R5 Retrotransposons Insert into a Family of Infrequently Transcribed 28S rRNA Genes of Planaria

William D. Burke, Daljit Singh, and Thomas H. Eickbush

Department of Biology, University of Rochester

A small (100 bp) region of the 28S rRNA gene has been shown to serve as the target site for the insertion of non–long terminal repeat (non-LTR) retrotransposons in both arthropods and nematodes. Here we characterize a lineage of non-LTR retrotransposons that inserts into this target site in the phylum Platyhelminthes. Dugesian planaria contain elements, named R5, that insert 8 bp upstream of the target site used by arthropod R2 elements. The complete sequence of this element from Girardia tigrina revealed that it encoded two open reading frames (ORFs). The second ORF contained reverse transcriptase and restriction enzyme–like endonuclease domains similar to those found in R2 and R4, the elements that insert into the 28S genes of nematodes. The closest relative of R5, however, was the element NeSL-1, which inserts into the spliced leader 1 exons of nematodes. The RNA genes of dugesian planaria are unusual in that they comprise two types of rDNA units that differ by 8%–10% in nucleotide sequence of the 18S and 28S coding regions. Type II units are transcribed in adult tissues at levels that are less than 1% of the type I units. R5 elements were only found inserted in the type II units, where presumably they cause less harm to the host. A second unusual aspect of the dugesian RNA genes is that the target site for the R5 insertion is duplicated 300 bp upstream of the original insertion site. R5 elements were identified at both sites. These findings expand the distribution of non-LTR elements that are specialized for insertion into the 28S gene and suggest that still more elements exist in other eukaryotic taxa. Attempts to trace the phylogeny of R5 did not offer sufficient resolution to determine whether R2, R4, and R5 represent the same lineage or whether they represent independent specializations for the 28S gene.

Introduction

Transposable elements are abundant components of most eukaryotic genomes. Although most elements transpose with little specificity, a significant number integrate into specific locations within a genome. The evolution of target specificity can be advantageous to both the host and the element. First, specific insertions into the genome can cause less harm than the inevitable disruption of gene expression brought about by random insertions. Second, specific insertion sites can give rise to greater opportunities for an element to control expression of its many copies. Presumably because of their long-term associations with host lineages, retrotransposons have evolved target specificity more frequently than transposons (Eickbush and Malik 2002). Long terminal repeat (LTR) retrotransposons have evolved site specificity by adapting their integrase to form specific interactions with chromosomal proteins. Well-characterized examples are the retrotransposon families of Saccharomyces cerevisiae, which insert either upstream of tRNA genes or within silent chromatin (Kirchner, Connolly, and Sandmeyer 1995; Devine and Boeke 1996; Zhu et al. 1999). In contrast, non-LTR retrotransposons have evolved site specificity by adapting the endonuclease that initiates the integration reaction to recognize specific DNA sequences (Xiong and Eickbush 1988; Feng, Schulmann, and Boeke 1998; Christensen, Pont-Kingdom, and Carroll 2000; Takahashi and Fugiwara 2002).

Two well-characterized site-specific non-LTR retrotransposons are the R1 and R2 elements of arthropods (Eickbush 2002). These elements insert into sites, 74 bp apart, in a highly conserved region of the 28S rRNA genes (fig 1A). The insertion of either element significantly reduces the transcription of the entire rRNA transcription unit (rDNA unit). Although they share similar insertion strategies, the elements are only distantly related, with R1 elements encoding an apurinic-like endonuclease (APE), while R2 elements encode an endonuclease with a restriction enzyme–like active site. Thus, these elements appear to have independently evolved target specificity for the 28S rRNA gene.

Phylogenetic analysis of R1 and R2 elements from organisms throughout the phylum arthropoda suggest that these elements have been inserting into the 28S rRNA genes since the origin of this phylum (Burke et al. 1998; Gentile et al. 2001). Within insects, the R1 lineage has given rise to elements that have changed their insertion specificity for sites within the rDNA unit or for sites outside the rDNA unit (Besansky et al. 1992; Kojima and Fujiwara 2003). A third rDNA element, R3, inserts a short distance upstream of the R2 site in some insects, but to date no complete element has been found (Kerrebrock, Srivastava, and Gerbi 1989; Burke et al. 1993). The only non-arthropod element identified to date which inserts into the rDNA unit is the R4 element of nematodes. R4 elements insert into the large subunit rRNA gene at a site midway between the R1 and R2 insertion sites and are similar in structure to R2 elements (Neuhaus et al. 1987; Burke, Müller, and Eickbush 1995).

In this report we describe another R2-like element, named R5, which inserts into the large subunit rRNA gene of the more primitive organism, planaria (phylum Platyhelminthes). An unusual aspect of this system is that the rDNA units of the planaria host are composed of two diverging families (Carranza, Baguna, and Riutort 1999). One family of units is actively transcribed, whereas the second family is only infrequently transcribed. R5 elements were found only in the rarely transcribed rDNA units.
Materials and Methods

Organisms and Nucleic Acid Extraction

*Girardia tigrina*, *Girardia dorotocephala*, and *Pro- cotyla fluviatilis* were obtained from Ward’s Biological Supplies (Rochester, N.Y.). Nucleic acid was isolated from 4–5 animals by the same procedures used in the isolation of genomic DNA from *Drosophila* (Perez-Gonzalez and Eickbush 2002), except that the worms were first ground into a powder in liquid nitrogen. After initial precipitation with ethanol, DNA was isolated by digestion of one half of the nucleic acid at 37°C for 40 minutes with 50 units RNase A (Ambion, Inc.). RNA was isolated from the remainder by digestion with 20 units RNase-free DNase I (Ambion, Inc.). The individual digestions were phenol extracted and ethanol was precipitated for a second time. The integrity and yield of the RNA and DNA preparations were checked on 1% agarose gels.

PCR Amplification and DNA Sequencing

A 1.0-kb region of the 28S rRNA genes was subjected to polymerase chain reaction (PCR) amplification using primers GGAAGTCGGC AAATTAGAT (ribo up420) and AAGAGCCGACATCGAAGGATC, which anneal to highly conserved regions in domains five and six of eukaryotic large subunit rRNA genes (nomenclature as in Hancock, Tautz, and Dover 1988). To confirm that the 28S rRNA genes with insertions corresponded to the type II rDNA units identified by Carranza, Baguna, and Riutort (1999), the primer TCTAGTCTAAGAAATGGC, which annealed specifically to the type II 18S rRNA sequences of *G. tigrina*, was used in combination with the primer CCGAGGAAAGCTCAAGTC, which annealed specifically to the 28S gene sequences with R5 insertions. Evidence for the R5 element in *G. tigrina* was first obtained by inverse PCR (Ochman, Gerber, and Hartl 1988) using Sau3a-digested genomic DNA and divergently oriented primers that anneal to 28S gene sequences downstream of the insertion region (Jakubczak, Burke, and Eickbush 1991). After this initial amplification recovered ~190 bp at the 3’ end of the element, the 3’ half of the element was obtained by PCR amplification using the primer GCNTWWGC NGAYGAY (N = any nucleotide, W = A or T, Y = T or C), which encodes the highly conserved amino acid sequence AF/YADD found in most eukaryotic large subunit rRNA genes.

![Diagram](https://academic.oup.com/mbe/article-abstract/20/8/1260/1081529)
non-LTR retrotransposons, and the primer GTTATAACATTAGAAGGAGGT, located in the region of the R5 recovered by inverse PCR. Based on the sequence of the 3’ half of R5, the primer CTGAGGCTTTTTGT GGGAGTGC located downstream of the AYADD region and directed toward the 5’ end of the element was used in combination with the primer TCTGCCCAGTGCCATGAATGTC, which anneals to the 28S gene 70 bp upstream of the R5 insertion site. To insure that full-length R5 elements were cloned, the reverse primer GCAAGAAATCGAAAACCT-CTGCTC which anneals to a sequence ~2.0 kb upstream of the AYADD region was used in combination with the ribo up420 28S primer. To clone multiple 3’ junctions of the R5 element, a primer near the 3’ end of the R5 element, ACCCGATATCACAATGGAATC, was used in combination with the primer CGTAAATCCATTCGAG CACG (ribo down60 REV). All PCR amplifications were for 30 cycles with 55°C annealing temperatures in 20 mM Tris-HCl (pH 8.3), 50 mM KCl, 0.2 mM dNTPs, 1 mM MgCl2, 0.25 μM of each primer, and 1.25 units of Taq DNA polymerase (Gibco-BRL) (Perez-Gonzalez and Eickbush 2002). PCR products to be sequenced were cloned into a modified mp18 vector (Burke, Müller; and Eickbush 1995). GenBank accession numbers for the 28S gene sequences are: G. tigrina type I, AY216702; G. tigrina type II, AY216703; G. dorotocephala type I, AY216704; G. dorotocephala type II, AY216705; P. flaviatilis AY216706, and for the G. tigrina R5 element, AY216701.

Reverse Transcriptase-PCR Reactions

For the reverse transcriptase (RT) reaction ~0.1 μg aliquots of total RNA from each species were annealed with the primer 5’-CAGTGGATTCTTTAGTCC-3’ (RT PCR REV) by first heating to 70°C for 5 min and then cooling to room temperature. Primer extension reactions were conducted in 30 μl volumes with 15 units of AMV reverse transcriptase (Promega) in 50 mM Tris-HCl (pH 8.3), 50 mM KCl, 5 mM MgCl2, 5 mM DTT, 0.5 mM spermidine and 250 μM of each dNTP for 1 hour at 42°C. The extension reaction was diluted 3-fold and 1 μl aliquots were PCR amplified as described above for genomic DNA using primer RT PCR REV and Ribo up420, except that the RT PCR REV primer was P32-end labeled (Perez-Gonzalez and Eickbush 2002) and only 18 cycles of amplification were conducted. The products were separated on high voltage denaturing 8% polyacrylamide gels. Dried gels were exposed to PhosphorImager cassette and the relative intensities of each band determined on a Molecular Dynamics Storm Analyzer using Imagequant 1.2.

Phylogenetic Analysis

All sequences used in the phylogenetic analysis were obtained from our previous reports (Malik, Burke, and Eickbush 1999; Burke et al. 2002) except for R5 and EhRLE from Entamoeba histolytica (Sharma et al. 2001). Protein domains of the elements were aligned using the multiple alignment options in CLUSTAL X (Thompson et al. 1997), followed by minor manual adjustments of gaps. Phylogenetic trees were generated by the Neighbor-Joining method using the PAM250 matrix of PHYLIP (Felsenstein 1993), 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 (Swofford 1999).

Results

Two Types of 28S rRNA Genes in Planaria

Previous experiments by Carranza and co-workers (Carranza et al. 1996; Carranza, Baguna, and Riutort 1999) have suggested that the rDNA units of species from the family Dugesidae can be divided into two distinct families (type I and type II). Species from the sister families Planariidae and Dendrocoelidae contain a single type of rDNA unit. Nucleotide divergence within each rDNA type from the dugesiid species is minimal, whereas divergence between the rDNA types in a species averaged 9% for the 18S rRNA coding region. Comparison of the ITS regions between the two rDNA types revealed numerous indels, which made alignment of the sequences difficult.

We have extended this analysis to the region of the 28S gene that includes the R1–R4 insertion sites in arthropods and nematodes (fig. 1A). Primers to conserved regions within domains 5 and 6 of the 28S gene were used to amplify 28S sequences from three species of planaria. This amplified region extends from 400 bp upstream of the R1–R4 insertion region to 600 bp downstream of this region. Multiple individually cloned products were then sequenced. Sequence comparison of the 5’ half of this region including the R1–R4 insertion sites is shown in figure 1B. Two distinct 28S gene sequences were found in Girardia tigrina and Girardia dorotocephala, but only one 28S gene sequence was recovered from the more distant Procotyla flaviatilis.

G. tigrina and G. dorotocephala are closely related species in the family Dugesidae (Carranza, Baguna, and Riutort 1999). The family association of P. flaviatilis is not known, but based on its 28S gene sequence divergence from the other two species, it is unlikely to be a member of the Dugesidae family. The level of nucleotide divergence between the type I and type II 28S genes from the same species was 7.2% (G. dorotocephala) and 8.0% (G. tigrina). Most of the nucleotide substitutions between type I and type II units in G. tigrina and G. dorotocephala were located within an expansion region D8 of the 28S gene (nucleotides 70–260 in fig. 1B). Expansion regions of rRNA genes are under reduced selective constraint and often contain length variation between species (reviewed in Hillis and Dixon 1991). A 23-bp insertion was found in the type II 28S genes of both G. tigrina and G. dorotocephala that corresponded to the duplication of a region located nearly 300 bp downstream (boxed regions in fig. 1B).

The type II 18S genes in Dugesidae were found to be accumulating nucleotide substitutions 2.3 times faster than type I repeats (Carranza, Baguna, and Riutort 1999). The 28S gene sequences reported here also suggested a faster rate of divergence for the type II units. In the ~1,000 bp
Type II genes were twice as abundant as the type I genes in DNA confirmed that type I and II 28S genes were present by PCR amplification of genomic DNA. Transcripts could be estimated. Shown in figure 2 are the types of 28S genes, the relative proportion of the two Ribo up420. Because of the 23-bp indel between the two primer (RT PCR REV) in combination with the primer transcription the cDNA was amplified using the same PCR REV, fig. 1 provided the means to directly compare the relative levels of transcripts from the two types of genes. The primer (RT PCR REV) with the latter P32 end-labeled. For each species: lane 1, complete RT-PCR reaction; lane 2, RT-PCR reaction but no reverse transcriptase added in the first step; lane 3, direct PCR amplification of genomic DNA. Products from transcripts of the type II genes are barely visible at levels less than 0.5% of that from type I genes. The DNA sequencing ladder of the m13 vector at right provides a size standard.

Previous attempts to monitor expression of the 18S gene from the type I and II units revealed no type II transcripts on Northern blots of RNA (Carranza et al. 1999; and data not presented). Our first evidence for the presence of an insertion in the 28S rRNA genes of G. tigrina was obtained by inverse PCR (see Materials and Methods). The G. tigrina insertion appeared to be a non-LTR retrotransposon because it ended in a series of short tandem repeats, similar to the tandem repeats found at the 3' end of some non-LTR retrotransposons (Eickbush and Malik 2002). To clone and sequence a complete R5 element, PCR amplification was conducted with a degenerate primer to the highly conserved AY/FADD protein motif within the reverse transcriptase domain of non-LTR retrotransposons and a primer to the sequenced region of the R5 element. To clone the 5' half of the element, a second set of PCR amplifications was conducted using primers to sequenced regions from the 3' half of the element and primers to the 28S gene upstream of the R5 insertion site. This approach revealed a second characteristic that R5 elements shared with non-LTR retrotransposons: many copies of R5 were 5' truncated (i.e., were missing variable lengths from their 5' end). For example, R5 copies were detected that were only 0.3 kb in length. The largest PCR fragments generated by amplifying genomic DNA with R5 primers directed to the 5' end of the element and upstream 28S gene sequences presumably revealed full-length elements (see Materials and Methods). The 5' and 3' junction sequences of multiple R5 copies with the 28S gene are summarized in figure 3. Surprisingly R5 insertions were located at two sites. One site was located 8 bp upstream of the R2 insertion site (nucleotide position 421 in fig. 1B). The 28S gene sequences flanking the 3' junction of the R5 insertions contained an A residue 5 bp downstream of the insertion site and a C residue 45 bp downstream of this site, indicating that the insertions were located in type II rDNA units. The second R5 insertion site was located 287 bp upstream of the first site (nucleotide position 134 in fig. 1). This upstream site was within the duplicated sequence of multiple R5 elements in both insertion sites.

Insertion Properties of the R5 Elements

The 5' and 3' ends of R5 elements in both insertion sites were composed of 33–80 bp of a simple repeat, Ag(T)ₙ, where n equals 2 to 8 nucleotides. Based on studies of the R2 and I element retrotransposition mechanisms

Fig. 2.—An RT-PCR assay showing the relative transcript levels of the type I and type II 28S genes. cDNA was first made by reverse transcription of total RNA using the primer RT-PCR REV (see fig. 1). This cDNA was then PCR amplified using the primers Ribo up420 and RT PCR REV with the latter P32 end-labeled. For each species: lane 1, complete RT-PCR reaction; lane 2, RT-PCR reaction but no reverse transcriptase added in the first step; lane 3, direct PCR amplification of genomic DNA. Products from transcripts of the type II genes are barely visible at levels less than 0.5% of that from type I genes. The DNA sequencing ladder of the m13 vector at right provides a size standard.
Comparison of the 5′ junction sequences of the R5 insertions with the uninserted 28S gene at each site indicated that one base pair was deleted during the process of insertion. However, because the T shown in figure 3 as indicated that one base pair was deleted during the process of insertions with the uninserted 28S gene at each site engage the RNA template for cDNA synthesis.

Two R5 insertions also generated additional deletions or replacements of the 28S gene upstream of the insertion (fig. 3A). Two base pair deletions of the target site with occasional more extensive deletions and replacements are also characteristic of R2 insertions in insects (George, Burke, and Eickbush 1996; Burke et al. 1999).

Like our findings with R1, R2, and R4, the level of nucleotide divergence of different R5 elements was low. With sequences from near the 3′ end of the element to enable different copies to be identified by their variable 3′ tails, the average level of nucleotide sequence divergence between 12 different R5 elements was 1.4% (range 0.4%–2.3%). Because our DNA was derived from 4 animals recently collected from the wild, this number is likely to represent the average level of R5 divergence within a population of G. tigrina. It should be noted that in our sequencing of different R5 copies we also detected a second family of R5 elements in G. tigrina that exhibited a 19.9% nucleotide divergence from the first family. Although a complete copy of this second family was not recovered, this second family was useful in establishing the ORF structure of the R5 elements (see below).

To determine what fraction of the 28S genes of G. tigrina were inserted with R5, we conducted a Southern blot in which we probed Rsal-digested genomic DNA with the 28S gene region downstream of both R5 insertion sites (fig. 4D, probe A). Uninserted 28S genes of type I gave rise to a 0.55-kb band, and uninserted genes of type II gave rise to a 0.77-kb band. As shown in fig. 4A, the relative proportion of these two rDNA types was similar to that revealed by the PCR analysis of genomic DNA in figure 2. R5 elements at the downstream insertion site of the type II 28S genes gave rise to a 1.0-kb band which is barely visible in figure 4A. This band was at a level only 1% that of the uninserted type II band. A band at 1.3 kb corresponding to R5 insertions at the upstream site was not visible on the blot.

As a more sensitive means to determine the relative level of R5 insertions in the upstream and downstream insertion sites, we conducted a PCR amplification using one primer to the 28S gene downstream of both insertion sites and a primer near the 3′ end of the R5 element (diagrammed in figure 4D). As shown in figure 4B, lane 1, R5 insertions in the upstream and downstream insertion sites gave rise to bands at approximately 1.1 and 0.8 kb, respectively. To confirm that the upper band corresponded to R5 insertions at the upstream 28S gene site, a second 28S primer that annealed upstream of the downstream site was used in combination with the R5 primer (~1.0-kb band in lane 2). This PCR analysis indicated that there were 5–10 times as many R5 elements located at the downstream site as at the upstream site, explaining why no inserted band was seen for the upstream site in figure 4A.

Finally, to determine what fraction of the R5 sequences in G. tigrina are located in the 28S genes, Splh–CiaI digested genomic DNA was probed with a segment of DNA from near the 3′ end of the R5 element (fig. 4D, probe C). The major 1.4-kb and 1.7-kb bands seen in this blot (fig. 4C) corresponded, respectively, to R5 insertions at the down and up sites of the type II 28S genes. Only a few faint upper bands (more visible on longer exposures) correspond to R5 elements with restriction polymorphisms, extreme 5′ truncations, or locations outside the rDNA units.

Structure of R5 and Its Relationship to Other Non-LTR Retrotransposons

The longest R5 elements recovered were 4.8 kb in length and encoded two ORFs (fig. 5). The 3′ untranslated region, excluding the variable length AG (T) n tail, was only 132 bp in length, whereas the 5′ UTR was only 8 bp in length; thus these ORFs occupied over 96% of the element’s length. The second R5 ORF encoded a central RT domain with fingers, palm, and thumb subdomains.
typical of other non-LTR retrotransposons (Burke, Malik, and Eickbush 1999). Located at the C-terminal end of the second ORF was a restriction-like endonuclease (EN) domain that was similar to that of R2 and R4 elements. Within the EN domain were two identifiable motifs, a C-X₇-C-X₄-H-X₄-C sequence that is potentially a nucleic acid binding motif and a PD-X₁₂-D motif which forms part of the active site of the enzyme (Yang, Malik, and Eickbush 1999). Most lineages of non-LTR elements are like R1 and encode an apurinic-like endonuclease (APE) upstream of the RT domain (Malik, Burke, and Eickbush 1999).

To determine the phylogenetic relationship of R5 to other non-LTR retrotransposons, the RT domain of R5 was aligned with that of other non-LTR retrotransposons. Based on this alignment, we used Neighbor-Joining algorithms to construct a phylogeny (fig. 6). The R5 element was found to be located in the same lineage (clade) as NeSL-1, an element in Caenorhabditis elegans which specifically inserts into the tandemly repeated spliced leader-1 genes (Malik and Eickbush 2000). Unfortunately, the phylogenetic relationship of the NeSL/R5 lineage relative to the other non-LTR clades was not resolved using either the RT domain (fig. 6) or the RT domain in combination with the EN domain (data not shown). In a previous attempt to resolve the non-LTR element phylogeny, it was suggested that the NeSL and the R4 clades were sister groups relative to the R2 clade (Burke et al. 2002). The bootstrap support for this relationship was low, however, and the addition of the divergent R5 element within the NeSL-1 clade and of the EhRLE element from Entamoeba histolytica (Sharma et al. 2001) to the R4 clade has unfortunately further reduced resolution of the phylogeny of those elements with C-terminal EN domains.

The structure of R5 elements is unusual compared to that of R2, R4, and NeSL-1, because they contain two
overlapping ORFs (fig. 5). All sequenced copies of the R5 element contained these two ORFs separated by this frameshift. In addition, as mentioned in the preceding section, a second family of R5 elements was discovered in G. tigrina. A 1.5-kb region spanning the region of the frameshift was sequenced from this second family. In comparing the second family with the sequence of the first family, we detected seven indels, yet both ORFs and the location of the frameshift are conserved. Because this frameshift has been conserved during the evolution of the two families, it is likely a feature of active R5 elements.

The identification of a frameshift in R5 may be relevant to an unusual feature of its nearest relative, NeSL-1 (Malik and Eickbush 2000). As shown in figure 5, NeSL-1 elements contain a single ORF that is somewhat longer than the combined lengths of the two R5 ORFs. Unlike all other non-LTR retrotransposons, the ORF of NeSL-1 contains a cysteine protease domain (PR) upstream of the RT domain. Sequences similar to a cysteine protease were not identified in R5. If the protease of NeSL-1 were to cleave its large ORF immediately downstream of this domain, as is typical for proteases within retrotransposons and retroviruses (Kirchner and Sandmeyer 1993), then the size of the protein containing the RT/EN would be similar to that encoded by R5 ORF2 as well as that of the R2 and R4 ORFs. The N-terminal region of R2 contains zinc-finger and c-myb DNA binding domains (gray vertical bars) involved in DNA binding (Burke et al. 1999; S. Christensen and T. Eickbush, in preparation). NeSL-1 elements also contain zinc-finger motifs at the N-terminal end of the NeSL-1 ORF, but no such motifs could be identified in R5 and R4.

R5 Elements Are Present in G. dorotocephala

We have conducted preliminary experiments to confirm that R5 elements are also present in G. dorotocephala. The 3′ half of the elements were obtained by PCR amplification using the 28S gene primer Ribodyn60 REV (fig. 1B) and the degenerate primer to the highly conserved SY/FADD motif of the RT domain (see Materials and Methods). A G. dorotocephala insertion element was identified that was inserted into the identical site as the R5 element of G. tigrina. The R5 element of G. dorotocephala ended in a series of GTT repeats, similar but not identical to the tandem repeats found at the 3′ end of R5 elements in the G. tigrina (fig. 3C). The G. dorotocephala R5 element was highly divergent from that of G. tigrina. Comparison of the protein-encoding regions

FIG. 5.—Diagram of the ORF structure of the R5 element compared to other non-LTR retrotransposons. ORFs are represented by rectangular boxes, with boxes at different levels indicating ORFs in different reading frames. Shaded regions indicate identified enzymatic domains: RT, reverse transcriptase; APE, apurinic-apyrimidinic endonuclease; PR, cysteine proteinase; EN, restriction-like endonuclease. Two conserved regions of the EN domain, a C-X_2-C-X_8-H-X_4-C domain (CCHC) and a PD-X_12-D domain (PD.D) which has been shown to be part of the active site of the endonuclease are indicated with the black bars (Yang, Malik, and Eickbush 1999). Gray bars at the amino-terminal end of the single ORF in NeSL-1 and R2 and at the C-terminal end of the R1 ORF1 indicate putative nucleic acid binding motifs.

FIG. 6.—Phylogenetic relationship of R5 to other non-LTR retrotransposons based on their reverse transcriptase domain. The phylogeny presented is derived by the Neighbor-Joining method based on the 420 amino acid residue positions that correspond to the fingers, palm, and thumb regions of the RT domain (Malik, Burke, and Eickbush 1999). The numbers next to each node indicate bootstrap values as percentages of 1,000 replicates. The analysis was conducted with the complete set of non-LTR retrotransposon sequences as they appear at EMBL (accession number ALIGN_0002311) with the addition of the R5 and EhRLE sequences. Only branches corresponding to elements with C-terminal restriction-like endonuclease (APE) domains are shown. All elements with APE endonuclease domains are located on the branch shown (dotted box). The topology of branches within this box is essentially the same as that in Malik, Burke, and Eickbush (1999). Each major grouping of elements (clades) was also supported by maximum parsimony methods (> 50% bootstrap values).
revealed only 25% identity (40% similarity), and their 3′ untranslated regions could not be unambiguously aligned. A series of Southern blots and PCR experiments of the type described in figure 4A and 4B were also conducted for G. dorotocephala (data not shown). The fraction of the 28S genes with R5 insertions in G. dorotocephala was again low (1%–2%). Unlike G. tigrina, no evidence was obtained for R5 insertions in the upstream (duplicated) target site of the G. dorotocephala 28S genes.

Discussion

The R5 retrotransposons of dugesiid planaria insert into the same region of the 28S gene as the R1–R4 elements of arthropods and nematodes (fig. 7). A DNA-mediated element, Pokey, has also been identified in the 28S genes of crustaceans (Penton, Sullender, and Crease 2002). The different copies of R5 in the rDNA loci of G. tigrina exhibit little nucleotide divergence, with most copies devoid of mutations that would cause premature termination or inappropriate frameshifts. This would suggest that like R1 and R2, R5 elements must continually retrotranspose to maintain their presence in the rDNA locus (Perez-Gonzalez and Eickbush 2001, 2002). However, unlike the previously described R elements, the percentage of the G. tigrina or G. dorotocephala rDNA units inserted with R5 is only around 1%.

The rRNA genes of dugesiid species are unusual in that they comprise two types of rDNA units that differ by 8%–10% in nucleotide sequence. The duplication or physical separation event that gave rise to these two families of rDNA units has been estimated to have occurred about 100 MYA (Carranza, Beguna, and Riutort 1999). The presence of two families of rRNA genes in the same species is not without precedent. There are two sets of tandemly arranged 5S rRNA genes in Xenopus laevis, one specific for somatic tissues and one specific for oocyte ribosomes (Petersen, Doering, and Brown 1980). More similar to the situation in planaria are the two sets of rDNA units in Plasmodium species (Dame, Sullivan, and McCutchan 1984). One set is expressed during the sporozite stage in the mosquito host, and the second set is expressed during the blood stage in the mammalian host (Gunderson et al. 1987). Perhaps the type II rDNA units of planaria are also expressed only in limited cell types or during limited developmental stages.

Based on the six 5′ junction sequences and the twelve 3′ junction sequences, G. tigrina R5 elements are preferentially inserted in the infrequently transcribed type II rDNA units. There are only three nucleotide differences between the type I and II 28S genes of G. tigrina in the 35 bp to either side of the insertion site. Thus it seems unlikely that sequence recognition by the R5 endonuclease could be responsible for the preference of insertions for the type II genes. A more likely basis for the R5 preference for inserting into type II genes is the chromatin structure of the insertion site. The R2 endonuclease can conduct the integration reaction when the target site is assembled into nucleosomes, but the reaction is highly dependent on the translational phase of the nucleosome (Ye et al. 2002). Thus it is possible that differences in the position of the nucleosomes, or higher order structures of the chromatin, inhibit insertion of R5 elements into the type I genes. An alternative explanation for the distribution of R5 elements is that these elements can insert into type I units but that such insertions are strongly selected against. When R elements insert into the rDNA units of insects and nematodes, transcription of the unit is severely inhibited (Long and Dawid 1979; Kidd and Glover 1981; Jamrich and Miller 1984; Neuhäusel et al. 1987). Because eukaryotes have more rDNA units than are needed at any one time for rRNA synthesis, the inactivation of a limited fraction of units may have little effect on the host’s fitness. However, if G. tigrina had no mechanism to turn off the transcription of R5-inserted units, then defective 28S rRNA would be produced, which could be highly damaging to the organism. As a result, R5 insertions in type I units may be strongly selected against and rapidly eliminated, whereas insertions in the infrequently transcribed type II units are less damaging and are able to accumulate.

A second unusual feature of R5 insertion is that because of a sequence duplication, R5 elements can insert into two sites approximately 300 bp apart. The duplicated sequence is identical in G. tigrina and G. dorotocephala.

![Diagram of transposable element insertions](https://example.com/diagram.png)
suggesting that this duplication arose in the common ancestor of these species. Although the duplication of a short DNA segment at a site 300 bp from the original site is difficult to explain by recombination, insights into the generation of this duplication may be obtained from studies of R2 retrotransposition. Short segments of the upstream 28S gene, usually 20–30 bp in length, are sometimes co-inserted along with the R2 sequence during its integration (Burke et al. 1999). These 5′ transduced 28S gene sequences have been suggested to result from reverse transcription of flanking 28S gene sequences present on the R2 RNA transcript.

Using this analogy to R2, we can suggest the following model for the generation of a duplicate R5 target site in the type II 28S genes of planaria. In the first step, a rare R5 retrotransposition event involving transduced 28S gene sequences occurred at a site ~300 bp upstream of the normal insertion site. This location may have been selected by the R5 endonuclease because this site contains sequence identity to the normal R5 target site. Using the current sequence of the type I genes to infer the probable sequence of this region before insertion, the upstream site contained an 8 of 10 match to the normal R5 insertion site (TGAATCTTGT versus TGAATCTCTT). In the second step, the R5 sequences were eliminated by recombination from the upstream site, leaving behind the duplicated 28S sequences. In the final step the continuing process of the concerted evolution of the rDNA locus spread the sequence duplication to all type II 28S genes of the species.

The discovery of R5 elements in planaria has important implications for the origin and evolution of several of the oldest non-LTR retrotransposon lineages. Both the phylogeny using the RT domain and the presence of a C-terminal EN domain suggest that R2, R4, and R5 represent three of the older clades of non-LTR retrotransposons. In contrast, R1 elements encode an APE endonuclease and are part of a younger lineage of non-LTR retrotransposons (Malik, Burke, and Eickbush 1999). Whereas one or more lineages of the 28S gene-specific R1 elements are present in all classes of arthropods (Burke et al. 1998), elements from the R1 clade have also been identified which specifically insert into telomeric repeats (TRAS and SART elements), into CA tandem repeats (WALDO elements), and into other specific locations of either the 18S or 28S genes (RT, R6, and R7 elements) (Besansky et al. 1992; Kubo et al. 2001; Kojima and Fujiwara 2003). Thus lineages of R1-like elements have evolved specificities for new sites both within and outside the rDNA units.

The presence in the 28S genes of a flatworm of non-LTR retrotransposons that are similar in structure to those of the R2 and R4 elements can have two possible explanations. First the elements could represent the same lineage and, like the R1 lineage, have shifted their insertion sites. This model implies that throughout animal history the RNA genes may have served as a breeding ground for non-LTR retrotransposons that have changed their insertion specificities. For example, elements within the R5 clade insert within the spliced leader exons of nematodes (NeSL-1 elements). The R4 clade contains elements that insert into tandem TA or TAA repeats—Dong elements in insects (Xiong and Eickbush 1993) and elements with no obvious target specificity, REX6 and EhRLE elements, in vertebrates and protozoans (Volf et al. 2001; Sharma et al. 2001). In the second model for the evolution of R2-like elements, R2, R4, and R5 could represent separate lineages of elements that have independently evolved specificity for the 28S gene. Unfortunately, attempts to define the phylogenetic relationship of the R2, R4, and R5 elements based on the sequence of their ORFs does not provide sufficient resolution to determine the relationship of the elements (see fig. 6). Evidence for or against these models should be obtained when more elements from these clades are identified, particularly in more primitive organisms. No matter which model is correct, the small region of the 28S gene shown in figure 7 is clearly a highly favored location for the insertion of transposons. The abundance and long history of these elements suggest that their life histories are intimately tied to the rDNA locus. They will thus provide clues and new tools for the study of both non-LTR retrotransposons and the rRNA genes themselves.

Acknowledgments

This work was supported by National Science Foundation grant MCB-9974606 to T.H.E. We thank Malka Korman, whose experiments originally suggested that an R-like element was inserted in the 28S genes of _G. tigrina_, and Thomas Prentice for assistance with the phylogenetic analysis. We also thank D. Eickbush for helpful comments on the manuscript.

Literature Cited


Pierre Capy, Associate Editor

Accepted April 1, 2003