are present in the genomes of diverse eucaryotes (DOOLITTLE *et al.* 1989; XIONG and EICKBUSH 1990). These retrotransposon classes are distinguished by whether or not they are flanked by long terminal direct repeats (LTRs) and are simply referred to as the LTR and non-LTR retrotransposons. The LTR retrotransposons are further composed of two distinct

element groups—the *Ty1/copia* and the *Ty3/gypsy* group elements—named after representative retrotransposon families from *Saccharomyces cerevisiae* and *Drosophila melanogaster* (XIONG and EICKBUSH 1990). The *Ty1/copia* and *Ty3/gypsy* groups are distinguished by the order of proteins encoded by their *pol* genes and similarities among the sequences of their reverse transcriptases.

We and others have previously assessed the distribution and diversity of *Ty1/copia* group elements among a wide variety of plant species (FLAVELL *et al.* 1992a,b; VOYTAS *et al.* 1992; HIROCHIKA and HIROCHIKA, 1993). This was accomplished using a polymerase chain reaction assay that specifically amplified *Ty1/copia* group reverse transcriptases. Species representing each Division of the plant kingdom were found to carry *Ty1/copia* group elements, and many species were found to harbor multiple diverse families of these retrotransposons. In cotton, for example, nine distinct lineages of *Ty1/copia* group elements were identified from the analysis of 89 partial reverse transcriptase sequences (VANDERWIEL *et al.* 1993). More recently, analysis of sequences in the DNA databases have revealed *Ty1/copia* retrotransposon insertions adjacent to 21 plant genes, further indicating the abundance of these elements in plants (WHITE *et al.* 1994). Although less well documented, examples of plant *Ty3/gypsy* group elements have also been reported (SMYTH *et al.* 1989; PURUGGANAN and WESSLER 1994), as have two examples of non-LTR retrotransposons (SCHWARZ-SOMMER *et al.* 1987; LEETON and SMYTH 1993).

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

¹Present address: Laboratorium Genetika, Universiteit Gent, Ledeganckstraat 35, B-9000 Gent, Belgium.

²Present address: Transkaryotic Therapies, Inc., 195 Albany St., Cambridge, MA 02139.

Retrotransposons have influenced plant genome organization by contributing to genome size. This occurs because retrotransposition is a replicative process and unlike the DNA transposons, retrotransposons do not excise. Rather, a single insertion can be transcribed to yield mRNA templates, which in turn are reverse transcribed to generate numerous progeny elements. The extent to which retrotransposons can contribute to genome bulk is well exemplified by lily, which has one of the largest characterized plant genomes (BENNETT and SMITH 1976). Two retrotransposon families have been identified in lily, the *del1* and *del2* elements, which have ~13,000 and 250,000 copies, respectively (SENTRY and SMYTH 1985; SMYTH *et al.* 1989). The *del2* elements alone account for ~4% of the lily genome. This amounts to ~10⁹ base pairs or 10 times the size of the entire *A. thaliana* genome (SMYTH *et al.* 1989; SMYTH 1991).

While the *A. thaliana* genome is small and lacks abundant interspersed repeats, it nonetheless carries numerous diverse *Ty1/copia* group retrotransposons (VOYTAS and AUSUBEL 1988; VOYTAS *et al.* 1990; KONIECZNY *et al.* 1991). We have previously identified and characterized 10 such families that are comparable in number and diversity to element families in species like cotton, which have considerably larger genomes (10⁸ bp *vs.* 5 × 10⁹ bp) (KONIECZNY *et al.* 1991; VANDERWIEL *et al.* 1993). Unlike their counterparts in other plants, the *A. thaliana* elements have not amplified to appreciable copy numbers. Each element family is represented by one or few insertions.

Here we present evidence for the presence of a class of retrotransposons not previously described in *A. thaliana*—the non-LTR retrotransposons. One such element, *Ta11-1*, was identified through analysis of DNA sequences flanking the *ABI3* gene. Sequence similarity between *Ta11-1* and related retrotransposons provided the basis for a polymerase chain reaction assay, which was used to characterize the number and diversity of *A. thaliana* non-LTR retrotransposons. Members of 17 additional element families were identified, including one that resides in the mitochondrial genome. As with the *A. thaliana* *Ty1/copia* group retrotransposons, the non-LTR retrotransposons are also all of low copy number. This suggests that at least in part, the small size of the *A. thaliana* genome is the consequence of the failure of retrotransposons to successfully amplify and colonize its genome.

MATERIALS AND METHODS

Plant material: The *A. thaliana* ecotypes represent wild populations that have been collected from around the world (KRANZ and KIRCHHEIM 1987): Kas, India; Co, Portugal; Sei, Italy; No, Germany; Mv, United States; Ll, Spain; Cvi, Cape Verde Islands; Fi, Finland; Ba, Great Britain; Hau, Denmark; Aa, Germany; Ge, Switzerland; Ms, Soviet Union; Ag, France; Mh, Poland. Columbia (Col) and Landsberg (La) are stan-

dard laboratory strains; Landsberg carries the recessive *erecta* mutation.

DNA manipulations: Standard methods were used for DNA manipulations, including the purification of plant, yeast and plasmid DNA, the preparation of Southern filters, the screening of yeast artificial chromosome (YAC) and lambda phage libraries, and enzymatic manipulation of cloned DNAs (AUSUBEL *et al.* 1987). All filter hybridizations were conducted at 65° by the method of CHURCH and GILBERT (1984). DNA fragments used as hybridization probes were separated on low melting agarose gels; gel slices containing desired DNA fragments were excised, melted and used directly for radiolabeling by random priming (Promega).

The presence of *Ta11-1* was initially suggested from the sequence of cosmid clone 4711, which was isolated from a Columbia genomic library (GIRAUDAT *et al.* 1992). To obtain the entire *Ta11-1* insertion as well as flanking sequences, a Columbia YAC library (GRILL and SOMERVILLE 1991) was screened using cosmid 4711 as a hybridization probe. A single clone was identified (EG15A3), from which a 7.2-kb *Bgl*II fragment was subcloned that was predicted to span the 5' end of *Ta11-1* (Figure 1). Because Southern hybridization analyses suggested that Landsberg did not carry a *Ta11-1* insertion, the empty target site was cloned using a Landsberg genomic lambda phage library (VOYTAS *et al.* 1990); hybridization probes included a 1.1-kb *Eco*RI/*Hind*III fragment 3' of *Ta11-1* (probe A, Figure 1) and an ~0.8-kb *Bgl*II/*Sal*I fragment near the 5' end of the insertion (probe B, Figure 1). The library screen yielded one phage clone with probe A (λDW1) and three unique clones with probe B (λDW2, λDW3, λDW4). λDW1 did not hybridize to phage isolated with probe B. This data, as well as additional DNA sequence of *Ta11-1*, suggested that YAC EG15A3 was a chimera, with the junction residing at the *Bam*HI site in *Ta11-1* (see RESULTS). The *bona fide* *Ta11-1* insertion, therefore, was cloned from a Columbia genomic lambda phage library (kind gift of J. MULLIGAN and R. DAVIS). The library was screened using a 1.8-kb *Hind*III fragment from λDW1 adjacent to the site of *Ta11-1* insertion (probe D, Figure 1).

PCR assay for reverse transcriptases: Two conserved amino acid sequence domains shared among non-LTR retrotransposon reverse transcriptases were used to design completely degenerate oligonucleotide primers (DVO144 = GGGATCCNGGNCNGAYGGNWT and DVO145 = GGAATTCGGNSWNARNGGRYMNCCYTG, where R = A + G; Y = C + T; M = A + C; S = G + C; W = A + T; N = A + G + C + T). Amplifications were carried out in 50-μl reactions that included 0.5 μg of Landsberg genomic DNA, 50 pmol of each primer, 2.5 units of *Taq* DNA polymerase and buffer supplied by the manufacturer (Gibco) with a final MgCl₂ concentration of 2.5 mM. Amplifications were conducted for 30 cycles (denaturation, 94° for 1 min; annealing, 47° for 1 min; extension, 72° for 2 min). The resultant 600-bp product was gel purified on low melting agarose gels and cloned into pT7Blue T-vector (Novagen). A total of 26 independent clones were sequenced.

DNA sequencing and sequence analysis: Templates for sequencing *Ta11-1* were generated by subcloning specific restriction fragments or through γδ-transposon mutagenesis (STRATHMANN *et al.* 1991) (TN1000, Gold Biotechnology). Oligonucleotides were also synthesized to prime some sequencing reactions. DNA sequences were determined by the DNA Sequencing Facility of Iowa State University or the fmol DNA sequencing system (Promega). Data analysis was performed using the GCG computer software (DEVEREUX *et al.* 1984). The phylogenetic tree was constructed by the neighbor-joining method (SAITOU and NEI 1987) using the program, MEGA (KUMAR *et al.* 1993). Amino acid sequences were used for the analysis; no distance correction formula was ap-

FIGURE 1.—Restriction endonuclease maps at the site of the *Ta11-1* insertion. The location of *Ta11-1* is shown with respect to the *ABI3* gene as well as the chromosome 3 morphological markers, *gl1* and *hy2*. All restriction endonuclease sites are shown for *EcoRI* (E), *BamHI* (B), *PstI* (P) and *KpnI* (K). Other relevant sites discussed in the Materials and Methods include *BglII* (Bg) and *XbaI* (X). The left-most boundary of the insert in cosmid 4711 is shown as a jagged line. The heavy line in YAC EG15A3 denotes chimeric DNA. The two lambda phage clones, λ DW15 and λ DW1, encompass the *Ta11-1* insertion in Columbia (Col) or the empty target site in Landsberg (La), respectively. Bars designated A, B, C and D indicate restriction fragments used as hybridization probes.

plied. All stop codons and frameshifts were eliminated from the data set. The tree was unrooted, and bootstrap analysis was conducted with 1000 replicates. GenBank accession numbers for DNA sequences described in this manuscript are as follows: *Ta11-1*, L47193; *Ta11-1* target site, L47211; *Ta12-Ta28*, L47175, L47181-92, L47194-5 and L47289-90.

RESULTS

Identification of *Ta11*, an *A. thaliana* non-LTR retrotransposon: The *A. thaliana ABI3* gene was cloned using a map-based strategy that entailed constructing a detailed map of closely linked restriction fragment length polymorphisms surrounding the *abi3* locus (GIRAUDAT *et al.* 1992). Through this analysis, *ABI3* was found to reside within a cosmid clone designated 4711. DNA sequence analysis of >23 kb of 4711 was performed to identify potential coding regions, and one open reading frame (ORF) was shown to encode *ABI3* (GIRAUDAT *et al.* 1992). In addition, a large ORF of 985 amino acids was found at one end of 4711 proximal to the *gl1* locus (Figure 1). Comparison of the amino acid sequence of this ORF to the DNA sequence databases revealed strong similarity to reverse transcriptases of non-LTR retrotransposons (see below), suggesting that this sequence is an insertion of a novel *A. thaliana* retrotransposon. We refer to this sequence as *Ta11-1* to indicate that it is the first insertion identified from the eleventh characterized *A. thaliana* retrotransposon family.

The ORF encoded by *Ta11-1* begins at the junction between vector and insert in cosmid 4711. To obtain the entire insertion, 4711 was used as a hybridization probe to screen a yeast artificial chromosome (YAC) library. One YAC was identified, EG15A3, from which a 7.2-kb *BglII* fragment was subcloned that extended the region 5' of the ORF by ~5.2 kb. Of this *BglII* fragment, 2.8 kb was sequenced, and the potential coding region was found to extend an additional 1.6 kb, to the *BamHI* site depicted in Figure 1.

The ends of non-LTR retrotransposons are difficult to discern because they are not flanked by terminal repeat sequences. Because *Ta11-1* was identified from the Columbia ecotype, Southern hybridizations were used to screen several other *A. thaliana* ecotypes to determine whether any lacked this insertion (data not shown). The common laboratory strain Landsberg was found to carry a number of restriction fragment length polymorphisms in this region when compared with Columbia, suggesting that either *Ta11-1* was absent or rearrangements at this site had occurred between the two strains. A genomic lambda phage library of Landsberg DNA was therefore screened using hybridization probes from either side of *Ta11-1* to clone the putative empty target site (Figure 1).

If the presence of *Ta11-1* in Columbia were due to a simple retrotransposon insertion, then the sequences immediately 5' and 3' of the ORF would be predicted to be juxtaposed in Landsberg. However, hybridization analyses revealed that phage clones isolated from Landsberg with 5' and 3' probes did not overlap and were separated by ≥ 10.7 kb. DNA sequence analysis further revealed that the *Ta11-1* ORF stopped at the *BamHI* site (Figure 1). Because *BamHI* was used to construct the YAC library, we considered that the YAC clone may be a chimera, consisting of sequences from two different regions of the *A. thaliana* genome. To determine if this was the case, *Ta11-1* was cloned from a second library, a lambda phage library of Columbia DNA. The region 5' of the *Ta11-1* in the phage clones differed from the corresponding region of the YAC clone, indicating that the YAC was a chimera. The *Ta11-1* coding region was found to extend ~1 kb 5' of the *BamHI* site (Figure 1).

Delimiting the ends of *Ta11-1*: The DNA sequence of *Ta11-1* and the presumed target site from Landsberg made it possible to delimit the ends of the element.

FIGURE 2.—The site of *Ta11-1* integration in Columbia (Col) and Landsberg (La). (A) Diagram depicting DNA organization at the site of insertion. Bars indicate the presumptive 15-bp target site duplication (a) and the adjacent 17 bp (b) that is present in both Col and La. The distance in bp is given between the beginning of *Ta11-1* and the first methionine (M) in ORF1 and between the stop codon (*) and the putative 3' target site duplication (a). (B) The DNA sequence of features depicted in A.

The Landsberg sequence contains an insertion of 197 bp not present in Columbia (Domain I, Figure 2). Because non-LTR retrotransposons create target site duplications ranging from 3 to 16 bp (HUTCHISON *et al.* 1989), sequences at the 5' junction of *Ta11-1* were compared with sequences downstream of the ORF. Fifteen base pairs immediately adjacent to *Ta11-1* (a, Figure 2) are repeated in direct orientation 284 bases downstream of the *Ta11-1* stop codon. The 3' repeat is immediately preceded by 11 adenine residues. Poly(A) stretches typically define the 3' ends of non-LTR retrotransposons and reflect the use of a polyadenylated mRNA as a template for reverse transcription (HUTCHISON *et al.* 1989; LUAN *et al.* 1993). The 3' repeat is flanked by 17 bases (b, Figure 2) that are present in Domain I of Landsberg (with one mismatch). In addition to this sequence, there is an additional 201 bp insertion at the 3' end of *Ta11-1* that is unique to Columbia (Domain II, Figure 2). While it is possible that Domain II is part of *Ta11-1*, there is no evidence of a target duplication adjacent to this sequence, and it is not preceded by poly(A). Until additional elements are characterized to confirm our interpretation of the element boundaries, we designate *Ta11-1* as the 6077 bp between the presumptive target site duplication (a, Figure 2). This interpretation suggests that the integration of *Ta11-1* resulted in the loss of Domain I and the addition of Domain II. Alternatively, *Ta11-1* and Domain II could have been deleted from Landsberg. Concomitant with this deletion was the gain of Domain I, perhaps through recombination with a *Ta11* element elsewhere in the genome.

***Ta11-1* encodes ORFs with similarity to non-LTR retrotransposons:** *Ta11-1* encodes two open reading frames, designated ORF1 and ORF2, of 487 and 1339 amino acids, respectively. These ORFs are separated by a -1 frame-shift (Figure 3), which is similar to the organization of ORFs encoded by most LTR retrotransposons and retroviruses, as well as some non-LTR retrotransposons (*e.g.*, LEETON and SMITH 1993; DANILEVSKAYA *et al.* 1994). For the LTR retroelements, the expression of the second reading frame is typically regulated by ribosomal frameshifting (JACKS 1990). The frameshift event is facilitated by ribosomal pausing, frequently provided by RNA secondary structure immediately downstream of the frameshift site. Two stem/loops are present in *Ta11-1* downstream of the frameshift site, which may serve such a function (data not shown).

Three amino acid sequence domains are present in ORF1 and ORF2 that show similarity to proteins encoded by non-LTR retrotransposons. Two of these are short cysteine motifs or finger domains, which are located in the *Ta11-1* ORFs at positions comparable with those of other non-LTR retrotransposons (MCCLURE 1991). The motif in ORF1 has the general structure CX₂CX₄HX₄C, and this motif is also found in *gag* proteins of LTR retrotransposons and retroviruses (COVEY 1986). Near the carboxy terminus of the second ORF is another finger domain that has the general structure of CX₁₋₃CX₇₋₈HX₄C (Figure 3). Although the function of these sequences is unknown, they are generally thought to be involved in binding nucleic acids (COVEY 1986; MCCLURE 1991).

The second and most striking region of sequence con-

FIGURE 3.—Conserved amino acid sequence domains in *Ta11-1*. (A) Diagram depicts structural organization of *Ta11-1*. ORF1 and ORF2 are represented by open boxes. Black and stippled boxes depict cysteine motifs and reverse transcriptase, respectively. Sites for *Bam*HI (B) and *Eco*RI (E) correspond to those shown in Figure 1. The arrowhead denotes the 3' end of *Ta11-1*, and the vertical bar denotes the 5' end of the element. (B) Alignment of conserved amino acid sequence domains. Black boxes designate residues conserved in five of the six sequences depicted. The LINE-1 and *R2Dm* elements do not have an ORF1 cysteine motif, and only one of two such motifs from *cin4* is depicted. For reverse transcriptase, I–VII denote the seven domains characteristic of all known reverse transcriptases (XIONG and EICKBUSH 1990). These domains are slightly truncated from the original description. An additional domain conserved among non-LTR retrotransposon reverse transcriptase is also shown preceding domain I. Asterisks indicate amino acid sequences used to design degenerate oligonucleotides for use in the polymerase chain reaction. GenBank accession numbers for retrotransposon sequences are: *cin4* (*Z. mays*) (Y00086), *del2* (*L. speciosum*) (Z17425), *Tx1*, (*Xenopus laevis*) (M26915), LINE-1 (*Homo sapiens*) (M80340) and *R2Dm* (*D. melanogaster*) (X51967).

servation consists of 315 amino acid residues near the middle of ORF2. This sequence shows significant similarity to reverse transcriptases, including the seven amino acid domains characteristic of all retroelement reverse transcriptases (Figure 3B) (XIONG and EICKBUSH 1990). The *Ta11-1* sequence is most closely related to a group of non-LTR elements that includes the mammalian LINE-1 elements, insect retrotransposons that reside in rDNAs, and the two non-LTR elements previously identified from plants, the *cin4* elements of *Zea mays* and the *del2* elements of *Lilium speciosum*. *Ta11-1* shares 38 and 25% amino acid identity to the *cin4* and *del2* reverse transcriptases, respectively.

A. thaliana has multiple diverse families of non-LTR retrotransposons: We have previously shown that plant genomes, including *A. thaliana*, contain numerous distinct families of LTR retrotransposons (KONIECZNY *et al.* 1991; VOYTAS *et al.* 1992). These LTR retrotransposons are members of the so-called Ty1/*cop* group elements, and were identified using a polymerase chain reaction assay that amplifies conserved regions of Ty1/*cop* group reverse transcriptases. Alignment of the *Ta11-1* reverse transcriptase with those from related non-LTR retrotransposons also revealed several regions of nearly invariant amino acid sequence (Figure 3). A PCR assay was developed based on these conserved

FIGURE 4.—Alignment of non-LTR retrotransposon reverse transcriptases identified by the polymerase chain reaction. Black boxes indicate amino acids conserved in 20 of the 22 aligned sequences. Asterisks over the sequences indicate the amino acids encoded by the primers. Asterisks imbedded in the sequence depict stop codons, periods represent gaps inserted to optimize alignment, and × denotes missing data. Reverse transcriptase domains I–IV are shown (see Figure 3) (XIONG and EICKBUSH 1990). Related non-LTR retrotransposon sequences are as in Figure 3.

sequences to assess *A. thaliana* non-LTR retrotransposon diversity.

The two amino acid sequence domains chosen for oligonucleotide design lie near the amino terminus of reverse transcriptase (Figure 3B). The 5′-most domain actually precedes the “core” reverse transcriptase (as defined by XIONG and EICKBUSH 1990), and this domain is well conserved among non-LTR retrotransposon reverse transcriptases. The second domain corresponds to domain IV (XIONG and EICKBUSH 1990), so the resultant PCR product spans three conserved amino acid domains to aid in the identification of novel reverse transcriptases. It should be noted that the PCR assay was not predicted to amplify Ta11-1, because the 5′ primer does not match the sequence of this element due to an amino acid substitution (Figure 3B).

Consistent with the spacing of the amino acid sequence domains, the PCR primers amplified a ~600-bp DNA fragment from Landsberg genomic DNA. This PCR product was cloned into a plasmid vector and DNA sequences were obtained for 26 independent clones. The nucleotide and derived amino acid sequences were compared with the DNA sequence databases. Twenty-two clones encoded ORFs with strong similarity to non-

LTR retrotransposon reverse transcriptases (Figure 4). Six of these ORFs had stop codons, three had a shift in reading frame, and one had both a stop codon and a frame shift. This suggested that these clones are derived from nonfunctional elements or that errors were induced during DNA amplification. All 22 clones showed a high degree of similarity to the three conserved amino acid domains included between the PCR primers (Figure 4). Two clones did not show any apparent reverse transcriptase similarity or similarity to sequences in the DNA databases, and two had the same primer at both ends, indicating that they were likely spurious amplification products.

Among the 26 clones, 17 distinct reverse transcriptases were identified that ranged in nucleotide similarity from 48 to 88%. Because of the high level of nucleotide differences between clones, we concluded that the 17 clones represent distinct retrotransposon reverse transcriptases. We designated these new retrotransposon families as Ta12-Ta28. Among the 26 clones, two clones were identified for each of Ta12, Ta13, Ta15, Ta17 and Ta22. The low level of redundancy among clones in the sample suggests that many more distinct non-LTR retrotransposons are resident in the *A.*

FIGURE 5.—Neighbor-joining tree of partial non-LTR retrotransposon reverse transcriptase sequences. The tree is unrooted, and branch lengths are proportional to the percent amino acid distance between sequences. Numerals adjacent to branches indicate percentage of bootstrap replicates supporting that branch.

thaliana genome that were not identified in this initial screen.

Relationships among the *A. thaliana* non-LTR reverse transcriptases were assessed using the neighbor-joining method (SAITOU and NEI 1987). Included in the data set were reverse transcriptases from other closely related non-LTR retrotransposons. The *A. thaliana* sequences formed a single, strongly supported clade by both neighbor-joining analysis (Figure 5) and parsimony analysis (not shown), indicating that they are more closely related to each other than to reverse transcriptases from other species. This is consistent with the vertical evolution of these sequences over most of their evolutionary history.

Each of the clones encoding a distinct reverse transcriptase was hybridized to Southern filters containing genomic DNA from a collection of 17 diverse *A. thaliana* ecotypes (e.g., Figure 6). All of the sequences were of low copy number and ranged from zero to no more than seven distinct insertions among the ecotypes (data not shown). In most cases, hybridization patterns were very similar among the ecotypes, suggesting that each had the same complement of retrotransposon insertions at the same chromosomal locations. The Co ecotype had the most retrotransposons, with five, six and seven members of the Ta20, Ta11 and Ta27 families, respectively (data not shown).

Ta17 resides in the mitochondrial genome: The Ta17 reverse transcriptase hybridized to a single restriction fragment among the ecotypes analysed. The intensity of the hybridization signal, however, was considerably greater than those observed with the other clones (Figure 6). This suggested that Ta17 was highly represented in the genomic DNA, and the hybridization signal was

similar to that observed with plastid or mitochondrial DNA probes (data not shown).

The DNA sequence of Ta17 was compared with the data that is accumulating as part of the effort to sequence the entire *A. thaliana* mitochondrial genome. A perfect match was found in the mitochondrial sequence confirming its organellar location (KNOOP *et al.* 1996). The Ta17 sequence has a frame shift and a stop codon, suggesting that it no longer encodes an active reverse transcriptase. Reverse transcriptases are known to be associated with some group II mitochondrial introns, particularly those of fungi (XIONG and EICKBUSH 1990). These reverse transcriptases, however, are clearly distinct from non-LTR retrotransposons (XIONG and EICKBUSH 1990). Because phylogenetic analysis places Ta17 well within the clade of other *A. thaliana* nuclear retrotransposons (Figure 5), we conclude that the presence of Ta17 in the mitochondria likely represents a transfer of sequences from the nucleus to the mitochondria, or a rare transposition event into the mitochondrial genome.

DISCUSSION

The Ta11 non-LTR retrotransposons: The non-LTR retrotransposons are among the most abundant class of interspersed repetitive sequences found in the genomes of higher eucaryotes. In plants, however, only two families of non-LTR retrotransposons have thus far been reported—the *cin4* elements of maize and the *del2* elements of lily (SCHWARTZ-SOMMER *et al.* 1987; LEETON and SMYTH 1993). Sequence analysis of a region near the *A. thaliana* *ABI3* gene revealed an open reading frame with similarity to non-LTR retrotransposon

FIGURE 6.—Southern hybridization analysis of *Ta11* and *Ta17*. Filters were prepared with genomic DNAs from 16 diverse *A. thaliana* ecotypes digested with *Bam*HI and *Bgl*II (La data not shown). Hybridization probes were from *Ta11-1* (probe C, Figure 1) (A) or the *Ta17* PCR product (B). The *Ta11-1*, *Ta11-2* and *Ta17* insertions are indicated.

reverse transcriptases. We have shown that this sequence is a member of an *A. thaliana* non-LTR retrotransposon family, which we have designated *Ta11*.

The *Ta11* insertion near *ABI3* (*Ta11-1*) has a structural organization similar to other characterized non-LTR retroelements. It has two open reading frames separated by a -1 frameshift, which encode three conserved amino acid sequence domains. These include two short cysteine motifs, which are present in ORF1 and ORF2 at comparable positions in other non-LTR retrotransposons. Although the function of these motifs remains unclear, it is likely that they bind nucleic acids during replication and integration (COVEY 1986; McCURE 1991). ORF2 also encodes a region of 315 amino acids that shows significant similarity to reverse transcriptase, including all seven conserved domains characteristic of retroelement reverse transcriptases (XIONG and EICKBUSH 1990).

Ta11-1 is not present in the Landsberg ecotype. Comparison of DNA sequence from Landsberg at the site of insertion to the Columbia sequence revealed what are apparently two insertion/deletion events in addition to the element. Although these sequence rearrangements make it difficult to definitively determine the ends of *Ta11-1*, we have identified a 15-bp direct repeat on either side of the ORFs that may represent a

target site duplication. The presumptive 3' target site duplication is immediately preceded by 11 adenine residues, similar to the 3' ends of other non-LTR retrotransposon insertions (HUTCHISON *et al.* 1989).

In contrast to most non-LTR retrotransposons, *Ta11-1* appears structurally intact. Non-LTR elements frequently have 5' deletions, which result from incomplete reverse transcription that is initiated at the poly(A) tract (LUAN *et al.* 1993). With the exception of the frame-shift, the open reading frame of *Ta11-1* is uninterrupted by stop codons for over 5.5 kb. This structural integrity suggests that *Ta11-1* has transposed relatively recently and has not accumulated numerous mutations since its insertion. In this regard, *Ta11* is similar to the LTR retrotransposons we have previously characterized in *A. thaliana*. For example, the *Ta3-1* insertion does not have any stop codons or frame-shifts in its >4.1 kb ORF (KONIECZNY *et al.* 1991). Definitive evidence on the functional integrity of *Ta11-1* will require determining whether or not it is capable of transposition. Nonetheless, the structural integrity of this element and other *A. thaliana* retrotransposons suggests that some may be transposition-competent.

An abundance of *A. thaliana* non-LTR retrotransposons: Alignment of the *Ta11* reverse transcriptase with those from related retrotransposons revealed several well-conserved amino acid sequence domains. A PCR assay was developed based on these conserved sequences to identify related *A. thaliana* retrotransposons. Of 26 cloned amplification products analyzed, 17 were found to encode putative protein products with significant similarity to non-LTR retrotransposon reverse transcriptases. These 17 sequences range in nucleotide identity from 48 to 88%, which is comparable with the divergence observed between distinct families of *A. thaliana* *Ty1/copia* retrotransposons (KONIECZNY *et al.* 1991). Each unique clone, therefore, was considered to represent a different retrotransposon family, designated *Ta12-Ta28*. Because the 17 reverse transcriptases were identified from the sequence of only 26 independent clones, it is unlikely that we have exhaustively identified all non-LTR retrotransposon sequences in the *A. thaliana* genome.

The PCR assay previously developed to amplify *Ty1/copia* group elements was based primarily on plant retrotransposon sequences and as a consequence, was found to only robustly amplify plant DNAs (VOYTAS *et al.* 1992). More exhaustive searches, however, have recently identified *Ty1/copia* group elements in a few nonplant lineages, including fish, reptiles and amphibians (FLAVELL and SMITH 1992; FLAVELL *et al.* 1995). The non-LTR reverse transcriptase assay described here is based on conserved amino acid sequences from both plant and animal elements, suggesting that they may be less taxonomically restricted. In support of this, we have amplified PCR products of the appropriate size from maize, cotton and human DNA; sequence analysis of the hu-

man clones indicated that they were, in fact, human LINE-1 non-LTR retrotransposons (data not shown).

Population dynamics of *A. thaliana* retrotransposons: Retrotransposon populations are continually in flux, and variations in copy number reflect the amplification and loss of specific element lineages from the genome. Such dynamics are well exemplified in *S. cerevisiae*, where the extensive genome characterization efforts have provided a more comprehensive picture of retrotransposon populations. *S. cerevisiae* element copy numbers range from 25 to 35 copies for the Ty1 family, 10–15 copies for Ty2, and 1–4 copies each of Ty3 and Ty4 (BOEKE and SANDMEYER 1992). Active Ty5 elements, however, appear to have been completely lost from the genome of this species (ZOU *et al.* 1995).

In contrast, all of the characterized *A. thaliana* retrotransposons are of low copy number. The 17 non-LTR elements reported here range from zero to no more than seven insertions among 17 diverse ecotypes analyzed. Similar low copy numbers were also observed for the 10 families of *A. thaliana* Ty1/*copia* group elements (KONIECZNY *et al.* 1991; VOYTAS *et al.* 1992). No retrotransposon families have yet been characterized in *A. thaliana* that are of appreciable copy number. Both empirical and theoretical studies of transposable element population dynamics suggest that element populations are maintained as a result of transpositional increase in copy number, balanced by some opposing force or forces (CHARLESWORTH and LANGLEY 1989). Regulation of rates of transposition, coupled with loss through recombination (see below) may be important contributing factors in the scarcity of these sequences in the *A. thaliana* genome.

It was surprising to find that one non-LTR element, Ta17, resides in the mitochondrial genome. Typical of higher plants, the *A. thaliana* mitochondrial genome is large (372 kb) (KLEIN *et al.* 1994), and the presence of Ta17 suggests that some of this DNA may have originated from mobile genetic elements. This is further supported by the previous documentation of a degenerate reverse transcriptase in the mitochondrial genome of the plant *Oenothera* (SCHUSTER and BRENNICKE 1987) and the identification of other retrotransposon reverse transcriptases in the *A. thaliana* mitochondrial genome (KNOOP *et al.* 1996). Other reverse transcriptases are known to reside in mitochondria and plastids as part of some group II introns, however their sequences are distinct from those of other non-LTR retroelements (XIONG and EICKBUSH 1990). Phylogenetic analysis indicates that Ta17 is more closely related to nuclear retrotransposons and therefore is likely not a component of a group II intron. Although Southern hybridization analysis did not reveal the presence of nuclear copies of Ta17, we favor the hypothesis that this element arose by a transfer event from the nucleus to the mitochondria or by a transposition event to the genome of this organelle.

Origins of *A. thaliana* retrotransposons: We have previously offered two scenarios to explain the presence of the numerous low copy retrotransposons in *A. thaliana* (KONIECZNY *et al.* 1991; see also CUMMINGS 1994): These elements may have been introduced into the genome from another organism by horizontal transfer and once resident, failed to proliferate. Alternatively, retrotransposons may have been abundant in the ancestor of *A. thaliana* and were lost from the genome through a process of genome reduction. We presently favor the second scenario based on several lines of evidence: 1) Ty1/*copia* elements have been identified in all plant lineages indicating that they are ancient components of plant genomes and therefore were likely present in the ancestor of *A. thaliana* (FLAVELL *et al.* 1992a; VOYTAS *et al.* 1992; HIROCHIKA and HIROCHIKA 1993). 2) Relationships among reverse transcriptases of *A. thaliana* and other plant non-LTR sequences are consistent with organismal relationships, suggesting that these sequences have been vertically transmitted. Although phylogenetic analysis of the *A. thaliana* Ty1/*copia* group elements implied that horizontal transfer may have played a role in their evolutionary history, the high level of divergence among the sequences and the limited number of characters analyzed (100 amino acids) made it difficult to convincingly argue for episodes of horizontal transfer. 3) Finally, because of the limited phylogenetic framework for the genus *Arabidopsis*, we have used cotton as a model to study plant Ty1/*copia* group evolution (VANDERWIEL *et al.* 1993). Species within the cotton genus (*Gossypium*) have well-defined organismal phylogenies, and exhaustive characterization of retrotransposons from these species indicated that retrotransposon evolution is completely consistent with vertical transmission; no evidence for horizontal transfer was uncovered.

If retrotransposons were lost from the *A. thaliana* genome, then the ancestor of *A. thaliana* likely had multiple active elements that evolved into the diverse families found today. The identification of transposition-competent retrotransposons will make it possible to determine if regulation of element activity has contributed to their present-day low copy number. In addition, the genome sequencing projects for *A. thaliana* and other plants will make it possible to evaluate more precisely the role played by retrotransposon amplification and loss in plant genome architecture.

We are grateful to SABINE SÜNKEL, VOLKER KNOOP and AXEL BRENNICKE for making available unpublished data. We would like to acknowledge DAVID NISSEN and BRYAN WULFEKUHLE for assistance in DNA cloning and sequencing, and JONATHAN WENDEL and RICH CRONN for assistance with the phylogenetic analysis. We are grateful to MIKE CUMMINGS for critically reading the manuscript. This work was supported by a grant from the Iowa State University Biotechnology Council to D.F.V. and a grant from Hoechst AG to Massachusetts General Hospital. This is Journal Paper No. J-16378 of the Iowa Agriculture and Home Economics Experiment Station, Ames, IA; Project No. 3120.

LITERATURE CITED

- AUSUBEL, F. M., R. BRENT, R. E. KINGSTON, D. D. MOORE, J. G. SEIDMAN *et al.*, 1987 *Current Protocols in Molecular Biology*. Greene Publishing Associates/Wiley Interscience, New York.
- BENNETT, M. D., and J. B. SMITH, 1976 Nuclear DNA amounts in angiosperms. *Proc. R. Soc. Lond. B Biol. Sci.* **274**: 227–274.
- BOEKE, J. D., and S. B. SANDMEYER, 1992 Yeast transposable elements, pp. 193–261 in *The Molecular Biology of the Yeast Saccharomyces: Genome Dynamics, Protein Synthesis, and Energetics*, edited by J. R. BROACH, J. R. PRINGLE, and E. W. JONES. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- CHARLESWORTH, B., and C. H. LANGLEY, 1989 The population genetics of *Drosophila* transposable elements. *Annu. Rev. Genet.* **23**: 251–287.
- CHURCH, G., and W. GILBERT, 1984 Genomic Sequencing. *Proc. Natl. Acad. Sci. USA* **81**: 1991–1995.
- COVEY, S. N. 1986 Amino acid sequence homology in *gag* region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. *Nucleic Acids Res.* **14**: 623–633.
- CUMMINGS, M. P. 1994 Transmission patterns of eukaryotic transposable elements—arguments for and against horizontal transfer. *Trends Ecol. Evol.* **9**: 141–145.
- DANILEVSKAYA, O., F. SLOT, M. PAVLOVA and M.-L. PARDUE, 1994 Structure of the *Drosophila* HeT-A transposon: a retrotransposon-like element forming telomeres. *Chromosoma* **103**: 215–224.
- DEVEREUX, J. P., P. HAEERLI 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.
- FLAVELL, A. J., and D. SMITH, 1993 A Ty1/ *copia* group retrotransposon sequence in a vertebrate. *Mol. Gen. Genet.* **233**: 322–326.
- FLAVELL, A. J., E. DUNBAR, R. ANDERSON, S. R. PEARCE, R. HARTLEY *et al.*, 1992a Ty1- *copia* group retrotransposons are ubiquitous and heterogeneous in plants. *Nucleic Acids Res.* **20**: 3639–3644.
- FLAVELL, A. J., D. B. SMITH and A. KUMAR, 1992b Extreme heterogeneity of Ty1- *copia* group retrotransposons in plants. *Mol. Gen. Genet.* **231**: 233–242.
- FLAVELL, A. J., V. JACKSON, M. P. IQBAL, I. RIACH and S. WADDELL, 1995 Ty1/ *copia* group retrotransposon sequences in amphibia and reptilia. *Mol. Gen. Genet.* **246**: 65–71.
- GIRAUDAT, J., B. M. HAUGE, C. VALON, J. SMALLE, F. PARCY *et al.*, 1992 Isolation of the *Arabidopsis* *ABI3* gene by positional cloning. *Plant Cell* **4**: 1251–1261.
- GRILL, E., and C. SOMERVILLE, 1991 Construction and characterization of a yeast artificial chromosome library of *Arabidopsis* which is suitable for chromosome walking. *Mol. Gen. Genet.* **226**: 484–490.
- HIROCHIKA, H., and R. HIROCHIKA, 1993 Ty1- *copia* group retrotransposons as ubiquitous components of plant genomes. *Jpn. J. Genet.* **68**: 35–46.
- HUTCHISON, C. A., S. C. HARDIES, D. D. LOEB, W. R. SHEHEE and M. H. EDGELL, 1989 LINES and related retrotransposons: long interspersed repeated sequences in the eucaryotic genome, pp. 593–617 in *Mobile DNA*, edited by D. E. BERG and M. M. HOWE. American Society for Microbiology, Washington, DC.
- JACKS, T., 1990 Translational suppression in gene expression in retroviruses and retrotransposons. *Curr. Top. Microbiol. Immunol.* **157**: 93–124.
- KLEIN, M., U. ECKERT-OSSENKOPP, I. SCHMIEDEBERG, P. BRANDT, M. UNSELD *et al.*, 1994 Physical mapping of the mitochondrial genome of *Arabidopsis thaliana* by cosmid and YAC clones. *Plant J.* **6**: 447–455.
- KNOOP, V., M. UNSELD, J. MARIENFELD, P. BRANDT, S. SÜNKEL *et al.*, 1996 *Copia*-, *gypsy*- and LINE-like retrotransposon fragments in the mitochondrial genome of *Arabidopsis thaliana*. *Genetics* **142**: 579–585.
- KONIECZNY, A., D. F. VOYTAS, M. P. CUMMINGS and F. M. AUSUBEL, 1991 A superfamily of retrotransposable elements in *Arabidopsis thaliana*. *Genetics* **127**: 801–809.
- KUMAR, S., T. KOICHI and M. NEI, 1993 MEGA: Molecular Evolutionary Genetics Analysis, Version 1.0, Pennsylvania State University, University Park, PA.
- KRANZ, A. R., and B. KIRCHHEIM, 1987 Genetic resources in *Arabidopsis*. *Arabidopsis Inform. Serv.* **24**.
- LEETON, P. R. J., and D. R. SMYTH, 1993 An abundant LINE-like element amplified in the genome of *Lilium speciosum*. *Mol. Gen. Genet.* **237**: 97–104.
- LEUTWILER, L. S., B. R. HOUGH-EVANS and E. M. MEYEROWITZ, 1984 The DNA of *Arabidopsis thaliana*. *Mol. Gen. Genet.* **194**: 15–23.
- LUAN, D. D., M. H. KORMAN, J. L. JAKUBCZAK and T. H. EICKBUSH, 1993 Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**: 595–605.
- MCCLURE, M. A., 1991 Evolution of retrotransposons by acquisition or deletion of retrovirus-like genes. *Mol. Biol. Evol.* **8**: 835–857.
- MEYEROWITZ, E. M., 1992 Introduction to the *Arabidopsis* genome, pp. 100–118 in *Methods in Arabidopsis Research*, edited by C. KONCZ, N.-H. CHUA, and J. SCHELL. World Scientific Publishing, Singapore.
- PRUITT, R. E., and E. M. MEYEROWITZ, 1986 Characterization of the genome of *Arabidopsis thaliana*. *J. Mol. Biol.* **187**: 169–183.
- PURUGGANAN, M. D., and S. R. WESSLER, 1994 Molecular evolution of *magellan*, a maize Ty3/*gypsy*-like retrotransposon. *Proc. Natl. Acad. Sci. USA* **91**: 11674–11678.
- SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**: 406–425.
- SCHUSTER, W., and A. BRENNICKE, 1987 Plastid, nuclear and reverse transcriptase sequences in the mitochondrial genome of *Oenothera*: is genetic information transferred between organelles via RNA? *EMBO J.* **6**: 2857–2863.
- SCHWARTZ-SOMMER, Z., L. LECLERCQ, E. GÖBEL and H. SAEDLER, 1987 *Cin4*, an insert altering the structure of the *A1* gene in *Zea mays*, exhibits properties of nonviral retrotransposons. *EMBO J.* **6**: 3873–3880.
- SENTRY, J. W., and D. R. SMYTH, 1985 An element with long terminal repeats and its variant arrangements in the genome of *Lilium henryi*. *Mol. Gen. Genet.* **215**: 349–354.
- SMYTH, D. R., 1991 Dispersed repeats in plant genomes. *Chromosoma* **100**: 355–359.
- 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.
- STRATHMANN, M., B. A. HAMILTON, C. A. MAYEDA, M. I. SIMON, E. M. MEYEROWITZ *et al.*, 1991 Transposon-facilitated DNA sequencing. *Proc. Natl. Acad. Sci. USA* **88**: 1247–1250.
- VANDERWIEL, P. L., D. F. VOYTAS and J. F. WENDEL, 1993 *Copia*-like retrotransposable element evolution in diploid and polyploid cotton (*Gossypium* L.). *J. Mol. Evol.* **36**: 429–447.
- 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 Ta1 retrotransposable element family of *Arabidopsis thaliana*. *Genetics* **126**: 713–721.
- VOYTAS, D. F., M. P. CUMMINGS, A. KONIECZNY, F. M. AUSUBEL and S. R. RODERMEL, 1992 *Copia*-like retrotransposons are ubiquitous among plants. *Proc. Natl. Acad. Sci. USA* **89**: 7124–7128.
- XIONG, Y., and T. H. EICKBUSH, 1990 Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**: 3353–3362.
- 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.
- ZOU, S., D. A. WRIGHT and D. F. VOYTAS, 1995 The *Saccharomyces* Ty5 retrotransposon family is associated with origins of DNA replication at the telomeres and the silent mating locus *HMR*. *Proc. Natl. Acad. Sci. USA* **92**: 920–924.

Communicating editor: J. CHORY