

Letter to the Editor

An Unusual Vertebrate LTR Retrotransposon from the Cod *Gadus morhua*

Margaret Butler, Timothy Goodwin, and Russell Poulter

Department of Biochemistry, University of Otago, Dunedin, New Zealand

The pursuit of genome sequencing projects has added impetus to the discovery and analysis of the repetitive DNA sequences that are present in all eukaryotes. Much interest has been focused on retroelements, elements propagated via an RNA intermediate, as these are the most abundant form of repetitive DNA in eukaryote genomes. One class of these elements is the retrotransposons, which can further be subdivided into two major groups: those that possess long terminal repeats (LTRs) and those that do not. The LTR retrotransposons contain open reading frames (ORFs) which encode the proteins (GAG and POL) required for the reverse transcription of the element mRNA and integration of the resulting cDNA into the host genome. The classification of LTR retroelements is currently under intense analysis, and several schemes have recently been proposed (Bowen and McDonald 1999; Hull 1999; Pringle 1999; Cook et al. 2000).

The phylogenetic analyses of LTR retrotransposons based on the predicted amino acid sequences of their reverse transcriptases (RTs) indicate that they fall into at least four major groups: the two extensively reviewed Ty1/copia and Ty3/gypsy groups, the BEL-like group that includes several newly discovered elements (Cook et al. 2000), and the retroviruses. The Ty1/copia retroelement can be distinguished from members of the other three groups not only on the basis of their RT sequences, but also on the basis of the order of amino acid motifs within their POL ORFs. The order of domains in the POL ORF of Ty1/copia elements is protease, integrase, RT/RNase H, while in Ty3/gypsy elements, Bel elements, and the retroviruses, the order is protease, RT/RNase H, integrase. It has been proposed that the difference in the orders of the POL domains be the main defining feature in the classification of LTR-containing retroelements (Bowen and MacDonald 1999; Hull 1999; Pringle 1999). Such a scheme is debatable, however, in that phylogenetic analyses suggest similar divergence between the BEL elements, the Ty3/gypsy elements, and the Ty1/copia elements.

Until fairly recently, it was assumed that retroviruses were confined to vertebrate hosts and that the LTR retrotransposons were present only in nonvertebrate eukaryotes. Lately, examples of retrotransposons with envelope domains characteristic of a retrovirus life cycle have been found in *Drosophila* and plants (Wright and

Voytas 1998), including plant elements of the Ty1/copia group (Peterson-Burch et al. 2000).

Conversely, LTR retrotransposons, or parts thereof, have been discovered in many species of vertebrates (Britten et al. 1995; Poulter and Butler 1998; Miller et al. 1999). While a small number of fragments of Ty1/copia elements have been found in fish and reptiles (Flavell et al. 1995; Roest Crolius et al. 2000), Ty3/gypsy-type retrotransposons are represented in a wide range of vertebrate classes (Miller et al. 1999), although none have yet been reported in birds or mammals.

Here, we describe a full-length LTR retrotransposon in a sequence from the Atlantic cod, *Gadus morhua*. This retrotransposon, which we call Gmr1, is unusual in that sequence comparisons clearly show that it is a member of the Ty3/gypsy group, but the order of the domains within its *pol* ORF is the same as that of Ty1/copia group elements. Analysis of additional vertebrate retrotransposons, including a previously undetected element in the sturgeon *Acipenser baeri*, suggests that there exists a new vertebrate retrotransposon lineage with an unusual POL domain order.

The DNA sequence of Gmr1, the *G. morhua* LTR retrotransposon is present between base pair 7520 and base pair 12788 in GenBank accession AF104899 (Widholm et al. 1999). This entry describes 14.976 Kb of the *G. morhua* immunoglobulin light-chain gCL5 gene cluster region. The retrotransposon lies between two immunoglobulin light-chain L1 regions. The 5' LTR extends from position 12788 to position 12385, and the 3' LTR extends from position 7925 to position 7520. The retrotransposon is 5,269 bp long, and its structure is illustrated in figure 1a. The LTRs differ in length by 2 bp and share 98% identity. Both LTRs start and end with a 5-bp inverted repeat, 5'-TGTGG ... CCACA-3', which is similar to the retroviral consensus (Temin 1980). Two base pairs downstream of the 5' LTR is an 18-bp primer-binding site (PBS)(fig. 1a). This PBS is a 17/18-bp match to the 3' end of a human tRNA^{Ala}. This suggests that a *Gadus* tRNA^{Ala} (not present in the database) is used to prime the minus-strand DNA synthesis of Gmr1. A run of 12 consecutive purine residues is found 2 bp upstream of the 3' LTR and likely serves as the priming site for plus-strand DNA synthesis of the retrotransposon. The element is not flanked by a short duplication of the genomic target site.

Analysis of the sequence between the LTRs for potential ORFs showed that all of the amino acid motifs expected in an LTR retrotransposon are present in either one of two reading frames (fig. 1a). Six reading frame changes are required to generate an uninterrupted ORF, however, suggesting that this particular copy is no longer functional. Initial BLAST searches with the predicted protein products of the cod element ORFs suggested that this retrotransposon was a member of the

Key words: vertebrate LTR retrotransposon, *Gadus morhua*, POL domain order.

Address for correspondence and reprints: Margaret Butler, Department of Biochemistry, University of Otago, P.O. Box 56, Dunedin, New Zealand. E-mail: margi@sanger.otago.ac.nz.

Mol. Biol. Evol. 18(3):443–447. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

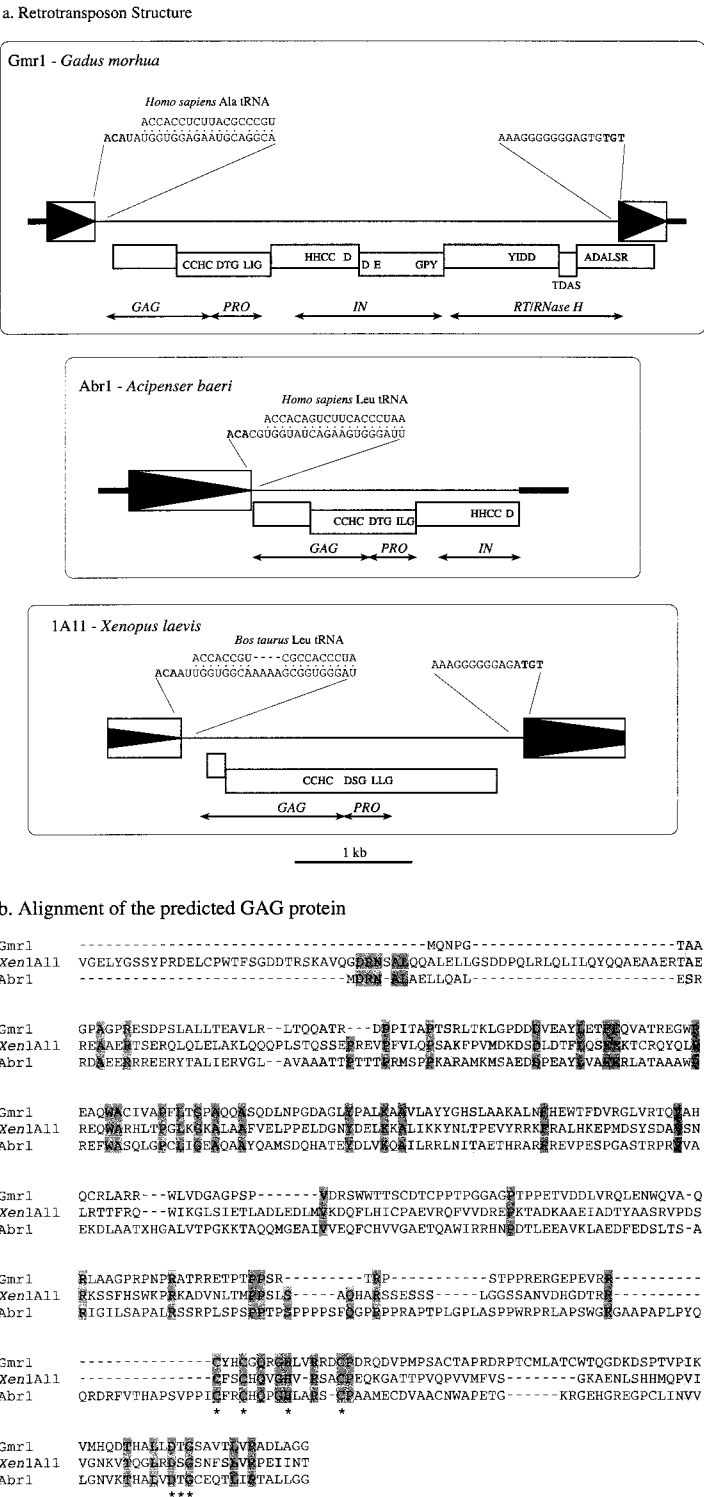


FIG. 1.—Gmr1 and similar retrotransposons found in vertebrates. *a*, Retrotransposon structure and POL domain order. Boxes with black triangles represent long terminal repeats (LTRs). Note that the 1A11 sequence was derived from a cDNA, and therefore the LTRs lack either the U5 or the U3 region. The last three bases of the left LTR (in bold) are shown, followed by the primer-binding site (PBS). Above the PBS are the complementary nucleotides at the 3' end of the corresponding tRNA. Also in bold are the first three bases of the right LTR, following the polypurine tract. Internal open reading frames (ORFs) are shown as shaded boxes. Offset boxes represent the predicted phase shifts necessary to maintain an ORF that will include the motifs frequently found in LTR retrotransposons. Motifs illustrated are as follows: CCHC, nucleic acid-binding; DS/TG, aspartic protease active site; I/LI/LG, C-terminal protease domain; HHCC, N-terminal region of IN; 'DDE,' core IN domain; GPY, C-terminal region of IN; YxDD, reverse transcriptase; TDAS/ADALSR, RNaseH motifs. The placement and extent of the protein domains of the POL ORF were determined by comparison with other full-length retrotransposons using the BLASTP, TBLASTN, and BLAST2 programs at the NCBI website (<http://www.ncbi.nlm.nih.gov>) (Altschul et al.1997). A common scale is shown at the bottom. *b*, Alignment of the predicted GAG amino acid sequences of the retrotransposons Gmr1, Abr1, and 1A11. Multiple alignments were created with CLUSTAL W, using the BLOSUM62 similarity matrix. Residues identical in all elements are highlighted in dark gray, similar residues in lighter gray. The Zn-finger (GAG) and DT/SG (protease) motifs are highlighted by asterisks.

Ty3/gypsy group. Examination of the cod sequence, however, revealed a most unusual feature. Unlike other Ty3/gypsy retrotransposons and retroviruses, the IN coding sequence of the cod element was upstream of the RT encoding sequence. This domain order is the same as that found in the Ty1/copia group. A similar structure can be discerned in a retrotransposon fragment from the sturgeon *Acipenser baeri* (fig. 1a). The relevant GenBank entry (AJ245365) describes a partial sequence of the immunoglobulin light-chain variable region in *Acipenser* (M. L. Lundqvist and L. Pilstrom, unpublished data). Within the ~8 kb of sequence preceding the IgLV gene, there is a recognizable PBS (tRNA^{Leu}), a GAG-encoding region (fig. 1b), then a protease-encoding region (PRO), followed by an IN-encoding region. There is, however, no apparent RT domain encoded within this *Acipenser* sequence. In a TBLASTN search of the nr database using as a query only the Gmr1 POL region shared by Gmr1 and Abr1, it can be seen that Abr1 is easily the closest known relative of Gmr1. For example, the match to the cod sequence in accession AF104899 has a score of 10^{-131} , the match to Abr1 has a score of 5×10^{-46} , whereas the next best match is to a *Drosophila buzzatii* element, Osvaldo, at 8×10^{-19} (Pantazidis, Labrador, and Fontdevila 1999).

The coding capacity of each of the cod element ORFs is outlined below. The 5' region of Gmr1 encodes the GAG protein, which is the structural component of the virus-like particle of the LTR retrotransposons. In Gmr1, the GAG protein contains a putative Zinc-finger RNA-binding site, CX₂CX₄HX₄C, which is found in many LTR retrotransposons. Apart from this motif, gag sequences are generally little conserved among different LTR retrotransposons. However, further sequences upstream of this motif in the cod element can be recognized as homologous to the corresponding regions of Abr1 from the sturgeon *Acipenser* and in a retroelement from *Xenopus* (1A11) already described by Greene et al. (1993) (fig. 1b). Gmr1, Abr1, and 1A11 also each contain, in a region 3' of the Zn-finger motif, a motif (DS/TG) resembling the active site of the aspartic protease domain of POL. The RNA-binding domain of the putative gag gene and the protease domain of the POL region in Gmr1 are encoded in the same reading frame without an intervening termination codon. The arrangements of the elements in *Acipenser* (Abr1) and *Xenopus* (1A11) are similar to that of Gmr1, with the Zn-finger and the protease encoded in the same ORF.

A phylogenetic analysis was conducted using multiple alignments of each of two POL domains (RT and IN) so that the relationship of Gmr1 to other LTR retrotransposons could be examined. The tree constructed using the seven motifs of the RT domain (Xiong and Eickbush 1990) is shown in figure 2. Gmr1 is clearly and robustly grouped among the Ty3/gypsy elements on the basis of the RT domain. Initial BLASTP searches had indicated that over the RT/RNaseH region, the retrotransposons most closely similar to Gmr1 were Osvaldo from *D. buzzatii*; Ted, the cabbage looper retrotransposon; 17.6 from *Drosophila melanogaster*, and Tom from

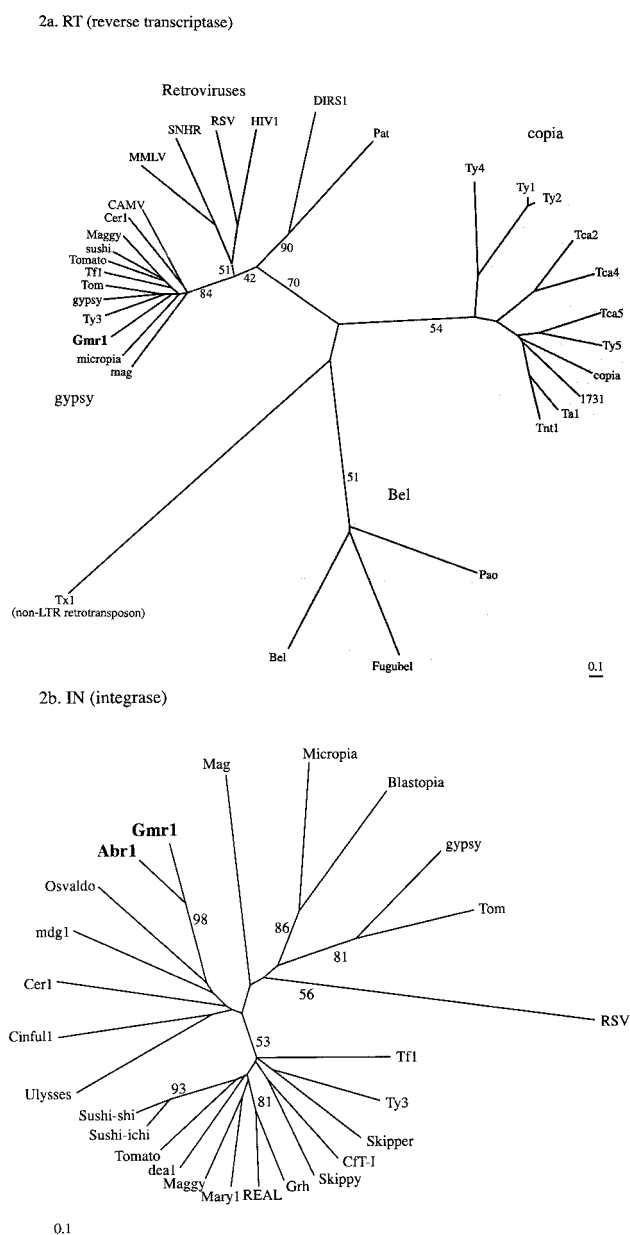


FIG. 2.—Phylogenetic analysis of Gmr1. The RT distance trees were constructed from alignments of the RT domain of a wide variety of retroelements representing the four major groups of LTR elements (a); these four major groups are highlighted. The IN distance tree was constructed from an alignment of a 5' region of the IN domain in Abr1 shared by a range of Ty3/gypsy elements and rooted using the RSV sequence (b). The alignments are available as supplementary material at the MBE website. Tree construction and bootstrap analyses were conducted using the PHYLIP package of programs (Felsenstein 1989). The percentage of bootstrap support is shown on the tree branches; distance shown is the categories distance of PROTDIST. Retrotransposon names and accession numbers for fungal elements are as follows: *Saccharomyces*—Ty1 S69982, Ty2 S45842, Ty3 M23367, Ty4 X67284, Ty5 U19263; *Schizosaccharomyces pombe*—Tf1 A36373; *Candida albicans*—Tca2 AF050215, Tca4 AF078809, Tca5 AF065434; *Magnaporthe grisea*—MAGGY L35053, Grh T18350; *Cladosporium fulvum*—CfT1 AF051915; *Alternaria alternata*—REAL BAA89272; *Fusarium oxysporum*—Skipper L34658; *Tricholoma*—MarY1 AB028236. Retrotransposon names and accession numbers for slime mold elements are as follows: *Dictyostelium discoideum*—DIRS1 M11340, Skipper AF049230. Retrotransposon names and accession numbers for invertebrate elements are as follows: *Panagrellus*

Drosophila ananassae. These are all Ty3/gypsy elements.

Trees constructed using an alignment of retroelement IN domains from Ty3/gypsy, Ty1/copia, Bel-like elements, and retroviruses also clearly placed Gmr1 within the Ty3/gypsy group (data not shown). In addition, a tree based on an alignment of the IN section present in Abr1 placed Gmr1 and Abr1 as close relatives (fig. 2b). Gmr1 and Abr1 appear to form a clade within the Ty3/gypsy group. This may indicate that their unusual POL domain order came from a unique event inside the Ty3/gypsy group.

A further feature of the phylogenetic analyses is that both the RT and the IN trees indicate that Gmr1 belongs within a group of Ty3/gypsy elements which is phylogenetically distinct from the group which contains sushi, the only full-length vertebrate Ty3/gypsy element previously described (Poulter and Butler 1998).

The classification of retrotransposons has been based mainly on the phylogenetic relationships generated by comparison of the amino acid sequence of the shared characteristic, the RT domain (Xiong and Eickbush 1990). Other characteristics, however, can also be used to distinguish one group from another: for example, the presence or absence of LTRs. The presence of an *env*-like domain was thought to be a distinguishing feature of vertebrate retroviruses, but it is now known that many Ty3/gypsy retrotransposons, some Ty1/copia retrotransposons, and some BEL-like elements have *env* genes. Another major distinction that has been used to classify LTR elements is the domain order in the POL ORF (Bowen and McDonald 1999). Ty3/gypsy and BEL elements share the POL arrangement of the vertebrate retroviruses, PRO-RT/RNaseH-IN. In contrast, Ty1/copia elements have the order PRO-IN-RT/RNaseH. On the basis of the order of POL domains, LTR retrotransposons can therefore be divided into two groups. There are at least two problems with this bipartite division. First, phylogenetic analyses suggest that Ty1/copia elements are more closely related to Ty3/gypsy and retroviral elements than are the BEL-like retrotransposons (fig. 2). The Ty3/gypsy-vertebrate retrovirus-Bel grouping therefore appears to be polyphyletic. The present analysis presents a second difficulty for this bipartite division of the LTR retrotransposons. The cod element,

Gmr1, and the sturgeon homolog, Abr1, belong with the Ty3/gypsy group of retrotransposons on the basis of RT and IN sequence similarity. However, their domain order in the POL ORF is that found in Ty1/copia retrotransposons. All other retroelements in which the IN domain is 5' of the RT/RNaseH domain have previously been shown to be members of the Ty1/copia phylogenetic group. Gmr1 and Abr1 do not fit easily into present schemes of LTR retrotransposon classification.

It may be suggested that the domain order in Gmr1 is simply due to some internal rearrangement subsequent to its integration. For several reasons, however, we believe that the structure shown in figure 1 represents the original form of Gmr1. The LTRs are almost identical, implying that the element was recently mobile. Gmr1 also contains all of the expected motifs of a functional element; that is, there are no essential parts missing from the IN and RT/RNaseH domains, and these domains are not internally rearranged. Indeed, only five nucleotide changes would be necessary to re-create an apparently intact element. Abr1, an element from the distantly related sturgeon, has an identical structure but has obviously been evolving independently for a sufficient length of time for all but the most highly conserved regions to diverge. It seems unlikely that Gmr1 and its closest relative would each suffer the same rearrangement subsequent to their integration, a type of rearrangement we have not encountered in any other retrotransposon. The PRO, RT/RNaseH, and IN domains are all separated by extensive spacer regions in LTR retrotransposons that would facilitate retention of functionality following internal rearrangements.

Another feature of interest is that Gmr1 appears to fall within a Ty3/gypsy lineage not previously encountered in vertebrates. This is supported by a recent analysis (Marin and Llorens 2000) which tentatively placed the RT sequence of Gmr1 within a group named the "Osvaldo" group by Malik and Eickbush (1999). Gmr1 and Abr1 therefore belong within a group which is phylogenetically distinct from the two LTR retrotransposon groups previously described from vertebrates, the Tf1/sushi (Poulter and Butler 1998) and mag/easel (Miller et al. 1999) groups. The Tf1/sushi group contains sushi and Hsr1 from the cave salamander *Hydromantes supramontis* (Marracci et al. 1996). Many vertebrate LTR retrotransposon fragments fall into the Tf1/sushi group (Miller et al. 1999). The mag/easel group contains easel, an LTR retrotransposon fragment from the chum salmon (Tristem et al. 1995), and some related fragments (Miller et al. 1999). Gmr1, which is one of the few full-length vertebrate LTR retrotransposon described to date, represents a distinct vertebrate retrotransposon lineage. Further elements from the Gmr1 lineage would assist phylogenetic analysis. As the analysis of retroelements continues, not only their abundance in genomes, but also the great plasticity apparent in their structure is becoming clearer. The discovery of a lineage of vertebrate Ty3/gypsy retrotransposons with a Ty1/copia-like POL domain order illustrates this plasticity. The structure of Gmr1 and related elements would almost certainly prevent their detection by methods employing redundant

←

redivivus—Pat X60774; *Drosophila*—copia P04146, gypsy P10401, micropia S02021, 1731 S00954, Tom S34639, BEL U23420, blastopia S38635, Osvaldo CAB39733, Ulysses S18211, mdg1 ST0430; *Bombyx mori*—mag S08405, Pao L09635; *Caenorhabditis elegans*—Cer1 U15406; Retrotransposon names and accession numbers for plant elements are as follows: *Ananas comosus*—Deal T07863; *Nicotiana tabacum*—Tnt1 S04273; *Arabidopsis thaliana*—Tal S05465; *Lycopersicon esculentum*—Tomato L95349; *Zea mays*—cinfu T14595, CAMV AAA62375; Retrotransposon names and accession numbers for vertebrate elements are as follows: *Fugu rubripes*—sushi-ichi AF030881, sushi-shi AF083221, Fugubel (suzu) AF108421 (Frame, Cutfield, and Poulter 2001); *Acipenser baeri*—Abr1 AJ245365; *Gadus morhua*—Gmr1 AF104899; non-LTR retrotransposon Tx1 from *Xenopus laevis*—M26915; Names and accession numbers for vertebrate retroviruses are as follows: SNHR (snakehead retrovirus, from a fish)—AAC54861, MMLV P03355, RSV AAC82561, HIV-1 K02013.

PCR primers corresponding to conserved sequences in the PRO and RT domains (Miller et al. 1999). It is therefore possible that the lineage may be widespread in vertebrates, given its occurrence in two divergent fish species and an amphibian.

LITERATURE CITED

- ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**:3389–3402.
- BOWEN, N. J., and J. F. McDONALD. 1999. Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retroviral-like elements. *Genome Res.* **9**:924–935.
- BRITTEN, R. J., T. J. MCCORMACK, T. L. MEARS, and E. H. DAVIDSON. 1995. Gypsy/Ty3-class retrotransposons integrated in the DNA of herring, tunicate, and echinoderms. *J. Mol. Evol.* **40**:13–24.
- COOK, J. M., J. MARTIN, A. LEWIN, R. E. SINDEN, and M. TRISTEM. 2000. Systematic screening of *Anopheles mosquito* genomes yields evidence for a major clade of Pao-like retrotransposons. *Insect Mol. Biol.* **9**:109–17.
- FELSENSTEIN, J. 1989. PHYLIP—phylogeny inference package (version 3.2). *Cladistics* **5**:164–166.
- FLAVELL, A. J., V. JACKSON, M. P. IQBAL, I. RIACH, and S. WADELL. 1995. Ty1-copia group retrotransposon sequences in amphibia and reptilia. *Mol. Gen. Genet.* **246**:65–71.
- FRAME, I. G., J. F. CUTFIELD, and R. T. M. POULTER. 2001. New BEL-like LTR-retrotransposons in *Fugu rubripes*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. *Gene* (in press).
- GREENE, J. M., H. OTANI, P. J. GOOD, and I. B. DAWID. 1993. A novel family of retrotransposon-like elements in *Xenopus laevis* with a transcript inducible by two growth factors. *Nucleic Acids Res.* **21**:2375–2381.
- HULL, R. 1999. Classification of reverse transcriptase transcribing elements: a discussion document. *Arch. Virol.* **144**:209–214.
- 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.
- MARIN, I., and C. LLORENS. 2000. Ty3/Gypsy retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. *Mol. Biol. Evol.* **17**:1040–1049.
- MARRACCI, S., R. BATISTONI, G. PESOLE, L. CITTI, and I. NARDI. 1996. Gypsy/Ty3-like elements in the genome of the terrestrial *Hydromantes* (Amphibia, Urodela). *J. Mol. Evol.* **43**:584–593.
- MILLER, K., C. LYNCH, J. MARTIN, E. HERNIOU, and M. TRISTEM. 1999. Identification of multiple Gypsy LTR-retrotransposon lineages in vertebrate genomes. *J. Mol. Evol.* **49**:358–366.
- PANTAZIDIS, A., M. LABRADOR, and A. FONTDEVILA. 1999. The retrotransposon Osvaldo from *Drosophila buzzatii* displays all structural features of a functional retrovirus. *Mol. Biol. Evol.* **16**:909–921.
- PETERSON-BURCH, B. D., D. A. WRIGHT, H. M. LATEN, and D. F. VOYTAS. 2000. Retroviruses in plants? *Trends Genet.* **16**:151–152.
- POULTER, R., and M. I. BUTLER. 1998. A retrotransposon family from the pufferfish (fugu) *Fugu rubripes*. *Gene* **215**:241–249.
- PRINGLE, C. R. 1999. Virus taxonomy—1999. The universal system of virus taxonomy, updated to include the new proposals ratified by the International Committee on Taxonomy of Viruses during 1998. *Arch. Virol.* **144**:421–429.
- ROEST CROLIUS, H., O. JAILLON, C. DASILVA et al. (12 co-authors). 2000. Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.* **10**:939–949.
- TEMIN, H. M. 1980. Origin of retroviruses from cellular movable genetic elements. *Cell* **21**:599–600.
- TRISTEM, M., P. KABAT, E. HERNIOU, A. KARPAS, and F. HILL. 1995. Easel, a gypsy LTR-retrotransposon in the Salmonidae. *Mol. Gen. Genet.* **249**:229–236.
- WIDHOLM, H., A. S. LUNDBACK, A. DAGGFELDT, B. MAGNADOTTIR, G. W. WARR, and L. PILSTROM. 1999. Light chain variable region diversity in Atlantic cod (*Gadus morhua* L.). *Dev. Comp. Immunol.* **23**:231–240.
- WRIGHT, D. A., and D. F. VOYTAS. 1998. Potential retroviruses in plants: Tat1 is related to a group of *Arabidopsis thaliana* Ty3/gypsy retrotransposons that encode envelope-like proteins. *Genetics* **149**:703–715.
- XIONG, Y., and T. H. EICKBUSH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**:3353–3362.

PEKKA PAMILO, reviewing editor

Accepted November 3, 2000