Xena, a Full-Length Basal Retropseudon From Tetraodontid Fish

Damian E. Dalle Nogare,* Melody S. Clark,† Greg Elgar,‡ Ian G. Frame,* and Russell T. M. Poulter*

*Department of Biochemistry, University of Otago, Dunedin, New Zealand; and †Fugu Genomics, Human Genome Mapping Project Resource Centre, Wellcome Genome Campus, Hinxton, Cambridge, U.K.

Introduction

Retrotransposons are ubiquitous eukaryotic mobile genetic elements that transpose via an RNA intermediate. These elements are divided into two broad groups, the LTR-retrotransposons, whose coding regions are flanked by direct long terminal repeat (LTR) sequences and the non-LTR retrotransposons, which lack such sequences. This distinction is supported by sequence similarity in the reverse transcriptase domain, the structure of the elements, and the proteins encoded. LTR-retrotransposons generally encode a Gag structural protein, which assembles into a virus-like particle (VLP) within which reverse transcription occurs (Mellor et al. 1985). The LTR-retrotransposons also have enzymatic functions required for the replication cycle, namely protease, reverse transcriptase, RnaseH, and integrase. Within this family there are at least three major subgroups, Ty1/Copia, Ty3/Gypsy, and Bel/suzu. The vertebrate retroviruses and DIRS elements might be included in the Ty3/Gypsy subgroup or might be considered as distinct subgroups. Non-LTR retrotransposons, on the other hand, encode a protein that binds RNA but does not assemble into a VLP. In addition, they encode a reverse transcriptase and often an endonuclease. The evolutionary relationship between LTR and non-LTR retrotransposons is unclear.

There are several described eukaryotic LTR retropseudons that do not easily fit into this classification. One such element, Penelope of D. virilis, was first described by Evgen’ev et al. (1997). It is thought to be responsible for a hybrid dysgenesis syndrome analogous to the P element system. The Penelope sequence contains direct LTR sequences, suggesting that it is an LTR-retrotranspon; however, Penelope lacks several of the defining characteristics of all known LTR retrotransposon groups, namely Gag-, protease-, and RnaseH-coding regions. For these reasons, Evgen’ev et al. refer to the Penelope element as belonging to the non-LTR retropseudon group (Evgen’ev et al. 2000b). They also described an apparent reverse transcriptase and possible integrase protein within the single open reading frame, the latter differing very considerably from all known retrotransposon and retroviral integrase sequences. The Penelope element has thus far defied classification into any known retropseudon group and remains something of an enigma, nothing resembling this unusual element having been described outside of the virilis subgroup of Drosophila. The position of this element within the evolutionary scheme of retropseudons as a whole is a particularly interesting but as yet unresolved question.

The Fugu fish, Takifugu rubripes, has a compact genome (400 Mbp) and gene content comparable to the much larger mammalian genomes. As a result it has emerged as a model system for vertebrate genomics (Brenner et al. 1993; Venkatesh, Gilligan, and Brenner 2000). Within this compact genome relatively uncorrupted LTR and non-LTR retrotransposons, retroviruses, and DNA transposons have been identified (Herniou et al. 1998; Poulter and Butler 1998; Poulter et al. 1999). As part of the initial mapping stages of the human genome project, some 6% of the Fugu genome was sequenced at the human genome mapping project (HGMP) (Elgar et al. 1999). Recently, an international consortium has undertaken to sequence the complete Fugu genome as part of the assembly and annotation phase of the human genome project. The identification of abundant repeat elements within this genome is of high priority. Critically, the presence of high copy number repetitive elements can hamper efforts during assembly of the final sequence, because of ambiguities in the overlap of different cosmid or BAC clones.
As part of this continuing comparative genomic characterization of Tetraodontid fish genomes, a BAC-end sequencing project and survey of the repetitive DNA content in a closely related species, *T. nigroviridis* (a freshwater pufferfish) was recently undertaken (Crollius et al. 2000). The *Tetraodon* analysis suggested the presence of a number of repetitive element families, some of which are also present in the Fugu fish.

Materials and Methods

Fugu cosmid clones were obtained from the HGMP resource center, Hinxton, Cambridge. The protocol for PCR amplification of cosmid DNA in a total volume of 50 µl consisted of 30 cycles of an annealing step of 1 min at 45°C, followed by extension at 72°C for 2 min and denaturation for 1 min at 95°C, with a 5-min final extension at 72°C. All PCR reactions used the Expand Hi-fidelity enzyme mix (Roche) and an Autogene II thermocycler. PCR products derived from these reactions were separated on a 1% agarose gel, visualized by ethidium bromide staining, and cleaned up for sequencing using a QIAQUICK PCR purification kit, according to the manufacturer’s instructions. Direct sequencing of PCR products was performed at the Center for Gene Research, University of Otago, using an ABI PRISM 377 automated DNA sequencer with dye-terminator chemistry. The oligonucleotide primer sequences used to amplify the complete Xena coding region were 5’-GCGAGCTCTTGGCTGTCCTCGAAGGAG-3’ and 5’-CTCTAGAACAGTCCAGTTGACAGAG-3’. Additional internal and inverse PCR amplification and sequencing primers were also used as required. Typical reaction mixtures contained 2 µl (0.1 µg) template DNA, 0.2 nmol of each primer, 1 µl dNTP solution containing 10 mM each dNTP, 5 µl 10× PCR buffer with MgCl₂, 0.5 µl of Expand Hi-Fidelity enzyme mix (3.5 U/µl), and 40.5 µl ddH₂O.

Protein sequence alignments were initially performed with ClustalX v 1.8.1 using the default parameters. Reverse transcriptase sequences for phylogenetic analyses were extracted from the GenBank database by BLAST searches (Altschul et al. 1990), and elements were grouped into previously described clades for initial alignment. These aligned groups were then aligned with other such groups in an iterative fashion, using the align profile option in ClustalX. After each iterative addition, the alignment was examined and any required adjustments made manually using Jalview 1.7.5.b (available http://www2.ebi.ac.uk/~michele/jalview/contents.html). Where further adjustments were necessary, alignments were anchored with previously described conserved residues (Xiong and Eickbush 1990) and adjusted by alignment of the intervening sequence using ClustalX. The full phylogenetic tree (239 reverse transcriptase sequences) was constructed from these data using the neighbor-joining method (Saitou and Nei 1987), implementing 100 bootstrap replicates with resampling with PAUP 4.0.b2a. The simplified phylogenetic tree (49 reverse transcriptase sequences) reflects a neighbor-joining analysis of a subset of this data with 1,000 bootstrap replicates. Consensus sequences from Tetraodon clones were assembled using the Contig Assembly Program (CAP) and the CAP3 algorithm (Huang and Madan 1999). Nucleotide sequence data for the Fugu elements (Xena-Fr) has been deposited in GenBank under accession numbers AF355375, AF355376, and AF355377. Consensus sequences used in this analysis are available from the authors’ website http://bioic111.otago.ac.nz:800/retrobase/home.htm.

Results

The Xena Sequence from *T. rubripes*

Xena, a Penelope-like retrotransposon, was first identified during shotgun sequencing of the Fugu-fish, *T. rubripes* at the HGMP. By searching the Fugu database with a Penelope query sequence, numerous matches to different cosmids were found, and a limited consensus sequence was constructed. Because of low database redundancy it was not, however, possible to extract from the database a complete element or even a full consensus sequence. Cosmids were screened by PCR for full-length Penelope-like elements using primers designed to the ORF sequences and the flanking repeat sequences. An unusual feature (shared with Penelope) is that the ORF extends into both the 5’ and 3’ flanking repeats. In view of this and other novel features, we refer to these repeats as PLTRs (Penelope-LTRs), as distinct from the LTRs of Ty1/copia, Ty3/Gypsy, and Bel/suzu elements. One cosmid, 008J20, carrying a putatively full-length element (AF355375) was selected for sequencing. The 008J20 element had identical direct PLTRs, and immediately preceding the 5’ direct PLTR there was an inverted PLTR fragment (see fig. 1A) identical in sequence to the PLTRs. A second selected element (AF355376) on cosmid 177B08 was incomplete, it carried a complete 3’ PLTR and a 5’ truncated ORF. Immediately preceding the boundary of this 5’ truncation there was an inverted, antisense fragment corresponding to the 3’ end of the ORF and the 3’ PLTR. A third selected element (AF355377), identified on a *T. rubripes* contig containing the utrophin gene, also carries a complete 3’ PLTR, a 5’ truncated ORF, and an inverted fragment preceding the 5’ truncation.

The full-length Xena element sequenced from cosmids 008J20 consists of a single 2,556-bp open reading frame within an element of total size 3,573 or 3,775 bp if the 5’ inverted PLTR fragment is included. This full-length element (fig. 1A) carries two identical 845-bp PLTR sequences in the direct orientation. The 5’ inverted fragment is 201 bp long and is missing the 34-bp terminal sequence present in both direct PLTRs. The open reading frame begins at an in-frame ATG just within (32 bp) the 5’ direct PLTR and extends 641 bp into the 3’ direct PLTR, leaving in the entire element only 172 noncoding nucleotides in each PLTR. Xena is thus an extremely compact retroelement, and this is reflected in the small size of this element when compared to both LTR and non-LTR retrotransposons. The only significant annotated BLAST match to the Xena ORF in the GenBank nr database is to Penelope, the highly unusual
Xena, a Penelope-Like Retroelement in *Fugu rubripes* 249

Fig. 1.—Full Length Xena element from *T. rubripes* and Xena fragment from *T. nigroviridis*. A. Full length Xena element from Fugu cosmid 008J20. YVDDT denotes the approximate position of the reverse transcriptase active site. B. Severely truncated element, showing inverted PLTR and flanking target site duplication.

A retrotransposon from the genomes of members of the *D. viridis* subgroup. This match is to a recognizable reverse transcriptase, but no protease or integrase coding regions are apparent in Xena or Penelope. The lack of these modules and the similarity of these elements (expect $3e^{-73}$ over the entire ORF at the protein level) suggest that these elements define a novel class of retrotransposons. Members of this class (Xena and Penelope) are recognizable to each other but are not closely related to any other known LTR or non-LTR retrotransposon.

The presence of an apparently uncorrupted ORF in Xena-Fr:AF355375 suggests that the element may be functional or recently functional. This belief is supported by the presence of identical PLTRs. The 201-bp inverted PLTR fragment of Xena-Fr:AF355375 is also identical to the sequence of the direct PLTRs. The other two Xena-Fr sequences show only slight divergence from Xena-Fr:AF355375, including some possible corruption in addition to the 5′ truncation. It is, however, still of interest to compare them with the complete element. The PLTRs of Xena-Fr:AF355376 are identical to those of Xena-Fr:AF355375. This is despite Xena-Fr:AF355375 and Xena-Fr:AF355376 being distinct elements from different cosmids. The single 5′ truncated PLTR of Xena-Fr:AF355377 is almost full length (820 bp compared with 845 bp for a full-length PLTR). This 820-bp PLTR sequence of Xena-Fr:AF355377 differs from that of the Xena-Fr:AF355375 PLTRs at eight positions, all of which are substitutions. Only one of these differences falls in the 648-bp coding regions of the PLTR (a synonymous change in a threonine codon), the other seven fall in the relatively short (172 bp) noncoding region of the PLTR. This pattern suggests the Xena sequences have been under stringent selective constraint in the coding region and that divergence was more acceptable in the noncoding region.

As described above, all three Xena elements isolated from *T. rubripes* are preceded by inverted Xena fragments. In the complete Xena-Fr:AF355375 element (from cosmid 008J20, fig. 1A), the inverted PLTR lacks the terminal 34 bp present in both the sense orientation direct PLTRs. This situation closely parallels that described in Penelope and is also found in the related Tetraodon elements (see below). In the sequence from 008J20, the complete element (inverted PLTR and direct PLTR flanked Xena element) is bracketed by perfect 9-bp target site duplications (TSD), reading GGATA-TAAT, suggesting that the element was integrated into the genome in this form, rather than being generated as the result of compound integration or recombination.

**Xena Elements from *T. nigroviridis***

Recently, a genome-wide BAC-end survey of the freshwater pufferfish, *T. nigroviridis*, was undertaken (Crolloiu et al. 2000). Using this resource, we were able to identify retrotransposons in *T. nigroviridis* (Xena-Tn) closely related to the Xena element from *T. rubripes* (Xena-Fr). The Xena elements from *Takifugu* and *Tetraodon* have high (>76% over the entire coding region) protein sequence identity (fig. 2), despite these two pufferfish lineages being separate for 18–30 Myr (Crnogorac-Jurcevic et al. 1997). A consensus sequence was built from these BAC end reads by iterative database searches using a compilation of overlapping clones. In Xena-Tn, as in Xena-Fr, the ORF begins within the 5′ PLTR, extends well into the 3′ PLTR, and contains a recognizable reverse transcriptase domain. No obvious indication of a protease or integrase coding region was found within this ORF. This structure is identical to that observed in the full-length Xena-Fr (fig. 1A). Comparison of the PLTR sequences of Xena-Fr and Xena-Tn shows strong conservation at the 5′ and 3′ end of the PLTRs, corresponding to the coding regions within the PLTRs. The Xena-Tn PLTR is, however, slightly shorter (761 bp) than the corresponding Xena-Fr PLTR (845 bp). This size difference is caused by an indel within the noncoding (poorly conserved) region of the PLTRs.

Searches of the Tetraodon and Fugu databases with the Xena-Fr ORF found numerous BLAST matches to the 3′ end of the ORF, whereas fewer matches were found to 5′ regions. Many Tetraodon sequences showed an abrupt internal termination of similarity to the Xena ORF, reflecting 5′ truncation of the Tetraodon elements. This truncation pattern is suggestive of a LINE-like replication system, where poor reverse transcriptase processivity during replication has been suggested as a cause of 5′ truncation (Kazazian and Moran 1998). As an example of these 5′ truncated *Tetraodon* sequences,
clone AL225633 matches (E value of $e^{-122}$) the Xena-Tn consensus ORF until nucleotide 800. This similarity abruptly ends at nucleotide 800, which corresponds to amino acid 567 of the ORF. To determine why this match did not continue over the entire 943-bp length of the clone, we examined the sequence from nucleotide 800–943. Aligning AL225633 with a Xena-Tn PLTR showed that this region, from residue 800–943, corresponds to a Xena PLTR; however, it is in the inverted, antisense orientation with respect to the rest of the first 800 nucleotides of the sequence. A similar situation appeared in numerous other 5' truncated Xena-Tn elements present on other clones, for example AL227991, AL272228, AL230293, AL311680, AL183863, and many others.

To further examine this phenomenon, we reasoned that if clones carrying a 5' truncated ORF were preceded by an inverted PLTR, there might exist even more severely 5' truncated clones with noncoding PLTR regions preceded by inverted PLTR fragments. Screening the gss database with the 3' boundary of the PLTR identified 624 clones matching PLTR sequence. This abundance of very severely truncated elements in Tetraodon contrasts with the situation in the Fugu genome. Of these 624 Tetraodon clones, 110 clones contained two PLTRs. Of these 110 dimeric elements, 109 contained the two PLTRs in an inverted orientation. These clones typically contain a variable length of PLTR ending with the 3' boundary sequence corresponding to the conceptual translation MTWMTENLHR(H/Q) (subsequently referred to as the 34-bp cap). The clones also contained an inverted PLTR fragment, which lacks the 34-bp cap. As an example, figure IB shows one such short inverted element from T. nigroviridis clone AL329115. A similar phenomenon is observed in Penelope, where one PLTR is often 34-bp shorter than the other, and in Xena-Fr, where the inverted PLTR lacks the terminal 34 bp found in the direct PLTRs.

These short elements with inverted PLTRs are flanked by TSD typical of retrotransposon insertion sites (fig. 3). This suggests that these elements were inserted into the Tetraodon genome in this configuration, rather than being the result of truncation, corruption, or compound insertion of elements. The exact length of this target site duplication varies, but it is generally between 8 and 15 bp (compared with 6 to 10 bp for the Penelope TSD and 9 bp for Xena-Fr). This variability in TSD length is similar to that observed in L1 non-LTR retrotransposons.

Given the paucity of sequences from the 5' section of the Tetraodon element in the database it was difficult to determine the left PLTR-ORF configuration. To accomplish this, we searched the gss database with a Xena-Fr sequence corresponding to the boundary of the left PLTR and the adjacent ORF. Only three clones were identified (AL350028, AL206347, and AL225104) which contained sequence covering a 5' PLTR and the adjacent ORF. In two of these clones (AL206347, AL350028), the 5' PLTRs were truncated and adjacent to inverted PLTRs. These two sequences are presumably derived from almost full-length elements truncated within the 5' direct PLTR. The sequence of clone AL225104 ends within the direct 5' PLTR, and it is therefore impossible to tell if the PLTR is truncated, this clone might therefore be from a complete element. The presence of these three sequences supports the belief that in Tetraodon, as in Fugu, full-length Xena elements have direct PLTRs flanking the ORF, and that the elements are preceded by an inverted sequence.
Other Penelope-Like Elements

In addition to the Xena-Tn elements, a second, rather different (31% identical, 46% similar to Xena-Tn over the central RT domain) Penelope-like retrotransposon is present in *T. nigroviridis* (for example, accession number AL295026). A similar divergent element is also present in Fugu (cosmid 041K15/045F03) but is highly corrupt. These two elements, which we call Callisto elements (Callisto-Tn and Callisto-Fr) are 61% identical and 75% similar over the central RT domain. Thus the Callisto elements are more similar to each other than they are to the Xena elements present in either species.

In addition to these lineages, we have found evidence in the databases for the presence of Penelope-like elements within the shrimp, *Penaeus monodon* (expect $e^{-15}$ [smallest sum probability]; [nr] AF077759), the sea urchin, *Strongylocentrotus purpuratus* (Cameron et al. 2000) (four matches with expect $e^{-10}$; [gss] AZ159544, AZ154071, AZ184610 and AZ183349), a range of cichlid fish (for example [nr] AF069088, $e^{-15}$), *Xenopus laevis* (for example EST BG513403, $5e^{-33}$), the fluke *Schistosoma mansoni* (for example EST AL974952, $e^{-12}$), and the roundworm *Trichuris muris* (for example ESTBG577753, $e^{-17}$). No Penelope-Xena–like elements were found amongst the unicellular eukaryotes.

A complete consensus sequence could be constructed for the *Tetraodon* Xena elements but only fragments are reported from these other organisms. Whereas the other matches were derived from genome or EST sequencing projects, the cichlid sequences were generated as part of a population study of cichlid species in Lake Victoria (Nagl et al. 1998) and are short (340 bp) fragments with low expect values in a BLAST search with Xena-Fr. These sequences establish the presence of Xena-like elements in diverse cichlid fish and suggest the possibility of using these mobile elements in population studies.

An alignment of these elements is shown in figure 2. Alignment of the protein sequences of all the three complete assembled elements (Penelope, Xena-Fr, Xena-Tn) and the partial elements identified the strongly conserved RT domains previously noted. No conserved motifs associated with other retrotransposon related proteins, including protease, integrase, or endonuclease were found. The possible conserved motifs within the 3’ region of the ORF alignment were supported by the cichlid and *Xenopus* sequences. However, BLAST or MAST (Bailey and Gribskov 1998) profile searches with this region still failed to give any significant ($e < 0.05$) database matches. The function of the non-RT coding regions thus remains speculative, although the apparent sequence conservation between distinct elements in this region suggests a critical function.

Phylogenetic Analysis

BLAST searches using the RT domain of Xena against the nr database did not return any significant matches with either LTR or non-LTR retrotransposons. The best BLAST matches outside the Penelope group were mobile group II intron reverse transcriptase sequences from *Escherichia coli* and *Sinorhizobium meliloti*. These low scoring matches ($e > 0.1$) probably represent low-level similarity to the central reverse transcriptase domain, rather than a close phylogenetic relationship between these elements. They do, however, serve to give an indication of the distance between Xena and all other known retroelements.

A phylogenetic analysis was performed using the seven conserved reverse transcriptase domains (Xiong and Eickbush 1990) and the neighbor-joining method (Saitou and Nei 1987). Figure 4 shows one distance tree based on the alignment with overlaid bootstrap values. This analysis strongly suggested that the Penelope-Xena elements comprise a distinct clade (bootstrap value 100% [1,000 replicates]) and suggests that they are a sister clade of the LTR-retrotransposon group (74%). This is supported by the extremely high bootstrap value for a non-LTR retrotransposon clade, excluding the Penelope-Xena elements (96%). The branching order within the LTR-retrotransposons, however, could not be resolved at all branch points based on this small dataset. In order to clarify the branching order a larger dataset was used to create a second tree, represented diagrammatically (collapsed) in figure 5. This tree confirms the distinct nature of the Penelope-Xena clade. It also suggests that these elements may lie basal to the other clades in the LTR-retrotransposon group, although the bootstrap support for this is not conclusive. Alternatively, the Xena elements may be best placed as a third distinct retroelement lineage. The alignments and complete tree are available from the authors’ website http://bioc111.otago.ac.nz:800/retrobase/home.htm.

Discussion

Element Distribution

In this report we describe a full-length Penelope-like vertebrate retroelement. This is an exemplar of a novel family of vertebrate retrotransposons, the Xena elements, found in two pufferfish species and in other fish. A related lineage of elements, termed Callisto, is also present in both pufferfish species. Both these novel lineages are related to the Penelope retrotransposon from *D. virilis*, suggesting the existence of an ancient and widely distributed Penelope-Xena retroelement lineage. This is the first report of a full-length Penelope-like element outside of the *D. virilis* group. Penelope-Xena–like elements are present in a wide range of taxa. Deuterostomes carrying Xena elements include *T. rubripes*, *T. nigroviridis*, a variety of cichlid fishes, and the tetrapod *X. laevis* (vertebrates), as well as the echinoderm *S. purpuratus*. The protostomes carrying Penelope-Xena...
elements are *D. virilis* and the shrimp *P. monodon*. In addition, Penelope-Xena elements are found in the pseudocoelomate roundworm *T. muris* and the acoelomate fluke *S. mansoni*. This wide distribution suggests an ancient origin for this group. Despite this apparently ancient origin, no Penelope-like elements could be found in the genomes of the invertebrates *Caenorhabditis elegans* and *Drosophila melanogaster* (both of which have been completely sequenced) or in any other organism currently within the available GenBank databases, including the recently released human genome draft sequence. This suggests extensive loss from many vertebrate and invertebrate lineages. The presence of Penelope in *D. virilis* and the absence of elements of this class in the closely related *D. melanogaster* implies that such a loss must have occurred recently, subsequent to the divergence of these two Drosophila species. Alternatively, this discontinuous distribution might suggest horizontal transmission. Recently Evgen’ev et al. have provided evidence for the recent invasion of the virilis subgroup by the Penelope transposable element (Evgen’ev et al. 2000a), although the significance of this in terms of the evolution and distribution of the Penelope-Xena group remains unclear. This recent invasion is accompanied by hybrid dysgenesis, and it is interesting to consider whether the Xena elements might be capable of causing hybrid dysgenesis in fish.

Open Reading Frames

The ORFs of the Penelope-Xena elements are highly unusual in terms of their sequence when compared with other known retroelements. Evgen’ev et al. suggested the presence of an N-terminal integrase protein...
Xena, a Penelope-Like Retroelement in *Fugu rubripes*

**Fig. 5.**—Further analysis of the relationship among known retroelement classes. This phylogenetic tree is also based on the central seven reverse transcriptase domains identified by Eickbush et al. and the neighbor-joining algorithm implemented using PAUP 4.0.b2a. The figure shows a representation of the collapsed groupings of a tree based on the alignment of a large number of retroelements. The tree is able to resolve some aspects of the branching order. Bootstrap support (100 replicates) for the terminal nodes reflecting each collapsed retroelement grouping is indicated in parentheses. The tree was rooted with RNA-dependant RNA polymerases. The complete tree is available from the authors’ website, http://bioc111.otago.ac.nz:800/retrobase/home.htm.

Based upon the presence of a DD(35)E motif thought to represent the conserved integrase active site. We find, using the new alignment with three complete elements (Xena-Fr, Xena-Tn, and Penelope, fig. 2) and various fragments no evidence for the presence of a conserved classical DD(35)E integrase motif. The putative N-terminal zinc finger motif identified in Penelope by Evgen’ev et al. is not conserved in the Fugu or Tetraodon Xena elements, and only one of the three proposed DDE active site residues is completely conserved. In addition, the distance between any suggested integrase active site residues is not conserved between elements because of indels. A specific function based upon sequence similarity could not be assigned to any region of the ORF other than the reverse transcriptase regions, although conserved cysteine and histidine residues are present C-terminal to the reverse transcriptase domain and may be involved in zinc finger–like structures.

**PLTRs and Inverted PLTRs**

The full length Xena-Fr element from *T. rubripes* cosmid 008J20 is flanked by identical direct PLTRs and preceded by an identical partial inverted PLTR sequence. There are three examples from *Tetraodon* showing the same relationship between a direct 5′ PLTR and ORF. The majority of *Tetraodon* Xena elements show 5′ truncation, often including all of the 5′ PLTR and much or all of the ORF. In these truncated elements it is obvious, even within single clone sequence reads, that the 5′ truncation is often preceded by an inverted sequence ending in a PLTR. The inverted PLTR and the direct PLTR in individual elements show very close sequence similarity. In elements preceded by inverted PLTRs (such as the full-length 008J20 and the profoundly truncated AL329F15), an interesting feature is that the direct PLTR(s) carry a 34-bp cap, whereas the inverted PLTR lacks this sequence boundary. A similar 34-bp cap has been noted in Penelope. The complete Fugu Xena element and many of the truncated dimeric Tetraodon elements (such as AL3291150) show long (8–15 bp) TSD, suggesting that the elements carried the inverted PLTRs at the time of integration.

Penelope-Xena elements have a distinct organization that probably reflects a novel replication system. It is likely, given their very close sequence similarity, that the inverted PLTR and the direct PLTR in individual elements are related as template-replicate. We suggest that the inverted sequences are generated during the reverse transcriptase–dependant replication of the elements. We further suggest that the 34-bp cap is deleted from the inverted PLTR during this replication cycle. Xena replication may be most similar to that of LINE elements, a possibility supported by the frequent 5′ truncations and other features such as the long target-site duplications and the absence of integrase and protease domains.
Phylogeny

Penelope has been seen as something of an anomaly amongst retrotransposon classes. We have confirmed the presence of Penelope-like elements in vertebrates and described one full-length element from Fugu (Xena-Fr). The relationship between the Penelope-Xena elements is supported by the similarity of the sequence of their RT domains. Additionally, the full-length Xena-Fr element displays many of the unusual structural features of Penelope. In Xena-Fr, as in Penelope, the ORF extends into both the 5' and 3' terminal repeats. Other unusual features shared with Penelope are the preceding inverted sequence of variable length and the 34-bp terminal cap found on the Xena-Fr direct PLTRs but not on the inverted PLTR. Penelope elements also carry a 34-bp cap on some terminal repeats but not others.

The presence of direct long terminal repeat sequences might suggest that Penelope-Xena elements should be grouped with the other LTR retrotransposons. Phylogenetic analysis based on the RT domain lends some support for this belief; Penelope-Xena elements show weak phylogenetic affinity with LTR retrotransposons. They grouped with the LTR elements 74% of the time (fig. 4) when one data set was employed and 66% when a much larger data set was used (shown as a collapsed representation in fig. 5).

There are, however, grounds for believing that Penelope-Xena elements should not be grouped with LTR elements. These include the 5' truncation of the elements, lack of conserved coding regions characteristic of LTR retrotransposons, and the absence of sequence features such as a tRNA primer binding site or purine tract. Xena elements may be more closely related to the non-LTR retrotransposon group. However, we find no obvious sign of an endonuclease protein which would normally be present in the coding region of a non-LTR retrotransposon. The highly unusual structure of the Penelope-Xena elements and the lack of compelling sequence similarity with any known retroelement class suggest that the Penelope-Xena elements are an ancient and distinct lineage of eukaryotic retroelements. Similarly, Evgen'ev et al. (2000a) found that no reverse transcriptase exhibited more than 21% amino acid identity with the Penelope RT and suggested that Penelope does not fit within any of the defined retrotransposon families. We suggest that it is likely that the Penelope-Xena elements are representatives of a novel, third retroelement clade distinct from both the LTR and non-LTR retrotransposons.

Note added in proof

Subsequent to the submission of this manuscript, two analyses (Lyozin et al. 2001; Volff et al. 2001) have established that Penelope elements carry sequence encoding a putative GIY-YIG endonuclease downstream from the gene for reverse transcriptase. This type of endonuclease has not been found in other retrotransposable elements.

Acknowledgments

The work reported in this publication was supported by the Marsden fund, Royal Society of New Zealand. D.E.D.N. was the recipient of an Otago School of Medical Sciences (OSMS) research scholarship during the course of this work. The authors would like to thank Timothy Goodwin for constructive criticism of the manuscript and Margaret Butler for practical aid and helpful discussion.

LITERATURE CITED


DIETHARD TAUTZ, reviewing editor

Accepted October 15, 2001