Syst. Biol. 55(6):875–885, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150601077683

Distribution and Phylogeny of *Penelope*-Like Elements in Eukaryotes

IRINA R. ARKHIPOVA

Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA; and Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA; E-mail: arkhipov@fas.harvard.edu

Abstract.—*Penelope*-like elements (PLEs) are a relatively little studied class of eukaryotic retroelements, distinguished by the presence of the GIY-YIG endonuclease domain, the ability of some representatives to retain introns, and the similarity of PLE-encoded reverse transcriptases to telomerases. Although these retrotransposons are abundant in many animal genomes, the reverse transcriptase moiety can also be found in several protists, fungi, and plants, indicating its ancient origin. A comprehensive phylogenetic analysis of PLEs was conducted, based on extended sequence alignments and a considerably expanded data set. PLEs exhibit the pattern of evolution similar to that of non-LTR retrotransposons, which form deep-branching clades dating back to the Precambrian era. However, PLEs seem to have experienced a much higher degree of lineage losses than non-LTR retrotransposons. It is suggested that PLEs and non-LTR retrotransposons are included into a larger eTPRT (eukaryotic target-primed) group of retroelements, characterized by 5′ truncation, variable target-site duplication, and the potential of the 3′ end to participate in formation of non-autonomous derivatives. [*Penelope*-like elements; retrotransposons; reverse transcriptase; GIY-YIG endonuclease.]

One of the biggest surprises that the whole-genome sequencing era has brought to us is the amount of nongenic, and especially transposable element–derived, sequences that comprise the bulk of chromosomal DNA in many eukaryotes. The power of genomics lies in the unbiased coverage of most of the genetic material in a given species, and although the choice of species to be sequenced is not so unbiased, it certainly expands genome studies to species not previously amenable to genetic and molecular manipulations. In the beginning of the present century, the comfortable feeling that the structure and behavior of transposable elements is well-understood has largely disappeared, as the chromosomal DNA of newly sequenced genomes began to reveal new classes of elements which did not happen to cause visible mutations in the pregenomics era, or were absent from the traditional model organisms and represented just odd exceptions from the known rules. Genomics brought to prominence the tyrosine recombinase-containing DIRS-like retrotransposons (Duncan et al., 2002; Goodwin and Poulter, 2004) and even more unconventional large DNA elements such as Helitrons and Polintons (Kapitonov and Jurka, 2001, 2006).

One of the previously underappreciated retrotransposon classes, the *Penelope*-like elements (PLEs), for a long time included only a single representative, *Penelope* (Evgen'ev et al., 1997). This element was discovered as the causative agent of visible mutations occurring during hybrid dysgenesis in *Drosophila virilis*, and, interestingly, still remains the only known mutagenic member of the class. Incorporation of PLEs as a novel class of retroelements became possible with the advent of genome and EST sequencing projects of nonmammalian genomes: the data provided by *Takifugu*, *Tetraodon*, and *Strongylocentrotus* genome projects, as well as the EST data from *Xenopus*, *Schistosoma*, *Trichuris*, and *Ancylostoma*, enabled identification of additional elements related to *Penelope* in these very distant taxonomic groups (Kapitonov and Jurka, 1999; Evgen'ev et al., 2001; Volff et al., 2001; Dalle Nogare et al., 2002). Several unusual structural features distinguish PLEs from the two well-known classes of reverse transcriptase (RT)-containing elements, i.e., LTR-retrotransposons and non-LTR retrotransposons (also called LINEs): the presence of the GIY-YIG endonuclease domain, the ability of some representatives to retain introns, and the partial-tandem organization (reviewed in Evgen'ev and Arkhipova, 2005; see also Fig. 1). PLEs occupy a very special place in the overall retroelement classification (Arkhipova et al., 2003): in the RT phylogeny, they do not belong to either LTR or non-LTR retrotransposon classes, but instead form a sister clade to telomerase reverse transcriptases (TERTs), a highly specialized class of non-mobile RTs which are responsible for maintenance of linear chromosome ends in most eukaryotes. Due to the enormous growth of genome databases, which are now expanding far beyond mammalian and bird genomes (both lacking PLEs so far, although their absence might be due to frequent loss), there is an emerging need for PLE classification into distinct groups, similar to that previously proposed for non-LTR and LTR retrotransposons (Malik et al., 1999; Malik and Eickbush, 1999; Eickbush and Malik, 2002).

A recent wave of applications of the so-called SINE/LINE method of inferring phylogenetic relationships (Shedlock and Okada, 2000; Shedlock et al., 2004; Ho et al., 2005; Kriegs et al., 2006) also prompted us to evaluate the relevant properties of PLEs in light of their usefulness for phylogenetic purposes, given their somewhat intermediate position between LTR and non-LTR retrotransposons. This method treats an insertion at a unique locus as a derived character, with the lack of insertion at this locus being regarded as the ancestral state. Analysis of multiple insertion-bearing loci can provide robust statistical support to the inferred phylogenetic relationships. The main properties of LINE/SINE insertions that make them suitable for phylogenetic inferences are their inability to excise, relatively random character of insertion, nonautonomous nature of most

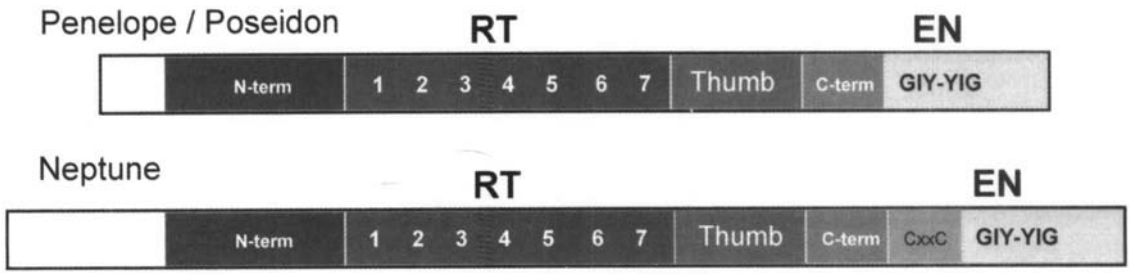

FIGURE 1. ORF structure of *Penelope*-like elements. The RT moiety consists of the core RT domain that includes the seven highly conserved motifs, followed by the thumb domain and the C-terminal extension. In *Penelope-Poseidon*–like elements, the GIY-YIG EN domain is immediately adjacent to the C-terminal extension; in *Neptune*-like and *Nematis*-like elements, a Zn finger–like domain appears between RT and EN. The long N-terminal extension is characteristic of both *Penelope*-like RTs and telomerase RTs. Some of the *Neptune*-like elements may also contain an additional upstream ORF, usually of simple amino acid composition.

copies, apparently rare horizontal transfers, and sufficiently large copy numbers. In addition, the respective host species (groups of species) should have sufficient tolerance to TE insertion, enabling it to persist at a given locus throughout the course of speciation events, until it is gradually erased by point mutations. The present study provides an updated overview of PLE phylogenetic distribution, assembles an extended data set that may hopefully serve as a reference point for building future PLE phylogenies, describes PLE cladistics, evaluates possible modes of their transmission, and discusses their potential for use as phylogenetic markers.

MATERIAL AND METHODS

Sequences were obtained by BLAST searches of publicly available genome databases and trace archives (www.ncbi.nlm.nih.gov, genome.jgi-psf.org, www.broad.mit.edu, www.tigr.org, www.venterinstitute.org, genome.wustl.edu). Assembled consensus sequences were deposited in Repbase (Jurka et al., 2005). Alignments were generated and edited using a ClustalW-based program AlignX from the Vector NTI Suite 7 (InforMax). The aligned data set was submitted to TreeBASE (http://www.treebase.org). Protein secondary structure prediction was done on the JPRED server (http://www.compbio.dundee.ac.uk/~www-jpred/). Phylogenetic studies and estimates of genetic divergences were performed with MEGA 3.1 (Kumar et al., 2004) and MrBayes 3.1.1 (Ronquist and Huelsenbeck, 2003). The trees were viewed and edited with TreeView (Page, 1996).

RESULTS AND DISCUSSION

Questions about PLE Distribution and Phylogenetic Relationships

Despite sharing certain features with non-LTR retrotransposons (5′ truncation, variable-length target site duplication) and LTR retrotransposons (the presence of "pseudo-LTRs," formed by a partial-tandem arrangement), phylogenetic studies reveal that PLEs do not belong to either of these classes (Arkhipova et al., 2003). It was therefore of interest to find out whether PLEs, which carry the GIY-YIG endonuclease domain not previously encountered in retroelements, exhibit patterns of evolution similar to that of non-LTR (LINE-like) retrotransposons, i.e., form deep-branching clades dating back to the Precambrian era, and are not prone to horizontal transmission. Such pattern would make them valuable phylogenetic markers, in addition to non-LTR retrotransposons, which are already widely used for this purpose (see above). Alternatively, they could be prone to horizontal transmission and interelement recombination, like many autonomous DNA transposons and full-length LTR retrotransposons (Robertson, 2002; Malik and Eickbush, 1999), and therefore be of limited use as phylogenetic markers.

It is a given in phylogenetic analysis that relationships are resolved more readily with inclusion of additional characters, as well as improved taxon sampling. However, the relative shortness of the core RT domain, which encompasses 300 amino acids including the seven most conserved sequence motifs (Xiong and Eickbush, 1990) and is often used to determine phylogenetic relationships between retroelements, limits the degree of resolution of PLE phylogenies. The GIY-YIG domain does not constitute a good alternative to RT-based phylogenies, being even shorter than RT (ca. 100 amino acids). The regions outside RT and EN domains of PLEs, however, exhibit a much lower degree of conservation, and, with insufficient datasets, previously failed to yield reasonably good alignments. It was therefore essential, by expanding the relatively small dataset of elements found in model species studied in the pre-genomics era, to generate a more robust alignment of two additional regions present in most PLEs, namely the N-terminal domain (200+ amino acids), which is found only in PLEs and TERTs, and the linker domain of variable length (ca. 100 to 150 amino acids) located between RT and EN.

Finally, an important question is whether phylogenetic distribution of PLEs is limited to animals, as may be inferred from previously published studies, or includes other eukaryotic kingdoms. A recent explosion in the number and diversity of eukaryotic genomes investigated at the level of whole-genome sequence analysis

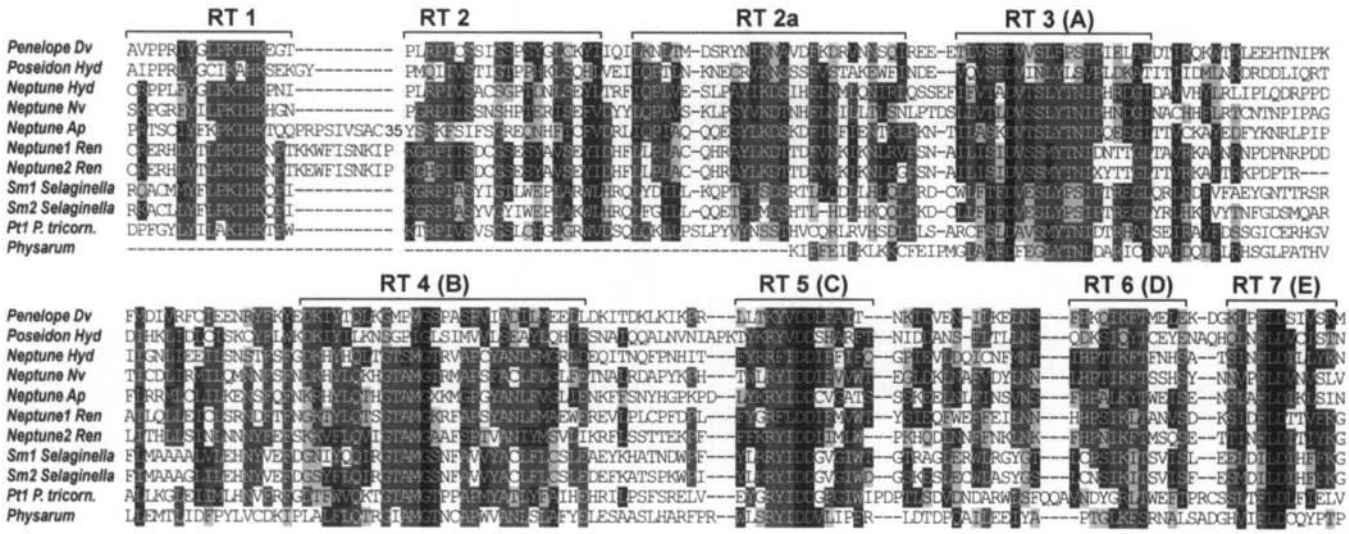

FIGURE 2. Amino acid sequence alignments of the seven conserved core RT motifs for *D. virilis Penelope* and the newly described *Penelope*-like RTs from the protist, plant, and basal metazoan taxa. Element names, accession numbers, and species assignment are listed in Table 1. The amino acid residues are shaded by the Vector NTI software as follows: identical, white letters/black background; block of similar, white letters/dark gray background; conservative, white letters/gray background; weakly similar, black letters/light gray background.

made it possible to search for PLEs in fully or partially sequenced genomes of animals, fungi, plants, and protists. Broadening the PLE diversity is essential for understanding their phylogenetic relationships, as it helps to improve multiple sequence alignment in regions that are less conserved than the core RT domain, and, by increasing taxon sampling, greatly reduces the effects of long-branch attraction for elements that could not previously be assigned to any particular clade. Representatives of most eukaryotic kingdoms were found to contain PLEs, supporting the hypothesis that PLEs and TERTs, which represent sister clades, originated from the common RT ancestor present in early eukaryotic organisms. Improved taxon sampling made it possible to conduct an analysis of phylogenetic groupings formed by PLEs from diverse taxa, inspired by the landmark study of non-LTR retrotransposons by Malik et al. (1999), which subdivided all known non-LTR retrotransposons into distinct clades.

PLEs are Found Predominantly in Animal Genomes

Protists, plants, and fungi.—Although no PLEs could be detected in any nonanimal genomes sequenced before 2005, the genome databases of several most recently sequenced protist, fungal and plant species yielded *Penelope*-like RTs: the slime mold *Physarum polycephalum* (Amoebozoa; Mycetozoa), the pennate diatom *Phaeodactylum tricornutum* (Heterokonta, or stramenopiles), the spike moss *Selaginella moellendorffii* (Viridiplantae; Streptophyta), the inky cap mushroom *Coprinus cinereus*, and the white rot fungus *Phanerochaete chrysosporium* (Opistokonta; Basidiomycota). [The trace archives of the marine alga *Emiliania huxleyi* (Haptophyceae) also appear to contain decayed PLE fragments, although contamination in this case cannot yet be ruled out.] No PLEs have yet been found in Rhizaria or Excavata (Simpson and Roger, 2004). An alignment demonstrating similarity between the core RT domains of the newly identified elements and other known PLEs is shown in Figure 2. However, all of the above retroelements do not carry a C-terminally located GIY-YIG EN domain, are present at a very low copy number, and apparently cannot be regarded as bona fide retrotransposons. Their properties, including association with telomeric repeats, will be described in detail in a separate manuscript. Hereafter, they are referred to as retroelements, as opposed to canonical or bona fide PLEs, which possess an associated EN domain and are therefore referred to as retrotransposons.

It may be hypothesized that the acquisition of the GIY-YIG EN domain by *Penelope*-like RTs present in early eukaryotes occurred only in the branch leading to metazoans, accounting for the absence of canonical PLEs in all of the protist, plant, and fungal genomes sequenced to date. A less likely alternative is a much earlier acquisition of the GIY-YIG domain in early eukaryotes, with subsequent loss of canonical *Penelope*-like retrotransposons in all kingdoms except Animalia.

Metazoa.—In genomic sequences deposited since the most recently published compilation of PLEs (Evgen'ev and Arkhipova, 2005; compiled as of September, 2004 submission date), additional PLEs were identified and/or assembled from the following animal species: the starlet sea anemone *Nematostella vectensis* (Cnidaria; Anthozoa), the silkworm *Bombyx mori* (Arthropoda), the sea lamprey *Petromyzon marinus* (Vertebrata; Hyperoartia), and the sequenced but as yet unnamed *Caenorhabditis sp.4* (Nematoda). Full-length or nearly full-length copies of elements previously known only as fragments were also identified and/or assembled from the freshwater hydra *Hydra magnipapillata* (Cnidaria; Hydrozoa),

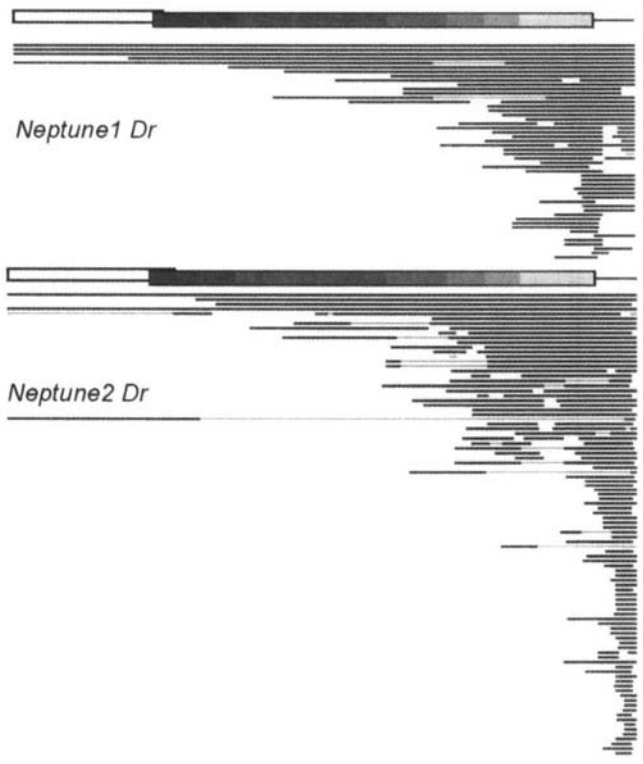

FIGURE 3. 5′ truncation in Penelope-like elements: Graphical output of a BLASTN search of the zebrafish genome assembly (http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=7955) with full-length sequences of two Neptune-like elements, Neptune1_Dr and Neptune2_Dr.

the planarian *Schmidtea mediterranea* (Platyhelminthes), the roundworms *Caenorhabditis remanei* and *Pristionchus pacificus* (Nematoda), and the purple sea urchin *Strongylocentrotus purpuratus* (Echinodermata). Partial copies were also assembled from trace archives of the following species: the sponge *Reniera sp.* (Porifera), the elkhorn corals *Acropora palmata* and *A. millepora* and the stony coral *Porites lobata* (Cnidaria; Anthozoa), the tuatara *Sphenodon punctatus*, and the green and brown anoles *Anolis carolinensis* and *A. sagrei* (Vertebrata; Lepidosauria). In addition, new PLEs from the tunicate *Oikopleura dioica* (Urochordata; Appendicularia) and *Perere10* from *Schistosoma mansoni* (Trematoda) were reported in the literature (Volff et al., 2004; DeMarco et al., 2005). The phylum Porifera thus becomes the most basal metazoan phylum in which bona fide PLEs have been identified, and the phylum Cndiaria already contains representatives of the two major PLE groups (see below).

The newly identified elements possess most of the features exhibited by canonical PLEs: a high degree of 5′ truncation, the presence of the GIY-YIG endonuclease domain, and the appearance of the full-length ORF only in copies preceded by the 3′ end of another copy (the "pseudo-LTR" structure). The main features of PLE ORF structure exemplified by the newly identified elements are summarized in Figure 1. The fact that most genomic copies are 5′ truncated underscores the similarity between PLEs and non-LTR retrotransposons, and also implies that PLEs would be equally unlikely to undergo excision, making them potentially useful as phylogenetic characters. The high degree of 5′ truncation also indicates that any *cis*-acting sequences required for reverse transcription and integration are expected to reside in the 3′ terminal part of the elements, as was previously shown for non-LTR retrotransposons (Luan and Eickbush, 1995). Figure 3 illustrates the degree of 5′ truncation for two *Neptune*-like elements in the zebrafish genome. It may be seen that the 3′ ends are significantly overrepresented; if an RNA polymerase III promoter becomes attached to any of such 3′ terminal fragments, it may have a potential to become a nonautonomous SINE-like element proliferating with the aid of an associated *Penelope*-like RT.

Alignment of N- and C-Terminal Domains of PLEs

Identification and assembly of multiple full-length copies of diverse PLEs makes it possible to extend the amino acid alignment to the N-terminal region (beyond the previously identified highly conserved DKG motif; Evgen'ev and Arkhipova, 2005), and to the linker domain of variable length connecting the RT and EN domains. To achieve this, the regions to the left and to the right of the last noticeable conserved motifs at the extremities of the RT domain were extracted and aligned separately. It may be seen that at least four additional conserved motifs appear in the N-terminal part of the PLE-encoded ORF (Fig. 4a). Furthermore, this region can also be aligned with the N-terminal domain of telomerases, strengthening the much closer relationship of PLEs with TERTs than with other retrotransposon-encoded RTs (Gladyshev and Arkhipova, submitted).

A secondary structure prediction for the region immediately downstream of the seven core RT domains consistently shows three conserved α-helices, a structure typical of the thumb domain of retroviral and group II intron RTs (Kohlstaedt et al., 1992; Blocker et al., 2005). Thus, although no sequence similarity is observed between PLEs and retroviral/group II intron RTs in this region, it is very likely that it indeed functions as a thumb domain, which in retroviruses is responsible for primer/template interactions and RT processivity.

Two additional conserved motifs, C1 and C2, may be discerned in the linker region between the thumb domain and the GIY-YIG EN domain (Fig. 4b). Interestingly, this region differs most dramatically between the two major phylogenetic groups of PLEs described below, *Neptune* and *Penelope-Poseidon*: in the *Neptune* group, the linker is C-terminally extended by 40 to 50 amino acids that include a well-conserved arrangement of cysteines with the CxxC core, which, in different elements, match various Zn finger domains in the SMART domain database, such as RING, ZnF_RBZ, ZnF_NFX, or ZnF_C2H2 (Fig. 4c). A somewhat shorter region (~20 amino acids) with the same arrangements of cysteines is found in PLEs from rhabditid and diplogasterid nematodes (Fig. 4c). It is tempting to speculate that the Zn-finger part of the linker region, positioned immediately upstream of the

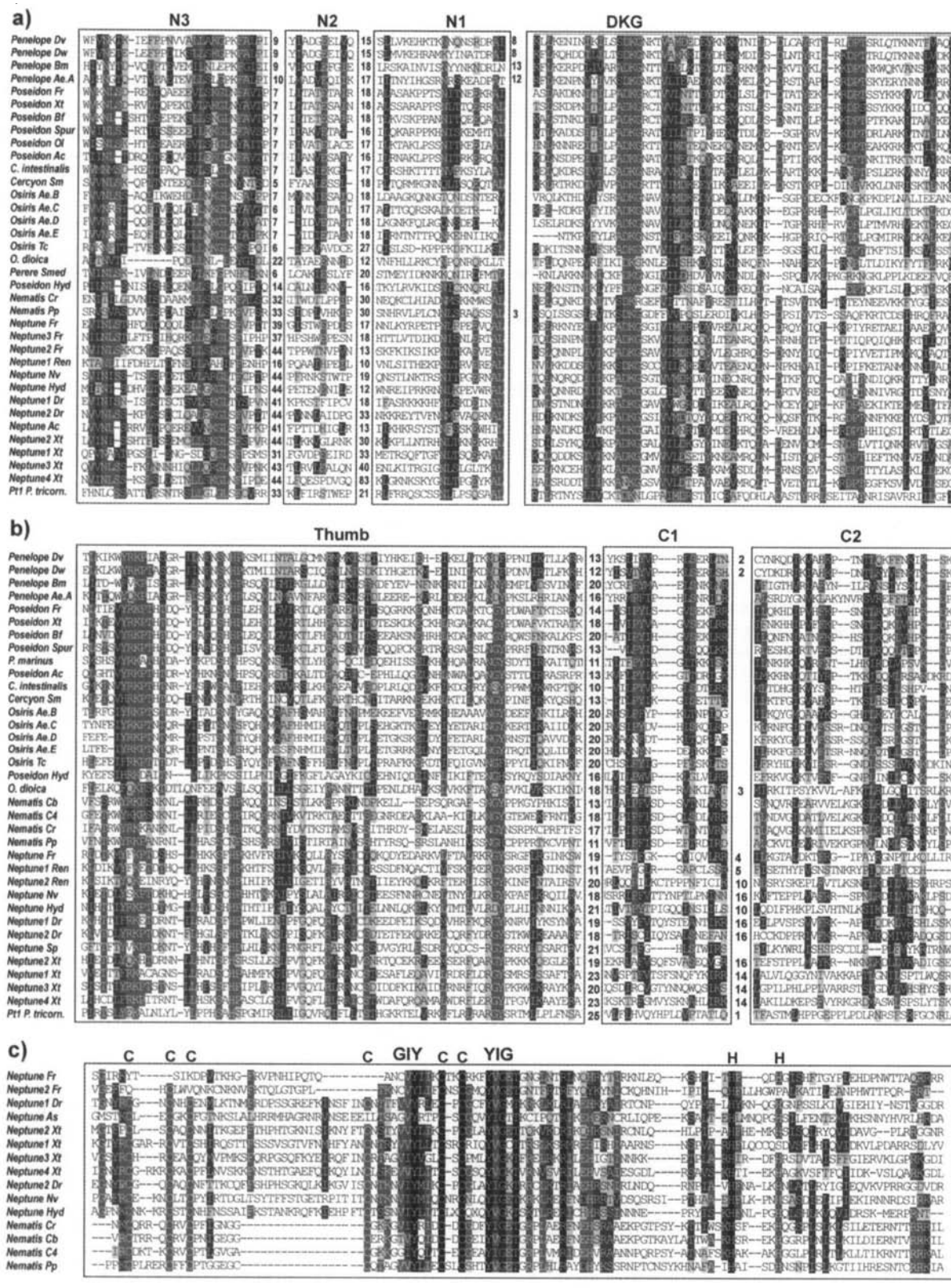

FIGURE 4. Amino acid sequence alignments of the regions immediately adjacent to the core RT domain: (a) the N-terminal domain; (b) the thumb domain; (c) the Zn-finger containing linker domain and the GIY-YIG domain of elements belonging to the *Neptune* and *Nematis* groups. Element names, accession numbers, and species assignment are listed in Table 1. The shading scheme is the same as in Figure 2. Only three N-terminal motifs, N1 through N3, are shown in (a); the fourth motif, N4, includes several basic residues and is not highly visible in intra-PLE comparisons but becomes evident in a combined alignment with telomerases. In (c), the first Zn finger in the linker appears only in *Neptune*-like and *Nematis*-like elements, while the second Zn finger (CC-HH) within the GIY-YIG domain is present in all EN-containing PLEs. The cysteine and histidine residues comprising the fingers are designated by letters C and H on the top.

GIY-YIG EN domain, could be somehow associated with the insertion bias (presumably EN-mediated) of the *Neptune*- and *Nematis*-like elements, which are mostly found adjacent to various microsatellite repeats (I.A., unpublished observations) and would therefore be less likely to cause insertional mutations.

Finally, an extremely high degree of conservation is observed for the C2H2 Zn finger domain, present in all EN-containing PLEs and unusually positioned right within the GIY-YIG domain, with the first two cysteines placed between the GIY and YIG moieties, and the two histidines within the middle domain C (Kowalski et al., 1999) (Fig. 4c). None of the other types of GIY-YIG endonucleases possess this unique arrangement of cysteines and histidines. First noted in four PLEs analyzed by Volff et al. (2001), it has now been found in the EN domains of virtually all of the canonical PLEs, suggesting its functional importance.

Major Clades Formed by PLEs

As mentioned above, early attempts to resolve phylogenetic relationships of PLEs were of limited success because of insufficient datasets, which included relatively few sequences and only the seven core RT motifs (~300 amino acids) supplemented with the thumb domain (~100 amino acids). If the acquisition of the GIY-YIG domain is regarded as a monophyletic event, resolution of phylogenetic relationships between PLEs may be improved if the entire PLE ORF is used for analysis. Because the separate phylogenies based on the RT and EN domains exhibit similar overall topologies (not shown), use of the entire ORF for inferring PLE phylogeny appears valid.

One of the first data sets, which included RT domains from six PLEs, divided the known elements into two major branches, with the *Neptune*-like elements in one branch and *Penelope-Poseidon* in the other (Volff et al., 2001). Extension of the aligned region from 300 to 400 to more than 700 amino acids results in significantly improved resolution of the PLE phylogeny. A few sequences from EST and genome databases, albeit incomplete, were also included to increase taxon sampling in cases where it could prevent long-branch attraction. The results of analysis of the extended dataset (50 taxa by 1000 characters, of which 760 are parsimony-informative) are presented in Figure 5. The tree was rooted with the *Phaeodactilum* and *Selaginella* PLEs; although these elements lack the EN domain and therefore do not provide a fully accurate rooting, it is nevertheless the best available option, because their placement appears basal to canonical PLEs in a combined RT phylogeny with telomerases (Gladyshev and Arkhipova, submitted). The phylogram in Figure 5 shows that all canonical PLEs can be subdivided into two major groups, *Neptune* and *Penelope-Poseidon*. A third, relatively minor group, named *Nematis*, consists of the nematode PLEs: although it sometimes tends to cluster with the Neptune group, as also supported by the presence of a Zn finger in the linker domain (Fig. 4c), its placement may vary when different methods are used, and they may in fact constitute a separate group. These large groups appear more or less equivalent in status to the five major groups distinguished in non-LTR retrotransposons by Eickbush and Malik (2002) (R2, L1, RTE, I, and Jockey, which are further subdivided into clades). Many species, such as *Hydra magnipapillata*, *Fugu rubripes*, *Oryzias latipes*, *Xenopus tropicalis*, and *Anolis carolinensis*, contain representatives of both major groups.

Continuing the analogy with the non-LTR retrotransposon classification of Eickbush and Malik, the large groups may be subdivided into distinct deep-branching clades, defined by the degree of sequence similarity and by inclusion of representatives of taxa that separated back in the pre-Cambrian era. At least six such major clades, and probably more, can be discerned. The *Neptune* group currently includes PLEs from sponges, cnidarians, fish, amphibians, and reptiles. It is quite possible that future expansions of the data set may result in designation of additional clades in this group. The *Poseidon-Penelope* group, in addition to the above-mentioned taxa, also contains elements from flatworms, mollusks, and arthropods. Members of the *Poseidon* clade are found in various deuterostomes, including echinoderms, cephalochordates, fish, amphibia, and reptiles. Addition of the *B. mori* PLE to the dataset helped to validate the *Penelope* clade, which also includes one of the PLE families from *A. aegypti*. The *S. mansoni Perere-10* element may be grouped with the planarian and mollusk PLE, forming the *Perere* clade. The placement of some elements that may be the only known members of a clade, such as PLEs from the cnidarian *Hydra magnipapillata* or the tunicate *Oikopleura dioica*, is uncertain, because of possible long-branch attraction. The *Nematis* clade from nematodes, as mentioned above, may form its own group. Overall, the tradition of naming clades after the first described representative appears most convenient: if any misplaced member of a named clade is later moved to a newly formed clade which receives better support, the original clade will still exist, because it will still contain the named original element. Such a system requires naming (and memorizing) only the first representative of a clade, which is of course much easier than naming all individual families.

The *A. aegypti* genome is notable for having a high diversity of PLE families: at least five major families can be discerned in this species, four of which form a novel clade named *Osiris*, and the fifth family belongs to the *Penelope* clade, as mentioned above. This finding is in line with the reputation of *A. aegypti* as a species exceptionally rich in TEs (Tu and Coates, 2003; Crainey et al., 2005).

Two additional early-branching clades, *Athena* and *Coprina*, in which all members possess only the RT domain and lack the EN domain, are not included into the present phylogeny, which is based on both RT and EN domains, and will be described in detail elsewhere (Gladyshev and Arkhipova, submitted).

The most plausible scenario of PLE evolution could include monophyletic acquisition by an ancestral *Penelope*-like RT of the GIY-YIG domain, probably with an N-terminal *Neptune*-like Zn finger domain, which was

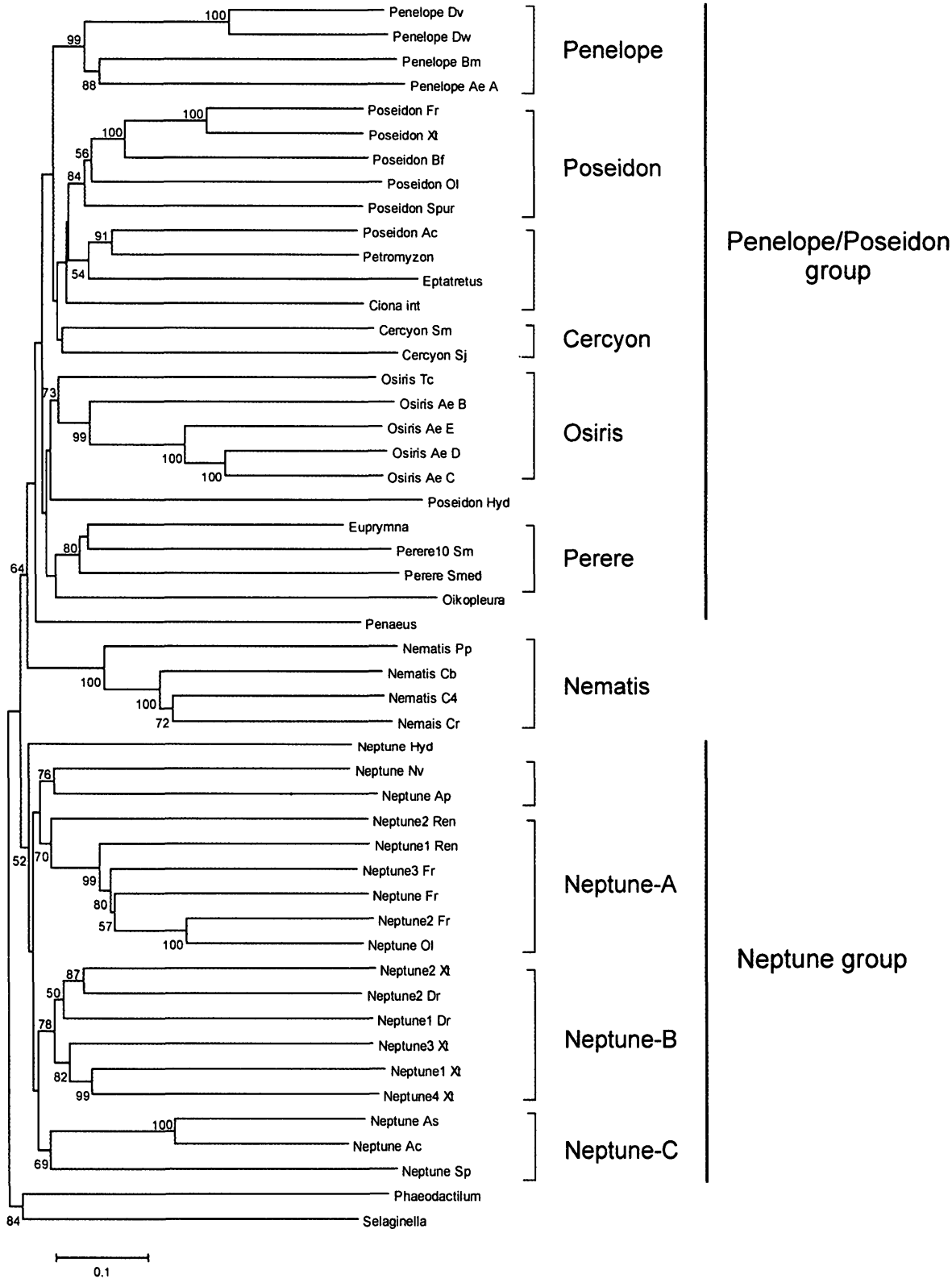

FIGURE 5. Phylogeny of *Penelope*-like elements based on the reverse transcriptase and endonuclease domains. The phylogeny is based on the data matrix consisting of 50 taxa by 1000 characters (including gaps), 760 of which are parsimony-informative. Shown is the 50% bootstrap consensus tree obtained by the neighbor-joining method (settings in MEGA 3.1: pairwise deletion, p-distance, uniform rates among sites, 1000 bootstrap replications), rooted with the RT sequences of protist and plant retroelements. The same tree topology is obtained by the minimum evolution method. Bootstrap support values exceeding 50% are shown. Each clade with more than 70% support value was named after its first described representative, as suggested by Malik et al. (1999). Clades with lower bootstrap support were left unnamed until more taxa become available.

TABLE 1. Phylogenetic distribution and source of PLE sequences from diverse eukaryotes. "Trace" indicates that the consensus sequence was assembled from trace reads and deposited in Repbase. If the clade assignment was insufficiently supported, the element was left unnamed until more data becomes available. Taxonomic placement was obtained from TaxBrowser (NCBI).

Phylum; class; (order)	Species	PLE name (if named)	Representative Accession No.
Bacillariophyta; Bacillariophyceae	*Phaeodactilum tricornutum*	*Pt1*	Trace
Mycetozoa; Myxogastria	*Physarum polycephalum*	*Pp1*	Trace
Streptophyta; Isoetopsida	*Selaginella moellendorffii*	*Sm1*	Trace
Basidiomycota; Homobasidiomycetes	*Coprinus cinereus*	*Coprina_Cc1*	AACS01000397
	Phanerochaete chrysosporium	*Coprina_Pc1*	AADS01000564
		Coprina_Pc2	AADS01000820
Porifera; Demospongiae	*Reniera sp. JGI-2005*	*Neptune1_Ren*	Trace
		Neptune2_Ren	Trace
Cnidaria; Hydrozoa	*Hydra magnipapillata*	*Neptune_Hyd*	Trace
		Poseidon_Hyd	Trace
Cnidaria; Anthozoa	*Nematostella vectensis*	*Neptune_Nv*	Trace
	Acropora palmata	*Neptune_Ap*	Trace
Platyhelminthes; Trematoda	*Schistosoma mansoni*	*Perere10_Sm*	BN000801
		Cercyon_Sm	BK000685
	Schistosoma japonicum	*Cercyon_Sj*	BU719939, BU804349
Platyhelminthes; Turbellaria	*Schmidtea mediterranea*	*Perere_Smed*	Trace
Nematoda; Chromadorea	*Caenorhabditis briggsae*	*Nematis_Cb*	CAAC01000421
	C remanei	*Nematis_Cr*	Trace
	C sp. 4	*Nematis_C4*	Trace
	Pristionchus pacifcus	*Nematis_Pp*	Trace
Mollusca; Cephalopoda	*Euprymna scolopes*	*Perere_Es*	DW286244
Arthropoda; Crustacea	*Penaeus monodon*		AF077579, DW042726
Arthropoda; Insecta	*Drosophila virilis*	*Penelope_Dv*	U49102
	D. willistoni	*Penelope_Dw*	AAQB01010750
	Bombyx mori	*Penelope_Bm*	BAAB01031930, BAAB01101440
	Tribolium castaneum	*Osiris1_Tc*	AAJJ01005571, AAJJ01007235
	Aedes aegypti	*Penelope_Ae.A*	AAGE02009673
		Osiris_Ae.B	AAGE02017473
		Osiris_Ae.C	AAGE02001225
		Osiris_Ae.D	AAGE02019145
		Osiris_Ae.E	AAGE02016919
Echinodermata; Echinoidea	*Strongylocentrotus purpuratus*	*Poseidon_Spur*	AAGJ02033054
Chordata; Urochordata	*Ciona intestinalis*		AABS01001109
	Oikopleura dioica		AY634216
Chordata; Cephalochordata	*Branchiostoma floridae*	*Poseidon_Bf*	BW883725
Chordata; Myxiniformes	*Eptatretus burgeri*		BJ646265, BJ655815
Chordata; Petromyzontiformes	*Petromyzon marinus*		AY577942
Chordata; Gnathostomata; Teleostei	*Fugu rubripes*	*Neptune_Fr*	CAAB01007056
		Neptune2_Fr	CAAB01007385
		Poseidon/Xena_Fr	CAAB01007635
	Oryzias latipes	*Neptune_Ol*	BAAF02119984
		Poseidon_Ol	BAAF02103523
	Danio rerio	*Neptune1_Dr*	NW_634120
		Neptune2_Dr	NW_634335
Chordata; Gnathostomata; Squamata	*Anolis sagrei*	*Neptune_As*	CF776418,776403 CK401349,722205
	Anolis caroinensis	*Neptune_Ac*	Trace
		Poseidon_Ac	Trace
	Sphenodon punctatus	*Neptune_Sp*	Trace
Chordata; Amphibia; Anura	*Xenopus tropicalis*	*Neptune1_Xt*	AC149155
		Neptune2_Xt	AC157689
		Neptune3_Xt	AC147888
		Neptune4_Xt	AC147895
		Poseidon_Xt	AC147885
Chordata; Mammalia; Insectivora	*Sorex araneus*	*Contamination is not ruled out	AALT01052298

considerably shortened in the nematode PLEs and completely eliminated in members of the *Penelope-Poseidon* group. Alternatively, one could entertain the less likely scenarios of polyphyletic acquisition of GIY-YIG domains by the three groups or insertion of the Zn finger domain into the linker region between the RT and EN domains in the *Neptune* and *Nematis* groups. Although the branching order of the major PLE clades depicted

in Figure 5 appears to agree with the first scenario, it should be kept in mind that these branches are so deep that confident resolution of their branching order may not be possible.

Modes of PLE Transmission

The overall PLE phylogeny is consistent with vertical transmission, accompanied by losses of PLE lineages from multiple species. This pattern was previously inferred for non-LTR retrotransposons, which are transmitted mostly vertically and experience occasional lineage losses (Eickbush and Malik, 2002). The evolutionary losses of PLE lineages appear to have occurred much more frequently than for non-LTR retrotransposons, as evidenced by their complete disappearance from many genomes. For example, PLEs were clearly lost from *C. elegans:* as judged by their presence in three other sequenced *Caenorhabditis* genomes, one may conclude that the lack of PLEs in *C. elegans* signifies their complete loss in this traditional model species. *C. briggsae* and *C. sp.4* are also apparently in the process of losing their PLEs, since in both cases all genomic copies are 5′ truncated, and only the *C. remanei* genome contains a full-length copy. Out of more than a dozen sequenced *Drosophila* genomes, only *D. virilis* and *D. willistoni* contain *Penelope*, which was previously shown to be present in five members of the virilis species group, in four of them only as remnants (Lyozin et al., 2001). PLE lineages in insects also experienced differential success: in terms of copy number, the *Penelope* clade flourished in *Bombyx mori*, but had limited success in *Aedes aegypti*, while members of the *Osiris* clade are well represented in shrimp and *Tribolium castaneum*, and diversified and reached thousands of copies in *Aedes* (I.A., unpublished observations). At the same time, the genomes of other sequenced insects—*Apis mellifera* and *Anopheles gambiae*—appear to be devoid of PLEs.

Because the PLE structure is most consistent with a TPRT-like mechanism of transposition, they are not expected to have a stable cytoplasmic DNA intermediate, making the possibility of their horizontal transmission between distant species very unlikely (Malik et al., 1999). Horizontal transfer between distant species would require a shuttle vector capable of transferring genetic information from nuclear DNA of one species to nuclear DNA of another species, such as a virus with a broad host range and the ability to integrate into genomic DNA of different hosts. In addition, other factors may also be at work, such as promoter compatibility with the new host. However, lateral transfer between more closely related species would be more likely, especially in related species with the potential for introgression.

Can PLEs be Regarded as non-LTR or LTR Retrotransposons?

The fact that PLEs exhibit properties usually regarded as characteristic features of either non-LTR or LTR retrotransposons often causes confusion by assigning them either to LTR retrotransposons (Volff et al., 2004) or LINEs (www.repeatmasker.org), correspondingly giving more weight either to the presence of LTR-like structures or to 5′ truncation and variable-length TSD indicative of the TPRT mechanism. Evidently, unambiguous assignment to either group is not possible on phylogenetic grounds: PLEs undoubtedly form a sister clade to TERTs and not to LTR or non-LTR retrotransposons (Arkhipova et al., 2003; Doulatov et al., 2005). Assignment of PLEs to LTR retrotransposons on the basis of a single character, i.e., the presence of "LTRs," would likely be erroneous: LTR retrotransposons arose relatively late in eukaryotic evolution (Malik and Eickbush, 2001), and acquired a complex replication cycle with a cytoplasmic reverse transcription stage, which places them apart from all other retroelements.

More consideration could be given to the similarity between PLEs and non-LTR retrotransposons, despite the fact that PLEs have "pseudo-LTRs." Because the TPRT mechanism would be typical not only of non-LTR retrotransposons and PLEs, but also of telomerases and group II introns (Zimmerly et al., 1995), it may be regarded as the ancestral universal mechanism, and the difference between non-LTR retrotransposons and PLEs essentially boils down to different types of endonucleases that cleave chromosomal DNA in order to generate a primer for TPRT: restriction enzyme-like or apurinic-apyrimidinic–like EN in non-LTR retrotransposons, and the GIY-YIG EN in PLEs. It may be suggested that PLEs and non-LTR retrotransposons are included into a larger eTPRT (eukaryotic target-printed) group of retrotransposons, characterized by 5′ truncation, variable target-site duplication, and the potential of the 3′ end to participate in formation of non-autonomous derivatives (see above).

Utility of PLEs as Phylogenetic Markers

As mentioned above, the main advantages of non-LTR (LINE) and SINE retrotransposable elements, which allow them to serve as essentially homoplasy-free phylogenetic markers, are their inability to undergo excision and the lack of integration preference (with the exception of certain non-LTR retrotransposons that code for a site-specific endonuclease). With the ancestral state defined as the absence of a particular insertion, the locus in question may be examined for the presence/absence of a retrotransposable element in this position (Shedlock and Okada, 2000). This approach is useful for determining relationships between relatively close species, which possessed retrotransposons that were active during the period when speciation occurred. In more distant species, especially in those with high overall rate of DNA turnover, the differences may be obscured by subsequent DNA loss and/or rearrangement.

If present in a given genome in multiple copies, PLEs do fulfill most of the criteria (inability to excise, relatively random character of insertion, nonautonomous nature of most copies, apparently rare horizontal transfers,

and sufficiently large copy numbers) for serving as phylogenetic markers, as an alternative or complement to the more numerous and diverse non-LTR retrotransposons. This is especially true for those species in which PLEs proliferated to a high copy number and left a substantial molecular fossil record, such as yellow fever mosquito, fish, amphibia, or reptiles. On the other hand, in many species PLEs are present at a very low copy number, often as a single master copy that gives rise to very few daughter copies, which makes them much less suitable for these purposes. In addition, it should be kept in mind that PLEs are often surrounded by other kinds of repetitive sequences, which represent a fluid genomic component and are not highly suitable for analysis of insertion polymorphisms.

It is also worth mentioning that, like other types of transposable elements, PLEs may give rise to highly repetitive satellite DNAs, which can by themselves be useful as phylogenetic markers. One of the first examples of non-drosophilid PLEs was a microsatellite sequence from shrimp (Xu et al., 1999; Volff et al., 2001), represented by a PLE fragment amplified to a large extent. A similar capture of mobile element fragments during minisatellite formation and amplification can be observed for virtually every class of mobile elements, including DNA transposons (Lopez-Flores et al., 2004), LINEs (Kapitonov et al., 1998), SINEs (Batistoni et al., 1995), and LTR retrotransposons (Kelly, 1994; Tek et al., 2005). On balance, the applicability of PLEs as sufficiently high copy number phylogenetic markers for every new species (excluding perhaps only birds and mammals) needs to be assessed for every individual species, via PCR screens based on degenerate primers and/or exploratory random sequencing of genomic/cDNA libraries. Indeed, our collaborative studies aimed at detecting PLEs in various invertebrate species via such PCR screens, with the purpose of exploiting them as molecular markers, yielded positive results in several relatively little studied invertebrates, such as the red swamp crayfish *Procambarus clarkii* and the black marsh fly *Plecia nearctica* (J. Doucet, personal communication).

CONCLUSIONS

This study represents the first comprehensive assessment of the phylogeny of *Penelope*-like elements, based on extended alignments and a considerably expanded dataset, which will hopefully serve as a reference dataset for building future PLE phylogenies as more sequence data accumulates. PLEs appear to exhibit essentially the same pattern of evolution as non-LTR retrotransposons, which form deep-branching clades dating back to the Precambrian era, although PLEs seem to have experienced a much higher degree of lineage losses than non-LTR retrotransposons. It is suggested that PLEs and non-LTR retrotransposons are included into a larger eT-PRT (eukaryotic target-primed) group of retroelements, characterized by 5′ truncation, variable target-site duplication, and the potential of the 3′ end to participate in formation of non-autonomous derivatives.

ACKNOWLEDGMENTS

I wish to thank Andrew Shedlock for his efforts in organizing the SSB symposium "Genome Analysis and Molecular Systematics of Retroelements" in Fairbanks and editing this special issue of *Systematic Biology*, as well as for stimulating discussions; and Cedric Feschotte and Vladimir Kapitonov for constructive criticisms of the manuscript. This work was supported by the U.S. National Science Foundation (MCB 0614142).

REFERENCES

Arkhipova, I. R., K. I. Pyatkov, M. Meselson, and M. B. Evgen'ev. 2003. Retroelements containing introns in diverse invertebrate taxa. Nat. Genet. 33:123–124.

Batistoni, R., G. Pesole, S. Marracci, and I. Nardi. 1995. A tandemly repeated DNA family originated from SINE-related elements in the European plethodontid salamanders (Amphibia, Urodela). J. Mol. Evol. 40:608–615.

Blocker, F. J., G. Mohr, L. H. Conlan, L. Qi, M. Belfort, and A. M. Lambowitz. 2005. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. RNA 11:14–28.

Crainey, J.L., C.F. Garvey, and C.A. Malcolm. 2005. The origin and evolution of mosquito APE retroposons. Mol. Biol. Evol. 22:2190–2197.

Dalle Nogare, D. E., M. S. Clark, G. Elgar, I. G. Frame, and R. T. Poulter. 2002. Xena, a full-length basal retroelement from tetraodontid fish. Mol. Biol. Evol. 19:247–255.

DeMarco, R., A. A. Machado, A. W. Bisson-Filho, and S. Verjovski-Almeida. 2005. Identification of 18 new transcribed retrotransposons in *Schistosoma mansoni*. Biochem. Biophys. Res. Commun. 333:230–240.

Doulatov, S., A. Hodes, L. Dai, N. Mandhana, M. Liu, R. Deora, R. W. Simons, S. Zimmerly, and J. F. Miller. 2004. Tropism switching in Bordetella bacteriophage defines a family of diversity-generating retroelements. Nature 431:476–481.

Duncan, L., K. Bouckaert, F. Yeh, and D. L. Kirk. 2002. Kangaroo, a mobile element from *Volvox carteri*, is a member of a newly recognized third class of retrotransposons. Genetics 162:1617–1630.

Eickbush, T. H, and H. S. Malik. 2002. Origin and evolution of retrotransposons. Pages 1111–1144 *in* Mobile DNA II (N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds.). Washington, DC: ASM Press.

Evgen'ev, M. B., and I. R. Arkhipova. 2005. Penelope-like elements–a new class of retroelements: Distribution, function and possible evolutionary significance. Cytogenet. Genome Res. 110:510–521.

Evgen'ev, M. B., H. Zelentsova, N. Shostak, M. Kozitsina, V. Barskyi, D. H. Lankenau, and V. G. Corces. 1997. Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in *Drosophila virilis*. Proc. Natl. Acad. Sci. USA 94:196–201.

Goodwin, T. J., and R. T. Poulter. 2004. A new group of tyrosine recombinase-encoding retrotransposons. Mol. Biol. Evol. 21:746–759.

Ho, H. J., D. A. Ray, A. H. Salem, J. S. Myers, and M. A. Batzer. 2005. Straightening out the LINEs: LINE-1 orthologous loci. Genomics 85:201–207.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110:462–467.

Kapitonov, V. V., G. P. Holmquist, and J. Jurka. 1998. L1 repeat is a basic unit of heterochromatin satellites in cetaceans. Mol. Biol. Evol. 15:611–612.

Kapitonov, V. V., and J. Jurka. 1999. bridge1_fr. Repbase Update, July 1999.

Kapitonov, V. V., and J. Jurka. 2001. Rolling-circle transposons in eukaryotes. Proc. Natl. Acad. Sci. USA 98:8714–8719.

Kapitonov, V. V., and J. Jurka. 2006. Self-synthesizing DNA transposons in eukaryotes. Proc. Natl. Acad. Sci. USA 103:4540–4545.

Kelly, R. G. 1994. Similar origins of two mouse minisatellites within transposon-like LTRs. Genomics 24:509–515.

Kohlstaedt, L. A., J. Wang, J. M. Friedman, P. A. Rice, and T. A. Steitz. 1992. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science 256:1783–1790.

Kowalski, J. C., M. Belfort, M. A. Stapleton, M. Holpert, J. T. Dansereau, S. Pietrokovski, S. M. Baxter, and V. Derbyshire. 1999. Configuration of the catalytic GIY-YIG domain of intron endonuclease I-TevI: Coincidence of computational and molecular findings. Nucleic Acids Res. 27:2115–2125.

Kriegs, J. O., G. Churakov, M. Kiefmann, U. Jordan, J. Brosius, and J. Schmitz. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. 4:e91.

Kumar, S., K. Tamura, and M. Nei. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinformat. 5:150–163.

Lopez-Flores, I., R. de la Herran, M. A. Garrido-Ramos, P. Boudry, C. Ruiz-Rejon, and M. Ruiz-Rejon. 2004. The molecular phylogeny of oysters based on a satellite DNA related to transposons. Gene 339:181–188.

Luan, D. D., and T. H. Eickbush. 1995. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. Mol. Cell Biol. 15:3882–3891.

Lyozin, G. T., K. S. Makarova, V. V. Velikodvorskaja, H. S. Zelentsova, R. R. Khechumian, M. G. Kidwell, E. V. Koonin, and M. B. Evgen'ev. 2001. The structure and evolution of Penelope in the virilis species group of *Drosophila*: An ancient lineage of retroelements. J. Mol. Evol. 52:445–456.

Malik, H. S., W. D. Burke, and T. H. Eickbush. 1999. The age and evolution of non-LTR retrotransosable elements. Mol. Biol. Evol. 16:793–805.

Malik, H. S., and T. H. Eickbush. 1999. Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons. J. Virol. 73:5186–5190.

Page, R. D. M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12: 357–358.

Robertson, H. M. 2002. Evolution of DNA transposons in eukaryotes. Page 1093–1110 *in* Mobile DNA II (N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, (eds). ASM Press, Washington, D.C.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Shedlock, A. M, and N. Okada. 2000. SINE insertions: Powerful tools for molecular systematics. Bioessays 22:148–160.

Shedlock, A. M., K. Takahashi, and N. Okada,2004. SINEs of speciation: Tracking lineages with retroposons. Trends Ecol. Evol. 19:545–553.

Simpson, A. G., and A. Roger. 2004. The real "kingdoms" of eukaryotes. Curr. Biol. 14:R693–R696.

Tek, A. L., J. Song, J. Macas, and J. Jiang. 2005. Sobo, a recently amplified satellite repeat of potato, and its implications for the origin of tandemly repeated sequences. Genetics 170:1231–1238.

Tu, Z., and C. Coates. 2004. Mosquito transposable elements. Insect. Biochem. Mol. Biol. 34:631–644.

Volff, J. N., U. Hornung, and M. Schartl. 2001. Fish retroposons related to the *Penelope* element of *Drosophila virilis* define a new group of retrotransposable elements. Mol. Genet. Genomics 265:711–720.

Volff, J. N., H. Lehrach, R. Reinhardt, and D. Chourrout. 2004. Retroelement dynamics and a novel type of chordate retrovirus-like element in the miniature genome of the tunicate *Oikopleura dioica*. Mol. Biol. Evol. 21:2022–2033.

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

Xu, Z., A.K. Dhar, J. Wyrzykowski, and A. Alcivar-Warren. 1999. Identification of abundant and informative microsatellites from shrimp (*Penaeus monodon*) genome. Anim. Genet. 30:150–156.

Zimmerly, S., H. Guo, P. S. Perlman, and A. M. Lambowitz. 1995. Group II intron mobility occurs by target DNA-primed reverse transcription. Cell 82:545–554.

First submitted 31 August 2006; reviews returned 28 September 2006; final acceptance 17 October 2006

Associate Editor: Andrew Shedlock