The Age and Evolution of Non-LTR Retrotransposable Elements

Harmit S. Malik, William D. Burke, and Thomas H. Eickbush

Department of Biology, University of Rochester

A comprehensive phylogenetic analysis was conducted of non-long-terminal-repeat (non-LTR) retrotransposons based on an extended sequence alignment of their reverse transcriptase (RT) domain. The 440 amino acid positions used included a region proposed to be similar to the ‘‘thumb’’ of the right-handed RT structure found in retroviruses. All identified non-LTR elements could be grouped into 11 distinct clades. Using the rates of sequence change derived from studies of the vertical inheritance of R1 and R2 elements in arthropods as a comparison, we found no evidence for the horizontal transmission of non-LTR elements. Assuming vertical descent, the phylogeny suggested that non-LTR elements are as old as eukaryotes, with each of the 11 clades dating back to the Precambrian era. The analysis enabled us to propose a simple chronology for the acquisition of different enzymatic domains in the evolution of the non-LTR class of retrotransposons. The first non-LTR elements were sequence specific by virtue of a restriction-enzyme-like endonuclease located downstream of the RT domain. Evolving from this original group were elements (eight clades) that acquired an apurinic-apyrimidic endonuclease-like domain upstream of the RT domain. Finally, four of these clades have inherited an RNase H domain downstream of the RT domain. The phylogenies of the AP endonuclease and RNase H domains were also determined for this report and are consistent with the monophyletic acquisition of these domains. These studies represent the most comprehensive effort to date to trace the evolution of a major class of transposable elements.

Introduction

Transposable elements have emerged as one of the principal driving forces behind the evolution of eukaryotic genomes (Charlesworth, Sniegowski, and Stephan 1994). One of the most abundant classes of transposable elements are the non-long-terminal-repeat (non-LTR) retrotransposable elements (also called LINE-like elements). These elements (hereafter referred to as non-LTR elements) insert by a simple mechanism in which an encoded endonuclease nicks (or cleaves) the chromosome, and the encoded reverse transcriptase utilizes this cut to prime reverse transcription of the cDNA copy directly onto the target site (Luan et al. 1993). The enzymatic machinery of non-LTR elements also appears to be utilized for the insertion of short interspersed nucleotide elements (SINEs) and processed pseudogenes—another large component of many eukaryotic genomes (Jurka 1997; Okada et al. 1977; Kazazian and Moran 1998). It is not only non-LTR-element abundance that has impacted eukaryotic genomes. Non-LTR elements are most closely related in organization and transposition mechanism to the bacterial and organellar group II introns (Cervio and Belfort 1996), which have in turn been proposed to be the progenitors of the nuclear spliceosomal introns (Cavalier-Smith 1991; Sharp 1991). The reverse transcriptases (RTs) of non-LTR elements also show a clear phylogenetic relationship and similar enzymatic activities to telomerase (Eickbush 1997; Nakamura et al. 1997). Thus, the influence of non-LTR elements and their relatives on the current size and structure of eukaryotic genomes has been significant and varied.

Many of the earlier characterizations of the non-LTR retrotransposons suffered from forced comparisons of their enzymatic domain architecture to those of LTR (or long-terminal-repeat-bearing) retrotransposons, a class that includes retroviruses. In addition, studies of the origins and evolution of non-LTR retrotransposons have been hindered by the lack of universal features. Non-LTR elements have been characterized as bearing either one or two open reading frames (ORFs); some, but not all, encode an apurinic/apyrimidic endonuclease, a ribonuclease H, and/or putative nucleic-acid-binding motifs. In addition, the mechanism of retrotransposition used by these elements results in numerous copies being 5′-truncated. With many genomic sequencing projects in various stages of completion, hundreds of copies of non-LTR elements are being discovered (for example, see recent reports on Caenorhabditis elegans; Malik and Eickbush 1998; Marin et al. 1998). There is an immediate need for a cogent scheme to organize and classify this widespread and disparate class of elements.

Phylogenetic analyses of non-LTR elements are restricted to the RT domain, the only domain common to all elements. Previous analyses of RT domains using the 178 amino acid positions that can be identified in all retroelements had indicated that the non-LTR class of elements is monophyletic with respect to all other retroelements (Xiong and Eickbush 1988a; 1990; Doolittle et al. 1989; Eickbush 1994). However, analyses based on these universal positions did not resolve the relationships of elements within the non-LTR class (Xiong and Eickbush 1988a; 1990; Doolittle et al. 1989; McClure 1991; Burke, Müller, and Eickbush 1995). In the most comprehensive study published to date, non-LTR elements could be divided into five subgroups (Tu, Isoe, and Guzova 1998). In addition to this lack of resolution, no individual lineage of retrotransposons had been studied in sufficient evolutionary depth to provide a time frame for the evolution of the non-LTR elements. This

Abbreviations: aa, amino acid; AP endonuclease, apurinic-apyrimidinic endonuclease; non-LTR, non-long-terminal-repeat; RT, reverse transcriptase.

Key words: retrotransposons, AP endonuclease, reverse transcriptase, RNase H, nucleotide divergence rates.

Address for correspondence and reprints: Thomas H. Eickbush, Department of Biology, Hutchison Hall, University of Rochester, Rochester, New York 14627. E-mail: eick@uhura.cc.rochester.edu.
latter problem has now been addressed by an analysis of two distinct lineages of non-LTR elements, R1 and R2, from a variety of diverse arthropod genomes (Burke et al. 1998). These studies of R1 and R2 strongly suggested a vertical inheritance in arthropods, thus providing a temporal yardstick with which to compare the evolution of other non-LTR elements.

In this report, we present a comprehensive phylogenetic analysis of the currently known non-LTR retrotransposons. We extend the sequence alignment of the common RT domain to include the proposed thumb region of the “right-hand” structure of retroviral RT domains (Kohlstaedt et al. 1992; Unge et al. 1994). The 440 amino acid positions in this alignment are sufficient to resolve non-LTR elements into 11 distinct clades that each date back to before the divergence of the major animal phyla.

Materials and Methods

Amino acid sequences of the non-LTR elements were downloaded from GenBank databases. The majority of these sequences revealed intact ORFs and conserved features. However, in six cases (CRE1, TRIM, BDDE, JAM1, G, ingi), ORFs had to be manually reconstructed due to frame shifts in the original sequence. In many cases, the ORFs had to be extended upstream of the first encountered methionine (M). The accession numbers of the various non-LTR elements represented here are as follows (numbering based on fig. 3): 1—X17078; 2—M62862; 3—M33009; 4—U19151; 5—X51967; 6—U13033; 7—U13035; 8—AF015685; 9—AF015819; 10—M16558; 11—AF015818; 12—L08889; 13—U29445; 14—U83119; 15—AF081114; 16—AB012223; 17—U93574; 18—AF055640; 19—AF055643; 20—M26915; 21—Y00086; 22—3047086; 23—X57034; 24—AB008896; 25—M63452; 26—Z86117; 27—AF025462; 28—U58755; 29—L76169; 30—AF018035; 31—X99080; 32—L25662; 33—X51968; 34—U13030; 35—AF015277; 36—L00945; 37—M19755; 38—D38414; 39—AF015489; 40—D85594; 41—M93690; 42—M93691; 43—X60177; 44—U87543; 45—X59239; 46—U73800; 47—M22874; 48—M38437; 49—U14101; 50—X06950; 51—X17551; 52—M17214; 53—Ring, Pfeifer, and Grigliatti (1996); 54—X77571; 55—AF012049; 56—AF012043; 57—U07847; 58—M91571; 59—M10822; 60—L79944; 61—M93689; 62—U03849; 63—U88211; 64—AB005891; 65—L08889; 66—Z82275; 67—U46668; 68—X60372; 69—M14954; 70—M28878; 71—X05710 (called TRS); 72—X83098. In those instances in which multiple members of a particular element family were sequenced from the same species, only the alignment programs were constrained at the putative active site residues, which are common to all endonucleases of this class. We employed both a multiple alignment and a profile alignment (of non-LTR AP domains to other prokaryotic and eukaryotic AP endonucleases) for this purpose. Alignments generated by these two methods were not significantly different. The alignments of the RT, AP Endonuclease, and RNase H domains have been deposited at EMBL with accession numbers DS36752, DS36736, and DS36735, respectively. Phylogenetic trees were generated by the neighbor-joining method (Saitou and Nei 1987) and maximum-parsimony heuristic options as implemented in PAUP*, version 4.0d64 (tree bisection-reconnection branch swapping, with maximum number of trees saved at each step limited to five). Bootstrapping was also carried out using PAUP*, version 4.0d64.

Results

Characterization of the RT Domain

Our previous analyses of RT sequences identified seven blocks of amino acid sequences common to all retroelements (Xiong and Eickbush 1988a, 1990; Eickbush 1994). However, for a more detailed look at the non-LTR retrotransposable elements, we wanted to expand our analysis to the entire RT domain of these elements. We had previously used a moving-window similarity index to compare nine R2 elements obtained from highly divergent arthropods (Burke et al. 1999). Homology modeling was then used to show that a 440-aa domain of R2 identified by this comparison could be folded into a three-dimensional structure similar to the “right-hand” (fingers-palm-thumb) domain of retroviral RTs (Burke et al. 1999). As shown in figure 1A, the RT region of R2 elements consisted of 11 blocks of high conservation—8 in the catalytic fingers/palm subdomain and 3 in the proposed thumb subdomain. While seven of these regions represented the sequence blocks identified earlier (peaks labeled 1–7), additional areas of equally high conservation were detected. These include the region “0,” previously termed the “Z” domain (Doolittle et al. 1989; McClure 1991), the “2A” domain, and two regions within the proposed thumb subdomain (peaks 8 and 9), which has previously been described as the “tether” domain (Doolittle et al. 1989; McClure 1991).

To determine if these 11 blocks of identity would be conserved across all non-LTR elements, we extended this moving-window index analysis to three other groups of non-LTR retrotransposons. Each of the groups selected contained at least six divergent elements, which were clearly monophyletic based on phylogenetic analysis using the original set of RT sequence blocks. These groups were (1) L1-like elements from slime molds, plants, and vertebrates (fig. 1B); (2) R1-like elements from arthropod species (fig. 1C); and (3) Jockey-like elements, also from arthropods (fig. 1D). As shown in figure 1, the 11 blocks of high conservation identified within the R2 group could also be found in the other three non-LTR groups. The best-conserved sequence
blocks are represented by peaks 2–4. The most variable sequence blocks are represented by peak 1, which appears as a shoulder to the larger peak 2 in R2 and Jockey-like elements, and peak 8, which appears as a double peak in R1 and Jockey-like elements. Finally, in Jockey-like elements, there is a region of conservation between peaks 3 and 4 which is not as well-defined in the other groups.

Little, if any, sequence similarity can be found between the different groups of non-LTR elements either upstream or downstream of the region defined by these 11 conserved peaks. Indeed, the ORFs of several groups of elements (RTE-like, Jockey-like, and CR1-like elements) actually ends with peak 9. We contend that the ~440-aa domain spanning these blocks of sequence homology constitutes the RT domain of non-LTR elements.

A complete RT alignment is shown in figure 2. Only 11 elements representing the distinct clades revealed by our study (see below) are shown to highlight the conserved residues in the RT domains of all non-LTR elements. A complete alignment of all elements is available at EMBL, DS36752. The first nine blocks of conserved sequences (peaks 0–7 in fig. 1) contain easily identified invariant residues that have been shaded in the figure. In the case of the thumb region, sequence conservation is high within each group but is reduced between groups. The conserved region represented by peak 8 contains no invariant residues but does include an extended region with numerous amino acids shared by a majority of the elements. In the case of peak 9, the non-LTR elements can be divided into two subgroups which contain different offset areas of maximum conservation (boxed regions). Thus, while there are no invariant residues in the thumb region of the RT domain, several blocks of residues are conserved across subsets of elements, indicating their suitability for phylogenetic analysis.

RT Domain Phylogenetic Analysis: Classification of the Non-LTR Elements

Using this updated definition of the RT domain, we attempted to reconstruct the phylogenetic relationships between the non-LTR elements that have been sequenced to date. Not all “reported” sequences could be used in the analysis, as some do not include entire RT domains. Other sequences represent highly degenerate elements that required repeated shifts of frame or bypassing of termination codons and therefore could not be included because their correct sequences could not be unambiguously defined. A 50% consensus neighbor-joining phylogram of 72 non-LTR elements is shown in figure 3. The tree was rooted using RT sequences of the bacterial and organellar group II introns, because their RT sequences are most closely related to the non-LTR sequences (Xiong and Eickbush 1990). In addition, group II introns contain sequences which can be considered homologous to all 11 conserved regions. A similar topology was also obtained using LTR-retrotransposon sequences as the outgroup; however, this rooting is formally less accurate, as several of the 11 conserved domains found in figures 1 and 2 are absent from the LTR group.

The non-LTR element phylogeny in figure 3 offers a greatly increased resolution of the relationship of the elements and thus allows us to propose a classification.
scheme for the entire class of non-LTR retrotransposons. It should be noted that this phylogeny is largely the same as that previously reported using only domains 1–7 (Xiong and Eickbush 1990; Tu, Iose, and Guzova 1998). Thus, including the intervening regions as well as domains 0, 8, and 9 only served to improve the support for the groupings within the phylogeny, rather than dictate the phylogeny itself. We propose the use of the term “clade” to represent those elements that (1) share the same structural features, (2) are grouped together with ample phylogenetic support, and (3) date back to the Precambrian era (see below). Based on this scheme, all non-LTR retrotransposons fall into 11 distinct clades with significant bootstrap support by neighbor-joining algorithms (bold numbers). The topology of the tree is identical using maximum-parsimony (heuristic) methods, although bootstrap supports are generally lower (numbers in italics). The only significant difference between the neighbor-joining and maximum-parsimony analyses is that the monophyly of the L1 and I clades is not well supported by the maximum-parsimony method (bootstrap support below 50%).

The earliest branching “CRE” clade includes non-LTR elements that are site-specific for mini-exon arrays in trypanosome genomes (Aksoy et al. 1990; Gabriel et al. 1990). Next to appear are the “R2” clade from arthropods (Burke, Calalang, and Eickbush 1997), the “R4” clade from nematodes and arthropods (Burke, Muller, and Eickbush 1995), and the “L1” clade (Hutchison et al. 1989) from a variety of plant, vertebrate, slime mold, and algal genomes. R2 elements are site-specific for the rRNA genes, while members from the R4 clade are site-specific for either the rRNA genes (R4) or simple repeats (Dong). Some representatives...
from the L1 clade are site-specific (Tx1, Zepp), while others appear to lack site-specificity (Ta11, L1). Unfortunately, the RT phylogeny does not resolve the individual relationships of these three clades of elements with the other non-LTR clades.

The remaining seven clades of non-LTR retrotransposons are monophyletic within the non-LTR elements. The earliest-diverging clade within these is the “RTE” clade (Youngman, van Luenen, and Plasterk 1996), with members in nematode, insect, and vertebrate genomes. Of the six remaining clades, three are the sister clades “Tad1” (Cambareri, Helber, and Kinsey 1994), “LOA” (Felger and Hunt 1992), and “R1” (Xiong and Eickbush 1988b) while two are the sister clades “CR1” (Silva and Burch 1989) and “Jockey” (Primagi, Mizrokhi, and Ilyin 1988). The Tad1 clade has representatives exclu-
sively from fungal genomes, while the LOA, Jockey, and R1 clades are exclusively present in arthropod genomes. Members of the R1 clade are site-specific either for rRNA genes (R1, RT1) or for telomeric repeats (TRAS, SART1). The CR1 clade has the broadest distribution, with elements identified in insects, nematodes, trematodes, and vertebrates. L2 elements in humans also appear to be members of this CR1 clade, but only highly defective copies have been sequenced to date (Smit 1996). The final non-LTR clade is the “I” clade. Representatives of this clade are from insects, molluscs, and trypanosomes and are clustered on the tree by the lowest bootstrap values (76 for neighbor joining and 45 for maximum parsimony). The I clade is thus the poorest-defined grouping of non-LTR elements, and it is possible that its members may eventually be divided into multiple clades as more non-LTR elements are sequenced from disparate genomes.

Mode of Transmission

Having defined the various lineages of non-LTR elements, the next question to be addressed is that of the ages of these lineages. Element distribution across eukaryotic taxa is not a reliable indicator of age because of the possibility of horizontal transfers between lineages or differential loss of elements from lineages. Indeed, the issue of vertical inheritance versus horizontal transfers of transposable elements is often plagued by the presence of many paralogous lineages that have been copropagated with varying degrees of success in a lineage. A comparison of two paralogs in related species would inflate the actual divergence time between orthologous elements (truly related by descent) in those species. Using such inflated estimates as “true” indicators of divergence might, in turn, lead to comparisons that are “less divergent than expected,” thus suggesting horizontal transfers. In this regard, we have been fortunate to obtain for two non-LTR clades, R1 and R2, representatives from diverse taxa within Arthropoda (Burke et al. 1993, 1998; Eickbush and Eickbush 1995; Eickbush et al. 1995, 1997; Lathe et al. 1995; Lathe and Eickbush 1997). We found no evidence to suggest horizontal transfers within the R1 and R2 clades; thus, these elements can provide divergence rates which can be compared with those of other non-LTR clades.

Our divergence-versus-age analysis is presented in figure 4, with the X-axis representing estimates of host divergence and the Y-axis representing amino acid divergence between the RT domains of the non-LTR elements. Comparisons within the R1 and R2 clades at seven different species divergence times provide a framework for this graph (solid circles). While paralogous lineages of R1 and R2 do exist in arthropods (Burke et al. 1993, 1998; Lathe et al. 1995), comparing only the most closely related lineages of R1 and R2 for each pair of species provides the best estimates of orthologous comparisons, as shown in the graph. The R1 and R2 lineages show a similar pattern of increased divergence with time that approaches saturation at approximately the time of arthropod divergence (>600 MYA).

![Figure 4](https://academic.oup.com/mbe/article-abstract/16/6/793/2925486/50%20of%20the%20domain%20that%20have%20been%20copropagated%20with%20varying%20degrees%20of%20success%20in%20a%20lineage.%20A%20comparison%20of%20two%20paralogs%20in%20related%20species%20would%20inflate%20the%20actual%20divergence%20time%20between%20orthologous%20elements%20(truly%20related%20by%20descent)%20in%20those%20species.%20Using%20such%20inflated%20estimates%20as%20“true”%20indicators%20of%20divergence%20might,%20in%20turn,%20lead%20to%20comparisons%20that%20are%20“less%20divergent%20than%20expected,”%20thus%20suggesting%20horizontal%20transfers.%20In%20this%20regard,%20we%20have%20been%20fortunate%20to%20obtain%20for%20two%20non-LTR%20clades,%20R1%20and%20R2,%20representatives%20from%20diverse%20taxa%20within%20Arthropoda%20(Burke%20et%20al.%201993,%201998;%20Eickbush%20and%20Eickbush%201995;%20Eickbush%20et%20al.%201995,%201997;%20Lathe%20et%20al.%201995;%20Lathe%20and%20Eickbush%201997).%20We%20found%20no%20evidence%20to%20suggest%20horizontal%20transfers%20within%20the%20R1%20and%20R2%20clades;%20thus,%20these%20elements%20can%20provide%20divergence%20rates%20which%20can%20be%20compared%20with%20those%20of%20other%20non-LTR%20clades.%20

Our divergence-versus-age analysis is presented in figure 4, with the X-axis representing estimates of host divergence and the Y-axis representing amino acid divergence between the RT domains of the non-LTR elements. Comparisons within the R1 and R2 clades at seven different species divergence times provide a framework for this graph (solid circles). While paralogous lineages of R1 and R2 do exist in arthropods (Burke et al. 1993, 1998; Lathe et al. 1995), comparing only the most closely related lineages of R1 and R2 for each pair of species provides the best estimates of orthologous comparisons, as shown in the graph. The R1 and R2 lineages show a similar pattern of increased divergence with time that approaches saturation at approximately the time of arthropod divergence (>600 MYA).

Thirteen other comparisons of sequence divergence versus host divergence time can be made with elements from the other non-LTR clades and are also shown in figure 4. Six of these estimates are from arthropod lineages (the Jockey, I, and LOA clades), one is from plants (the L1 clade), and six are from vertebrates (the L1 and CR1 clades). In the case of the arthropod comparisons (solid squares), all points fall near that predicted from the R1 and R2 data. The single plant comparison (open triangle) falls above the arthropod curve and probably represents a paralogous comparison. All vertebrate comparisons (open squares) fall below the arthropod comparisons, suggesting that non-LTR sequence evolution in vertebrates is slower than that in arthropods. All other possible comparisons between the 72
elements in figure 3 either fall above the arthropod curve, suggesting that they are paralogous comparisons, or represent comparisons of species separated by more than 600 Myr, where both divergence and time estimates have little resolution.

It has been shown for both R1 and R2 elements that, when corrected for codon bias, the rates of nucleotide sequence evolution within the protein-encoding regions are of the same order as those of nuclear genes (Eickbush et al. 1995; Lathe et al. 1995). A similar conclusion has been derived for the evolution of the L1 lineage in mammals (Hardies et al. 1986; Usdin et al. 1995), with the L1 elements even being used to date rodent speciation events (Verneau, Catzellis, and Furano 1998). Thus, the different rates of evolution in the invertebrate and vertebrate lineages appear to reflect general differences in rates of evolution of their respective genomes (reviewed in Li 1993). Although an average global rate appears to apply for both the vertebrate and invertebrate lineages, rate differences have previously been noted over short periods of evolution for R1 elements in the melanogaster species subgroup in Drosophila (Eickbush et al. 1995) and in L1 elements from Peromyscus (Casavant et al. 1998).

It should be noted that the data from three previous reports of possible horizontal transmission of non-LTR elements have also been placed on the graph in figure 4, comprising (1) Jockey elements between Drosophila melanogaster and Drosophila funebris (Mizrokhi and Mazo 1990) (solid square at 40 Myr), (2) CR1 elements between a schistosome and a vertebrate (Drew and Brindley 1997) (open square at 600 Myr), and (3) RTE elements between cow and viper genomes (Kordis and Gubensek 1995, 1998) (X at 250 Myr). Remarkably, these three points fall on or near the appropriate curve and are therefore consistent with a vertical means of descent.

In conclusion, our analysis allows us to make two important conclusions. First, there is no reliable evidence to date for the horizontal transfer of non-LTR elements between eukaryotes during the past 600 Myr. Second, based on our still-crude estimates of sequence evolution, all 11 clades of non-LTR elements date back to before the Cambrian period. The different rates of sequence evolution for non-LTR elements in arthropods and vertebrates, and probably in other host lineages, should become more precisely defined as additional elements are sequenced. These rate estimates will serve as increasingly accurate quantitative indicators with which to evaluate the role of vertical versus horizontal transmission in the distribution of non-LTR elements.

Phylogeny of Two Other Non-LTR-Element Domains

While the RT domain is the only protein component shared by all non-LTR elements, two additional catalytic domains are shared by subsets of these elements: an apurinic/pyrimidinic (AP) endonuclease domain (Martin et al. 1995; Feng et al. 1996) and an RNase H domain (Fawcett et al. 1986; Blesa and Martinez-Sebastian 1997). The AP endonuclease is believed to generate the DNA cleavage that initiates the reverse transcription reaction at the chromosomal target site (Feng et al. 1996; Moran et al. 1996), while the RNase H domain would presumably eliminate the RNA template after reverse transcription. The AP endonuclease domain is present in elements from eight clades of non-LTR elements (the exceptions are the CRE, R2, and R4 clades). We also conducted a phylogenetic analysis of this domain to estimate its origin within the non-LTR lineages and to help corroborate the non-LTR phylogeny based solely on the RT domain. While evidence to date suggests that an AP endonuclease domain is present in all members of the eight non-LTR clades, 12 elements from these clades that are present on the RT tree could not be positioned on the AP tree. In most of these cases, the AP domain, located upstream of the RT domain, was not present in the 5'-truncated elements sequenced.

The phylogeny in figure 5 was rooted using cellular AP endonucleases involved in the general mechanisms
of DNA repair in eukaryotes and prokaryotes. The phylogeny of the non-LTR-element AP endonucleases is in agreement with the RT phylogeny (fig. 3). Unfortunately, the AP phylogeny has less resolution than the RT phylogeny, because the AP endo domain is smaller (230 aa) and less conserved than the RT domain. As a consequence of this lower resolution, the two clades with the weakest bootstrap support on the RT phylogeny (the L1 and I clades) are not resolved as distinct clades on the AP phylogeny. The only apparent violation of the RT phylogeny with the AP phylogeny is in the relationship of the LOA and R1 elements. LOA elements appear as a sister clade to R1 on the RT tree but fall within the R1 clade on the AP tree. It is uncertain at present whether there is sufficient resolution in this phylogeny to postulate a domain swap between these two clades of insect elements.

The AP phylogeny indicates that among the eight non-LTR clades containing this endonuclease, the L1 clade is the oldest, followed by the RTE clade. The acquisition of the AP endonuclease by the non-LTR lineage from a host repair machinery appears to be of ancient origin, because it is not possible to resolve whether this domain was obtained from a prokaryotic or a eukaryotic source. We conclude that the acquisition of the AP endonuclease by non-LTR elements was a single event that occurred early in the evolution of eukaryotes (see also Discussion). Once introduced, the domain has been stably maintained in all subsequent lineages.

An RNase H domain downstream of the RT domain can be found in non-LTR elements only from the LOA, I, R1, and Tad clades. Our identification of the RNase H domain in the L1 clade (McCline 1991, 1993). All members of the LOA and I clades contain this domain (the RNase H sequences are not available from BGR). In the Tad clade, both Mgr583 and Cgt13 contain the domain, while the Tad1 element itself does not (the Mars1 sequence is incomplete). This differential retention of the RNase H domain is even more striking in the R1 clade. R1 elements in the non-insect arthropods (tarantula and centipede) contain RNase H domains, while those in insects do not. Within the silkworm telomere-specific elements, TRAS1 elements contain the RNase H domain, while SART1 elements do not, even though both are present in the same host species. As shown in figure 6, with only 130 aa positions and many fewer taxa, the resolution of the RNase phylogeny is even lower than that of the AP endonuclease phylogeny. While there is little resolution within the non-LTR elements, the phylogeny is sufficient to suggest a monophyletic origin of the RNase H domain within the non-LTR lineage. Unlike the AP domain, for which it was not possible to resolve a prokaryotic or eukaryotic origin, the RNase H domain appears to be derived from a eukaryotic source. This is consistent with the later introduction of the RNase H domain in non-LTR elements relative to the AP endonuclease domain.

Discussion

Non-LTR retrotransposable elements were the last group of eukaryotic mobile elements to be identified (Weiner, Deininger, and Efstratiadis 1986), probably because they lack an easily defined, consistent structure, and many, if not most, copies contain 5' truncations. In recent years, the basic features of non-LTR element structure and mode of integration have been established. In this report, we have attempted to characterize the evolution of this group of elements. There are several reasons why our phylogenetic analysis of these elements has significantly higher resolution than do those of earlier reports. First, a large number of taxa are now available from a wide range of organisms. Second, the number of amino acid positions used in the phylogenetic analysis has been expanded to 440 positions. Third, additional enzymatic domains within the non-LTR elements have been found and confirmed by mutagenesis and expression experiments. Finally, detailed studies of the evolution of individual lineages have enabled the construction of time lines to monitor the evolution of elements.
other lineages. The increased resolution afforded by these advances and the vertical mode of evolution of these elements suggest that non-LTR elements have a simple, highly tractable history. Indeed, that history may now be better known than that of any other class of transposable elements.

Vertical Inheritance of Non-LTR Elements

All non-LTR elements can be organized into 11 distinct clades. Four of these clades (L1, RTE, CR1, and I) have wide distributions in eukaryotes, while the remaining clades are confined to either one or two taxonomic groups. No evidence of horizontal transfer was detected within or between these clades. While the absence of evidence does not mean evidence of absence, our contention is that the strict vertical inheritance seen for R1 and R2 can be extended to the other clades of non-LTR elements. These evaluations of non-LTR inheritance are limited to events that have occurred during the past 600 Myr, because the RT amino acid divergence within non-LTR elements approaches saturation around this time (fig. 4). If the assumption of vertical descent is projected back through the entire history of non-LTR elements, then the presence of trematode, vertebrate, and insect elements in one of the more recently derived clades (CR1), and the presence of common elements in early protists (Trypanosome) and metazoans in another clade (I) implies that many, if not all, of the non-LTR clades date back to close to the origin of the eukaryote crown group. Ultimate resolution of age and distribution of the distinct non-LTR lineages will require the recovery of more elements in divergent organisms. At present, our collection of non-LTR elements is significantly biased toward insect elements. Eight of the 11 clades have insect representatives, and 4 of the clades (R2, LOA, R1, and Jockey) are defined at present with only arthropod representatives. As additional genomic sequences become available from more organisms, particularly the extensive intergenic regions obtained as part of genomic sequencing projects, new clades will undoubtedly be discovered and broader distributions for the existing clades will be found.

The apparent lack of horizontal transfers among non-LTR elements is in stark contrast to what has been observed for other transposable elements (reviewed in Kidwell 1993). In prokaryotes, rates of horizontal transfer of segments of the genome itself are too high to expect a stable vertical inheritance of transposons (Lawrence and Ochman 1998). In the case of eukaryotes, several spectacular cases of transspecies and even transphyla lateral transfers have been observed for the DNA-mediated elements P and mariner (Clark and Kidwell 1997; Hartl, Lohe, and Lozovskya 1997; Robertson 1997). In the case of the LTR-containing (or viral) retrotransposons, phylogenetic analyses have noted that the sequences of such elements in plants, fungi, and insects are more closely related than are the non-LTR elements in similar comparisons (Xiong and Eickbush 1990). Occasional horizontal transfers have been found in SURL elements in a recent study of over 40 echinoderm species (Gonzalez and Lessios 1999). Indeed, the presence of envelope-like domains in many members of the gypsy/Ty3 group provides a potential mechanism for cross-species infection (Song et al. 1994; Wright and Voytas 1998).

Why are non-LTR elements so stable when other transposable elements, particularly the DNA-mediated elements, do not appear to be? First, non-LTR elements utilize a “replicative” mechanism of integration, which enables purifying selection to retain active copies in a lineage. DNA-mediated elements, on the other hand, utilize a predominantly “nonreplicative” mechanism of integration and are therefore largely not subject to purifying selection. All copies of an element within a host will eventually succumb to mutational inactivation. Thus, DNA-mediated elements may be under selective pressure to evolve horizontal transfer mechanisms (Lohe et al. 1995; Hartl, Lohe, and Lozovskya 1997).

A second reason for the different inheritance patterns of DNA-mediated elements and non-LTR elements is that the mechanism of non-LTR integration may preclude jumps between species. Non-LTR elements rely on a target-primed reverse transcription reaction, in which the cDNA strand is reverse-transcribed from an RNA template directly onto a chromosomal target site (Luan et al. 1993). As a consequence, the only extrachromosomal copies of a non-LTR element genome that could be directly transferred between species is in the form of RNA rather than the more stable DNA intermediates of other transposable elements. It is also generally suggested that the most frequent mechanism for the transfer of mobile elements between species is for the elements to insert into viral genomes, which can then infect another species. The required insertions into extrachromosomal viral genomes may be easy for DNA-mediated elements but not for the non-LTR elements. Complete integration of a non-LTR element appears to require a much more extensive cooperation of the cellular DNA repair/replication machinery than is required for other transposable elements (see discussion in George, Burke, and Eickbush 1996). This is a result of the apparent lack of an efficient mechanism encoded by the element to ensure the attachment of the newly synthesized cDNA strand to the DNA sequences upstream of the target site and for the synthesis of the second strand. This dependency on DNA repair is reflected in the observation that most non-LTR elements have highly variable effects on the target site, as well as frequent 5′ truncations. If the DNA repair/replication mechanism required for non-LTR integration is not closely associated with the production of extrachromosomal viral genomes, then non-LTR elements may be unable to integrate into these genomes. Thus, non-LTR elements may be deprived of the major vectors used by other transposable elements for their horizontal transfers.

LTR retrotransposable elements also use a replicative mechanism of integration and thus should not depend on horizontal mobility for long-term survival. However, because LTR elements make double-stranded DNA intermediates, which are then inserted into DNA by an integrase similar to that of DNA-mediated elements (Capy et al. 1996), LTR elements would be read-
Fig. 7.—Summary of non-LTR retrotransposable element evolution. The phylogeny is a schematic diagram of that shown in figure 3, with only the branch points (and neighbor-joining bootstrap support) leading to the 11 major clades of non-LTR elements indicated. The tree was rooted using the RT sequences of group II introns. Shown at right are schematic diagrams of the ORF structures of representative elements from each clade. ORF structures for more than one element from a clade are presented if they have significant differences. Open boxes indicate ORFs. Shaded boxes indicate enzymatic domain encoded in each element. RT = reverse transcriptase; APE = apurinic/apyrimidinic endonuclease, REL-endo = restriction enzyme-like endonuclease, and RNH = RNase H domain. The REL-endo domain is described elsewhere (unpublished data). Vertical bars within the ORF represent the location and number of cysteine-histidine motifs that are believed to represent nucleic-acid-binding domains. The two ORFs present in many elements are offset whether or not they are of the same frame.

Acquisition of Non-LTR Domains

As diagrammed in figure 7, our phylogeny of non-LTR elements suggests a simple scenario for the evolution of the various catalytic domains and structural motifs present in this diverse class of elements. Members of the oldest clade of non-LTR elements, the CRE clade, encode a single ORF with the RT domain located near the carboxyl-terminal end. The RT domain is flanked by regions in which Cys and His residues are positioned in a manner consistent with nucleic-acid-binding motifs (Berg and Shi 1996). Non-LTR elements of the R2 and R4 clades also encode a single ORF, but the regions upstream of the RT domain are shorter and more uniform. The most distinctive attribute of these three clades of non-LTR elements is their high degree of target sequence specificity, i.e., either rRNA genes or spliced leader exons.

We have recently shown that the endonuclease for the R2 element is located downstream of the RT domain (unpublished data). The active site of this endonuclease contains conserved residues that are similar to those of certain type II restriction enzymes. All members of the CRE, R4, and R2 clades contain these highly conserved active-site residues located approximately 40 aa downstream of a Cys-His nucleic-acid-binding motif. We propose that the original non-LTR elements were sequence-specific in their insertions by virtue of this restriction-enzyme-like endonuclease (REL-endo). That the first non-LTR elements were site-specific is also consistent with their likely evolution from the group II introns of bacterial and organellar genomes. Certain group II introns also encode a restriction-enzyme-like endonuclease downstream of their RT domain, although these endonucleases are of the H-N-H type (Gorbalenya 1994; Shub, Goodrich-Blair, and Eddy 1994).

The REL-endonuclease domain encoded by the original non-LTR elements was replaced in subsequent lineages with an AP endonuclease acquired from the DNA repair machinery of the host cell. Based on the AP endonuclease phylogeny, this replacement took place early in the evolution of non-LTR elements, most likely in the lineage leading to the L1 clade of non-LTR elements. While all non-LTR elements containing the AP endonuclease domain have lost the critical active-site residues associated with the REL-endonuclease, a carboxyl-terminal domain containing a potential nucleic-acid-binding motif is retained in most non-LTR elements. Thus, while the catalytic domain responsible for DNA cleavage has shifted upstream of the RT domain, the carboxyl-terminal domain may still participate in the DNA binding of these elements. For example, the putative nucleic-acid-binding motif associated with this domain in ingi elements has expanded in number. On the other hand, the non-LTR elements of the RTE, Jockey, and CR1 clades do not encode a carboxyl domain (fig. 7), and therefore most aspects of DNA binding are probably associated with the AP endonuclease domain. For most non-LTR elements, the acquisition of the AP endonuclease domain resulted in the loss of target site specificity. However, a few elements still retain in a site-specific manner, including Tx1, Zep, and DRE from the L1 clade and the entire R1 clade. In some of these cases, the AP endonuclease domain itself has acquired site specificity (Feng, Schumann, and Boeke 1998; S. Christensen and D. Carroll, personal communication).

The acquisition of the AP endonuclease domain in the non-LTR lineage also coincided with the appearance of a second ORF in front of the major RT-encoding ORF. This additional ORF (ORF1) has been retained by all non-LTR clades containing the AP endonuclease, with...
the possible exception of the RTE clade (see Malik and Eickbush [1998] for further discussion of this clade). ORF1 of mammalian L1 elements has been shown to be required for retrotransposition (Moran et al. 1996). The presence of closely spaced nucleic-acid-binding motifs similar to that of retroviral gag proteins (Berg and Shi 1996) and the finding that ORF1 proteins from both L1 and I elements can bind RNA (Hohjoh and Singer 1996; Dawson et al. 1997; Kolosha and Martin 1997) suggests a nucleic-acid-binding role for ORF1. Because the L1 clade lacks these gag-like motifs but contains a leucine zipper missing in the other clades (Holmes, Singer, and Swergold 1992), there have either been major alterations of the ORF1 protein in the various non-LTR clades or two separate acquisition events (one in L1 and one for the remaining clades). Sequence conservation within ORF1 is too low to be of use in the resolution of this issue.

The only other enzymatic domain that has been found in certain non-LTR elements is RNase H (Fawcett et al. 1986; Blesa and Martinez-Sebastian 1997). When present, this domain is immediately downstream of the RT domain, as it is in retroviruses and in all LTR retrotransposons. Only four non-LTR clades contain RNase H domains (Tad, R1, LOA, and I). The variable retention of this domain within certain of these clades (e.g., the R1 elements of arachnids have an RNase H domain but those of insects do not) suggests that in most instances, cellular RNase H activity can readily substitute for the activity occasionally encoded by the element. Based on our phylogenetic analysis, acquisition of the RNase H domain occurred sometime after the AP endonuclease addition. While it is not clear whether the acquisition of the AP endonuclease domain was from a prokaryotic or a eukaryotic source (fig. 5), the non-LTR element RNase H domain is clearly more similar in sequence to eukaryotic sources (fig. 6).

In summary, we have provided evidence that the non-LTR retrotransposable elements have evolved by vertical transmission. This mode of descent has enabled us to trace the evolution of this abundant, diverse group of elements. We hope this analysis will provide a framework on which future non-LTR elements can be classified and perhaps named. The placement of these new elements within the non-LTR phylogeny will undoubtedly confirm or refute the strict vertical evolution we have suggested. The discovery of even rare horizontal events will be critical to our understanding of non-LTR elements. Such transfers will directly impact our understanding of non-LTR elements within the non-LTR phylogeny will undoubtedly be critical to our understanding of non-LTR retrotransposons. Only four non-LTR clades contain RNase H (Fawcett et al. 1997; Blesa and Martinez-Sebastian 1997). When present, this domain is immediately downstream of the RT domain, as it is in retroviruses and in all LTR retrotransposons. Only four non-LTR clades contain RNase H domains (Tad, R1, LOA, and I). The variable retention of this domain within certain of these clades (e.g., the R1 elements of arachnids have an RNase H domain but those of insects do not) suggests that in most instances, cellular RNase H activity can readily substitute for the activity occasionally encoded by the element. Based on our phylogenetic analysis, acquisition of the RNase H domain occurred sometime after the AP endonuclease addition. While it is not clear whether the acquisition of the AP endonuclease domain was from a prokaryotic or a eukaryotic source (fig. 5), the non-LTR element RNase H domain is clearly more similar in sequence to eukaryotic sources (fig. 6).

In summary, we have provided evidence that the non-LTR retrotransposable elements have evolved by vertical transmission. This mode of descent has enabled us to trace the evolution of this abundant, diverse group of elements. We hope this analysis will provide a framework on which future non-LTR elements can be classified and perhaps named. The placement of these new elements within the non-LTR phylogeny will undoubtedly confirm or refute the strict vertical evolution we have suggested. The discovery of even rare horizontal events will be critical to our understanding of non-LTR elements. Such transfers will directly impact our confidence that the current broad distribution of individual lineages of non-LTR elements indicates that the entire group may date back to the very origin of eukaryotes.

Acknowledgments

This work was supported by N.S.F. grant MCB-9601198 to T.H.E. We thank Danna Eickbush and Cesar Perez-Gonzalez for many useful comments on the manuscript. We thank David Duvernell for providing us with the unpublished sequences of the swimmer elements.

LITERATURE CITED

Aksoy, S., S. Williams, S. Chang, and F. F. Richards. 1990. SLACS retrotransposon from Trypanosoma brucei gambiense is similar to mammalian LINEs. Nucleic Acids Res. 18:785-92.


Matsuoka, Y., and K. Tsunewaki. 1999. Evolutionary dynamics of Ty1-copia group retrotransposons in grass shown...
Pierre Capy, reviewing editor
Accepted February 17, 1999