

Identification of a non-LTR retrotransposon from the gypsy moth

K. J. Garner and J. M. Slavicek

USDA Forest Service, Northeastern Research Station,
Delaware, Ohio, U.S.A.

Abstract

A family of highly repetitive elements, named LDT1, has been identified in the gypsy moth, *Lymantria dispar*. The complete element is 5.4 kb in length and lacks long-terminal repeats. The element contains two open reading frames with a significant amino acid sequence similarity to several non-LTR retrotransposons. The first open reading frame contains a region that potentially encodes a polypeptide similar to DNA-binding GAG-like proteins. The second encodes a polypeptide resembling both endonuclease and reverse transcriptase sequences. All members of the LDT1 element family sequenced thus far have poly-A tails or A-rich tails of 12–18 nucleotides in length, but lack a poly-A addition signal in the expected location. The amplification of retrotransposon insertion junction regions in different gypsy moth individuals indicates that polymorphisms exist at some of the insertion sites, suggesting that this element is or was, until recently, capable of transposition.

Keywords: Gypsy moth, *Lymantria dispar*, non-LTR retrotransposon.

Introduction

Retrotransposons are common mobile genetic elements, which have been described in many eukaryotic organisms. Retrotransposons fall into two major classes, based on the presence or absence of long terminal repeats (LTRs), as well as sequence homology (Xiong & Eickbush, 1988b). The non-LTR retrotransposons are also known as LINE-like elements (Singer & Skowronski, 1985), poly(A)-type retrotran-

sposons (Boeke & Corces, 1989), or retroposons (McClure, 1991). Many non-LTR retrotransposons have been described in insects, including the Doc (O'Hare *et al.*, 1991), F (Di Nocera & Casari, 1987), I (Fawcett *et al.*, 1986) and jockey (Priimägi *et al.*, 1988) elements of *Drosophila melanogaster*, the T1Ag (Besansky, 1990) and Q (Besansky *et al.*, 1994) elements of *Anopheles gambiae*, and the R1Bm (Xiong & Eickbush, 1988a) and R2Bm (Burke *et al.*, 1987) families of ribosomal DNA insertions in *Bombyx mori*.

Gypsy moths (*Lymantria dispar*) are currently widespread forest pests in the north-eastern United States and the adjacent regions of Canada. Population markers have been sought to distinguish the North American gypsy moths introduced from Europe in 1869 from those recently introduced from Asia (Bogdanowicz *et al.*, 1993; Pfeifer *et al.*, 1995; Garner & Slavicek, 1996; Schreiber *et al.*, 1997). During characterization of an amplified DNA polymorphism in the gypsy moth, a sequence similarity was found with a non-LTR retrotransposon in the silkworm *Bombyx mori*. This paper describes the cloning and sequence analysis of a complete element of this gypsy moth retrotransposon family, designated LDT1 (*Lymantria dispar* transposable element 1). Several incomplete elements were also characterized. To our knowledge, this is the first non-LTR retrotransposable element described in the gypsy moth.

Analysis of these elements may prove to be useful for the development of a transformation vector for use in biological control (Pfeifer & Grigliatti, 1996). In addition, variability of retrotransposon insertion sites within the species may provide useful genetic markers for the identification of gypsy moth strains or populations.

Results

Identification and cloning of a gypsy moth retrotransposon

During the characterization of RAPD-PCR markers for the identification of gypsy moth strains, a gypsy moth RAPD-PCR-amplified fragment called A15_{N1800} was

Received 10 March 1998; accepted 10 August 1998. Correspondence: James M. Slavicek, USDA Forest Service, Northeastern Research Station, 359 Main Road, Delaware, OH 43015, U.S.A. E-mail: jslavicek/ne.de@fs.fed.us

sequenced. When this sequence was used to search the GenBank database, a partial non-LTR retrotransposon sequence in an intron of the alpha amylase gene of *B. mori* (accession number U007847) was found to be similar. Additional isolates of the *B. mori* retrotransposon have been described under the names BMC1 (Ogura *et al.*, 1994) and L1Bm (Ichimura *et al.*, 1997). The region that is similar to the *B. mori* element within A15_{N1800} extends 1489 basepairs and ends in an 18-basepair poly-A tail. Within this region the encoded amino acid sequences of the two species are 49% identical. The translated gypsy moth A15_{N1800} sequence also matches numerous reverse transcriptase sequences in GenBank.

Hybridization of the A15_{N1800} *Bam*HI/*Hind*III 0.7 kb fragment to a blot of *Eco*RI-digested gypsy moth genomic DNA (not shown) resulted in a smear indicating that this sequence is highly repetitive in the gypsy moth genome. The degree of hybridization to the probe was similar for the DNA from moths from Asia and North America. In order to characterize a complete copy of this repetitive element, a genomic library was constructed and screened by hybridization with the

A15_{N1800} *Bam*HI/*Hind*III 0.7-kb subfragment. Numerous positive clones were obtained, but many were found to contain only partial retrotransposon sequences truncated at the 5' end. For example, isolate LDT1-2 comprises approximately 1700 basepairs, which are homologous to the *B. mori* BMC1/L1Bm retrotransposon sequence, based on a comparison with translated open reading frames (ORFs). In order to identify a phage clone containing a complete retrotransposon, a consensus map was developed using the A15_{N1800} and LDT1-2 restriction maps as starting points. The map is shown in Fig. 1. The first restriction fragment used to probe genomic DNA blots was a 0.6-kb *Sa*I restriction fragment from LDT1-2. This probe hybridized to a 2.2-kb *Bam*HI fragment in genomic DNA and to a fragment of the same size in the phage clone containing LDT1-4, which later proved to be a complete element. The LDT1-4 element was restriction mapped and a 2.3-kb *Bam*HI fragment was found upstream of the 2.2-kb *Bam*HI fragment. The 2.3-kb *Bam*HI fragment was detected in genomic DNA. However, restriction fragments upstream of the 2.3-kb fragment in LDT1-4 did not hybridize to discrete bands in restriction enzyme-

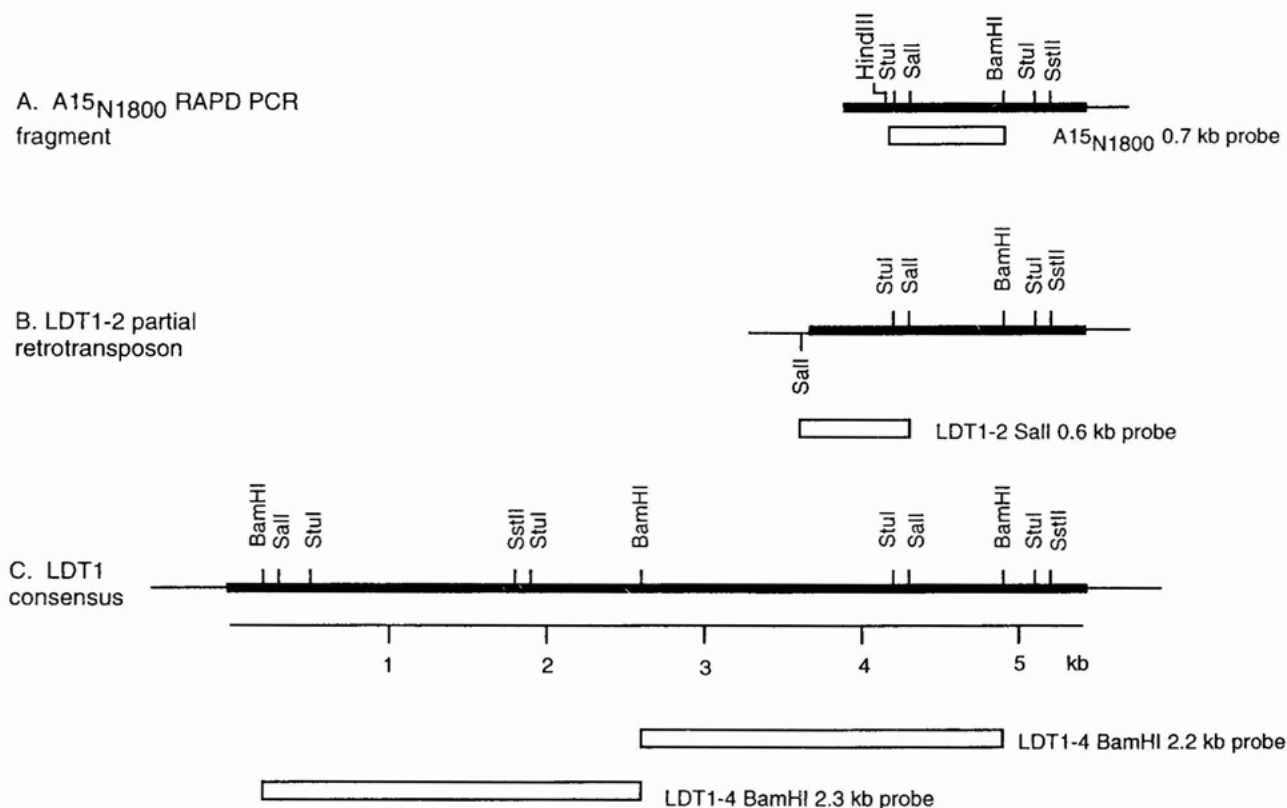


Figure 1. Consensus mapping of the LDT1 element. Heavy lines indicate the extent of the LDT1 sequences. (A) Restriction map of the PCR fragment A15_{N1800} containing a partial LDT1 element. (B) Restriction map of the partial retrotransposon LDT1-2. The *Sa*I site shown under the line is outside of the LDT1 sequence. (C) LDT1 consensus restriction map based on genomic mapping. The 5' endpoint of the retrotransposon is based on the sequence of LDT1-4, which matches the consensus restriction map.

digested genomic DNA. The diagnostic genomic fragments that were observed hybridizing to the above probes were the 2.2 and 2.3-kb *Bam*HI fragments, a 4.0-kb *Sa*I fragment, a 3.3-kb *Sst*II fragment, and 1.3, 2.3 and 1.0-kb *Stu*I fragments. All these fragments were found using hybridization in phage clones containing the elements LDT1–4 and LDT1–20. The elements LDT1–21 and LDT1–25 were found to contain fragments corresponding to the 5' end of LDT1–4, although the 3' ends of these elements were truncated during the cloning process. Comparison of sequences from these four elements was used to locate the 5' end of the LDT1 element, as described below.

Sequence characterization of LDT1

Two clones, containing the elements LDT1–4 and LDT1–21, were found to have restriction maps matching the complete consensus map. The retrotransposon in LDT1–4 was sequenced in its entirety. In addition, the 5' end regions of LDT1–21 and of two other elements, LDT1–20 and LDT1–25, were sequenced to determine the retrotransposon 5' end site (not shown). The sequenced regions of LDT1–20 and LDT1–25 extended $\approx$ 600 bp downstream from the start of the retrotransposon, and all three sequences differ from each other at several positions. When a consensus sequence (not shown) was derived from these three sequences and compared to each sequence individually, the LDT1–4, LDT1–20 and LDT1–25 sequences differed from the consensus by 0.5%, 0.3%, and 3.9%, respectively.

The complete retrotransposon sequence of the LDT1–4 element is shown in Fig. 2. Long terminal repeats were not observed; however, 11-basepair direct repeats (with one mismatch) were found flanking the retrotransposon insertion site (see Fig. 4 below). Beginning 737 basepairs from the beginning of the element, the retrotransposon sequence contains two long ORFs, the first of which could potentially encode a 537-amino acid sequence with homology to retrotransposon nucleic acid-binding proteins. Conserved nucleic acid-binding 'C-rich' domains are double-underlined in Fig. 2. The second ORF appears to begin with the second nucleotide past the ORF1 stop codon, and could encode a 998-amino acid sequence with significant similarity to endonuclease and reverse transcriptase sequences. Conserved endonuclease and reverse transcriptase domains are single underlined and dot underlined, respectively, in Fig. 2. Although the two ORFs of many retrotransposons overlap by a few nucleotides, if the LDT1 ORF2 extended further upstream, it would contain two in-frame stop codons before another start codon appears. However, it is possible that a frameshift could allow

translation of a second ORF that overlaps the first. Ninety basepairs after the putative termination codon of ORF2 is an 18 nucleotide poly-A string; however, no poly-A addition signal was seen within the 3' untranslated region of this element or of any partial elements that were sequenced. The poly-A tails of these elements range from 13 to 19 nucleotides in length (eight 3' end sequences were analysed).

The polypeptides potentially encoded by the LDT1 retrotransposon share similarity with several sequences found in Genbank. As mentioned above, a high degree of similarity was found to a retrotransposon family from *B. mori*. The DNA sequences of these two elements are $\approx$ 50% identical within the ORFs. The amino acid sequences potentially encoded by the LDT1 and *B. mori* element ORFs are also $\approx$ 50% identical. The LDT1 sequence also has regions of similarity to several dipteran non-LTR retrotransposons. The LDT1 reverse transcriptase is 30.0% identical to that of the *Drosophila* element Doc (O'Hare *et al.*, 1991), 31.3% identical to the reverse transcriptase of the jockey element (Priimägi *et al.*, 1988) and 29.3% identical to that of the F-element (Di Nocera & Casari, 1987; Kerber *et al.*, 1996). Other dipteran elements with reverse transcriptase sequences related to LDT1 are non-LTR retrotransposons from the midge *Chironomus thummi* (31.0% identical) (Blinov *et al.*, 1993) and the mosquito *Aedes aegypti* (30.0% identical) (Mouchès *et al.*, 1992). Lepidopteran retrotransposons such as the *B. mori* ribosomal DNA-associated R1Bm (21.1% identical) (Xiong & Eickbush, 1988a) and R2Bm (21.8% identical) (Burke *et al.*, 1987) and the telomere-associated TRAS1 (23.6% identical) (Okazaki *et al.*, 1995) also share limited sequence similarity with LDT1. The GAG-related and endonuclease sequences of these elements share short regions of similarity as well.

The first ORF in LDT1 could potentially encode an amino acid sequence related to the GAG protein of retroviruses and to similar sequences from retrotransposons. Although overall homology between these sequences is low, retrotransposon ORF1 sequences generally contain cysteine-rich motifs with the consensus CX₂CX₄HX₄C where X can be any amino acid (Berg & Shi, 1996). These motifs resemble the zinc finger sequences found in numerous nucleic acid-binding proteins. The nucleocapsid fragment of the GAG protein from HIV, HTLV and other retroviruses has been shown to associate with zinc (Bess *et al.*, 1992). The LDT1 sequence contains three C-rich motifs at the same relative positions as in several other non-LTR retrotransposons. An alignment of the C-rich sequences of the LDT1 ORF1 with the L1Bm, Doc, jockey and the F-element Fw non-LTR retrotranspo-

CATTCCGCGTCGGAGACTCGTAGCGCAGGACGTGCGCCTCGCTGCTCTCCATTTCGCGCGCTCACCCCGCACCCGCCCCCTGCGGTCTGCGTTCTCG 100
 CTCGCTCTCGCTCGCTCTCGCAGCGAACCACGCGTCGCGTCAAGCCGCCATCTTGTTCACGCTCTGTGGCCGTTTCAGTGGCCCCGCTCCGGGTGGCCGTGT 200
 ATACGGTTGTCTGTGGCAGTTTCGCTGTTTGGACGCAGGTGTTGCCGATCCCAACTGCACGAGTCGACTTTTCAAGACGTCGAGTGCAGATTCTTACC 300
 GTTGGTAGATAAAAGTCAACCTATTTTGGCCCCCTCAGGGTAGAAAAACCATATGTTTGGTCCCGGCGAAACCTCACGGTGTCAACGAAGCCGGGTG 400
 AGGCAAGGCCCGATTGCTGCGGAGCCCCCTCAGGGCAGAGCTAAACATCGGAAGGGGTAGAGCATCCGCCCTGGGGGATCTCTAGTCGGACCTA 500
 ATCTACGGCTTTTCAACATTTTCTGCGAAGGCCAAGCGCCGAGCAGTAGGAACCTATCGCACATTAAGATTGGCGAAGGATTACAAGCTTAGGGCA 600
 CACTAATTAAGTACTCATCTCCCGACCTGCACCCGGTTGGGAGGGACACGCACCACACACAACCCAGGGTATCCACCCAGGGCGCTAGGTTTTTTACA 700
 TCCCTCTTAGCTATCGTTTTTTTCCGTTAGCCACGATGGCGGAAAAAACCATCACTATAGACGCCCTTCTGGCAATTCAGAAATCAAAGCTTATATCG 800
 M A E K T I T I D A L L A I P E I K A Y I 21
 CTCAACATTTCGGAGCGATAACGATCCTTCTCCGCAACCCAAACCGCTCGATCCACGCCCGAGCGCTCGAACCGAACTACCGCCGCTCGGTACGGC 900
 A Q H F G A I T I L P P Q P K P A R S T A Q P L E P E L P A V G T A 55
 GATGGAACGAGCGGCTTGTCTCTCTCAACTCGTCTTCTCTCTCTCTCTCGAGTCGGAAGATTCTGTCATATGAGGGCTCATTATTGAGGTCCGC 1000
 M E T E P V L S S S N S S S S S S S S E S E D S S N E G S F I E V R 88
 TCTCGCAACGAGCAAAAGGACCAAAACACCAAGAGGCCCCCTTAAGACCCCAAAATCAAGCAGACTGACTCGGCAACAACATCTGCCGAGCCC 1100
 S R K R S K R T K N T K E A P S K T P K I K Q T D S A T S A A A 121
 CGCTTGAACGACCGCCACGGCCGCTCGGCCCGCCAGTCGCTCACAGGCAGCCTAACCGCTGCCCCACCACTACGGACAGCAATACCGTCGCCCC 1200
 P L A T T A T A A P A P P V V S Q A A L T A A P P P T D S N T V A P 155
 ACCGACGGTCAAAAAATCACTCGCTCCGCCGCCCTATATATTAGGGACAAGGACAAGTGGCTAGCCATCTCCGCATGGCTAAACACTAACAGGATTTTCG 1300
 P T V K K S L A P P P L Y I R D K D K W L A I S A W L N T N R I S 188
 TATAATCTGCGAAATCCACCCCAAGGTATACGAGTGGACCTTCACTCGTCAGCCGACCCGATCTGTATCAAAATGCTAGGGAATCAACAGATTC 1400
 Y K S A K S T P Q G I R V D L H S S A D H R S V S K M L G N Q Q I 221
 CCTACCATCTACACCTCCCTGAAGCAAACTTCTGCGTGTGTAATACGCAACGTACCGAGGGAATAGAGACCAAGAGATTCTAGAAATCTCTTAA 1500
 P Y H T Y T L P E A K L L R A V I R N V P R E I E T K E I L E S L K 255
 AACCCAGGATCTACCGGTGTAGAAGTACACCGGATGATCCGGGTAGGGGTGCTTATCCCTCAACATGATACTCGTCTGCCTTACAAATAACGCAGAG 1600
 T Q D L P V V E V H R M I R G R G R Y P L N M I L V C L T N N A E 288
 GGTAAGGCATTTTAAATTAAGACCATCTCGGTCTATCTGGGTCTCGGTTAGGCCCCACACAAAAATGGCAACCTGGCACAATGCCACAAGTGCC 1700
 G K G I F K I K T I C G L S G V S V E P P H K N G N L A Q C H K C 321
 AACTGTATGGACAATCGTCAAAAACTGCTTTGCCCGTCTAGGTGCGTTAAGTGCCTAGGTGACCATCACACCTCTCAGTGTGAACGGCCAAAGGACAT 1800
 Q L Y G Q S S K N C F A R P R C V K C L G D H H T S Q C E R P K D I 355
 CTCCTCTGCAAGAGCCACCAGCGTGTGTTCTCTGTGGAGAATACGGACACCCGGCTAACTACCGGGCTGCCAAGAGCGCCGAGGAGTTAGTCCGC 1900
 S L C K E P P A C V L C G E Y G H P A N Y R G C P R A P R R L V R 388
 CAGCCCAACACAAACGGCAAGGCCTTGTACTATAACAAGACCTTCGTCACGCCCCCTTACCAACCCACAATGCGTGGGCTCGCCCTCTGTAAACAGCA 2000
 Q P N T N G K A L Y Y N K T F V P A P L P T H N A W A R P L L N S 421
 AGGAGGCTTTCCCAACCTGCCCCAACGGCAGCCCGCTGCCCCACCCGCAATCTCAACTCAAACCAACAACCGGGCCAAACAGCCCATGGTGG 2100
 K E A F P T L P N R Q P A A P P A T S Q L K P P T T G P N P T G G 455
 TAATCCCAAGCCCCCTCAGGGCGCCCTATTCTATCGTCGTCGTAAGCGCGCCCCCAAGGCCCAAGGACTAGATCCAGATCTAGCCCTGGTG 2200
 N P K P P Q G A P I P I V V A K S A P P K A P Q G L D P D L A L V 488
 GCCAATCTCGCAGCTCAATCAATTTACGGAAATCCGCGAAATTGCAGCCGAATTAAGAAGGAACGAGGGAACAACCTCGCCCTCTTAATACCGCAA 2300
 A N F A A Q I N F T E I R E I A A E L R R N E G N N L A L L N T A 521
 TTAAATTTTCCCCACGTTAGAGCGTCTAGGCTCGCTTAAATTCAAATAAATGGATAGGTACACTCCAGGCAGAGTTAAACCCAGACACCTCAAAATAG 2400
 I K F S P T L E R L G S L K F K ter 537
 M D R Y T P G R V K P R H L K I 16
 GGACATTCAATGCAATAGCCTAACCCAACAGAAGGATGAAGTAGGCACATTCCTTCGGGAACACCAGCTCGACATCTTGCTAGTACAAGAGACGTTCT 2500
 G T F N A N S L T Q Q K D E V G T F L R E H Q L D I L L V Q E T E L 50
 TAAACCATTTAACAAGGACCCTAGAGTCGCCAATATAATAGTTCGCAACGATAGGACAACCTCACCAATGGGCGGAACGCTTATCTACTACAAAGG 2600
 K P F N K D P R V A N Y N I V R N D R T T S P M G G T L I Y Y K R 83
 TCGCTGCATTGCACACCTGTGGATCCGCCCCCTCTGATGTACATTGAAGCTTCAATTTGACGGCTAGCCCTCTCAGGCCATCAGCCTATCACCCCTGTAT 2700
 S L H C T P V D P P L M Y I E A A S I C R L A L S G H Q P I T L V 116
 CCGCATATCTCTCCCAATAGAGAATTCGAAACAGTTGGTTGAAAAAGATCTAAAGCCCTTGCTAGAAATTGGGAGGCGCGCTCATCTATTGGCGGTGA 2800
 S A Y L S P N R E H S K Q L T V E A K D L K A L L E L G G A V I I G G D 150
 CCTTAACGCCAAGAACCTCTTGGAGCTGTCAAAGTTCCAATAAGAGGGGCAAGACTTTAGAGCTATTCTGCTGACAGACTGGGATTTGATGTCATAGCG 2900
 L N A K N T S W S C Q S S N K R G K T L E L F A D R L G F D V I A 183
 CCCATAGAGCGACACACTTCCCTTTAAGCTTTCGCACAGGCGGACATACTCGATATCTTCTTCTCAAAAACATAAACCTTCGTTTATGTTCAATAG 3000
 P I E P T H F P F N V S H R P D I L D I E L L K N I N L R L C S I 216
 AAGTTCAACACGAGTTAGACTCAGATACCGCCCGGTGACTCTGGAATTAGCCTCCCGCACCCCGGCCCTCAGGCCACAATAACGAAAGGACCAAGAT 3100
 E V Q H E L D S D H R P V T L E L A S R T P G P S G H N N E R T K I 250

Figure 2. Complete DNA sequence of LDT1–4. Deduced amino acid sequences are shown using the single letter code. C-rich regions in ORF1 are double-underlined. Endonuclease and reverse transcriptase domains in ORF2 are single-underlined and dot-underlined, respectively. The location of PCR primers KG39 and KG40 are indicated.

A

		C	C	H	C		C	C	H	C		C	C	H	C	
LDT1	318	CHKCQLYGQSSKNC	5aa	CVKCLGDHHTSQ--C	14aa	CVLCGEYGHHPANYRGC	379									
L1Bm	232	CYRCQLYGHSAARNC	5aa	CVKCLGDHATLD--C	12aa	CVLCLKQGHHPANYRGC	291									
Jockey	405	CQRCQIFGHSKNYC	5aa	CGKCSGPHMTGFALC	6aa	CINCGGD-HVSTDKSC	459									
Doc	399	CTNCQEYGHHTKAYC	5aa	CVVCSEPHTTAN--C	11aa	CSNCGEK-HTANYRGC	456									
Fex	365	CTNCQEYGHTRSYC	5aa	CVVCGDLHDSKQ--C	11aa	CNNCGGN-HTANYRGC	422									

B

		N		E		Y	
LDT1	1	MDRYTPGRVKPRHLKIGTFNANSL	15aa	LDILLVQETF	23aa	PMGGTLIYYK	29aa
L1Bm	1	MAQKGRGKPYSLTLAFYNANGL	15aa	IDILLVQETC	23aa	SKGGTAIYYR	29aa
JOCKEY	1	MTQP-TLKIGLWNARGL	15aa	IDVMLTTETH	24aa	RGGSAVIKS	27aa
Tcruzi	93	-----DIEQNPGPIAVLQMNVSCL	12aa	ADIIAIQETW	24aa	GGGVAVLVRK	31aa
L1Hs	1	MTGSNSHITILTININGL	17aa	PSVCCIQETH	25aa	KAGVAILVSD	27aa

		D N		D		SDH	
LDT1		AVIIGGDLNAKNTSWS	24aa	PIEPHTFPFNVSHRPDILDIFLLK	15aa	DSDHRPVTLELAS	235
L1Bm		SVILAGDLNCHHTRWN	25aa	PPTPTCYPYNILRPSTIDLALLR	15aa	DSDHRPVVMQLGR	229
JOCKEY		KFIAGGDYNAKHAWWG	24aa	TGEPTFYSYNPLLTSPALDFFITC	15aa	SSDHLPLAVLHA	223
Tcruzi		PLLLCGDFNMHHPQWE	28aa	PGEITYAR--GTRERSCIDLTWSK	13aa	LSDHYVLTFTLHQ	320
L1Hs		HTLIMGDFNTPLSTLD	34aa	-TEYTFFS-APHHTYSKIDHILGS	16aa	LSDHSAIKLELRI	239

C

		1		2	
		++K+		RP+++ K+ +	
LDT1	507	AALNNNIFPAQWKEAVVIGIHKPGKR	6aa	YRPISLLRGLGKIYERIVLSRLRTFAEAN	32aa
L1Bm	490	AAMTHCIFPAVWKEADVIGIHKPGKP	6aa	YRPISLLSTIGKIYERLLRKRLWDFPTAN	32aa
Jockey	483	SVLDVGYPFKAWKSASIIIMHKTGKT	6aa	YRPTSLPSLGKIMERLILNRLLTCKDVT	32aa
Doc	476	AIKLGYPFQKWKSTIVMIPKPGKD	6aa	YRPISLLTCLSKLFEKMLLRISPHLRIN	32aa
Fex	479	AITKLGYPFQRWKMMKIIMIPKPGKN	6aa	YRPISLLSCISKLFEKCLLIRLNQHQTYH	32aa

		3		4	
		D FD + + +		POG ++ L + +	
LDT1		--MNRYTGVFLDVAKAFDKVWHEGLIYKLYLLKIP	24 aa	TSHPIKAGVPQGSVLSPLFTLYTSDMPKFNVRH	
L1Bm		RRKQIPTGASFFDIAKAFDKVWHNGLIYKLYNMGVP	24 aa	RPRQLTAGVPQASALSPLLSLYINDIPRSPETH	
Jockey		--NKEYAVGAFLDIQQAQFDRVWHPGLLYKAKRLFP-	24 aa	SIKPIAAGVPQGSVLGPTLYSVFASDMPHTPVT	8aa
Doc		--HREYCTAIFLDVAQAFDRVWLDGLLFKIIKLLP-	24 aa	RDCAIEAGVPQGSVLGPILYTLTADFPIDYNLT	
Fex		--YREYCTAVFLDVSQAQFDKVWLDGLMFKIKISLP-	24 aa	TVHTIEAGVPQGSVLGPTLYLIYTADIPTNSRLT	

		5		6		7	
		+ADD +		+ + K+		+LG	
LDT1		KALYADDTALFCAGRSP	6aa	LQTAVTALANWFRNWRLEINPEKSQA	28aa	WTRQAKYLGVTL	780
L1Bm		LALFADDTAIYYSCKRM	6aa	LQIAVATLGQWFRKWRIDINPTKSAA	40aa	WAPKVYLGVTL	777
Jockey		IATYADDTAVLTKSKI	6aa	LQEYLDFAQQWAENWVRINAECAN	22aa	HHQAYKLGITL	757
Doc		TSTFADDTAILSRSKCP	6aa	LSRHLTSVERWLADWRISINVQKCKQ	22aa	QADEVTYLGVHL	742
Fex		VSTFADDTAILSRSRSP	6aa	LALYLIDIKKWLSDWRIKVNEQKCKH	22aa	KADEVTYLGVHL	745

Figure 3. Amino acid alignments of conserved regions of LDT1 ORFs 1 and 2 with related retrotransposon sequences. Amino acid positions are shown before and after sequences. (A) ORF1 cysteine-rich sequences with highly conserved residues indicated above the alignment. (B) Endonuclease conserved domains. Residues conserved among numerous retrotransposons (Feng *et al.*, 1996) are indicated above the alignment. Putative active site residues are marked with asterisks. (C) ORF2 sequence region containing seven conserved reverse transcriptase domains. Invariant (letters) and chemically conserved (+) residues observed in numerous retrotransposons are indicated above the alignment (Xiong & Eickbush, 1988b). Sequences included in the alignments are L1Bm (Ichimura *et al.*, 1997), Jockey (Priimägi *et al.*, 1988), Doc (O'Hare *et al.*, 1991), Fex (Di Nocera & Casari, 1987), L1Tc from *T. cruzi* (Martin *et al.*, 1995), and human L1 (Hattori *et al.*, 1986).

shows an alignment of conserved endonuclease sequence regions from LDT1 and several other non-LTR retrotransposons.

Similarly, comparison of a large range of retroviruses and retrotransposons has resulted in the identification of seven conserved reverse transcriptase sequence domains (Xiong & Eickbush, 1988b, 1994). The LDT1 ORF2 sequence can be aligned with other retro-element sequences in the regions of these domains and shows good agreement with highly conserved residues (Fig. 3C). The Y/FADD sequence (positions 696–699) is characteristic of non-LTR retrotransposons (Xiong & Eickbush, 1988b). The homologous region in HIV has been shown by site-directed mutagenesis to be required for reverse transcriptase function (Larder *et al.*, 1987).

Analysis of LDT1 genomic insertion sites

Some classes of non-LTR retrotransposons show insertion site specificity, integrating within 28S ribosomal genes (Burke *et al.*, 1987; Xiong & Eickbush, 1988a; Jakubczak *et al.*, 1990) or at chromosome ends (Okazaki *et al.*, 1995; Sheen & Levis, 1994). However, comparison of six cloned LDT1 retrotransposon insertion sites shows no sequence conservation. None of the LDT1 insertions that were sequenced were flanked by ribosomal RNA-encoding sequences or areas of short repeats that would be characteristic of telomeres. The LDT1 insertions appear to be randomly dispersed. An 11–13 bp repeat, probably a duplication of host DNA at the insertion site was seen in five out of six cases where the sequence was determined at both endpoints of a LDT1 element (Fig. 4). Three of these insertions, LDT1–23 A, B and C, were within one phage clone and were arranged in a cluster of head-to-tail tandem copies. All three elements were truncated and differed in length at the 5' end. In this cluster, the same sequence (differing at 1 or 2 positions out of 12) was repeated flanking and in between each copy.

The copy number of the LDT1 element was not determined because the genome size of the gypsy moth is not known. However, the density of complete and partial LDT1 elements in the genome was estimated by probing genomic library phage plaques with probes from the 5' end (nucleotides 34–398), middle (nucleotides 2128–2947) and 3' end (nucleotides 4898–5398) regions of the complete retrotransposon. A total of 2283 phage plaques were probed, and 33, 48 and 457 plaques hybridized to the 5' end, middle and 3' end probes, respectively. The average size of the phage clone inserts was 18 kb. Therefore, the approximate density of complete or nearly complete retrotransposons, extending to within the 5' end probe region is ≈ 1 per 1.2 million basepairs. The density of elements hybridizing to the 3' end probe, most of which would be truncated, would be almost 14 times higher, or ≈ 1 per 90 kb. This estimate is probably low because hybridizing phage clones could contain more than one element. This estimate also assumes the library is complete, although this has not been demonstrated. The relatively low number of phage plaques hybridizing to the middle probe indicates that the majority of truncated elements extend less than halfway through the length of the complete element.

Analysis of LDT1 genomic polymorphisms in Asian and North American gypsy moth populations

Several examples of polymorphisms at transposable element insertion sites have been described. Such polymorphisms may be useful as phylogenetic or population markers (Batzner *et al.*, 1994; Shimamura *et al.*, 1997). In order to look for polymorphisms at LDT1 element insertion sites in gypsy moths, PCR primers were designed based on sequences flanking the insertion junction for five complete or partial LDT1 elements. These primers were used to amplify DNA samples from gypsy moths from several different locations in North America and Asia. In three out of

```

LDT1-4      tttgctgtctattcaatCATTCC...ACACACA18gtctattcattcagtt

LDT1-21    ataattgttcataagccaacgaaCATTCC...ACACACA17gttcataagccagtcgctg

LDT1-23 (3 tandem partial elements)

                                ttcagcgaatcatcgaataGCTACA...LDT1-23B
LDT1-23B   ...ACACACA16gatcaccgaataTCTCTG...LDT1-23C
LDT1-23C   ...ACTCACA17tgatcacngaaTTGCGCA...LDT1-23A
LDT1-23A   ...ACACACA16gatcatcgaatacaaccg

```

Figure 4. Short duplicated sequences flanking LDT1 insertions. LDT1 sequences are in bold uppercase type. Flanking sequences are in lowercase type. Duplicated sequences are underlined. Clone LDT1–23 contains three head-to-tail tandem insertions.

the five cases, PCR using these primers resulted in nonspecific amplification characterized by multiple bands, many of which were present in all the individuals tested (not shown). However, two primer sets amplified fragments primarily of the predicted size in some but not all of the moth samples. A primer pair designed to flank the 3' junction site of the partial LDT1 retrotransposon in RAPD fragment A15_{N1800} and another primer set flanking the 5' end of the LDT1-2 partial retrotransposon amplified fragments that were polymorphic within the gypsy moth population. The identity of these amplified fragments was confirmed by cloning and partial sequence analysis (not shown). Approximately fifty Asian and fifty North American moth samples were analysed with each primer set (Table 1). The A15_{N1800} primer set amplified the expected 851 bp fragment in 71% of North American samples and 9% of Asian samples. The LDT1-2 primers amplified the expected 430 bp fragment in 69% of North American moth samples and 9% of Asian samples. While the overall percentages of positive samples was similar for the two insertion loci in the different North American populations and identical in the Asian population, the two insertions were not always seen in the same individual moths, indicating that the two insertion site loci are not linked.

The insertion site polymorphisms suggest that the LDT1 element has recently been or may now be active, therefore Northern blot analysis was used in an attempt to detect transcription of the element. Blots were made using RNA from whole adults and from gonads from fifth instar larvae and pupae (male and female). Also tested were RNA from gut tissue from

fifth instar larvae (undetermined sex) and from an *L. dispar* ovarian cell line, 652Y. The RNA blots were hybridized with the 5' end and middle LDT1 probes, but no discrete transcripts were detected from any of the tissue sources.

Discussion

The gypsy moth transposable element LDT1 was identified as a non-LTR retrotransposon based on sequence similarity to the reverse transcriptase of several such elements. Further characterization of LDT1 has confirmed that it has many features typical of this group, including the lack of long terminal repeats, a short duplication of target site DNA flanking the insertion, frequent truncation at the 5' end, and a poly-A or A-rich tail at the 3' end. The LDT1 element has two ORFs, the first of which encodes an amino acid sequence homologous to GAG-related proteins. The second ORF encodes an amino acid sequence similar to endonuclease and reverse transcriptase sequences, with conserved residues in the reverse transcriptase characteristic of the non-LTR retrotransposons.

The LDT1 elements characterized in this study have 13–19 bp poly-A tails but, unlike many other non-LTR retrotransposons, lack sequences resembling consensus poly-A addition signals near the poly-A tract. However, a few other retrotransposons lack typical poly-A tails or polyadenylation signals. Some elements have the usual AATAAA signal but unusual A-rich tails, such as (ATGAA)_n (Besansky, 1990) or (TAA)_n (Besansky *et al.*, 1994). The I element of *Drosophila melanogaster* (Fawcett *et al.*, 1986) and the LOA element of *Drosophila silvestris* (Felger & Hunt, 1992) both lack poly-A addition signals and end with (TAA)_n tails. The RNA termination and tailing mechanisms may differ in various families of non-LTR retrotransposons. The reverse transcriptase of the *B. mori* retrotransposon R2 has been shown to add nontemplated nucleotides to the 3' end of experimentally shortened R2 transcripts before initiating reverse transcription (Luan & Eickbush, 1995). The resemblance of some retrotransposon tails to telomere sequences has raised the theoretical possibility that telomerase may be involved in the tailing and possibly the reverse transcription of these elements (Besansky, 1990).

While the LDT1 sequence lacks a poly-A addition signal at the expected position, the sequence AATAAA appears at three locations within the LDT1 coding regions, beginning at nucleotides 2346, 3810 and 5193. The poly-A addition signal at position 2346 overlaps the putative stop codon TAA of ORF1, raising the theore-

Table 1. Results of PCR amplification of retrotransposon insertion site fragments.

	A15 _{N1800}	LDT1-2
<i>Asian</i>		
China		
Beijing	0/4 0/4	
Hubei	0/4	0/4
Liaoning	0/5	0/5
Shandong	1/4	0/4
Japan	0/10	0/10
Russia		
Bellyk	1/10	5/10
Mineralni	3/18	0/18
Total	5/55 (9%)	5/55 (9%)
<i>North American</i>		
Connecticut	3/5	3/5
Massachusetts	7/9	8/9
Michigan	3/8	5/8
New Jersey	8/8	8/8
North Carolina	10/16	6/16
West Virginia	6/6	6/6
Total	37/52 (71%)	36/52 (69%)

tical possibility that a LDT1 transcript could end near this position. In addition, sequences resembling TATA box signals appear at positions 2301–2306 and 2337–2342, providing a possible signal for initiation of a distinct transcript containing ORF2. However, the authors know of no reports to-date of retrotransposons having two transcripts. No experimental evidence exists for this possibility in the gypsy moth, because discrete LDT1 transcripts have not been detected in any tissue or stage of development tested so far (not shown).

Numerous transposable element insertion site polymorphisms have been documented. In addition to numerous mutations caused by transposable element insertions, such as the *Drosophila* white-one locus mutation caused by a Doc insertion (Driver *et al.*, 1989), insertional polymorphisms have been found which do not affect gene function. These may be useful as phylogenetic markers, because the chances of two independent insertions occurring at the same location in different taxa is extremely low. Polymorphisms in *Alu* element insertion sites have been used to trace human evolutionary history (Batzer *et al.*, 1994), and two families of short interspersed elements (SINEs) have been used to infer the phylogeny of whales (Shimamura *et al.*, 1997). Once inserted, retrotransposable elements are assumed to be stable, since no mechanism for excision is known to exist. Because of their stability, transposable elements could be useful as population markers within species if the insertions have been recent enough that not all individuals have been affected.

In order to look for polymorphic LDT1 insertion sites in the gypsy moth, several primer pairs were designed with one primer within the LDT1 element and the other in the flanking host DNA. When used to amplify DNA samples from individual gypsy moths, many of these primer pairs amplified multiple fragments or a smear of DNA, which may have resulted from the flanking DNA primer targeting another repetitive sequence. However, two primer pairs, one set flanking the A15_{N1800} insertion junction and the other flanking the LDT1–2 partial retrotransposon, consistently amplified fragments of the expected size in some but not all gypsy moth samples. This suggests that the retrotransposon at these sites were inserted recently enough that not all members of the ancestral population were affected. On the other hand, although no known mechanism exists for precise excision of retrotransposons, it is possible that these loci may be polymorphic due to a random deletion affecting the region containing the retrotransposon or the flanking primer. It is also possible that point mutations preventing PCR primer binding have occurred. Further work is

needed to determine the exact nature of these polymorphisms.

Whatever the basis for the PCR polymorphisms, when ≈ 100 gypsy moth individuals were tested with each primer set, clear differences in amplification frequency were seen between the Asian and North American populations. Both markers were present in $\approx 70\%$ of North American moths and 9% of Asian moths; however, their distribution in subpopulations was different. Although these two markers are more common in the North American population, an LDT1 probe hybridized with similar intensity to genomic DNA from the Asian and North American moths, indicating that overall the two populations have comparable numbers of LDT1 insertions. Cloning of LDT1 elements from a member of a population of interest could lead to selection of LDT1 insertions unique to that population. Such markers could be useful for tracing the migration patterns of gypsy moths and for determining sources of introductions into new areas.

Experimental procedures

Origin and preparation of probe DNA

A DNA fragment containing a partial retrotransposon was obtained by random amplification of polymorphic DNA by polymerase chain reaction (RAPD-PCR). Gypsy moth DNA was amplified under standard RAPD-PCR conditions (Williams *et al.*, 1990) using the primer OPA-15 (5' TTCCGAACCC 3', Operon Technologies, Inc., Alameda, CA, USA). An 1800-bp fragment designated A15_{N1800} was generated from DNA from several individual North American gypsy moths, but not from Asian gypsy moth DNA samples. The fragment was gel-purified and cloned into the plasmid vector pCRII (Invitrogen Corporation, Carlsbad, CA, USA). A 700-bp *Bam*HI/*Hind*III subfragment of the A15_{N1800} fragment containing only retrotransposon sequences was gel purified and used as a probe for screening the genomic library.

Genomic library construction and screening

Gypsy moth DNA used for the construction of a genomic library was prepared from New Jersey Standard Strain first and second instar larvae following a modification of the method of Ish-Horowicz *et al.* (1979). Larvae were starved overnight and homogenized in liquid nitrogen using a mortar and pestle. The powder was mixed with suspension buffer (10 mM Tris-HCl pH 7.5, 60 mM NaCl, 10 mM EDTA) and heated to 65 °C. After resuspension, an equal volume of post-grind (lysis) buffer (200 mM Tris-HCl pH 9, 30 mM EDTA, 2% w/v SDS) was added to the homogenate and Proteinase K was added to a final concentration of 200 μ g/ml. The digestion solution was incubated at 65 °C for 30 min. Sodium acetate was added to a final concentration of 0.3 M. The solution was extracted with an equal volume of buffered phenol, then with an equal volume of chloroform/isoamyl alcohol (24 : 1). The DNA was precipitated with two volumes of ethanol, pelleted, and resuspended in water. The DNA was then treated with RNase (50 μ g/ml) at 37 °C

for 2 h, and re-extracted with phenol and chloroform followed by ethanol precipitation and resuspension.

A genomic library was constructed in the lambda phage vector EMBL4 following the procedures described by Maniatis *et al.* (1982). Genomic DNA from New Jersey Standard Strain gypsy moths was partially digested with *Sau3A*I and ligated to *Bam*HI-digested lambda phage arms. The average genomic insert size was 18 kb.

Consensus mapping

Genomic DNA digested with restriction enzymes was blotted and probed with a series of probes, starting with the 700 bp *Bam*HI/*Hind*III probe from the A15_{N1800} fragment. Restriction fragments in genomic DNA hybridizing to the probe were identified. Comparison of these with restriction sites within the partial retrotransposon in fragment A15_{N1800} allowed a restriction map to be constructed. Phage clones with fragments matching this restriction map were selected. From these clones, additional upstream probes were selected and used to determine the consensus map of the entire retrotransposon. Phage clones with maps matching the entire consensus map were selected for further analysis.

Genomic Southern blots

For consensus mapping, 2–5 µg of genomic DNA were digested with restriction enzymes and run on a 0.8% agarose gel. After electrophoresis the gel was blotted on to Nytran hybridization membrane (Schleicher & Schuell Inc, Keene, NH, USA) following the manufacturer's directions. Probe DNA fragments were radioactively labelled using the Gibco BRL Nick Translation System. Unincorporated dNTPs were removed using a Sephadex G-50 spun column. The probes were hybridized to the blot following the membrane manufacturer's directions.

DNA sequencing

DNA sequences were determined using the Sequenase sequencing kit (United States Biochemical Co., Cleveland, OH, USA). In most cases double-stranded plasmid DNA was used for sequencing. Fragments that yielded poor sequences from double-stranded DNA were cloned into M13 mp18 and mp19 vectors in order to generate single stranded DNA.

DNA and protein sequence analysis

DNA sequences were assembled and analysed using AssemblyLign and MacVector 4.1 (Eastman Chemical Company, New Haven, CT, USA). Amino acid sequences were aligned using CLUSTAL W (Thompson *et al.*, 1994). The alignment procedure used a gap opening penalty of 10.0, a gap extension penalty of 0.05, and the BLOSUM series of weight matrices (Henikoff & Henikoff, 1992). Amino acid identity scores were calculated by dividing amino acid matches by total aligned amino acids (gaps were excluded). The OLIGO Primer Analysis software (National Biosciences Inc., Plymouth, MN, USA) package was used to select primer sequences for sequencing and PCR, and to analyse sequences for possible internal secondary structure or primer interactions.

LDT1 element quantification

Approximately 2000 phage plaques containing gypsy moth genomic DNA were plated and the plaques were blotted on to nylon membranes. These membranes were probed with

probes from the 5' end (nucleotides 34–398), middle (nucleotides 2128–2947), and 3' end (nucleotides 4898–5398) of a complete retrotransposon. The plaques hybridizing to each probe were counted.

DNA amplification methods

Amplification of A15_{N1800} and LDT1–2 retrotransposon insertion site sequences was carried out by PCR in 15 µl reactions containing 10 ng genomic DNA, 50 mM KCl, 10 mM Tris-HCl pH 9.0 and 2.0 mM MgCl₂, 0.4 µM primer, and 0.2 units of Taq DNA polymerase (Boehringer Mannheim, Indianapolis, IN, USA). Reactions were topped with mineral oil and amplified in a Perkin Elmer thermal cycler for 2 min at 94 °C (initial denaturation step) followed by forty cycles of 94 °C for 1 min, 50 °C or 55 °C for 1 min, and 72 °C for 1 min. The A15_{N1800} primers were annealed at 55 °C, and the LDT1–2 primers were annealed at 50 °C. DNA amplification was analysed by electrophoresis in 1.2% agarose gels using Tris-borate buffer followed by ethidium bromide staining of the DNA.

The sequences of the primers used to amplify the A15_{N1800} insertion junction are as follows: KG3: 5' TTCCGAACCCGTGAC-CAGAGT 3', and KG40: 5' AGACACTWCAAAATCGTTTC 3' (W = A + T). The sequences of the primers flanking the LDT1–2 junction are: KG39: 5' GCTCGGAARCCAACTGTGC 3' (R = A + G) and KG43: 5' AGACCTCATAGCCGATTGTG 3'.

DNA samples for insertion site analysis

Asian gypsy moth samples analysed with retrotransposon insertion site primers were collected in China (Beijing, Hubei, Liaoning, and Shandong), Japan (Hokkaido, Kashiwada, Kuki-saki, Namiki, and Sakuragaoka), and Russia (Mineralni in Primor'ye Territory, and Bellyk in Krasnoyarsk Territory). North American moths were collected from Connecticut, Massachusetts (Norfolk Co.), Michigan (Manistee, Newaygo, Oceana, and Ottawa Co.), North Carolina (Currituck and Dare Co.), and West Virginia (Monongalia Co.). Additional individuals were selected from the New Jersey Standard Strain, a laboratory strain originating from New Jersey.

RNA isolation

RNA was prepared using the method of Chomczynski & Sacchi (1987). Northern blots were carried out using formaldehyde gel electrophoresis as described in Maniatis *et al.* (1982).

GenBank accession numbers

The nucleotide sequences described in this paper have been assigned GENBANK accession numbers AF081101 (A15N1800), AF081102 (LDT1–2), and AF081103 (LDT1–4).

Acknowledgements

This research was supported by funds provided by the U.S. Forest Service. We thank Drs Tom Pfeifer and Ed Dougherty for their critical reviews of this manuscript. We are also grateful for the comments from three anonymous reviewers.

References

- Batzer, M.A., Stoneking, M., Alegria-Hartman, M., Bazan, H., Kass, D.H., Shaikh, T.H., Novick, G.E., Ioannou, P.A., Scheer, W.D., Herrera, R.J. and Deininger, P.L. (1994) African origin of human-specific polymorphic *Alu* insertions. *Proc Natl Acad Sci USA* **91**: 12288–12292.
- Berg, J.M. and Shi, Y. (1996) The galvanization of biology: a growing appreciation for the roles of zinc. *Science* **271**: 1081–1085.
- Besansky, N.J. (1990) A retrotransposable element from the mosquito *Anopheles gambiae*. *Mol Cell Biol* **10**: 863–871.
- Besansky, N.J., Bedell, J.A. and Mukabayire, O. (1994) Q: a new retrotransposon from the mosquito *Anopheles gambiae*. *Insect Mol Biol* **3**: 49–56.
- Bess, J.W., Powell, P.J., Issaq, H.J., Schumack, L.J., Grimes, M.K., Henderson, L.E. and Arthur, L.O. (1992) Tightly bound zinc in human immunodeficiency virus type 1, human T-cell leukemia virus type I, and other retroviruses. *J Virol* **66**: 840–847.
- Blinov, A.G., Sobanov, Y.V., Bogachev, S.S., Donchenko, A.P. and Filippova, M.A. (1993) The *Chironomus thummi* genome contains a non-LTR retrotransposon. *Mol Gen Genet* **237**: 412–420.
- Boeke, J.D. and Corces, V.G. (1989) Transcription and reverse transcription of retrotransposons. *Annu Rev Microbiol* **43**: 403–434.
- Bogdanowicz, S.M., Wallner, W.E., Bell, J., O'Dell, T.M. and Harrison, R.G. (1993) Asian gypsy moths (Lepidoptera: Lymantriidae) in North America: Evidence from molecular data. *Ann Entomol Soc Am* **86**: 710–715.
- Burke, W.D., Calalang, C.C. and Eickbush, T.H. (1987) The site-specific ribosomal insertion element Type II of *Bombyx mori* (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol Cell Biol* **7**: 2221–2230.
- Chomczynski, P. and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* **162**: 156–159.
- Di Nocera, P.P. and Casari, G. (1987) Related polypeptides are encoded by *Drosophila* F-elements, I factors, and mammalian L1 sequences. *Proc Natl Acad Sci USA* **84**: 5843–5847.
- Driver, A., Lacey, S.F., Cullingford, T.E., Mitchelson, A. and O'Hare, K. (1989) Structural analysis of Doc transposable elements associated with mutations at the *white* and *suppressor of forked* loci of *Drosophila melanogaster*. *Mol Gen Genet* **220**: 49–52.
- Fawcett, D.H., Lister, C.K., Kellett, E. and Finnegan, D.J. (1986) Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINES. *Cell* **47**: 1007–1015.
- Felger, I. and Hunt, J.A. (1992) A non-LTR retrotransposon from the Hawaiian *Drosophila*: the LOA element. *Genetica* **85**: 119–130.
- Feng, Q., Moran, J.V., Kazazian, H.H. Jr and Boeke, J.D. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905–916.
- Feng, Q., Schumann, G. and Boeke, J.D. (1998) Retrotransposon R1Bm endonuclease cleaves the target sequence. *Proc Natl Acad Sci USA* **95**: 2083–2088.
- Garner, K.J. and Slavicek, J.M. (1996) Identification and characterization of a RAPD-PCR marker for distinguishing Asian and North American gypsy moths. *Insect Mol Biol* **5**: 81–91.
- Hattori, M., Kuhara, S., Takenaka, O. and Sakaki, Y. (1986) L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. *Nature* **321**: 625–628.
- Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**: 10915–10919.
- Ichimura, S., Mita, K. and Sugaya, K. (1997) A major non-LTR retrotransposon of *Bombyx mori*, L1Bm. *J Mol Evol* **45**: 253–264.
- Ish-Horowicz, D., Pinchin, S.M., Schedl, P., Artavanis-Tsakonis, S. and Mirault, M. (1979) Genetic and molecular analysis of the 87A7 and 87C1 heat-inducible loci of *Drosophila melanogaster*. *Cell* **18**: 1351–1358.
- Jakubczak, J.L., Xiong, Y. and Eickbush, T.H. (1990) Type I (R1) and Type II (R2) ribosomal DNA insertions of *Drosophila melanogaster* are retrotransposable elements closely related to those of *Bombyx mori*. *J Mol Biol* **212**: 37–52.
- Kerber, B., Fellert, S., Taubert, H. and Hoch, M. (1996) Germ line and embryonic expression of Fex, a member of the *Drosophila* F-element retrotransposon family, is mediated by an internal cis-regulatory control region. *Mol Cell Biol* **16**: 2998–3007.
- Larder, B.A., Purifoy, D.J.M., Powell, K.L. and Darby, G. (1987) Site-specific mutagenesis of AIDS virus reverse transcriptase. *Nature* **327**: 716–717.
- Luan, D.D. and Eickbush, T.H. (1995) RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol Cell Biol* **15**: 3882–3891.
- Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- Martin, F., Marañón, C., Olivares, M., Alonso, C. and López, M.C. (1995) Characterization of a non-long terminal repeat retrotransposon (L1Tc) from *Trypanosoma cruzi*: Homology of the first ORF with the Ape family of DNA repair enzymes. *J Mol Biol* **247**: 49–59.
- McClure, M.A. (1991) Evolution of retrotransposons by acquisition or deletion of retrovirus-like genes. *Mol Biol Evol* **8**: 835–856.
- Mouchès, C., Bensaadi, N. and Salvado, J.-C. (1992) Characterization of a LINE retrotransposon dispersed in the genome of three non-sibling *Aedes* mosquito species. *Gene* **120**: 183–190.
- O'Hare, K., Alley, M.R.K., Cullingford, T.E., Driver, A. and Sanderson, M.J. (1991) DNA sequence of the Doc retrotransposon in the *white-one* mutant of *Drosophila melanogaster* and of secondary insertions in the phenotypically altered derivatives *white-honey* and *white-eosin*. *Mol Gen Genet* **225**: 17–24.
- Ogura, T., Okano, K., Tsuchida, K., Miyajima, N., Tanaka, H., Takada, N., Izumi, S., Tomino, S. and Maekawa, H. (1994) A defective non-LTR retrotransposon is dispersed throughout the genome of the silkworm, *Bombyx mori*. *Chromosoma* **103**: 311–323.
- Okazaki, S., Ishikawa, H. and Fujiwara, H. (1995) Structural analysis of TRAS1, a novel family of telomeric repeat-associated retrotransposons in the silkworm, *Bombyx mori*. *Mol Cell Biol* **15**: 4545–4552.
- Pfeifer, T.A. and Grigliatti, T.A. (1996) Future perspectives on insect pest management: Engineering the Pest. *J Invertebr Pathol* **67**: 109–119.
- Pfeifer, T.A., Humble, L.M., Ring, M. and Grigliatti, T.A. (1995) Characterization of gypsy moth populations and related species using a nuclear DNA marker. *The Canadian Entomologist* **127**: 49–58.
- Priimägi, A.F., Mizrokhi, L.J. and Ilyin, Y.V. (1988) The *Drosophila* mobile element *jockey* belongs to LINES and contains coding

- sequences homologous to some retroviral proteins. *Gene* **70**: 253–262.
- Schreiber, D.E., Garner, K.J. and Slavicek, J.M. (1997) Identification of three RAPD-PCR markers for distinguishing Asian and North American gypsy moths. *Ann Entomol Soc Am* **90**: 667–674.
- Sheen, F.-M. and Levis, R.W. (1994) Transposition of the LINE-like retrotransposon TART to *Drosophila* chromosome termini. *Proc Natl Acad Sci USA* **91**: 12510–12514.
- Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M., Munechika, I. and Okada, N. (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* **388**: 666–670.
- Singer, M.F. and Skowronski, J. (1985) Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. *Trends Biochem Sci* **10**: 119–122.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties, and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.
- Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res* **18**: 6531–6535.
- Xiong, Y. and Eickbush, T.H. (1988a) The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal repeat retrotransposons. *Mol Cell Biol* **8**: 114–123.
- Xiong, Y. and Eickbush, T.H. (1988b) Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. *Mol Biol Evol* **5**: 675–690.
- Xiong, Y. and Eickbush, T.H. (1994) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* **9**: 3353–3362.