A Novel Active L1 Retrotransposon Subfamily in the Mouse

John L. Goodier, Eric M. Ostertag, Kevin Du, et al.

Genome Res. 2001 11: 1677-1685
Access the most recent version at doi:10.1101/gr.198301

Letter

A Novel Active L1 Retrotransposon Subfamily in the Mouse

John L. Goodier,[1] Eric M. Ostertag, Kevin Du, and Haig H. Kazazian, Jr.[1]

Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA

Unlike human L1 retrotransposons, the 5′ UTR of mouse L1 elements contains tandem repeats of ~200 bp in length called monomers. Multiple L1 subfamilies exist in the mouse which are distinguished by their monomer sequences. We previously described a young subfamily, called the T_F subfamily, which contains ~1800 active elements among its 3000 full-length members. Here we characterize a novel subfamily of mouse L1 elements, G_F, which has unique monomer sequence and unusual patterns of monomer organization. A majority of these G_F elements also have a unique length polymorphism in ORF1. Polymorphism analysis of G_F elements in various mouse subspecies and laboratory strains revealed that, like T_F, the G_F subfamily is young and expanding. About 1500 full-length G_F elements exist in the diploid mouse genome and, based on the results of a cell culture assay, ~400 G_F elements are potentially capable of retrotransposition. We also tested 14 A-type subfamily elements in the assay and estimate that about 900 active A elements may be present in the mouse genome. Thus, it is now known that there are three large active subfamilies of mouse L1s; T_F, A, and G_F, and that in total ~3000 full-length elements are potentially capable of active retrotransposition. This number is in great excess to the number of L1 elements thought to be active in the human genome.

Non-LTR LINE1 (L1) elements are non-long terminal repeat (non-LTR) retrotransposons that are capable of autonomous retrotransposition and have expanded to large copy numbers in mammalian genomes. The mouse genome contains >100,000 L1s comprising ~10% of the genomic DNA (Hutchison et al. 1989). Most L1 elements are inactive, a consequence of 5′-end truncation, inversion, or mutation. The consensus sequence for full-length mouse L1s is ~7 kb and, like human L1s, has two open reading frames (ORFs) and a 3′ poly(A) tail, and is flanked by short target site duplications (TSDs). Unlike human L1s, mouse L1s have a bipartite 5′ UTR consisting of tandemly repeated sequences of ~200 bp called monomers, which are situated upstream of single-copy, nonmonomeric sequence. By linking monomers to reporter genes, it has been shown that they possess promoter activity and that increasing the number of monomers increases the level of transcription (Severynse et al. 1992; DeBerardinis and Kazazian 1999).

Phylogenetic analyses by Adey et al. (1994a) suggest that mouse L1 evolution has been dominated by a single lineage which has spawned several subfamilies of L1s differing in sequence at their 5′ ends. A-type subfamily and F-type subfamily members have monomer lengths of 208 bp and 206 bp, respectively (Fanning 1983; Loeb et al. 1986; Padgett et. al. 1988). Members of the more ancient V subfamily probably lack a 5′-repeated sequence (Jubier-Maurin et al. 1992). For some time it was thought that all F elements and most A elements had been rendered inactive by mutation, but that a small subset of A elements remained active in the genome. Evidence supporting this hypothesis includes the findings that some A-type elements are highly similar to each other, possess two intact ORFs, and are transcribed. However, their capacity for retrotransposition was not directly assayed (Shehee et al. 1987; Schichman et al. 1992; Martin 1995).

Unexpectedly, two disease-causing insertions, $L1_{spa}$ and $L1_{orl}$, were found to be members of a large, young, and expanding subfamily of mouse L1s (Kingsmore et al. 1994; Takahara et al. 1996; Naas et al. 1998). The monomer sequence belonging to members of this T_F subfamily was on average 72% identical to the consensus monomer sequence of F-type L1s. Based on the results of a retrotransposition assay of 11 randomly cloned T_F-type L1s in cell culture, we estimated that a majority of full-length T_F elements are potentially active (DeBerardinis et al. 1998).

Here we present evidence that a previously unreported subfamily of L1 elements exists in high copy number in the mouse genome and that many members of this novel subfamily are likely capable of retrotransposition.

RESULTS

A Unique 5′ UTR Distinguishes a New Mouse L1 Subfamily

Examination of the GenBank database revealed the presence of a previously undescribed subfamily of mouse L1 elements remarkable for their 5′ UTR regions. Like those of the previously described T_F elements (DeBerardinis et al. 1998), the 5′ monomer arrays of L1s from this new subfamily are related to, but clearly distinct from, the monomers of F subfamily members. We have called these G_F elements and have extracted from the database 17 members for characterization (Table 1).

The 17 G_F elements contain a variable number of monomers, averaging 2.4 with a maximum of 6. We aligned the sequences of 34 full-length monomers to obtain a consensus sequence 206 bp in length and 69% identical to the reconstructed F monomer consensus sequence described by Adey et al. (1994b) and 67% identical to the 212-bp T_F monomer con-

[1]Corresponding authors.
E-MAIL jgoodier@mail.med.upenn.edu, kazazian@mail.med.upenn.edu; FAX (215) 573–7760.
Article published on-line before print: *Genome Res.,* 10.1101/gr.198301.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.198301.

Table 1. G_F-Type L1 Sequences Used in This Study

L1 name	Accession #	GI #	Database[a]	Coordinates[b]	Intact ORF1/ORF2	# full length monomers[c]	# truncated monomers[c]	% identity with body consensus[d]	Active w/wo CMV promoter
G_F13	AF146793	7684609	nr	114636–121199	Y/Y	2.2	2	99.5	Y/N?
G_F21	AC021631	13310926	htgs/nr	68991–62229	Y/Y	3.3	2	99.5	Y/Y
G_F23	AF229843	6862583	nr	42031–48206	N/Y	0.7	1	98.6	
G_F26	AC002406	2981248	nr	19967–26229	N/N	1.2	1	96.1	
G_F27	AC026767	12643015	htgs/nr	118223–110609	N/Y	2.2	2	99.4	N/N?
G_F36	AC003996	2772523	nr	20312–27252	Y/N	1.4	3	96	
G_F43	AL450393	11876100	htgs	7656–13977	Y/Y	1.1	2	99.6	
G_F46	AL049866	7630118	nr	22185–15089	Y/N	5.1	1	99.3	
G_F56	AC058786	11120825	nr	200048–206662	Y/N	2.4	2	99.6	
G_F62	AC068252	13357510	htgs	161311–154202	Y/Y	6.1	1	99.6	Y/Y
G_F68	AC006508	4454569	nr	167020–160414	N/N	2.2	2	97.7	
G_F71	AC079221	10122065	htgs	58136–64915	Y/N	1.1	1	98.4	
G_F78	AC007978	12000475	nr	42762–49862	Y/N	5.1	1	99.4	
G_F82	AC083912	10697433	htgs	925–7430	Y/Y	2	2	99.2	
G_F84	AC087134	11610860	htgs	179278–172998	N/N	1.7	0	92.6	
G_F251	AF259071	8571430	nr	61240–55080	Y/N	1.1	0	96.4	
G_F253	AF259073	8575573	nr	6526–256	Y/N	1.7	0	93.4	

[a]Database refers to the non-redundant or high-throughput genomic sequence databases of GenBank.
[b]Coordinates are the first and last nucleotides in the GenBank record delimiting the L1.
[c]See Figure 1B for description of monomer structure.
[d]Identity with G_F consensus sequence beginning 250 bp upstream of the start of ORF1.

sensus (DeBerardinis and Kazazian 1999). Individual monomers ranged in size from 204 to 207 bp. The 3′ end of the G_F consensus monomer sequence shown in Figure 1A corresponds to the point at which the monomer array joins the nonmonomeric 5′ UTR region 250 bp upstream of the start of G_F ORF1. The monomer array of a T_F element begins 258 bp upstream of T_F ORF1. The nonmonomeric 5′ UTRs of the G_F consensus and T_F consensus sequences are 85% identical.

The organization of the 5′ UTR of G_F elements is atypical of mouse L1s. The nonmonomeric region of the 5′ UTR of all G_F elements contains a 47-nt sequence which is 82% identical to a portion of the F monomer consensus (nts 132–178) (Fig. 1B). The monomer arrays are organized in four distinct patterns. At the 3′ end of many arrays are 5′ truncated monomers which exist singly (patterns II and III) or as a pair (pattern I) and which consist of the last 64 bps (nts 143–206) of a full-length monomer. Pattern III also has two additional truncated monomers upstream of the first full-length monomer. All of the abbreviated monomers are the same length and only 77%–84% identical to the consensus G_F monomer (69%–75% identical to the F consensus). Most G_F elements fall within patterns I and II whose full-length monomers are greater than 95% identical to the G_F consensus (~70% identical to the F consensus).

DeBerardinis and Kazazian (1999) noted that many T_F L1s were truncated at their 5′ ends in the vicinity of a predicted binding site for the transcription factor YY1 (GCCATCTT, Fig. 1A). A YY1 site is also present in F monomers and in the human L1 5′ UTR (Becker et al. 1993). Twenty-six of the 31 T_F sequences examined were truncated within 25 bp upstream or downstream of the monomer YY1 binding sequence (DeBerardinis and Kazazian 1999; J.L. Goodier, unpubl.). All G_F monomers have a single nucleotide change which alters this sequence (GCCTTCTT), and this may explain why G_F elements do not tend to truncate near this site (only 3 of 17 elements truncated within 25 nts of this sequence).

ORF1 Length Polymorphism Region (LPR)

All mouse L1 subfamilies contain an LPR within the N-terminal one-third of ORF1, consisting of tandemly repeated blocks of 66 bp and/or 42 bp that do not interrupt the reading frame of the protein. These LPRs were defined by Schichman et al. (1992) and Adey et al. (1994a), and the system of nomenclature proposed by Mears and Hutchison (2001) is summarized in Figure 2A. Most A-type elements have 66-42 bp (Group I) or 66-42-42 bp (Group II) repeat structures, although some also have a 66-66 bp (Group III) pattern. F-type elements belong to Groups I–IV. Although almost all T_F elements have a Group II repeat structure (DeBerardinis et al. 1998; J.L. Goodier, unpubl.), Mears and Hutchison (2001) identified a novel T_F Group I element. Thirty-five percent (6/17) of G_F L1s also belong to Group II. However, the majority of G_F elements have an extra 42-nt repeat (i.e., 66-42-42-42 nt), a pattern not previously reported for any mouse L1 element. We have therefore placed G_F elements in a new LPR group, Group V. The four LPR repeats of the G_F consensus sequence are aligned in Figure 2B.

Features of the Body of GF Elements

We derived a nucleotide consensus sequence for the bodies of the 17 G_F members of our dataset beginning at the nonmonomeric 5′ UTR. The bodies of the 17 G_F elements are 92.6%–99.6% identical to their consensus sequence (Table 1). In contrast, 11 randomly cloned T_F elements were on average 99.8% identical to their consensus sequence (DeBerardinis et al. 1998). This suggested that different subgroups of G_F elements exist and that at least some of these are older than T_F elements. We therefore performed neighbor-joining phylogenetic analyses which included both the G_F members of our dataset and F, A, and T_F elements of the testset used in the analysis by Mears and Hutchison (2001). Results using a portion of the nonmonomeric 5′ UTR defined as region α by

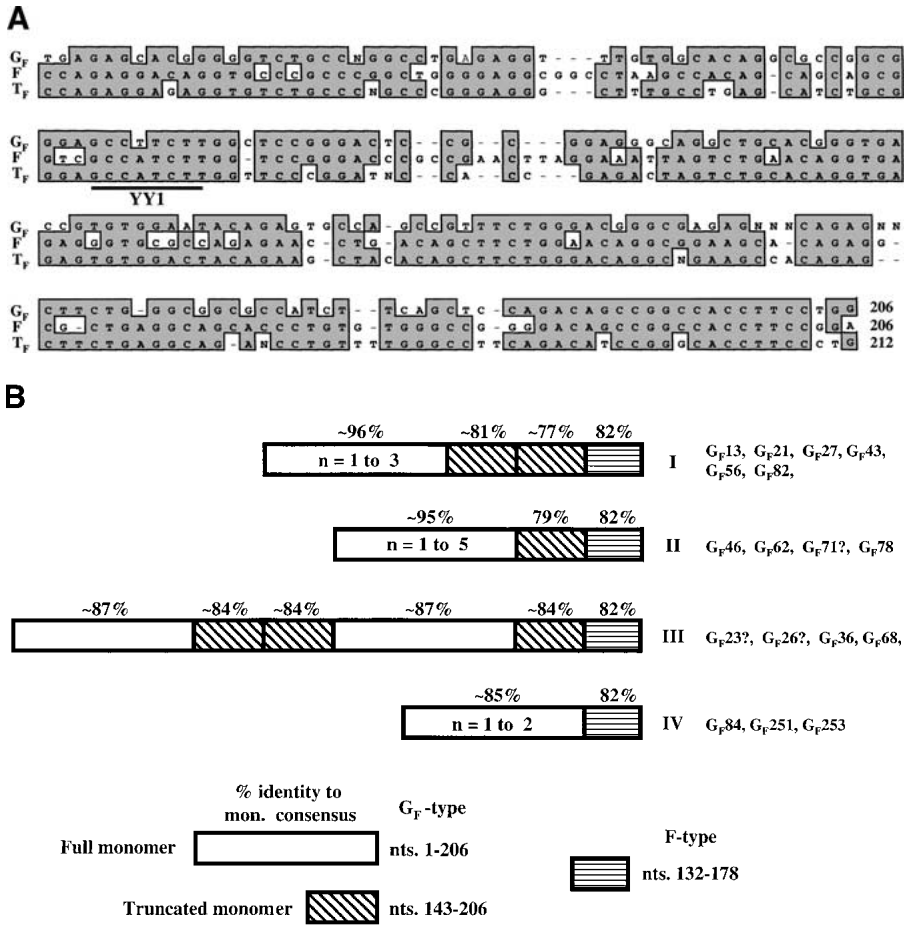

Figure 1 (*A*) Alignment of the 206-bp G_F monomer consensus sequence, the 206-bp reconstructed F-consensus monomer (Adey et al. 1994b), and the 212-bp T_F monomer consensus (DeBerardinis and Kazazian 1999). The location of the consensus YY1-binding sequence is underlined. (*B*) G_F element variants (labeled I– IV) with different patterns of 5′ UTR monomer organization. The fragment shown with horizontal stripes in each element is in the nonmonomeric 5′ UTR and is 82% identical to sequence in the F-consensus monomer. G_F monomers are either full-length (204–207 bp) or truncated (~64 bp). The monomeric regions of some clones (indicated by "?") are too short to be certain of their variant group.

Mears and Hutchison (2000) and corresponding to nucleotides 1539–1740 of L1MdA2 (Loeb et al. 1986) are shown in Figure 3A.

Three distinct G_F clades (labeled G-I, G-II, and G-III) are evident and are supported by 1000 rounds of bootstrap analysis. Most of the elements (9 of 17) belong to clades G-I and are on average 99.1% identical to each other and 94.0% identical to the T_F consensus (DeBerardinis et al. 1998). These elements also have monomers closest in sequence to the consensus and 5′ UTR structures as depicted in Figure 1B, patterns I and II. The existence of a subset of clade G-I, designated G-I_2, is weakly supported across the length of the element. Members of clade G-II are on average 97.9% identical to each other, and members of clade G-III are most dissimilar (94.8%) and have 5′ UTR patterns III and IV (Fig. 1B). Only G_F68 (Clade G-II) and members of clade G-III belong to ORF1 LPR Group II; all others belong to the novel LPR Group V (66-42-42-42 bp). Two members of clade G-III, G_F36 (Fnb1) and G_F26 (Fmu1), were previously defined as representatives of a young clade of F elements (F-II2) by Mears and Hutchison (2001). While these authors did not describe the monomer structure of these elements, it is now evident that the monomers are more closely related to the G_F (87% identity) than the F (74% identity) monomer consensus.

Phylogenetic analysis of a portion of the 3′ UTR (nts 6701–7333 of L1MdA2 and region θ as defined by Mears and Hutchison 2001) supports clades G-I and G-II which were established by the 5′ UTR analysis (Fig. 3B). However, clade G-III splits into two subgroups whose members have undergone separate recombination events. As proposed by Mears and Hutchison (2001), G_F36 (Fnb1) and G_F26 (Fmu1), together with G_F251, have been formed by recombination with either an A-type Group II LPR element or a young F-type element donating the 3′ end. On the other hand, G_F84 and G_F253 group with older F-type elements at their 3′ ends. Phylogenetic analysis of the set of subsequences spanning the entire L1 used by Mears and Hutchison (2001) reveal that the former recombination breakpoint occurred in the central portion of ORF2 (regions ε, nts 2798–4115, or ξ, nts 4116–5455), whereas the latter occurred closer to the N-terminus of ORF2 (region ε). Clades G-I and G-II are supported by analyses of regions spanning the length of the L1, and these clades do not contain recombinant members. Included in the analysis of region θ are L1s designated as members of a new "Z subfamily" by Hardies et al. (2000). These L1s are not G_F elements and, as previously reported by Mears and Hutchison (2001), are in fact members of the T_F subfamily.

GF Elements Transduce 3′ Flanking DNA

Retrotransposing L1s can occasionally transduce nonL1 DNA flanking their 3′ ends to new genomic locations. This is a consequence of the inherent weakness of the L1 polyadenylation signal which during transcription is occasionally bypassed in favor of a stronger signal downstream. We previously identified (in the mouse genome database) two transduction events involving A-type elements which in both instances mobilized 3 kb of sequence (Goodier et al. 2000). Two G_F elements in our dataset have transduced flanking DNA (Fig. 4). Element G_F71 transduced 558 nts, while the 3′ flank of element G_F27 is the final product of two consecutive transduction events, an initial movement of 24 nts followed by a second retrotransposition event and the transduction of an additional 1033 nts.

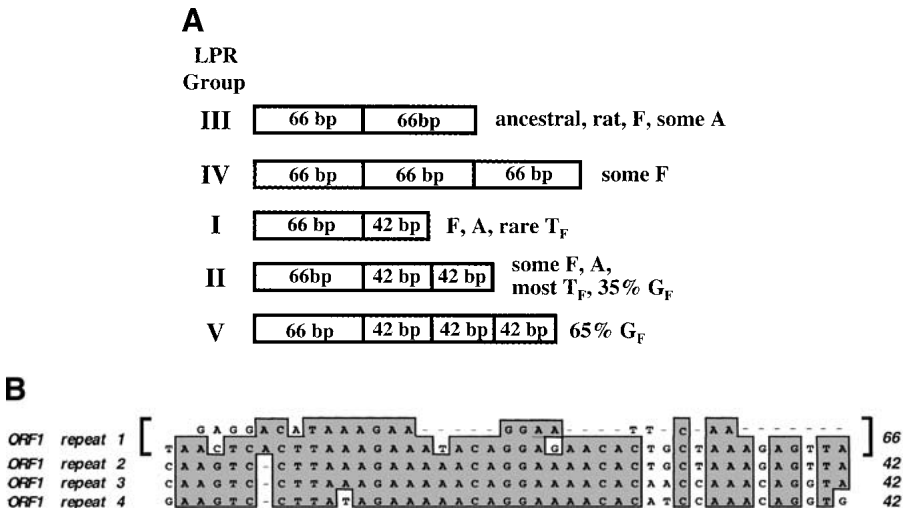

Figure 2 (*A*) Classification scheme for mouse L1 ORF1 length polymorphism region (LPR) groups, as defined by Mears and Hutchison (2001). LPRs consists of tandem arrays of 66-bp and/or 42-bp blocks, and each begins with a 66-bp block. The 42-bp block is homologous with the 3′ portion of the 66-bp repeat. The novel G_F variant has been designated LPR Group V. (*B*) Alignment of the ORF1 LPR repeats from the G_F consensus sequence. As shown in the first line of the alignment, the 5′-most 24 bp of the 66-bp repeat has partial homology with the 42-bp repeat.

GF Elements Have Been Recently Active in the Mouse

The high degree of sequence similarity among some of the G_F elements implies that these elements may have recently dispersed within the mouse genome. To test this, we examined the genomes of *Mus* subspecies and laboratory strains for the presence or absence of three G_F elements (Fig. 5). We designed PCR primers flanking the elements which would amplify a 7.5-kb filled site or a 450-bp empty site. None of the G_F elements were detected in *Mus spretus* or *Mus musculus* genomes. G_F13 was detected in laboratory strains only, whereas G_F21 was present in *M. m. castaneus* but absent from some laboratory strains. Surprisingly, G_F46 was absent from all samples tested. However, it was possible to amplify G_F46 DNA from the bacmid library clone sequenced for the GenBank entry. The library was generated from an embryonic stem cell library of strain 129/Sv. We confirmed the presence of the L1 sequence in this clone by PCR amplification and sequencing.

Some A and GF Subfamily Members Have Retrotransposition Capability

We previously tested A-type L1s in the cell culture retrotransposition assay (DeBerardinis et al. 1998). From phage library clones we PCR-amplified ten A-type elements without their monomers and cloned these into a retrotransposition vector under transcriptional control of a CMV promoter (Moran et al. 1996). One of the ten cloned elements supported retrotransposition. However, the possibility existed that PCR errors introduced during cloning could have inactivated otherwise potentially active elements. Therefore, we directly isolated eight A-type elements (including four elements tested in the previous experiment and four new elements) from bacteriophage and inserted each without their monomers into a retrotransposition vector in which both the monomer array of the T_F element $L1_{spa}$ and the CMV promoter directed expression. In all, two (A101 and A102) of a total of 14 different A elements from both experiments (14%) were capable of supporting retrotransposition (Table 2). Although transcription of these A-type elements was not driven by their endogenous promoters, Severynse et al. (1992) showed that a single A monomer is sufficient to direct transcription. Since A101 possesses 3.5 monomers and A102 possesses at least two monomers, it is probable that these two elements are capable of transcription and retrotransposition in vivo. Saxton and Martin (1998) estimated that the diploid genome of mouse strain 129/Sv contains 6500 full-length A-type elements. If 14% are active, then over 900 potentially active A elements reside in the genome.

Seven of the 14 A-type elements tested in the retrotransposition assay have a Group II ORF1 LPR (66-42-42 nts). Both active elements have a Group I (66-42 nts) LPR and belong to a younger transcribed A-type subset whose members possess >99.5% identity with each other in their 5′ UTR and ORF1 regions (Schichman et al. 1992). RNP-58B, a cDNA copy of an L1 RNA isolated by Martin (1995) from a ribonucleoprotein (RNP) particle, is also a member of this subset. L1 RNPs are comprised of ORF1 protein bound to L1 RNA and are thought to be intermediates in the retrotransposition process. The ORF1 sequences of the active elements A101 and A102 are identical, while RNP-58A differs at a single position (R22C). Only four residues in the ORF2 of RNP-58A differ from those of both A101 and A102: S86L in the endonuclease domain, T359K situated between the endonuclease and reverse transcriptase domains, and R761W and V/A928D, which are not conserved among mammalian species. Therefore, it is likely that RNP-58A is also an active L1 element.

Using PCR, we directly amplified DNA of three G_F elements with two complete ORFs from the genomes of several *Mus m. domesticus* strains. We also tested G_F27, which has a frameshift mutation 16 amino acid residues from the end of ORF1, but found this element to be inactive in human 143B cells and barely active when driven by a CMV promoter in HeLa cells. All elements belonged to clade G-I (Fig. 3A). They were cloned with their entire 5′ UTRs intact into retrotransposition vectors both with and without an exogenous CMV promoter and tested for retrotransposition activity. All three of the cloned elements with intact ORFs were capable of retrotransposition in cell culture when linked with a CMV promoter (Table 2). G_F21 and G_F62 were clearly active when driven by their endogenous promoters alone, while G_F13 may be active at a very low level. G_F21 and the A-type element, A102, are the most active mouse L1s tested in the retrotransposition assay to date. After screening a mouse strain 129/OlaHsd genomic library with a G_F monomer probe, we estimated that ~1500 full-length G_F elements reside in the diploid mouse genome. This number was confirmed by probing GenBank's high-throughput genomic sequence (htgs) mouse database with amino acid sequence from the N-terminus of the G_F consensus ORF1. It is likely that 440 full-length G_F elements have two complete ORFs (since 5 of the 17 members of our dataset have intact ORFs). These data suggest that ~400

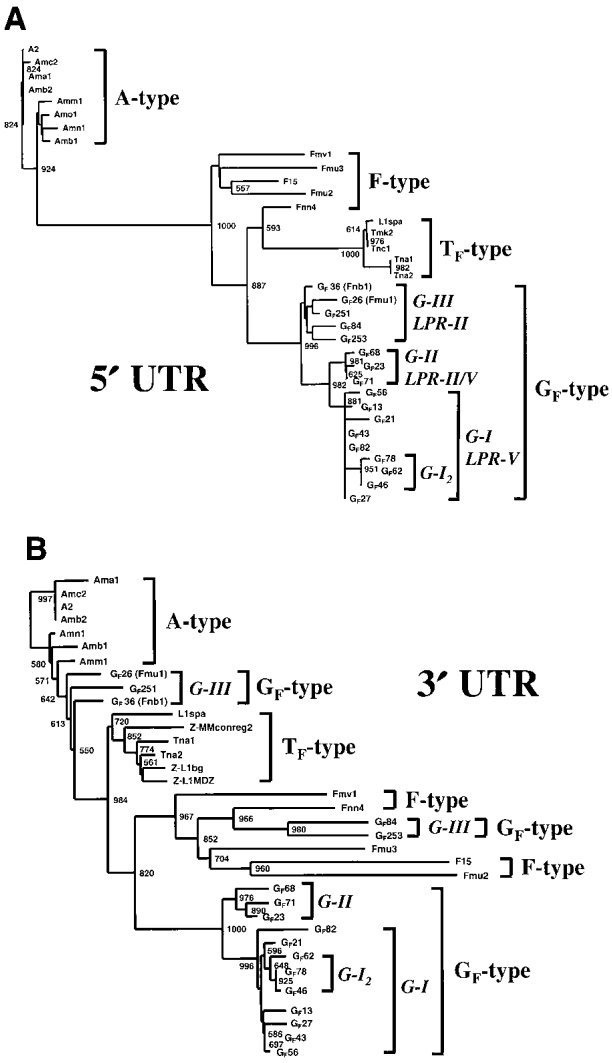

Figure 3 (*A*) Phylogenetic analysis of that portion of the 5′ UTR of mouse L1s defined as region α by Mears and Hutchison (2001) and corresponding to nucleotides 1539–1740 of the A-type element L1Md-A2. The unrooted tree was constructed by the nearest neighbor-joining method (Saitou and Nei 1987) and includes the 17 members of the G_F dataset and members of the dataset used by Mears and Hutchison (2001). G_F clades identified by the analysis are named G-I, G-II, and G-III, and their corresponding LPR groups are indicated. Significant bootstrap values (of 1000 replicates) are shown at the nodes. (*B*) Phylogenetic analysis of 3′ UTR region corresponding to nts 6701–7333 of L1MdA2. Also included in this analysis are L1s classified by Hardies et al. (2000) as members of the "Z subfamily": MMconreg2, $L1_{bg}$ (Perou et al. 1997), and L1MdZ (Kraft et al. 1992). These group with T_F-type elements.

potentially active full-length G_F elements may exist in the mouse genome.

DISCUSSION

G_F elements belong to a novel subfamily of mouse L1s with unique monomeric 5′ UTR sequence. At different times in evolution, murine L1s have acquired novel 5′ sequences, perhaps by recombination or by the capture of upstream promoter sequence. The shuffling of L1 fragments, whether by recombination or template switching, is a leit motif in the evolution of murine subfamilies. In *Rattus norvegicus*, the youngest subfamily, L1mlvi2, acquired its ORF1 from an ancestral L1 (Hayward et al. 1997). Flood et al. (1998) identified unusual hybrid L1s in the mouse which possess fused 3′ sequence homologous to a fragment in the first intron of C_ε immunoglobulin. Five hybrid elements have poly(A) signals and A-rich 3′ ends, and one has TSDs, evidence of retrotransposition. (Since the C-terminus of ORF2 was displaced by the fusion, transcomplementation by ORF2 protein from an intact L1 may have allowed this hybrid element to multiply). In the mouse, Saxton and Martin (1998) and Mears and Hutchison (2001) have shown that T_F elements are recombinant, deriving 5′ UTR sequence from an "older" F-type element and 3′ sequence from a "young" F-type or an A-type element. Similarly, by phylogenetic reconstruction we show that 5 of the 17 G_F elements in our dataset are recombinant. Three of these elements, G_F36, G_F26, and G_F251, are members of a subclade which Mears and Hutchison (2001) designated F-II_2 and estimated to be 1.4–2 million years old.

In addition to a novel 5′ UTR, a majority of G_F elements also contain a unique length polymorphism in ORF1. The LPR is contained within the N-terminal one-third of ORF1, a region which is highly divergent in sequence among different mammalian L1 families. In this region, human L1 ORF1 contains a putative leucine zipper domain implicated in protein–protein interaction (Holmes et al. 1992; Hohjoh and Singer 1996). Mouse elements do not contain a leucine zipper. However, Martin et al. (2000) predicted that the N-terminal region of mouse ORF1 is capable of forming a coiled-coil structure, and they demonstrated that this region mediates protein multimerization. Furthermore, employing two-hybrid assays,

G_F71 - 558 nt transduction

5′ TSD
aagaatgtgtttgtagacagccggccaccttcctgg.....6204 nt.....

L1 poly(A) signal
gagctaaatacct**AATAAA**aatggaaaaaaaatcagcaaa.....460 nt.....
Downstream poly (A) signal
cacattctctggta**AATAAA**ggcatacaagatccccccagttatatata*taaaaa***aaga**
3′ TSD
atgtgtttgtaga

G_F27 - 24 + 1033 nt transductions

5′ TSD
aagactttgtagcctgcggtgccatcttcagctcca.....6496 nt .....
1st downstream poly (A) signal
L1 poly(A) signal
gaaatacct**AATAAA**aaatgggaaaaaaaaaag***AATAAA***tatcacatattc*aaaaa*

*aaaaaaaaaaaaaaaaaa*gaaacctagcacagtgg.....957 nt .....
2nd downstream poly (A) signal
atgtatataaagaataaa**AATAAA**actaaaaaagaaaatgaaaaaaaaaa**aagact**

3′ TSD
ttgtagcct

Figure 4 Two G_F elements, G_F27 and G_F71, have transduced DNA flanking their 3′ ends to new genomic locations. Consecutive retrotransposition events have created the composite transduction of G_F27. TSDs are shown in bold lowercase letters, and putative hexanucleotide poly(A) signals are in bold uppercase letters.

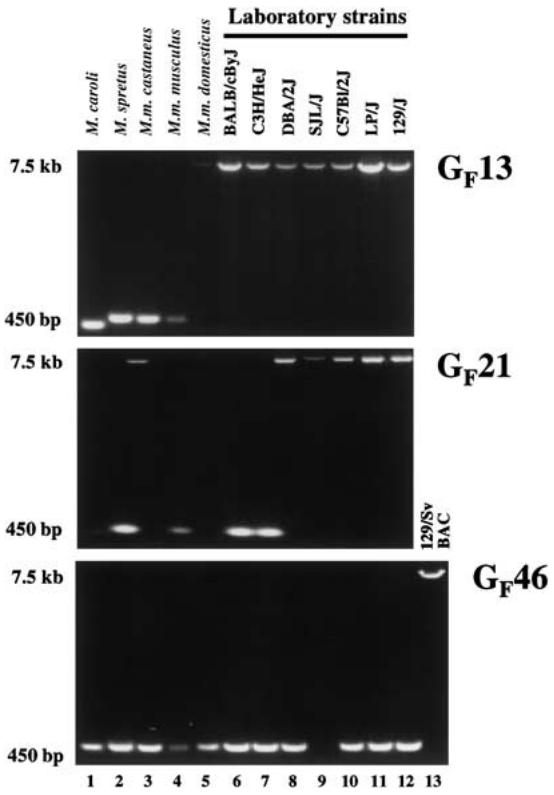

Figure 5 Some G_F elements are recent insertions. Shown are PCR reaction products which verify the presence (7.5-kb product) or absence (450-bp product) in the genomes of individual mice of different subspecies and laboratory strains. A faint, ~7.5-kb product is present in lane *5* (*M. m. domesticus*) of G_F21 which we have not been able to confirm as L1-related by sequencing. In the case of G_F46, no PCR product is seen in lane *9* (strain SJL/J), although in other reactions we have been able to amplify an obvious 450-bp product. The 7.5-kb product labeled 129/Sv BAC (lane *13* of G_F46) was amplified from bacmid DNA of CITB/Research Genetics clone #437P9.

those authors showed that ORF1 protein of the T_F element $L1_{spa}$ dimerizes more strongly than does protein of an A-type element, and they speculate that this might be due to the longer LPR region of the T_F ORF. It would be interesting to determine whether those G_F elements with the long Group V LPR show even stronger ORF1 interactions. However, the finding that two very active A-type elements have the short Group II LPR indicates that LPR length does not correlate with retrotransposition capacity.

Three ORF1 protein variants, p41, p43, and p43.5, have been detected in mouse F9 cells by Kolosha and Martin (1995). A-type and T_F-type elements have been identified with the first two forms, respectively. Phosphatase treatment suggested that the 43.5 kD protein was a phosphorylated form of p43. Furthermore, the 43.5 kD form is too small to represent ORF1 protein expressed from G_F elements with the Group V LPR (expected size 44.5 kD). Previous RT-PCR analyses also failed to detect ORF1 sequence with a Group V LPR (Schichman et al. 1992; Kolosha and Martin 1995), although the primers used in those studies likely would not have detected G_F RNA.

It would appear that the dominant active elements in the mouse genome are members of the T_F subfamily which have accounted for at least five of the seven known mutagenic insertions. These include $L1_{spa}$ (Kingsmore et al. 1994; Mulhardt et al. 1994), $L1_{orl}$ (Takahara et al. 1996), and insertions into the *beige* (Perou et al. 1997), *black-eyed white* (Yajima et al. 1999), and *disabled1* (Kojima et al. 2000) genes. The *disabled* gene L1 lacks a poly(A) tail, 3′ end, and TSDs, and thus may not be a bona fide retrotransposition event. The remaining insertions into the sodium channel gene *Scn8a* (Kohrman et al. 1996) and the copper-transporting ATPase gene *Atp7a* (Cunliffe et al. 2001) are too short to assign to a subfamily (Kohrman et al. 1996). However, the number of elements characterized to date is still small, and we expect that A and G_F insertions will eventually be detected. Phylogenetic analysis of several G_F elements reveals their relatively recent expansion in the *Mus* genus. All three G_F elements tested here were absent from the *M. spretus* and *M. m. musculus* (CZECH II/Ei) genomes. *M. spretus* diverged from the *M. musculus* lineage 1–3 million years ago (Thaler 1986), and *M. musculus* subspecies shared a common ancestor 350,000 years ago (She et al. 1990). We have also shown that element G_F46 is a de novo insertion specific to a particular mouse substrain or perhaps even a single individual.

Three distinct L1 subfamilies possessing unique promoter regions are simultaneously capable of retrotransposition in the mouse genome. Approximately 400 G_F elements and 900 A elements are potentially active in the mouse. DeBerardinis et al. (1998) estimated that mouse strain 129/OlaHsd contains ~4800 full-length T_F elements, while Saxton and Martin (1998) reported that the diploid genome of strain 129/Sv contains 2500 full-length elements. Examination of GenBank's mouse htgs database with amino acid sequence from the N-terminus of ORF1 suggested ~3000 as the number of T_F elements in the mouse genome. Assuming that 64% of T_F elements are retrotranspositionally competent (DeBerardinis et al. 1998), ~1800 of these elements may be present in the diploid genome. Thus the total number of potentially active L1 retrotransposons from the A, T_F, and G_F subfamilies in the diploid genome is roughly 3000. This is in sharp contrast to 40–70 active L1s in the human genome, most belonging to a single subfamily, Ta (Skowronski et al. 1988; Sassaman et al. 1997).

The large number of potentially active elements in the mouse accounts for the much higher rate of mutation due to L1 retrotransposition. It has been estimated that L1 insertions make up ~2.5% of spontaneous mutations in the mouse, and only 0.07% in humans (Kazazian and Moran 1998). Therefore, it is probable that retrotransposition has been a factor in driving the high rate of evolution in the mouse (She et al. 1990).

METHODS

Isolation of L1s, Subcloning and Retrotransposition Assay

Using G_F monomer sequence to search GenBank entries, we examined the highest scoring contigs in both the nonredundant (score >150) and high-throughput genomic sequence (score >200) databases and extracted from these 17 G_F elements (Table 1). Only full-length elements having obvious TSDs were considered.

To isolate full-length A elements, we screened a phage genomic library of the strain 129/OlaHsd embryonic stem cell line E14TGa with a monomer from the A-type element L1Md-A2 (Loeb et al. 1986). A-type elements were isolated from the phage and cloned by two methods. First, we amplified ten

Table 2. G_F Elements Retrotranpose in Human TK$^-$ 143B Cells and HeLa Cells

Name	Strain Clone #)	Retrotransposition frequency			
		143B cells		HeLa cells	
		with CMV	w/o CMV	with CMV	w/o CMV
G_F13	120/Ola (19–37)		1		<1
	C57Bl/2J (14–25)		0		<1
	129/Ola (19)	12		26	
	C57Bl/6J (15)	3		6	
	C57Bl/6J (14)	6		4	
	129/Sv^{+C}/$^{+P}$ (22)	0		nk	
G_F21	C57Bl/6J (18–17)		1482		720
	C57Bl/6J (35–10)		1418		1683
	C57Bl/6J (18–36)	1150		1523	
	C57Bl/6J (35–37)	727		2942	
	C57Bl/6J (35–92)	394		2240	
	SJL/J (86–39)	380		816	
G_F27	C57Bl/6J (55–62)		0		0
	C57Bl/6J (55–48)	0		2	
G_F62	LP/J (92–59)		35		1
	LP/J (92–45)	65		69	
A101		332		792	
A102		682		3767	
T_F9.1		557		1173	
$L1_{spa}$		121		352	
$L1_{spa}$ ORF2 D709Y		0		0	
$L1_{RP}$		1884		~7300	
JM105 ORF2 D709Y		0		0	

Shown for each construct is the average retrotransposition frequency (G418^R foci/10^6 hygromycin-resistant cells) for at least three independent transfections. Cloned G_F constructs were obtained by PCR amplification of DNA from a single individual from several different laboratory strains of mouse. Clonal differences in activity may be due to PCR errors. DNA from strains 129/Ola and 129/Sv^{+C}/$^{+P}$ was obtained from pluripotent embryonic stem cells (ATCC CRL-1821 (ES-E14TG2a) and ATCC CRL-1934 (ES-D3), respectively). The T_F subfamily member, T_F9, is described in DeBerardinis et al. (1998). $L1_{RP}$, a human L1 discovered as a de novo insertion in the RP2 gene of a retinitis pigmentosa patient, is the most active L1 that has so far been tested in the cell culture assay (Kimberland et al. 1999). As negative controls, we used reverse-transcriptase defective alleles of the T_F element, $L1_{spa}$ (ORF2 D709Y), and the human element, JM105 (ORF2 D709Y) (Moran et al. 1996; Naas et al. 1998).

elements by PCR (primers: L1ANOT5P, CGTACGCGGCCGC TGGTTCGAACACCAGATATCTGGG; L1ABSTZ3P, ATACG TATACATTTCCAATGCTATACCAAAAG). These were directly cloned into our retrotransposition vector under the control of a CMV promoter (Moran et al. 1996). Secondly, we excised eight A-element sequences from purified phage DNA (four were duplicates of those cloned by PCR amplification and four were unique) and inserted them into the retrotransposition construct in the same manner as described for cloning T_F elements (DeBerardinis et al. 1998). Briefly, we swapped SmaI-SfiI fragments (including 296 bp of the A-type nonmonomeric 5′ UTR and 495 bp of the 3′ UTR) for the corresponding fragment of the T_F element $L1_{spa}$. Since the resulting constructs contained both a CMV promoter and $L1_{spa}$ monomers, we could assay A-type element ORF protein activity only. Complete sequences of the two active A elements have been deposited in GenBank (accession numbers AY053455 and AY053456).

To test G_F elements, we selected from our dataset four elements, three with two intact ORFs, and we PCR-amplified these from genomic DNA of several mouse strains using primers closely flanking the TSDs. These elements were cloned into the retrotransposition vector with or without a CMV promoter. All cell culture retrotransposition assays were performed in either human 143B TK$^-$ osteosarcoma (ATCC# CRL-8303) or HeLa cells as described (Moran et al. 1996). In two previous papers (Naas et al. 1998; DeBerardinis et al. 1998) we reported testing T_F element retrotransposition in mouse LTK$^-$ cells. It recently came to our attention that the cells assayed were not mouse LTK$^-$ cells but rather human TK$^-$ 143B cells. We have determined that L1s are also active in LTK$^-$ cells but at a level two orders of magnitude lower than reported.

Analysis of GF Polymorphism

Mouse genomic DNA was obtained from the Jackson Laboratories. Primers used for genomic amplification of the following elements were as follows. G_F13: GTCTGCGTAAGGCCT GTGCTTGC (1AF1465P) and GCAAGTTTGATCTTCACCAT CAGG (2AF1463P); G_F21: TTCCTGATATGAAGCCTATG TACC (3AC02163P5) and TCTCTGAATGTTACATGATTTGGC (4AC02163P3); G_F46: GCCTGTGCTCTAAATCGCCAACAC (5AL049ES5P) and AGAGAAGTACCTGCGTGGCCCACC (6AL049ES3P). The *M. spretus* sample is SPRET/Ei, the *M. m. castaneus* sample is CAST/Ei, the *M. m. musculus* sample is CZECH II/Ei, and the *M. m. domesticus* sample is WSB/Ei. Amplifications were performed with the Expand Long PCR system (Roche), and PCR conditions were optimized for the generation of both empty site (~450 bp) and filled site (7.1–7.6 bp) products using an MJ Research PTC-200 Peltier Thermal Cycler. DNA mixing studies showed that our PCR conditions could detect in a single reaction both empty and filled sites (data not shown). Despite this ability, no individual mice heterozygotic for the presence and absence of an L1 were discovered.

Many PCR products were gel purified and directly sequenced to confirm the presence or absence of the L1. The L1 G_F46 was also amplified from the bacmid clone originally used to generate the GenBank sequence entry (CITB/Research Genetics clone #437P9, originating from mouse strain 129/Sv).

Estimation of GF Copy Number

We screened the mouse genomic phage library with the monomer array from element G_F46. The array was PCR-amplified from the bacmid clone DNA (CITB/Research Genetics clone #437P9) using the 5′ primer, TAAGGAATTCCATC TATTTCGAGGGGGTAAAG (3AL049ECOR), and 3′ primer, TAAGCTCGAGTCCCAGAAGCTGTGTTGCTTTG (2PROBE-GXHO). The GenBank entry (GI# 7630118) shows G_F46 to have 5.1 monomers. However, our PCR product was 600 bp longer than expected, suggesting that G_F46 in reality contains about eight monomers and that the GenBank sequence was misassembled. We confirmed that our probe T_F contained no non-L1 DNA by end-sequencing and by analysis with restriction enzymes cutting only once within each monomer. The PCR product was cloned into *Xho*I/*Eco*RI sites of pBS KS − and reexcised for use as a hybridization probe to screen nylon filters lifted from five plates containing about 5×10^4 plaques each. Hybridization and washing conditions were high-stringency under standard conditions (Sambrook and Russell 2001). We confirmed by slot blot analysis that the probe would not cross-hybridize to the T_F element $L1_{spa}$, an A-monomer probe from L1Md-A2 (accession #M13002; Loeb et al. 1986), or to an F-monomer probe from Padgett et al. (1988) but would hybridize to clones G_F21 and G_F62. Knowing the average insert size of the library (16 kb) and the mouse diploid genome size (6×10^9), we estimated the G_F copy number to be ~1500 elements.

We also examined GenBank's high-throughput genomic (htgs) mouse database to confirm our estimate of the number of full-length G_F or T_F elements, by using the first 63 amino acid residues of consensus ORF1 as query sequence in a `TBLASTN` search. This region of ORF1 was used because it contains both G_F and T_F subfamily-specific residues and because L1s that contain the 5′ end of the ORF1 coding region are likely to represent full-length elements. At the time the searches were performed, the htgs database represented 9.6% of the total mouse genome. After counting the number of sequence hits, we extrapolated to determine the total number of G_F and T_F elements in the diploid genome (1500 and 3000 elements, respectively).

DNA Sequence and Phylogenetic Analyses

Sequences were aligned with `MacVector 6.53` (Oxford Molecular Group) or `ClustalW` 1.8 and consensus sequences were determined with `MacVector`. Phylogenetic analyses were performed with the `ClustalW` program using the neighbor joining algorithm of Saitou and Nei (1987) with exclusion of gaps. Significance was determined by 1000 bootstrap analyses. Unrooted maximum parsimony analyses were also performed as confirmation (using the `DNAPARS` program in `PHYLIP` ver. 3.5c; Felsenstein 1993). Trees produced by the phylogenetic analyses were viewed and manipulated with `TreeView`.

ACKNOWLEDGMENTS

We thank R.J. DeBerardinis for critical reading of the manuscript, E. Luning-Prak for mouse genomic DNA, and K. Kaestner for supplying the mouse genomic library. The work was supported by a grant from the NIH to H.H.K and a Howard Hughes Medical Institute Predoctoral Fellowship to E.M.O.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adey, N.B., Schichman, S.A., Graham, D.K., Peterson, S.N., Edgell, M.H. and Hutchison, III, C.A. 1994a. Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. *Mol. Biol. Evol.* **11:** 778–789.

Adey, N.B., Tollefsbol, T.O., Sparks, A.B., Edgell, M.H., and Hutchison, III, C.A.. 1994b. Molecular resconstruction of an extinct ancestral promoter for mouse L1. *Proc. Natl. Acad. Sci.* **91:** 1569–1573.

Becker, K.G., Swergold, G.D., Ozato, K., and Thayer, R.E. 1993. Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Hum. Mol. Genet.* **2:** 1697–1702.

Cunliffe, P., Reed, V., and Boyd, Y. 2001. Intragenic deletions at *Atp7a* in mouse models for Menkes disease. *Genomics* **74:** 155–162.

DeBerardinis, R.J. and Kazazian, Jr., H.H. 1999. Analysis of the promoter from an expanding mouse retrotransposon subfamily. *Genomics* **56:** 317–323.

DeBerardinis, R.J., Goodier, J.L., Ostertag, E.M., and Kazazian, Jr., H.H. 1998. Rapid amplification of a retrotransposon subfamily is evolving the mouse genome. *Nat. Genet.* **20:** 288–290.

Fanning, T.G. 1983. Size and structure of the highly repetitive BamHI element in mice. *Nucleic Acids Res.* **11:** 5073–5091.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) Version 3.5c, Distributed by the author, Dept. of Genetics, University of Washington.

Flood, W.D., Rogozin, I.B., and Ruvinsky, A. 1998. A novel subfamily of LINE-derived elements in mice. *Mamm. Genome* **9:** 881–885.

Goodier, J.L., Ostertag, E.M., and Kazazian, Jr., H.H. 2000. Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9:** 653–657.

Hardies, S.C., Wang, L., Zhou, L., Zhao, Y., Casavant, N.C., and Huang, S. 2000. LINE-1 (L1) lineages in the mouse. *Mol. Biol. Evol.* **17:** 616–628.

Hayward, B.E., Zavanelli, M., and Furano, A.V. 1997. Recombination creates novel L1 (LINE-1) elements in *Rattus norvegicus*. *Genetics* **146:** 641–654.

Hohjoh, H. and Singer, M.F. 1996. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J.* **15:** 630–639.

Holmes, S.E., Singer, M.F., and Swergold, G.D. 1992. Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. *J. Biol. Chem.* **267:** 19765–19768.

Hutchison, C.A., Hardies, S.C., Loeb, D.D., Shehee, W.R., and Edgell, M.H. 1989. LINEs and related retroposons: Long interspersed sequences in the eucaryotic genome. In *Mobile DNA* (eds. D.E. Berg and M.M. Howe), pp. 593–617. ASM Press, Washington, D.C.

Jubier-Maurin, V., Cuny, G., Laurent, A.-M., Paquereau, L., and Roizes, G. 1992. A new 5′ sequence associated with mouse L1 elements is representative of a major class of L1 termini. *Mol. Biol. Evol.* **9:** 41–55.

Kazazian, Jr., H.H. and Moran, J.V. 1998. The impact of L1 retrotransposons on the human genome. *Nat. Genet.* **19:** 19–24.

Kimberland, M.L., Divoky, V., Prchal, J., Schwahn, U., Berger, W., and Kazazian, Jr., H.H. 1999. Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Hum. Mol. Genet.* **8:** 1557–1560.

Kingsmore, S.F., Giros, B., Suh, D., Bieniarz, M., Caron, M.G., and Seldin, M.F. 1994. Glycine receptor beta-subunit gene mutation in spastic mouse associated with LINE-1 element insertion. *Nat. Genet.* **7:** 136–141.

Kohrman, D.C., Harris, J.B., and Meisler, M.H. 1996. Mutation detection in the *med* and *medJ* of the sodium channel Scn8a. *J. Biol. Chem.* **271:** 17576–17581.

Kojima, T., Nakajima, K., and Mikoshiba, K. 2000. The *disabled1* gene is disrupted by a replacement with L1 fragment in yotari mice. *Mol. Brain Res.* **75:** 121–127.

Kolosha, V.O., and Martin, S.L. 1997. In vitro properties of the first ORF protein from mouse LINE-1 support its role in

ribonucleoprotein particle formation during retrotransposition. *Proc. Natl. Acad. Sci.* **94:** 10155–10160.

Kraft, R., Kadyk, L., and Leinwand, L.A. 1992. Sequence of organization of variant mouse 4.5 S RNA genes and pseudogenes. *Genomics* **12:** 255–566.

Loeb, D.D., Padgett, R.W., Hardies, S.C., Sheehee, W.R., Comer, M.B., Edgell, M.H., and Hutchinson, III, C.A. 1986. The sequence of a large L1Md element reveals a tandemly repeated 5′ end and several features found in retrotransposons. *Mol. Cell. Biol.* **6:** 168–182.

Martin, S.L. 1995. Characterization of a LINE-1 cDNA that originated from RNA present in ribonucleoprotein particles: Implications for the structure of an active mouse LINE-1. *Gene* **53:** 261–266.

Martin, S.L., Li, J., and Weisz, J.A. 2000. Deletion analysis defines distinct functional domains for protein–protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1. *J. Mol. Biol.* **304:** 11–20.

Mears, M.L. and Hutchison, III, C.A. 2001. The evolution of modern lineages of mouse L1 elements. *J. Mol. Evol.* **52:** 51–62.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, Jr., H.H. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87:** 917–927.

Moran, J.V., DeBerardinis, R.J., and Kazazian, Jr., H.H. 1999. Exon-shuffling by L1 retrotransposition. *Science* **283:** 1530–1534.

Mulhardt, C., Fischer, M., Gass, P., Simon-Chazottes, D., Guenet, J.L., Kuhse, J., Betz, H., and Becker, C.M. 1994. The spastic mouse: Aberrant splicing of glycine receptor beta subunit mRNA caused by intronic insertion of L1 element. *Neuron* **13:** 1003–1015

Naas, T.P., DeBerardinis, R.J., Moran, J.V., Ostertag, E.M., Kingsmore, S.F., Seldin, M.F., Hayashizaki, Y., Martin, S.L., and Kazazian, Jr., H.H. 1998. An actively retrotransposing, novel subfamily of mouse L1 elements. *EMBO J.* **17:** 590–597.

Padgett, R.W., Hutchinson, III, C.A., and Edgell, M.H. 1988. The F-type 5′ motif of mouse L1 elments: A major class of L1 termini similar to the A-type in organization but unrelated in sequence. *Nucl. Acids Res.* **16:** 739–749.

Perou, C.M., Pryor, R.J., Naas, T.P., and Kaplan, J. 1997. The *bg* allele mutation is due to a LINE1 element retrotransposon. *Genomics* **42:** 366–8.

Saitou, N. and Nei, M. 1987. The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Ecol.* **4:** 406–435.

Sambrook, J. and Russell, D.W. 2001. *Molecular Cloning, A Laboratory Manual*. Cold Spring Harbor Laboratory Press, New York.

Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., and Kazazian, Jr., H.H. 1997. Many human L1 elements are capable of retrotransposition. *Nat. Genet.* **16:** 37–43.

Saxton, J.A. and S.L. Martin. 1998. Recombination between subtypes creates a mosaic lineage of LINE-1 that is expressed and actively retrotransposing in the mouse genome. *J. Mol. Biol.* **280:** 611–622.

Schichman, S.A., Severynse, D.M., Edgell, M.H., and Hutchison, III,C.A. 1992. Strand-specific LINE-1 transciption in mouse F9 cells originates from the youngest phylogenetic subgroup of LINE-1 elements. *J. Mol. Biol.* **224:** 559–574.

Severynse, D.M., Hutchinson, III, C.A., Edgell, M.H. 1992. Identification of transcriptional regulatory activity within the 5′ A-type monomer sequence of the mouse LINE-1 retroposon. *Mamm. Genome* **2:** 41–50.

She, J.X., Bonhomme, F., Boursot, P., Thaler, L., and Catzeflis, F. 1990. Molecular phylogenies in the genus *Mus*: Comparative analysis of electrophoretic, scnDN5, and mtDNA RFLP data. *Biol. J. Linnean Soc.* **41:** 83–103.

Shehee, W.R., Chao, S., Loeb, D.D., Comer, M.B., Hutchison, III, C.A., and Edgell, M.H. 1987. Determination of a functional ancestral sequence and definition of the 5′ end of A-type mouse L1 elements. *J. Mol. Biol.* **196:** 757–767.

Skowronski, J., Fanning, T.G., and Singer, M.F. 1988. Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol. Cell. Biol.* **8:** 1385–1397.

Takahara, T., Ohsumi, T., Kuromitsu, J., Shibata, K., Sasaki, N., Okazaki, Y., Shibata, H., Sato, S., Yoshiki, A., Kusakabe, M., et al. 1996. Dysfunction of the Orleans reeler gene arising from exon skipping due to transposition of a full-length copy of an L1 sequence into the skipped exon. *Hum. Mol. Genet.* **5:** 989–993.

Thaler, L. 1986. Origin and evolution of mice: An appraisal of fossil evidence and morphological traits. *Curr. Topics Microbiol. Immunol.* **127:** 3–11.

Yajima, I., Sata, S., Kimura, T., Yasumoto, K. Shibahara, S., Goding, C.R., and Yamamoto, H. 1999. An L1 element intronic insertion in the *black-eyed white* (*Mitfmi-bw*) gene: The loss of a single Mitf isoform responsible for the pigmentary defect and inner ear deafness. *Hum. Mol. Genet.* **8:** 1431–1441.

Received May 24, 2001; accepted in revised form July 25, 2001.