Ty3/*Gypsy* Retrotransposons: Description of New *Arabidopsis thaliana* Elements and Evolutionary Perspectives Derived from Comparative Genomic Data

Ignacio Marín and Carlos Lloréns

Instituto Cavanilles de Biodiversidad y Biología Evolutiva and Departamento de Genética, Universidad de Valencia, Spain

We performed a comprehensive analysis of the evolution of the *Ty3*/*Gypsy* group of long-terminal-repeat retrotransposons (also known as *Metaviridae*). Exhaustive database searches allowed us to detect novel elements of this group. In particular, the *Arabidopsis thaliana* and *Drosophila melanogaster* genome sequencing projects have recently disclosed a large number of new *Ty3*/*Gypsy* sequences. So far, elements of three different *Ty3*/*Gypsy* lineages had been described for *A. thaliana.* Here, we describe six new lineages, which we have called *Tit-for-tat1, Tit-for-tat2, Gimli, Gloin, Legolas,* and *Little Athila.* We confirm that plant *Ty3*/*Gypsy* elements form two main monophyletic groups. Moreover, our results suggest that at least four independent ancestral lineages existed before the monocot-dicot split, about 200 MYA. Twelve sequences from *D. melanogaster* that may correspond to new elements are also described. Some of these sequences are similar to those of *Osvaldo* and *Ulysses,* two elements of the *Osvaldo* clade that had never before been described for *D. melanogaster.* Comparative analyses of multiple organisms, some of them with completely sequenced genomes, show that the number of lineages of *Ty3*/*Gypsy* elements is very variable. Thus, while only 1 lineage is present in *Saccharomyces cerevisiae,* at least 6 exist in *Caenorhabditis elegans,* at least 9 are present in the *A. thaliana,* and perhaps 20 are present in *D. melanogaster.* Finally, we suggest that the presence of a chromodomain-containing integrase, a feature of some closely related *Ty3*/*Gypsy* elements of fungi, plants, and animals, may be used to define a new *Metaviridae* genus.

Introduction

Several lines of evidence have shown the close relationship between retroviruses and some long terminal repeat (LTR) retrotransposons. Phylogenetic analyses based on reverse transcriptase (RT) domain sequences have demonstrated that most LTR-containing retrotransposons belong to one of two subgroups, traditionally called *Ty1*/*Copia* and *Ty3*/*Gypsy.* RTs of *Ty3*/*Gypsy* elements and retroviruses were shown to be very similar (Xiong and Eickbush 1990). These results agreed with the fact that the structure of most *Ty3*/*Gypsy* elements resembles that of retroviruses, while *Ty1*/*Copia* elements are significantly different. Particularly, in both retroviruses and *Ty3*/*Gypsy* elements, the *pol* gene domains are in the order [protease–RT–ribonuclease H (RH)–integrase (IN)], while in *Ty1*/*Copia* elements the IN domain appears N-terminal to the RT and RH domains. Moreover, it was found that a few *Ty3*/*Gypsy* elements (e.g., *Drosophila melanogaster Gypsy*) had a third open reading frame (ORF), putatively encoding an envelope (ENV) protein. These *env*-containing elements were thus structurally identical to retroviruses (reviewed in Eickbush 1994).

Structural and functional data converged when it was shown that the *Gypsy* element of *D. melanogaster* was able in some circumstances to function as a retrovirus (Song et al. 1994; Kim et al. 1994). This result established the convenience of classifying LTR retrotransposons as viruses. In the most recent virus taxonomy, LTR-containing retroelements are classified into two main families, Pseudoviridae (corresponding to the *Ty1*/*Copia* subgroup) and Metaviridae (*Ty3*/*Gypsy* elements). The Metaviridae are further split according to the presence of the *env* gene (genus *Errantivirus*) or its absence (genus *Metavirus*) (reviewed in Pringle 1998, 1999; Hull 1999).

Various studies have analyzed in depth the evolution of *Ty3*/*Gypsy* elements using either the slowly evolving RT domain sequences or several *pol* domains at the same time, progressively including more sequences as they became available (Xiong and Eickbush 1990; Springer and Britten 1993; Eickbush 1994; Wright and Voytas 1998; Malik and Eickbush 1999; Pantazidis, Labrador, and Fontdevila 1999). In a recent study, Malik and Eickbush (1999) used phylogenetic analyses of the RT, RH, and IN domains to operatively define eight clades of *Ty3*/*Gypsy* elements. It is unclear whether all eight of those clades correspond to ancient classes of *Ty3*/*Gypsy* retrotransposons or some of them are relatively recent, because the inner branches that relate the clades in the phylogenetic tree essentially form a polytomy.

With respect to the most common approach, using just the slowly evolving RT domain sequences, the analysis of multiple domains has the obvious advantage of increasing the amount of information. However, it also has the drawback that many elements for which only partial (often RT domain) sequences are available cannot be included. This has two effects: (1) complete clades of *Ty3*/*Gypsy* retrotransposons may be missed, and (2) the phylogenetic range of a particular clade may be underestimated. Moreover, if elements of recombinant origin were present, they would be difficult to detect due to lack of resolution of the trees obtained independently with the sequences of each domain (especially those obtained from RH and IN sequences, which evolve rapidly). Thus, the comparison of trees obtained

Key words: *Gypsy*, genome sequencing, *Arabidopsis*, *Drosophila*, evolution.

Address for correspondence and reprints: Ignacio Marín, Departamento de Genética, Universidad de Valencia, Calle Doctor Moliner, 50, Burjassot 46100, Valencia, Spain. E-mail: ignacio.marin@uv.es.

Mol. Biol. Evol. 17(7):1040–1049. 2000

combining several domains but having a limited number of elements versus those obtained using only RT domain sequences but having many more elements is advisable. In this study, we analyzed all of the *Ty3*/*Gypsy* RT domain sequences currently available, with an emphasis on detecting and comparing the many new sequences generated by the genome sequencing projects. Particularly interesting are the results obtained for plant elements, results that allow us to describe the evolution of *Ty3*/*Gypsy* retrotransposons in plant species for the last 200 Myr. General conclusions about the success of this group of elements in different organisms were also obtained. Finally, we argue for additional taxonomic criteria to classify the Metaviridae.

Materials and Methods

We used RT protein sequences of known *Ty3*/*Gypsy* elements as queries to search online against the nonredundant database at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nhm.gov/). The programs TBLASTN and BLASTP (Altschul et al. 1997) were used for these searches. The limits of the RT domain were defined according to Xiong and Eickbush (1990), and thus corresponded to the most conserved part (''core'') of this domain. With the output of these multiple searches, we built a preliminary database of RT sequences formed by about 40 different elements, with representatives of all of the clades defined by Malik and Eickbush (1999). Sequences with frameshifts in the core of the RT domain were excluded from this or later analyses. Next, we proceeded exhaustively by iteratively searching with each of the sequences of our database against the nonredundant, dBEST, and month (up to August 1999) databases at NCBI, using the program TBLASTN. In this way, we generated a second database with around 150 sequences, in which we detected several duplicates, which were eliminated. We then included six additional RT domain sequences for use as outgroups (the *Drosophila* non-LTR retrotransposon *jockey*), because they represented the two groups of viruses closest to *Ty3*/*Gypsy* elements (the retrovirus HIV-1 and the caulimovirus CaMV) or simply because of their problematic phylogenetic positions (sequences of three LTR retrotransposons that are not assigned to the *Ty1*/*Copia* or *Ty3*/*Gypsy* groups: *Prt1*, *Dirs*, and *Pat*; see Eickbush 1994). This work is based on our final database, completed in late August 1999.

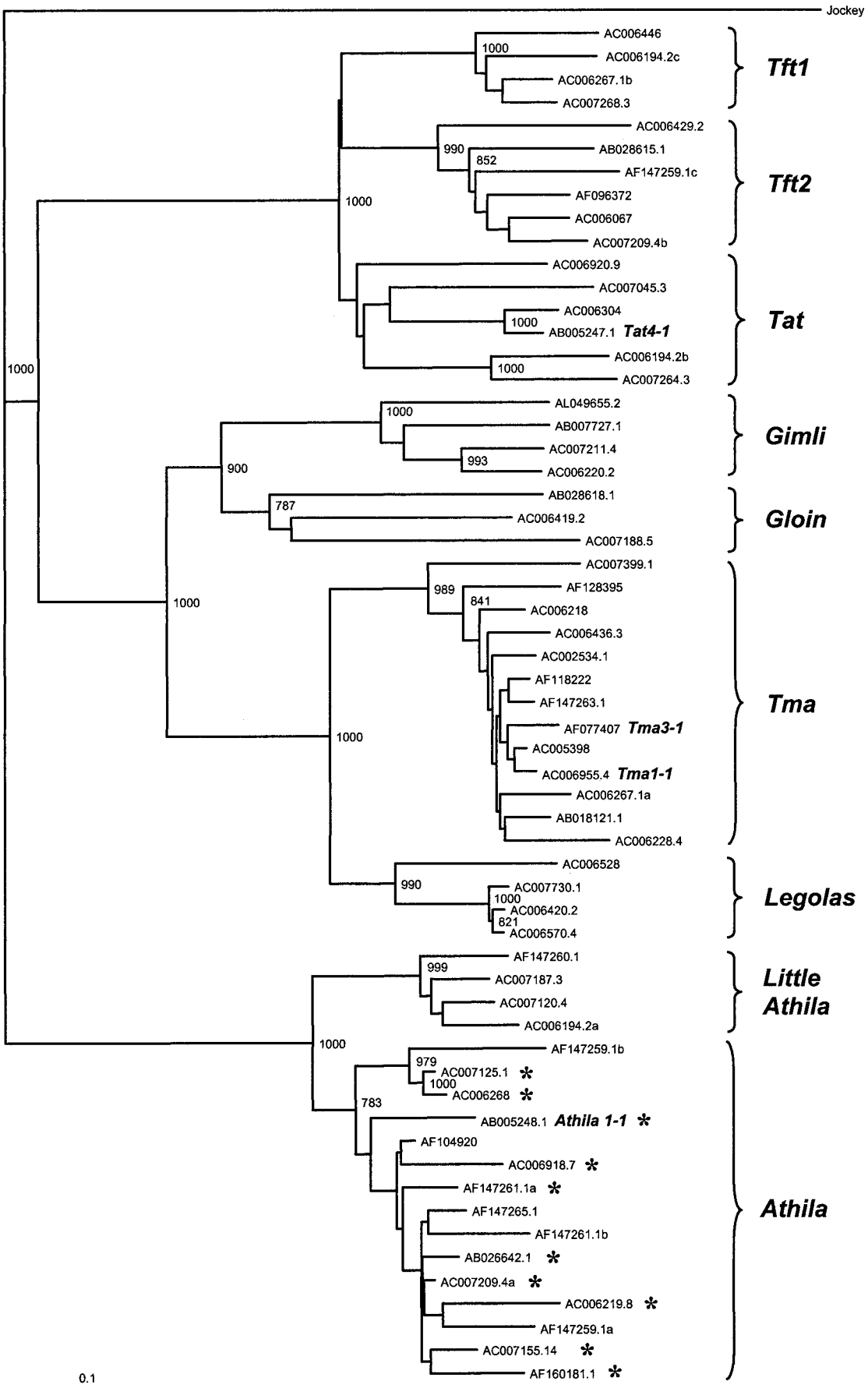
Methods to obtain the multiple alignments, phylogenetic trees (based on the neighbor-joining algorithm; Saitou and Nei 1987), and bootstrapping values for the branches were those described previously (Marín et al. 1998), except that in this work the program ClustalX (Thompson et al. 1997) was used instead of CLUSTAL W (Thompson, Higgins, and Gibson 1994). The program GeneDoc (Nicholas and Nicholas 1997) was used to edit the sequences for manual refinement of the multiple alignments. The multiple alignments on which the phylogenetic trees shown in figures 1 and 3 are based are available at the EMBL European Bioinformatics Institute web pages (ftp://ftp.ebi.ac.uk/pub/databases/embl/align), with accession numbers DS41404 and DS41405, respectively. In order to establish the structure of the *Arabidopsis thaliana* elements that we describe for the first time in this study, we used (1) the local alignment program LALIGN (designed by W. R. Pearson, based on Huang and Miller [1991]; online at the Genestream Network Server, http://xylian.igh.cnrs.fr/) to determine the lengths and sizes of the LTRs; (2) TBLASTN to establish the relationships among the different copies of a same element; (3) the ORF Finder tool, also implemented at NCBI, to determine whether the ORFs of the elements were truncated or intact; and (4) BLASTP and TBLASTN to determine, by comparison with previously described elements, the positions of the different protein domains presented in figure 2. Phylogenetic trees were obtained using TreeView (Page 1996).

Results

Sequences of at Least Nine Different *Ty3*/*Gypsy* Lineages are Present in the *Arabidopsis thaliana* Genome

The complexity and novelty of our findings regarding *A. thaliana Ty3*/*Gypsy* retrotransposons deserves a full section. In figure 1, we present the phylogenetic tree obtained using 59 different RT domain sequences of *Ty3*/*Gypsy* elements in *A. thaliana.* The sequence of the non-LTR retrotransposon *Jockey* was used as the outgroup. From now on, we will use an operative definition of ''lineage'' as a group of retrotransposon sequences found in a particular species that are identical or very similar. In general, it is found that copies belonging to the same lineage have structural features that distinguish them from those of other lineages, particularly LTRs with characteristic sizes and sequences. Using this definition, we found that so far only three *Ty3*/*Gypsy* lineages had been characterized for *A. thaliana.* Those three lineages were, respectively, called *Athila, Tma,* and *Tat* (Peleman et al. 1991; Pelissier et al. 1995; Wright and Voytas 1998). However, as can be seen in figure 1, although several copies belonging to those three lineages are present in the databases, they do not comprise the whole range of variation of *A. thaliana Ty3*/*Gypsy* retrotransposons. On the contrary, at least six other lineages, highly supported by bootstrap data and represented by multiple copies, are present. Structural data confirm that at least nine different lineages are present in *A. thaliana.* In particular, element length and LTR lengths and sequences are different for members of the nine main branches found in our tree (see below). A third ORF was found in most *Athila* sequences (asterisks in fig. 1) but is absent in all the other elements of this species.

We now summarize the results of the analyses for the nine *A. thaliana* lineages. In particular, for the new lineages defined in figure 1, the copies structurally most similar to active elements are described.


Jockey
AC006446
1000
AC006194.2c
AC006267.1b
AC007268.3
Tft1
AC006429.2
990
AB028615.1
852
AF147259.1c
1000
AF096372
AC006067
AC007209.4b
Tft2
AC006920.9
AC007045.3
AC006304
1000
AB005247.1 Tat4-1
1000
AC006194.2b
1000
AC007264.3
Tat
AL049655.2
1000
AB007727.1
AC007211.4
993
AC006220.2
900
Gimli
AB028618.1
787
AC006419.2
AC007188.5
Gloin
AC007399.1
989
AF128395
841
AC006218
1000
AC006436.3
AC002534.1
AF118222
AF147263.1
AF077407 Tma3-1
1000
AC005398
AC006955.4 Tma1-1
AC006267.1a
AB018121.1
AC006228.4
Tma
AC006528
990
AC007730.1
1000
AC006420.2
821
AC006570.4
Legolas
AF147260.1
999
AC007187.3
AC007120.4
AC006194.2a
Little Athila
1000
AF147259.1b
979
AC007125.1 *
1000
AC006268 *
783
AB005248.1 Athila 1-1 *
AF104920
AC006918.7 *
AF147261.1a *
AF147265.1
AF147261.1b
AB026642.1 *
AC007209.4a *
AC006219.8 *
AF147259.1a
AC007155.14 *
AF160181.1 *
Athila
0.1

Tit-for-Tat1 and *Tit-for-Tat2* (*Tft1* and *Tft2*) are represented by four and six sequences, respectively. The name is inspired by the fact that they are close relatives of *Tat.* They form two well-supported monophyletic groups according to their RT domain sequences (fig. 1), but it is unclear at present whether active *Tft1* or *Tft2* elements exist in *A. thaliana.* None of the *Tft1* copies available seem to be functional. The structurally best conserved element is the one found in the sequence with accession number AC006194.2. The element contained in this sequence is 8,449 bp long (nucleotides 20655–12207) and has 96% identical, 1,327-bp-long LTRs. These LTRs are totally unrelated to those in *Tat.* The putative ORFs of this copy contain several frameshifts and many stop codons (fig. 2).

None of the six *Tft2* sequences are intact enough to precisely describe the structure of the element, particularly their LTRs. It is thus possible that *Tft2* is actually a particular *Tft1*-like lineage, inactivated long ago. The *Tft2* elements would be at least 6,600 bp long. We used the RT of the element present in sequence AC006067 as the canonical copy for the analyses presented in figure 3 (see below).

We conservatively kept the name *Tat* (Peleman et al. 1991) for all of the RT domain sequences that are closely related to that of the canonical copy *Tat4-1* (Wright and Voytas 1998) and are, at the same time, excluded from the *Tft* lineages described above (fig. 1). However, considering that the topology of the phylogenetic tree is supported by low bootstrap values, this group might be artificial. It is unclear whether those five sequences correspond to one or to several lineages. The *Tat4-1* copy described by Wright and Voytas (1998) is structurally very different from the *Tft1* copy described above. First, it is substantially longer, spanning around 12 kb. Moreover, it has a long 3′ noncoding region that is not present in *Tft1.* Finally, it has shorter LTRs, about 400 bp long.

Gimli and *Gloin* (named after Tolkien 1954) are the smallest *Ty3*/*Gypsy* elements found so far in *Arabidopsis.* They are closely related and represented by four and three sequences, respectively (fig. 1). The *Gimli* element found in sequence AL049655.2 may be active. It is 5,221 bp long (nucleotides 77110–82330 in AL049655.2) with 341-bp-long LTRs that are 98% identical. These LTRs are unrelated to those of other elements, including *Gloin.* Two contiguous ORFs of 868 and 593 amino acids are apparent in this sequence. However, both ORFs are in the same frame, and thus it is possible that they both actually form a single continuous ORF about 1,500 amino acids long. However, such a single ORF would contain a stop codon between the RT and RH domains (fig. 2). At the end of the integrase, a chromodomain is observed (detailed in fig. 4). Isolated *Gimli* LTRs have not been found.

For canonical *Gloin* element, we took the one found in sequence AC007188.5. It is 5,409 bp long, spanning nucleotides 22525–27933 in that sequence. It has 359-bp-long LTRs that are 95% identical and unrelated to those in other elements. No similar isolated LTRs are apparent in the databases. In this copy, two frameshifts, affecting the RT and IN domains, are observed. This last domain also contains a stop codon (fig. 2). It is therefore unlikely that it corresponds to an active element. It also has a chromodomain-containing integrase (fig. 4).

The name *Tma* defines a clearly monophyletic group of similar sequences, including the canonical copies *Tma1-1* and *Tma3-1* (Wright and Voytas 1998). Wright and Voytas (1998) showed that *Tma* elements are about 7.8 kb long and have LTRs that are 1.15 kb long. We found that these elements also have a chromodomain-containing integrase (fig. 4).

Legolas (name also from Tolkien 1954) is a relative of *Tma* (fig. 1). The copy found in AC006570.4 has the typical structure of an active *Ty3*/*Gypsy* retrotransposon (fig. 2). The element is 7,740 bp long (nucleotides 45994–38255 in AC006570.4), with 1,347-bp-long LTRs that are 99% identical. It has two ORFs of 498 and 1,215 amino acids, respectively. At the end of the first ORF, homology to the *gag* proteins is found, including three C_2HC zinc fingers. The second ORF con-

←

FIG. 1.—Phylogenetic tree showing the relationships among the *Ty3*/*Gypsy* RT domain sequences found in the *Arabidopsis thaliana* genome. The neighbor-joining method (Saitou and Nei 1987) was used to obtain this tree and the one presented in figure 3. For simplicity, bootstrap values lower than 750 are not shown. Names refer to the accession numbers of the nucleotide sequences that encode the corresponding reverse transcriptases (RTs). When they correspond to previously described elements, the names of these elements are detailed. Where more than one retroelement was present in the same sequence, letters have been added at the end of the accession number to distinguish them. Asterisks refer to the presence of an ORF3 in some *Athila* elements. From top to bottom in the figure, locations of the RT domain sequences are as follows (the two numbers refer to the N- and C-termini of the RT domain in the corresponding nucleotide sequence): *Tft1* sequences—AC006446, nucleotides 7755–8297; AC006194.2c, 16307–15765; AC006267.1b, 90577–91119; AC007268.3, 67972–68514. *Tft2* sequences—AC006429.2, 61097–60558; AB028615.1, 52508–53050; AF147259.1c, 95640–95098; AF096372, 67703–67161; AC006067, 122537–121997; AC007209.4b, 25614–25072. *Tat* sequences—AC006920.9, 37751–37209; AC007045.3, 35604–35551; AC006304, 9709–10251; AB005247.1 (*Tat4-1*), 58721–59263; AC006194.2b, 24421–23879; AC007264.3, 54714–55256. *Gimli* sequences—AL049655.2, 79337–79876; AB007727.1, 16108–15569; AC007211.4, 65405–65944; AC006220.2, 18939–19478. *Gloin* sequences—AB028618.1, 69729–70268; AC006419.2, 108155–107616; AC007188.5, 24908–25447. *Tma* sequences—AC007399.1, 24339–23800; AF128395, 87950–87411; AC006218, 35783–36322; AC006436.3, 93107–92568; AC002534.1, 23122–22583; AF118222, 40305–40844; AF147263.1, 24870–25409; AF077407 (*Tma3-1*), 29025–28564; AC005398, 46868–47407; AC006955.4 (*Tma1-1*), 94814–95353; AC006267.1a, 30273–30812; AB018121.1, 7082–6543; AC006228.4, 73188–72649. *Legolas* sequences—AC006528, 46313–45774; AC007730.1, 32531–33070; AC006420.2, 9046–8507; AC006570.4, 42366–41827. *Little Athila* sequences—AF147260.1, 49216–49812; AC007187.3, 64875–64279; AC007120.4, 6207–5611; AC006194.2a, 2353–1784. *Athila* sequences—AF147259.1b, 38293–38889; AC007125.1, 99805–99209; AC006268, 91998–92594; AB005248.1 (*Athila1-1*), 28901–29497; AF104920, 7865–7269; AC006918.7, 71212–70616; AF147261.1a, 35131–34535; AF147265.1, 25161–24565; AF147261.1b, 45866–46462; AB026642.1, 12470–11874; AC007209.4a, 17510–16914; AC006219.8, 77740–78336; AF147259.1a, 22996–22400; AC007155.14, 60937–60341; AF160181.1, 23554–24150.

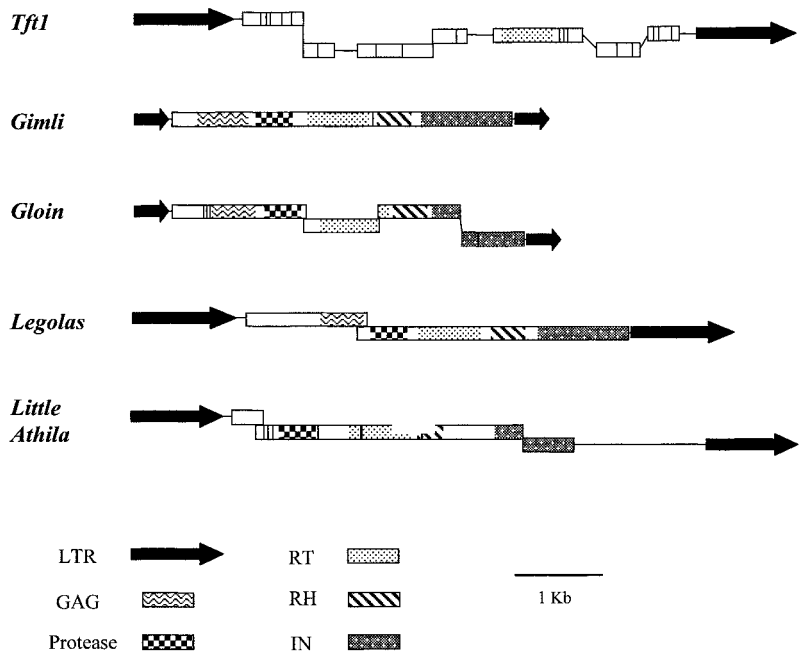


FIG. 2.—Structure of canonical copies. The five lineages that we describe for the first time and for which complete copies have been obtained are represented (*Tft2* full-length elements have not yet been found; see *Results*). Boxes indicate putative open reading frames. Vertical lines correspond to stop codons. Offset boxes show changes in reading frame. The positions of *gag* and *pol* protein domains are indicated, with the exception of the *Tft1* element, for which multiple frameshifts and stop codons affect the putative coding regions, and only the location of the intact reverse transcriptase (RT) domain is represented. The positions of the protein domains have been delimited by comparison with the coding regions of other *Ty3*/*Gypsy* elements.

tains the *pol* domains. Its integrase also has a chromodomain (fig. 4). The structure of *Legolas* is very similar to that of the *Tma* elements described above.

Little Athila (new element) is represented by four RT domain sequences, as shown in figure 1. The name refers to the fact that this element is closely related to *Athila* but shorter, lacking an ORF3. It is unclear whether active *Little Athila* elements exist. For the canonical copy, we took the one present in AF147260.1. This copy is 8,658 bp long (nucleotides 46089–54746 in AF147260.1) with asymmetrical LTRs (1,133 and 1,148 bp long, respectively; 92% identical). This is an inactive copy. A frameshift in the IN domain and stop codons in the RT and RH domains were observed (fig. 2). Moreover, its RH domain is C-terminally truncated. Isolated LTRs almost identical to the ones present in this copy are found in the databases (not shown). They are totally unrelated to *Athila* LTRs.

We have reserved the name *Athila* (Pelissier et al. 1995; Wright and Voytas 1998) for the very large group of elements, many of them with *env*-containing ORF3, that are included in an apparently monophyletic branch (fig. 1). However, whether one or several closely related *Athila*-like elements are present deserves further study, because these sequences are substantially heterogeneous. Together with the previously characterized copy *Athila1-1* (Wright and Voytas 1998), we included in our analyses (see fig. 3) a copy that is central in the tree, located in sequence AC007209.4 (nucleotides 8571–22061), which we called *Athila4.*

Plant *Ty3*/*Gypsy*Elements Can Be Grouped into Two Classes, but at Least Four Lineages of These Elements Predate the Monocot/Dicot Split

Figure 3 shows the phylogenetic tree obtained using all of the sequences detected in our exhaustive database screenings. Only the canonical representatives of the *A. thaliana* lineages described in the previous section were included. We will now summarize the main results deduced from this tree for plant elements. In the next section, we describe some new findings for other species.

The tree presented in figure 3 is very similar to those found, with a more limited set of elements, by Wright and Voytas (1998) and Malik and Eickbush (1999). As first described by Wright and Voytas (1998), all plant elements fall into two branches, forming what we call classes A and B. Class B corresponds to elements with chromodomain-containing integrases (Malik and Eickbush 1999; see also fig. 4). Class B elements are part of the only group in the whole tree that has a phylogenetic range spanning fungi, animals, and plants and at the same time is well-supported by bootstrap data. The precise relationship of class A plant elements with animal or fungal elements is unclear. The group formed by the animal *Mag* and *SURL* elements is the closest one to class A elements, but the bootstrap values supporting this association are low (277 out of 1,000 replicates).

A major result for plant elements is that there are several branches of the tree shown in figure 3 that com-

prise elements in both monocot and dicot species. As can be seen in figure 3, class A is divided into two groups. One of the groups (ancestral lineage I) is itself divided into two subgroups, containing, respectively, elements of monocot species (*Grande* from *Zea diploperennis*, *Cinful* from *Zea mays*, and *Retrosor1* from *Sorghum bicolor*) and of the dicot *A. thaliana* (*Tat*, *Tft1*, and *Tft2*). Elements of a second group, formed by the retrotransposons *Diaspora* (*Glycine max*), *Cyclops1* (*Pisum sativum*), and *Cyclops-Vicia* (*Vicia faba*) and the *A. thaliana* elements *Athila* and *Little Athila* have been found so far only in dicot species. Assuming no horizontal transmission, these results suggest that at least one class A lineage was already present before the monocot-dicot split.

Class B elements have a more complex history. Data suggest that at least three different lineages were present before the monocot-dicot split (fig. 3). One of these lineages (ancestral lineage II) comprises 11 elements, some from monocot species (*Lilium henryi* element *Del1*; *S. bicolor* elements *Retrosor2* and *Retrosor3*; *Oryza sativa* elements *Retrosat2* and *RIRE3* and *Retrosor-2*-like sequence found in AP000364.1; *Ananas comosus Dea1*) and some from dicots (*A. thaliana* elements *Legolas* and *Tma*; *P. sativum* element *Peabody*). In addition to such large group, there are two other class B ancestral lineages (III and IV in fig. 3), with two and four elements, respectively, that seem to correspond to other types of *Ty3*/*Gypsy* elements existent before the monocot-dicot split. One branch contains the new element *Galadriel* from the dicot *Lycopersicon esculentum* (also named after Tolkien [1954] by J. Jones [John Innes Centre, Norwich, England] and the authors) and *Monkey* from the monocot *Musa acuminata*. The second comprises element *Reina* (from the monocot *Z. mays*), the *A. thaliana* (dicot) elements *Gimli* and *Gloin* and, most interestingly, the only element so far found in conifers, the *Ifg7* element from *Pinus radiata*. The close relationship of *Ifg7* and the *A. thaliana* retrotransposons *Gimli* and *Gloin* suggests that elements of this branch existed before the *Coniferophyta*/*Magnoliophyta* separation.

Therefore, assuming no horizontal transmission, our data support the hypothesis that at least four lineages of *Ty3*/*Gypsy* elements coexisted in the last common ancestor of monocots and dicots, roughly 200 MYA. In fact, considering the topology we have obtained and that information is still fragmentary, especially for monocot species, the number of lineages in this ancestor may well have been much larger than four.

Other Interesting New Elements and Comparisons with Previous Evolutionary Studies

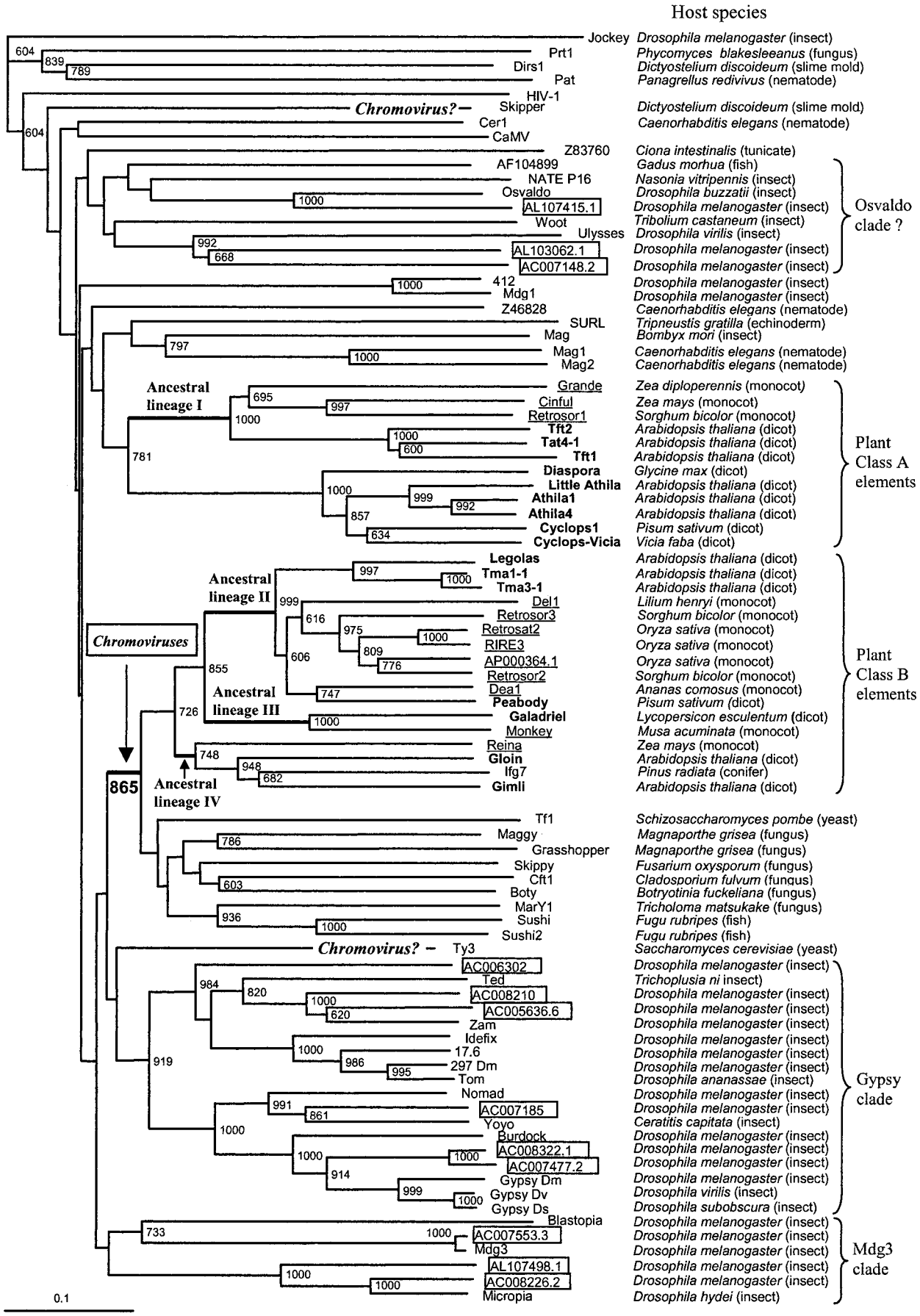
Apart from the *A. thaliana* elements described above, several of the sequences presented in figure 3 are analyzed from a phylogenetic point of view for the first time in this study. The most interesting additions are concentrated in a particular branch of the tree that corresponds to what Malik and Eickbush (1999) called the *Osvaldo* clade. So far, only three insect elements (*Drosophila buzzatii Osvaldo*, *Drosophila virilis Ulysses*, and *Tribolium castaneum Woot*) were included in this clade. However, the *D. melanogaster* genome project has unearthed sequences that are closely related to both *Osvaldo* and *Ulysses*, two elements discovered in species of the *Drosophila* genus but never before detected in *D. melanogaster*. Moreover, our data suggest that there may be representatives of the *Osvaldo* clade in many other species, perhaps even including chordates (the fish *Gadus morhua* contains an RT sequence closely related to that of *Osvaldo*; see fig. 3). The rest of the *D. melanogaster* sequences are close relatives of previously known elements that belong to the *Gypsy* and *Mdg3* clades defined by Malik and Eickbush (1999) (fig. 3).

An important conclusion that derives from our data is that the number of different *Ty3*/*Gypsy* lineages is much larger for some species than for others. In *D. melanogaster* this number is very high, around 20. We reported in this study that *A. thaliana* contains at least nine *Ty3*/*Gypsy* lineages. We also characterized four in *C. elegans* (see fig. 3), and, in a study published after our analyses were finished, Bowen and McDonald (1999) detected sequences that may correspond to two other lineages, closely related to the *Mag*-like and Z46828 sequences, respectively. Finally, *S. cerevisiae* has only one lineage, namely, *Ty3*.

Two differences of our tree from those of previous studies concern the positions of the elements *Ty3* and, particularly, *Skipper*. *Ty3* is included by Malik and Eickbush (1999) together with the chromodomain-containing elements to define the *Ty3* clade. Our results, as well as those by Wright and Voytas (1998), situate *Ty3* outside of the group formed by the chromodomain-containing elements. More pronounced is the difference that concerns the *Dictyostelium discoideum* element *Skipper* (Leng et al. 1998). This is an element with chromodomain-containing integrase and, in Malik and Eickbush's (1999) study using sequences of the RT, RH, and IN domains together, appears closely related to the other chromodomain-containing elements. However, RT domain sequences alone situate *Skipper* outside of the tree formed by the other *Ty3*/*Gypsy* elements.

Discussion

Information provided by genome sequencing projects is qualitatively changing our understanding of the evolution of eukaryotic mobile elements by avoiding the limitations imposed by nucleic acid hybridization or PCR techniques to detect related sequences. Several authors have taken advantage of the completion of the *S. cerevisiae* and *C. elegans* genome sequencing projects to perform comprehensive studies of particular groups of mobile elements (Oosumi, Garlick, and Belknap 1996; Kim et al. 1998; Malik and Eickbush 1998; Marín et al. 1998; Wright and Voytas 1998; Bowen and McDonald 1999; Jordan and McDonald 1999*a*, 1999*b*). In this work, we performed a similar analysis for *Ty3*/*Gypsy* elements in multiple organisms. Some of the most interesting results concern plant species. Our study establishes two important general conclusions. First, the model species *A. thaliana* has many different types of


Host species
Jockey Drosophila melanogaster (insect)
Prt1 Phycomyces blakesleeanus (fungus)
Dirs1 Dictyostelium discoideum (slime mold)
Pat Panagrellus redivivus (nematode)
HIV-1
Chromovirus? Skipper Dictyostelium discoideum (slime mold)
Cer1 Caenorhabditis elegans (nematode)
CaMV
Z83760 Ciona intestinalis (tunicate)
AF104899 Gadus morhua (fish)
NATE P16 Nasonia vitripennis (insect)
Osvaldo Drosophila buzzatii (insect)
AL107415.1 Drosophila melanogaster (insect)
Woot Tribolium castaneum (insect)
Ulysses Drosophila virilis (insect)
AL103062.1 Drosophila melanogaster (insect)
AC007148.2 Drosophila melanogaster (insect)
Osvaldo clade ?
412 Drosophila melanogaster (insect)
Mdg1 Drosophila melanogaster (insect)
Z46828 Caenorhabditis elegans (nematode)
SURL Tripneustis gratilla (echinoderm)
Mag Bombyx mori (insect)
Mag1 Caenorhabditis elegans (nematode)
Mag2 Caenorhabditis elegans (nematode)
Ancestral lineage I
Grande Zea diploperennis (monocot)
Cinful Zea mays (monocot)
Retrosor1 Sorghum bicolor (monocot)
Tft2 Arabidopsis thaliana (dicot)
Tat4-1 Arabidopsis thaliana (dicot)
Tft1 Arabidopsis thaliana (dicot)
Diaspora Glycine max (dicot)
Little Athila Arabidopsis thaliana (dicot)
Athila1 Arabidopsis thaliana (dicot)
Athila4 Arabidopsis thaliana (dicot)
Cyclops1 Pisum sativum (dicot)
Cyclops-Vicia Vicia faba (dicot)
Plant Class A elements
Legolas Arabidopsis thaliana (dicot)
Ancestral lineage II
Tma1-1 Arabidopsis thaliana (dicot)
Tma3-1 Arabidopsis thaliana (dicot)
Del1 Lilium henryi (monocot)
Retrosor3 Sorghum bicolor (monocot)
Chromoviruses
Retrosat2 Oryza sativa (monocot)
RIRE3 Oryza sativa (monocot)
AP000364.1 Oryza sativa (monocot)
Retrosor2 Sorghum bicolor (monocot)
Ancestral lineage III
Dea1 Ananas comosus (monocot)
Peabody Pisum sativum (dicot)
Galadriel Lycopersicon esculentum (dicot)
Monkey Musa acuminata (monocot)
Reina Zea mays (monocot)
Gloin Arabidopsis thaliana (dicot)
Ifg7 Pinus radiata (conifer)
865
Ancestral lineage IV
Gimli Arabidopsis thaliana (dicot)
Plant Class B elements
Tf1 Schizosaccharomyces pombe (yeast)
Maggy Magnaporthe grisea (fungus)
Grasshopper Magnaporthe grisea (fungus)
Skippy Fusarium oxysporum (fungus)
Cft1 Cladosporium fulvum (fungus)
Boty Botryotinia fuckeliana (fungus)
MarY1 Tricholoma matsukake (fungus)
Sushi Fugu rubripes (fish)
Sushi2 Fugu rubripes (fish)
Chromovirus? Ty3 Saccharomyces cerevisiae (yeast)
AC006302 Drosophila melanogaster (insect)
Ted Trichoplusia ni insect)
AC008210 Drosophila melanogaster (insect)
AC005636.6 Drosophila melanogaster (insect)
Zam Drosophila melanogaster (insect)
Idefix Drosophila melanogaster (insect)
17.6 Drosophila melanogaster (insect)
297 Dm Drosophila melanogaster (insect)
Tom Drosophila ananassae (insect)
Nomad Drosophila melanogaster (insect)
AC007185 Drosophila melanogaster (insect)
Yoyo Ceratitis capitata (insect)
Burdock Drosophila melanogaster (insect)
AC008322.1 Drosophila melanogaster (insect)
AC007477.2 Drosophila melanogaster (insect)
Gypsy Dm Drosophila melanogaster (insect)
Gypsy Dv Drosophila virilis (insect)
Gypsy Ds Drosophila subobscura (insect)
Gypsy clade
Blastopia Drosophila melanogaster (insect)
AC007553.3 Drosophila melanogaster (insect)
Mdg3 Drosophila melanogaster (insect)
AL107498.1 Drosophila melanogaster (insect)
AC008226.2 Drosophila melanogaster (insect)
Micropia Drosophila hydei (insect)
Mdg3 clade
0.1

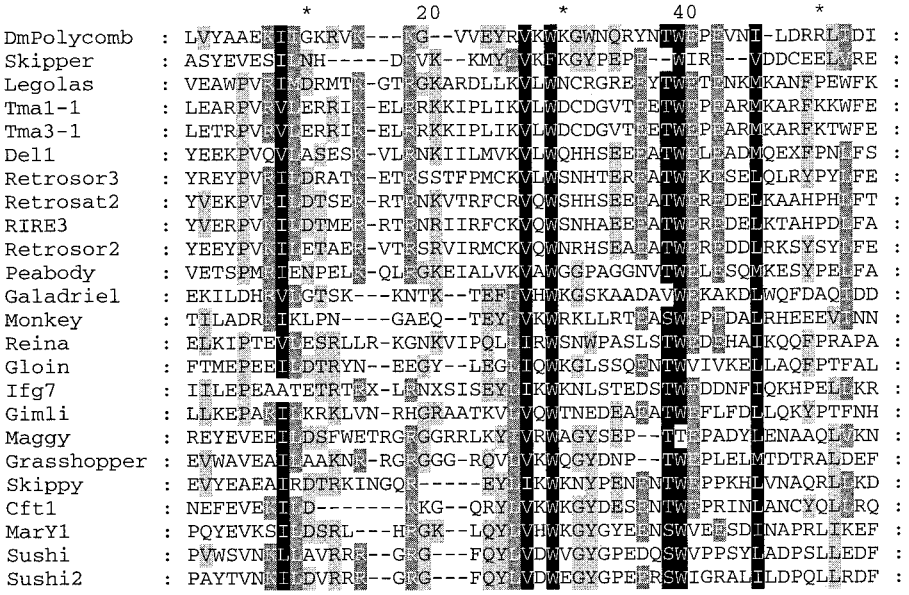


FIG. 4.—Multiple alignment of the chromodomains found at the C-terminal end of the integrases of several *Ty3*/*Gypsy* retrotransposons. The chromodomain of the *Polycomb* gene of *Drosophila melanogaster* is shown for comparison. All of the ''chromoviruses'' in figure 3 contain a chromodomain, with the exception of four sequences (*Oryza sativa* AP000364.1; *Ananas comosus Dea1*; *Schizosaccharomyces pombe Tf1*; *Botryotinia fuckeliana Boty*) in which this domain is not found.

Ty3/*Gypsy* retrotransposons. A second important conclusion of our study is that, assuming horizontal transmission has been very infrequent or absent, the origin of several independent lineages of plant *Ty3*/*Gypsy* retrotransposons can be traced back to before the monocot-dicot split, about 200 MYA. We cannot rule out horizontal transfer, but support for such process requires finding very similar elements in distant species (e.g., Daniels et al. 1990; Lohe et al. 1995; Robertson and Lampe 1995; Gonzalez and Lessios 1999; Jordan, Matyunina, and McDonald 1999), something that so far is not evident when elements from dicots and monocots are compared (see the topology and lengths of the branches in fig. 3). Therefore, our data suggest that the genome of the last common ancestor of monocots and dicots was, from the point of view of *Ty3*/*Gypsy* elements, quite similar to that of *A. thaliana,* containing several active, distantly related elements. The presence in other monocot (e.g., *S. bicolor*, *Z. mays*) and dicot (e.g., *P. sativum*) species of elements belonging to the two main classes of plant *Ty3*/*Gypsy* elements (fig. 3) suggests that the existence of multiple lineages of these retrotransposons is a general feature of modern angiosperms. The complexity of plant element evolution deserves further study. In particular, none of the elements of one of the lineages that we have described (that including *Grande*, *Tat* and their relatives) was considered by Malik and Eickbush (1999), so they may form an independent, ninth clade of *Ty3*/*Gypsy* retrotransposons.

The *A. thaliana* results parallel those found for the *D. melanogaster* genome, in which also many different, perhaps 20, types of *Ty3*/*Gypsy* retrotransposons coexist (fig. 3). However, in other genomes, there are limited numbers of successful *Ty3*/*Gypsy* lineages. In particular, in the completely sequenced genome of *S. cerevisiae*, only one lineage (*Ty3*) has been found. *Ty3*/*Gypsy* elements seem to also be rare in vertebrates. Figure 3 shows two independent lineages in fishes (*Sushi* from the pufferfish *Fugu rubripes,* and the *Osvaldo*-like sequence in *G. morhua*). *Sushi*-related sequences are also present in amphibians and reptiles, and a third lineage might exist also in fishes (Miller et al. 1999). However, in spite of abundant available data, none has been found so far in mammalian genomes. Considering the relatively small sizes of the genomes of *A. thaliana* and *D. melanogaster* with respect to vertebrates, we can rule

←

FIG. 3.—Phylogenetic tree obtained using the reverse transcriptase (RT) domain sequences of all *Ty3*/*Gypsy* retrotransposons. Only the canonical copies of the *Arabidopsis thaliana* elements described in our study are included. Several outgroups are also included (see MATERIALS AND METHODS). All bootstrap values over 600 are detailed. We highlight the significant value for the branch that groups chromodomain-containing elements and their relatives (''chromoviruses''; see *Discussion*). Notice the locations in the tree of two other putative chromoviruses (*Skipper* and *Ty3*) that are outside of this branch. The plant element classes A and B, as well as those branches that correspond to ancestral lineages (I–IV) predating the monocot/dicot split, are also indicated. Elements from monocot species are underlined and those from dicot species are in bold letters. New *Drosophila melanogaster* sequences are boxed, and it is shown that they are all included in three of the clades (*Osvaldo*, *Gypsy*, and *Mdg3*) defined by Malik and Eickbush (1999). The question mark added to the *Osvaldo* clade reflects the fact that it is still formally undetermined whether some of the sequences that we show together belong to this clade or not. The accession numbers for new sequences are included. A complete list of the accession numbers, along with the multiple alignment from which this figure was obtained, is available at the EMBL European Bioinformatics Institute web pages (see *Materials and Methods*).

out the explanation that the success of *Ty3*/*Gypsy* elements is correlated with large genome sizes. It is more likely that genome-specific factors determine the success or failure of these retrotransposons (see Labrador and Corces 1997).

The presence of a third ORF putatively encoding an ENV protein in distantly related elements (so far, *Gypsy* and some of its close relatives, *Cer1*, *Athila*, and *Osvaldo*) that belong to four different clades of the *Ty3*/*Gypsy* group (Malik and Eickbush 1999) can be explained by losses of the *env* gene in many different lineages or by recent independent acquisitions of an ORF3 by certain elements. In favor of this latter hypothesis, an important result is the finding of an ORF3 in a member of the *Ty1*/*Copia* group (Laten, Majumdar, and Gaucher 1998) that moreover encodes a protein lacking certain common motifs present in the ENV protein or retroviruses or *Gypsy* and its relatives (Lerat and Capy 1999). Bowen and McDonald (1999) also detected a third ORF in a *C. elegans* LTR retrotransposon (*Cer7*) that belongs to a group of elements phylogenetically situated between the *Ty1*/*Copia* and *Ty3*/*Gypsy* groups. If multiple independent acquisitions of an ORF3 have occurred, the classification of the Metaviridae, based on the *Errantivirus* (with *env*)—*Metavirus* (without *env*) dichotomy, should be reconsidered. On the other hand, chromodomain-containing elements form the only group of *Ty3*/*Gypsy* elements that is well supported by two independent lines of evidence (RT-based phylogeny: see fig. 3; presence of chromodomain in their integrases: see fig. 4) and whose origin seems to predate the plant/fungal/animal split. Although horizontal transmission has been proposed to explain the wide phylogenetic range of these elements (Poulter and Butler 1998; Miller et al. 1999), the negative evidence invoked (i.e., absence of these elements in invertebrates) is hardly compelling. The simplest hypothesis is that the acquisition of a chromodomain was a very old, unique event and that elements with this characteristic have been vertically transmitted, being eventually lost in some lineages. Therefore, we propose giving genus status to the monophyletic group formed by chromodomain-containing elements and their closest relatives (that lack this domain as a result of a secondary loss). We propose the name *Chromovirus* for this genus.

Whether it is convenient to classify the elements *Ty3* and *Skipper* as chromoviruses is unclear. Malik and Eickbushs (1999) analyses placed *Ty3* and *Skipper* together with the chromodomain-containing elements. However, *Ty3* lacks a chromodomain, and RT sequences alone (Wright and Voytas 1998 and this study) do not support this association. On the other hand, *Skipper* has a chromodomain, but our RT-based analysis places it outside of the tree formed by the rest of *Ty3*/*Gypsy* elements (fig. 3). The contradictory results for *Skipper* when the RT domain alone are considered versus those when several *pol* protein domains are considered can be explained in three ways: (1) *Skipper* has a recombinant origin: The RT sequences would derive from certain highly divergent elements, while the IN domain would come from a typical chromodomain-containing element (see Jordan and McDonald 1999*a* for an example of such a of process in retrotransposons); (2) *Skipper* RT sequences are evolving at an abnormally fast rate; and (3) *Skipper* integrase has acquired a chromodomain independently of the other elements. Thus, *Ty3* and *Skipper* may provisionally be considered chromoviruses, following Malik and Eickbush's (1999) results, but this classification should be reconsidered when more data are obtained.

Acknowledgments

We thank Mariano Labrador for comments on a previous version of this paper. Our manuscript was greatly improved by the comments of Thomas Eickbush and two anonymous reviewers.

LITERATURE CITED

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

BOWEN, N. J., and J. F. MCDONALD. 1999. Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. Genome Res. **9**:924–935.

DANIELS, S. B., K. R. PETERSON, L. D. STRAUSBAUCH, M. G. KIDWELL, and A. CHOVNICK. 1990. Evidence for horizontal transmission of the P transposable element between *Drosophila* species. Genetics **124**:339–355.

EICKBUSH, T. H. 1994. Origin and evolutionary relationships of retroelements. Pp. 121–157 *in* S. S. MORSE, ed. The evolutionary biology of viruses. Raven Press, New York.

GONZALEZ, P., and H. A. LESSIOS. 1999. Evolution of sea urchin retroviral-like (SURL) elements: evidence from 40 echinoid species. Mol. Biol. Evol. **16**:938–952.

HUANG, X., and W. MILLER. 1991. A time-efficient, linear-space local similarity algorithm. Adv. Appl. Math. **12**:337–357.

HULL, R. 1999. Classification of reverse transcribing elements: a discussion document. Arch. Virol. **144**:209–214.

JORDAN, I. K., and J. F. MCDONALD. 1999*a*. Phylogenetic perspective reveals abundant *Ty1*/*Ty2* hybrid elements in the *Saccharomyces cerevisiae* genome. Mol. Biol. Evol. **16**: 419–422.

———. 1999*b*. Tempo and mode of *Ty* element evolution in *Saccharomyces cerevisiae*. Genetics **151**:1341–1351.

JORDAN, I. K., L. V. MATYUNINA, and J. F. MCDONALD. 1999. Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. Proc. Natl. Acad. Sci. USA **96**: 12621–12625.

KIM, A., C. TERZIAN, P. SANTAMARIA, A. PELISSON, N. PURD'HOMME, and A. BUCHETON. 1994. Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **91**:1285–1289.

KIM, J. M., S. VANGURI, J. D. BOEKE, A. GABRIEL, and D. F. VOYTAS. 1998. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res. **8**:464–478.

LABRADOR, M., and V. G. CORCES. 1997. Transposable elements–host interactions: regulation of insertion and excision. Annu. Rev. Genet. **31**:381–404.

LATEN, H. M., A. MAJUMDAR, and E. A. GAUCHER. 1998. *SIRE-1*, a *copia/Ty1*-like retroelement from soybean, en-

codes a retroviral envelope-like protein. Proc. Natl. Acad. Sci. USA **95**:6897–6902.

LENG, P., D. H. KLATTE, G. SCHUMANN, J. D. BOEKE, and T. L. STECK. 1998. *Skipper,* an LTR retrotransposon of *Dictyostelium.* Nucleic Acids Res. **26**:2008–2015.

LERAT, E., and P. CAPY. 1999. Retrotransposons and retroviruses: analysis of the envelope gene. Mol. Biol. Evol. **16**: 1198–1207.

LOHE, A. R., E. N. MORIYAMA, D. A. LIDHOLM, and D. L. HARTL. 1995. Horizontal transmission, vertical inactivation, and stochastic loss of *mariner*-like transposable elements. Mol. Biol. Evol. **12**:62–72.

MALIK, H. S., and T. H. EICKBUSH. 1998. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol. Biol. Evol. **15**:1123–1134.

———. 1999. Modular evolution of the integrase domain in the *Ty3*/*Gypsy* class of LTR retrotransposons. J. Virol. **73**: 5186–5190.

MARÍN, I., P. PLATA-RENGIFO, M. LABRADOR, and A. FONTDEVILA. 1998. Evolutionary relationships among the members of an ancient class of non-LTR retrotransposons found in the nematode *Caenorhabditis elegans.* Mol. Biol. Evol. **15**:1390–1402.

MILLER, K., C. LYNCH, J. MARTIN, E. HERNIOU, and M. TRISTEM. 1999. Identification of multiple *Gypsy* LTR-retrotransposon lineages in vertebrate genomes. J. Mol. Evol. **49**: 358–366.

NICHOLAS, K. B., and H. B. NICHOLAS JR. 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the authors (www.cris.com/~ketchup/genedoc.shtml).

OOSUMI, T., B. GARLICK, and W. R. BELKNAP. 1996. Identification of putative nonautonomous transposable elements associated with several transposon families in *Caenorhabditis elegans.* J. Mol. Evol. **43**:11–18.

PAGE, R. D. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12**:357–358.

PANTAZIDIS, A., M. LABRADOR, and A. FONTDEVILA. 1999. The retrotransposon *Osvaldo* from *Drosophila buzzatii* displays all structural features of a functional retrovirus. Mol. Biol. Evol. **16**:909–921.

PELEMAN, J., B. COTTYN, W. VAN CAMP, M. VAN MONTAGU, and D. INZE. 1991. Transient occurrence of extrachromosomal DNA of an *Arabidopsis thaliana* transposon-like element, *Tat1*. Proc. Natl. Acad. Sci. USA **88**:3618–3622.

PELISSIER, T., S. TUTOIS, J. DERAGON, S. TOURMENTE, S. GENESTIER, and G. PICARD. 1995. *Athila,* a new retroelement from *Arabidopsis thaliana.* Plant Mol. Biol. **29**:441–452.

POULTER, R., and M. BUTLER. 1998. A retrotransposon family from the pufferfish (fugu) *Fugu rubripes.* Gene **215**:241–249.

PRINGLE, C. R. 1998. Virus taxonomy—San Diego 1998. Arch. Virol. **143**:1449–1459.

———. 1999. Virus taxonomy—1999. The universal system of virus taxonomy, updated to include the new proposals ratified by the International Committee on Taxonomy of Viruses during 1998. Arch. Virol. **144**:421–429.

ROBERTSON, H. M., and D. J. LAMPE. 1995. Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. Mol. Biol. Evol. **12**:850–862.

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

SONG, S. U., T. GERASIMOVA, M. KURKULOS, J. D. BOEKE, and V. G. CORCES. 1994. An *env*-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus. Genes Dev. **8**:2046–2057.

SPRINGER, M. S., and R. J. BRITTEN. 1993. Phylogenetic relationships of reverse transcriptase and Rnase H sequences and aspects of genome structure in the *Gypsy* group of retrotransposons. Mol. Biol. Evol. **10**:1370–1379.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN, and D. G. HIGGINS. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**:4876–4882.

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.

TOLKIEN, J. R. R. 1954. The lord of the rings. George Allen and Unwin, London.

WRIGHT, D. A., and D. F. VOYTAS. 1998. Potential retroviruses in plants: *Tat1* is related to a group of *Arabidopsis thaliana Ty3*/*Gypsy* retrotransposons that encode envelope-like proteins. Genetics **149**:703–715.

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

THOMAS H. EICKBUSH, reviewing editor

Accepted March 8, 2000