



## L1 integration in a transgenic mouse model

Daria V. Babushok, Eric M. Ostertag, Christine E. Courtney, et al.

*Genome Res.* 2006 16: 240-250

Access the most recent version at doi:[10.1101/gr.4571606](https://doi.org/10.1101/gr.4571606)

---

**References** This article cites 75 articles, 27 of which can be accessed free at:  
<http://genome.cshlp.org/content/16/2/240.full.html#ref-list-1>

### License

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Cold Spring Harbor Laboratory Press

# L1 integration in a transgenic mouse model

Daria V. Babushok,<sup>1</sup> Eric M. Ostertag,<sup>1,2,3</sup> Christine E. Courtney,<sup>1</sup> Janice M. Choi,<sup>1</sup> and Haig H. Kazazian Jr.<sup>1,4</sup>

<sup>1</sup>Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; <sup>2</sup>Department of Pathology and Laboratory Medicine, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; <sup>3</sup>Transposagen Biopharmaceuticals, Inc., Philadelphia, Pennsylvania 19104, USA

To study integration of the human LINE-1 retrotransposon (L1) in vivo, we developed a transgenic mouse model of L1 retrotransposition that displays de novo somatic L1 insertions at a high frequency, occasionally several insertions per mouse. We mapped 3' integration sites of 51 insertions by Thermal Asymmetric Interlaced PCR (TAIL-PCR). Analysis of integration locations revealed a broad genomic distribution with a modest preference for intergenic regions. We characterized the complete structures of 33 de novo retrotransposition events. Our results highlight the large number of highly truncated L1s, as over 52% (27/51) of total integrants were <1/3 the length of a full-length element. New integrants carry all structural characteristics typical of genomic L1s, including a number with inversions, deletions, and 5'-end microhomologies to the target DNA sequence. Notably, at least 13% (7/51) of all insertions contain a short stretch of extra nucleotides at their 5' end, which we postulate result from template-jumping by the L1-encoded reverse transcriptase. We propose a unified model of L1 integration that explains all of the characteristic features of L1 retrotransposition, such as 5' truncations, inversions, extra nucleotide additions, and 5' boundary and inversion point microhomologies.

[Supplemental material is available online at [www.genome.org](http://www.genome.org).]

The long interspersed nucleotide element-1 (L1) retrotransposon enjoyed tremendous evolutionary success in colonizing eukaryotic genomes (Kazazian Jr. 2004); its roughly 500,000 copies comprise ~17% of human DNA (Lander et al. 2001). The full-length 6-kb L1 encodes a 5' UTR containing an internal promoter, two proteins—ORF1, a nucleic acid-binding protein with chaperone activity (Hohjoh and Singer 1996, 1997; Martin and Bushman 2001), and ORF2, a protein with endonuclease (EN) and reverse transcriptase (RT) activities (Mathias et al. 1991; Feng et al. 1996), and a 3' UTR ending with a poly(A) tail (Fig. 1A). Both ORF1 and ORF2 proteins are required for retrotransposition (Moran et al. 1996). Once transcribed and translated, L1 RNA is copied into the genome by target-primed reverse transcription (TPRT) (Luan et al. 1993; Cost et al. 2002). During TPRT, one strand of host DNA is cleaved by the EN to expose a free 3'-hydroxyl, which is then used by the RT as a primer in reverse transcription, copying the L1 RNA template directly into the host genome. To complete integration, the second strand of host DNA must be cleaved, RNA removed, the newly synthesized strand copied, and breaks resolved. The mechanism by which these later steps are accomplished is only beginning to be understood, with recent biochemical data from a related non-LTR retrotransposon R2 of the silkworm, *Bombyx mori*, suggesting that a second R2 protein is involved in cleavage, RNA displacement, and synthesis of the second strand (Bibillo and Eickbush 2002a; Christensen and Eickbush 2005).

Following essentially random integration, endogenous L1s are thought to be lost over time due to strong negative selection leading to their uneven genomic distribution (Boissinot et al. 2001, 2004; Ovchinnikov et al. 2001; Medstrand et al.

2002). This is reflected in their overrepresentation in GC-poor genomic regions (Lander et al. 2001; Ovchinnikov et al. 2001; Boissinot et al. 2004), under-representation within 5 kb of genes (Medstrand et al. 2002), and the selective loss of longer (>500 bp) L1s from recombining regions of the genome (Boissinot et al. 2001).

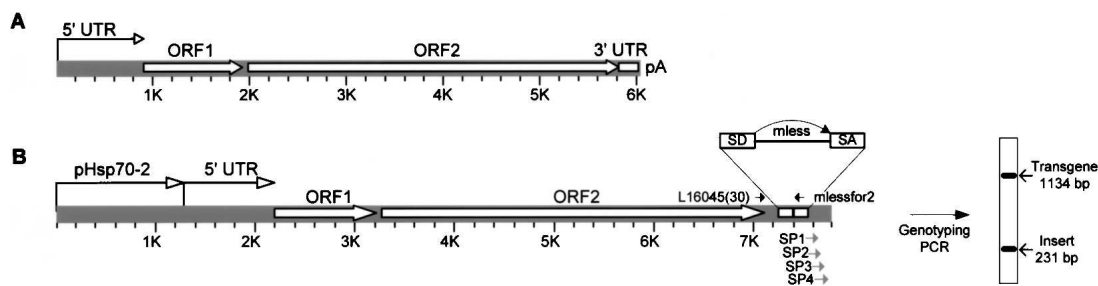
The L1's 5' end variability and highly repetitive nature make cloning insertions difficult and have prohibited analysis of a large number of de novo integrants. In contrast to several thousands of analyzed integrants for retroviruses and DNA transposons (Bushman et al. 2005; Yant et al. 2005), the total number of completely characterized de novo L1 insertions is under 200. These include 15 known L1 insertions causing human disease (Ostertag and Kazazian Jr. 2001a; Chen et al. 2005) and ~150 integrants characterized in transformed cell lines (Moran et al. 1996; Gilbert et al. 2002, 2005; Symer et al. 2002). Nonetheless, even this small number of insertions revealed several key insights into the mechanism and potential impact of L1 retrotransposition, including the prevalence of microhomologies at the 5' end of de novo inserts (Symer et al. 2002; Gilbert et al. 2005) and a realization that ~10% of retrotransposition events may produce large deletions of host DNA sequence at the integration site (Gilbert et al. 2002, 2005; Symer et al. 2002). However, as noted previously (Symer et al. 2002), known defects in DNA stability and repair pathways in cultured cells create a concern that some features of retrotransposition observed in cell culture may not be generalizable. Indeed, some integration phenomena common in cultured cells, such as the creation of deletions and large target-site duplications, may occur much less frequently in an organism (Szak et al. 2002; Vincent et al. 2003; Chen et al. 2005; Han et al. 2005; Ho et al. 2005).

To better understand L1 integration, a study of de novo L1 insertions in vivo is needed. Unfortunately, previous transgenic mouse models of L1 retrotransposition (Ostertag et al. 2002; Prak

#### <sup>4</sup>Corresponding author.

E-mail [kazazian@mail.med.upenn.edu](mailto:kazazian@mail.med.upenn.edu); fax (215) 573-7760.

Article published online ahead of print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.4571606>.



**Figure 1.** Structures of L1 element and L1-mless transgene. (A) Human full-length L1 contains a 5'UTR, ORF1, ORF2, 3'UTR, and a poly(A) tail. (B) L1<sub>LRE3</sub> is tagged in its 3'UTR with a markerless retrotransposition cassette (mless) and is driven by the mouse Hsp70-2 promoter. Mless contains the human  $\gamma$ -globin intron in the same orientation as L1 (SD and SA denote splice donor and acceptor sites, respectively), flanked by splice acceptors from the human Bcl-2 gene (not shown). After retrotransposition, the intron is spliced, allowing de novo insertions to be distinguished from the original transgene by PCR with oligos L16045(30) and mlessfor2 (black arrows). The expected band pattern from this "genotyping PCR" is shown on the right. Gray arrows SP1–SP4 indicate insertion-specific oligos used in TAIL-PCR.

et al. 2003) had disappointingly low retrotransposition frequencies, allowing characterization of only a few novel insertions. Here, we describe a transgenic mouse model with a higher rate of retrotransposition, in which a more active human L1 element is tagged with a markerless cassette, allowing for PCR-based detection of shorter insertions. One such transgenic line displayed a high frequency of L1 insertions, occasionally several insertions per mouse. We mapped the genomic locations of 51 retrotransposition events, and completely characterized the structures of 33. Based on their structural features, we propose an L1 integration model that accounts for all common features of L1 retrotransposition.

## Results

### L1<sub>LRE3</sub>-mless transgenic mice are somatic mosaics for de novo retrotransposition events

To increase the retrotransposition frequency in transgenic mice, we modified our previous transgene to include a more active L1 (L1<sub>LRE3</sub>) (Brouha et al. 2002, 2003) tagged with a short markerless cassette, permitting PCR-based genotyping of smaller insertions (>400 bp) (Fig. 1B). The L1-mless transgene was driven by a mouse heat-shock protein 70-2 promoter (pHsp70-2) known to produce strong transgene expression in meiosis of male spermatogenesis (Dix et al. 1996). Nine independent L1<sub>LRE3</sub>-mless transgenic mouse lines were produced, and the line (founder number 27, H27) with the highest retrotransposition frequencies in F<sub>1</sub> offspring was studied further.

We screened H27 F<sub>1</sub> offspring for the presence of transgenes and insertions using "genotyping PCR" (Fig. 1B). Of 23 F<sub>1</sub> offspring of H27, 16 contained transgene (~70%), at least nine of which were accompanied by one or more insertions. The apparent high rate of transgene inheritance (>50%) persisted in F<sub>2</sub> offspring, and led us to suspect the presence of multiple integrated transgenes in the founder. We were able to characterize the genomic locations of two distinct transgene arrays by TAIL-PCR, and found that H27 founder contained transgenes integrated in three distinct locations, i.e., chromosome 11 (transgene-11), chromosome 18 (transgene-18), and an unknown location (transgene-Un). Individual transgenes showed independent segregation and dramatic differences in retrotransposition capability, likely due to position effects, i.e., transgene-11 displayed the highest retrotransposition frequencies with some insertions inherited independent of transgene (Fig. 2), transgene-

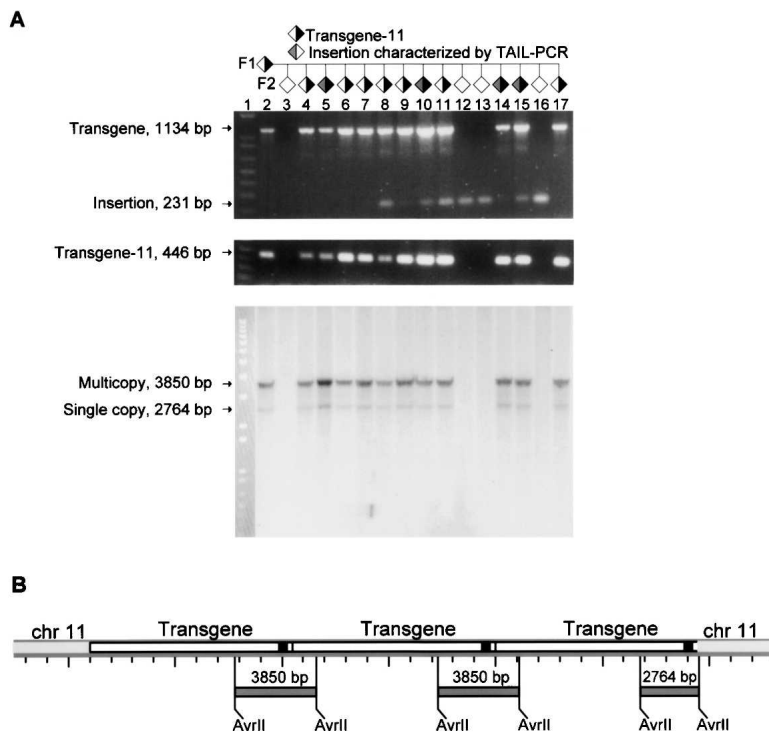
18 showed intermediate frequencies, while transgene-Un produced little retrotransposition.

To determine the genomic locations of L1 retrotransposition events, DNA was obtained from tails of transgenic mice and used in TAIL-PCR reactions (Liu and Whittier 1995; Liu et al. 1995) to amplify the L1 3' end together with its flanking host DNA. TAIL-PCR allowed detection of shorter insertions than would have been possible with positive selection (EGFP or *neo* retrotransposition cassettes) or genotyping PCR. This is because TAIL-PCR does not require the 5'-truncated elements to contain a spliced intron junction for detection, thus lowering the detection limit to 214 nt. Sequences of amplified products were compared with the mouse genome to obtain chromosomal locations.

Using this procedure, we determined the locations of 51 distinct retrotransposition events from 41 different mice. For some mice, two or three distinct insertions were recovered. We have not yet isolated shared insertions from F<sub>2</sub> siblings as expected for insertions inherited from parent F<sub>1</sub>s. Southern blot revealed that individual de novo insertions were present at less than one copy per cell, a result indicative of somatic retrotransposition events (Fig. 2). Segregation of insertion from transgene observed by genotyping PCR in the absence of insertions detectable by Southern blot (Fig 2A, lanes 12,13,16) was confirmed by independent PCR analyses with several primer sets using independently obtained DNA. Because insertions are below the level of detection by Southern blot and because of the tremendous difficulties encountered in amplifying insertions by nested PCR, we feel that while there can be multiple somatic insertion events in a given mouse, each individual insertion is likely rare in the animal.

### New L1 retrotransposition events modestly favor intergenic regions

In 48 cases, we unambiguously mapped the insertion site. Two of the remaining three elements were in young mouse LINES and one was present only in an incompletely sequenced clone in the Celera database with no available annotation. The 48 identifiable insertions are distributed widely throughout the genome (Fig. 3A). The variation in integration rates for individual chromosomes was not statistically significant, except for the increase in integration into chr. 18 ( $P = 0.017$ ). This may be partially explained by its lower gene density and possibly higher frequency of AT-rich L1 endonuclease target sites, and is similar to the uneven distribution of Ta-1 elements in the human genome (Boissi-



**Figure 2.** Characterization of a pedigree segregating transgene-11. (A) F<sub>1</sub> mouse and its 15 offspring (top pedigree schematic) were genotyped by PCR with L16045(30) and mlessfor2 oligos (top gel) to reveal 11/15 F<sub>2</sub>s carrying transgene (1134 bp) and 7/15 carrying de novo inserts (231 bp). Minus DNA control was negative; de novo insertion presence was confirmed with different oligos and independently isolated tail DNA stocks (data not shown). All transgene inheritance was accounted for by transgene-11-specific PCR (middle gel, 446 bp). Southern blot (bottom gel) of 10 µg of tail DNA, digested with AvrII, and probed with a <sup>32</sup>P-labeled 376-bp probe showed multicopy (3850 bp) and single-copy (2764 bp) fragments expected for AvrII digest of transgene-11, shown in B. No other fragments were detected, suggesting that the frequency of de novo insertions is less than one copy/cell. Four F<sub>2</sub>s contained inserts characterized by TAIL-PCR (gray shading in pedigree schematic). A 1-kb Plus DNA Ladder is shown in lane 1 of all gels. (B) A hypothetical multicopy transgene-11 array with relevant AvrII sites and the 376 bp probe (black rectangle).

not et al. 2004). The closest distance between two insertions was ~971 kb, with no known functional similarities between integration sites.

Approximately 15% (7/48) of insertions landed in RefSeq transcription units (Pruitt et al. 2005) (Table 1), slightly lower than expected for random integration ( $P = 0.040$ ). A similar modest preference for intergenic regions ( $P = 0.044$ ) was obtained when insertion coordinates were compared with the “known gene” track of the UCSC genome browser (Karolchik et al. 2003, 2004), and Mammalian Gene Collection Full ORF mRNAs (MGC) (Strausberg et al. 2002) (not statistically significant). Two-thirds of insertions (32/48) were within annotated Genscan genes (Burge and Karlin 1997), a proportion indistinguishable from random integration. Lower exon prediction sensitivity and specificity (0.70) (Rogic et al. 2001) of Genscan gene prediction algorithm likely obscured the slight bias of L1 integration into intergenic regions using this gene annotation.

All insertions into genes were intronic, and there was no significant difference in the frequency of L1 integrants oriented in the same or opposite direction with respect to the gene. We observed no correlation with position around the transcription unit (e.g., transcription start sites) or within introns. Integration sites were also distributed randomly with respect to broad classes of repeats and CpG islands (Table 1). Perhaps with a larger sample

size, a modest preference for some of these features may be detected.

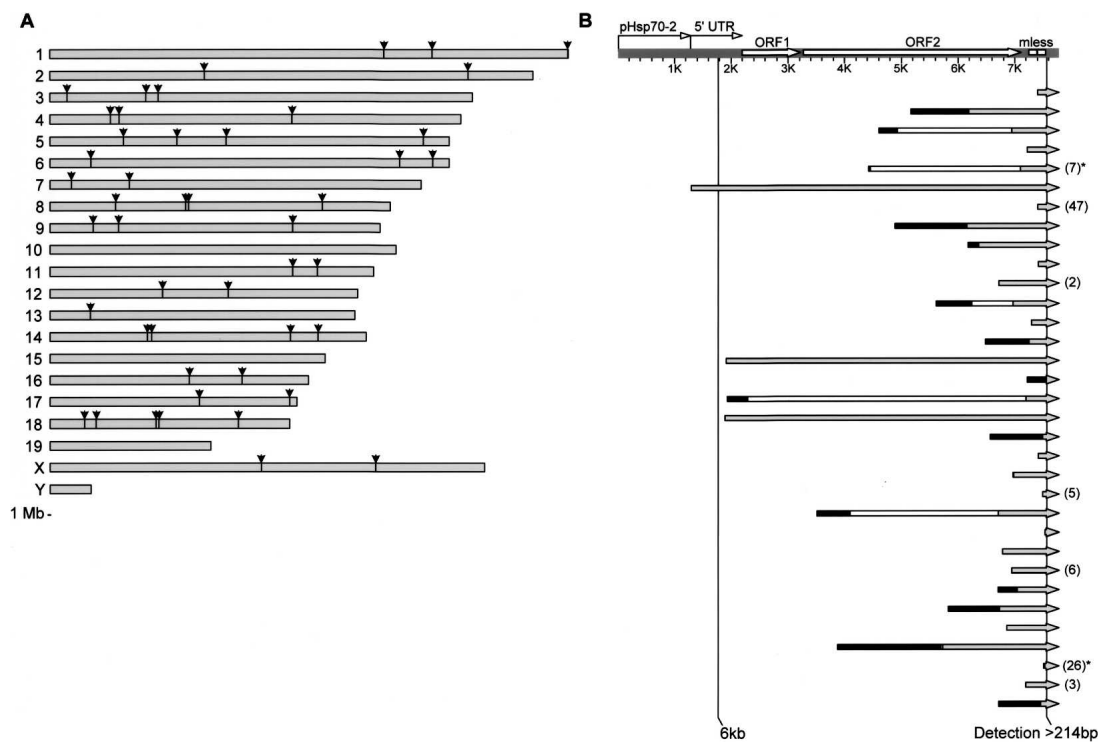
### De novo inserts have characteristic features of L1 retrotransposition

To analyze integration-site specificity, we aligned the host DNA from 10 nt upstream to 10 nt downstream of the insertion sites (Fig. 4A). Analysis of the nucleotide frequencies at each of these 20 positions revealed a highly stereotypical pattern of L1 EN cleavage (Jurka 1997; Morrish et al. 2002), with a consensus sequence of 5'-TTTT'AA-3'. Thus, the recovered integrants represent true L1 endonuclease-dependent events.

To determine the complete structures of new integrants, multiple series of nested PCRs were performed using sets of oligos specific to genomic DNA flanking the 5' ends of individual insertions and to the L1<sub>LRE3</sub>-mless construct. We were able to determine the complete structures of 33 insertions (Fig. 3B). The majority of integrants were 5' truncated; over 81% (27 of 33) contained <1/3 of the full-length L1. Their lengths ranged from 245 to 6481 bp and displayed a bimodal distribution with a large peak of very short, and a smaller peak of long, ~6-kb integrants, in qualitative agreement with the length distribution of young genomic L1s (Fig. 4B) (Myers et al. 2002; Pavlicek et al. 2002; Szak et al. 2002; Boissinot et al. 2004). It is likely that while TAIL-PCR allowed us to detect short integrants, we are still missing integrants under 214 bp. Similarly, PCR-

based detection may be limiting the recovery of the 5' ends of longer insertions, resulting in an apparent underrepresentation of full-length (FL) and nearly FL elements (3/33 or 9%) as compared with the expected 28%–34% FL L1<sub>Ta</sub> elements in the human genome (Boissinot et al. 2000; Myers et al. 2002; Pavlicek et al. 2002; Symer et al. 2002). Alternatively, it is possible that FL insertions are subject to negative selection and are underrepresented due to selective cell loss.

Finer analysis revealed stereotypical features of L1 retrotransposition (Table 2; Fig. 3B). Approximately 91% (30/33) were 5' truncated, and 51.5% (17/33) were both 5' truncated and inverted. Of 17 inversions, two resulted in a precise junction of direct and inverted segments, one contained a short duplication, and 14 contained a deletion ranging from 1 to 4901 bp. The observed frequency of inversion/deletion events (82%) is comparable to that in cultured cell studies (82%–84%) (Gilbert et al. 2002, 2005), but is higher than the ~51% reported for genomic L1s (Szak et al. 2002); the lower frequency in genomic L1s can be explained by the exclusion from that study of elements with deletions >51 nts. In our study, two of the elements contained a dual inverted fragment, likely generated by two consecutive twin priming attempts (see Extra Nucleotide Formation). These are distinct from previously reported multiple inversions comprised of adjoined segments of opposite orientation (Pavlicek et al.



**Figure 3.** Genomic distribution and structural features of de novo insertions. (A) Using TAIL-PCR, 51 de novo integration sites were determined, 48 of which were uniquely mapped to the mm6 assembly of the mouse genome. Depicted are the locations of 47 inserts (one mapped to an unplaced location and is not shown). 1-Mb scale reference is shown at the bottom. (B) The structures of 33 fully characterized de novo inserts. A hypothetical full-length (FL) insert is shown at the top, with vertical lines indicating the 6-kb endogenous FL element and the 214-bp detection limit in our study. De novo integrants are shown on separate lines, aligned to the FL element. (Direct fragment) Gray rightward arrow; (inverted fragment) black rectangle; (deletion of sequence in inverted elements) white rectangle; (extra 5' nt) numbers in parentheses; (dual inversions) asterisk. Three elements mobilized ~6 kb of sequence, one is FL and two are nearly FL.

2002; Szak et al. 2002), some of which could have been created during L1 integration into a pre-existing genomic L1 (Szak et al. 2002). The observed proportion of inverted integrants exceeds the 8%–35% inversions in the human genome (Boissinot et al. 2000; Goodier et al. 2000; Myers et al. 2002; Symer et al. 2002; Szak et al. 2002; Salem et al. 2003), and 16%–30% in cultured cells (Gilbert et al. 2002, 2005; Symer et al. 2002). More frequent inversions in our study may be caused by the relative overrepresentation of shorter elements, the reduced ability to detect inverted L1s in the genome, or as yet unknown differences intrinsic to human and mouse hosts. Among the previously char-

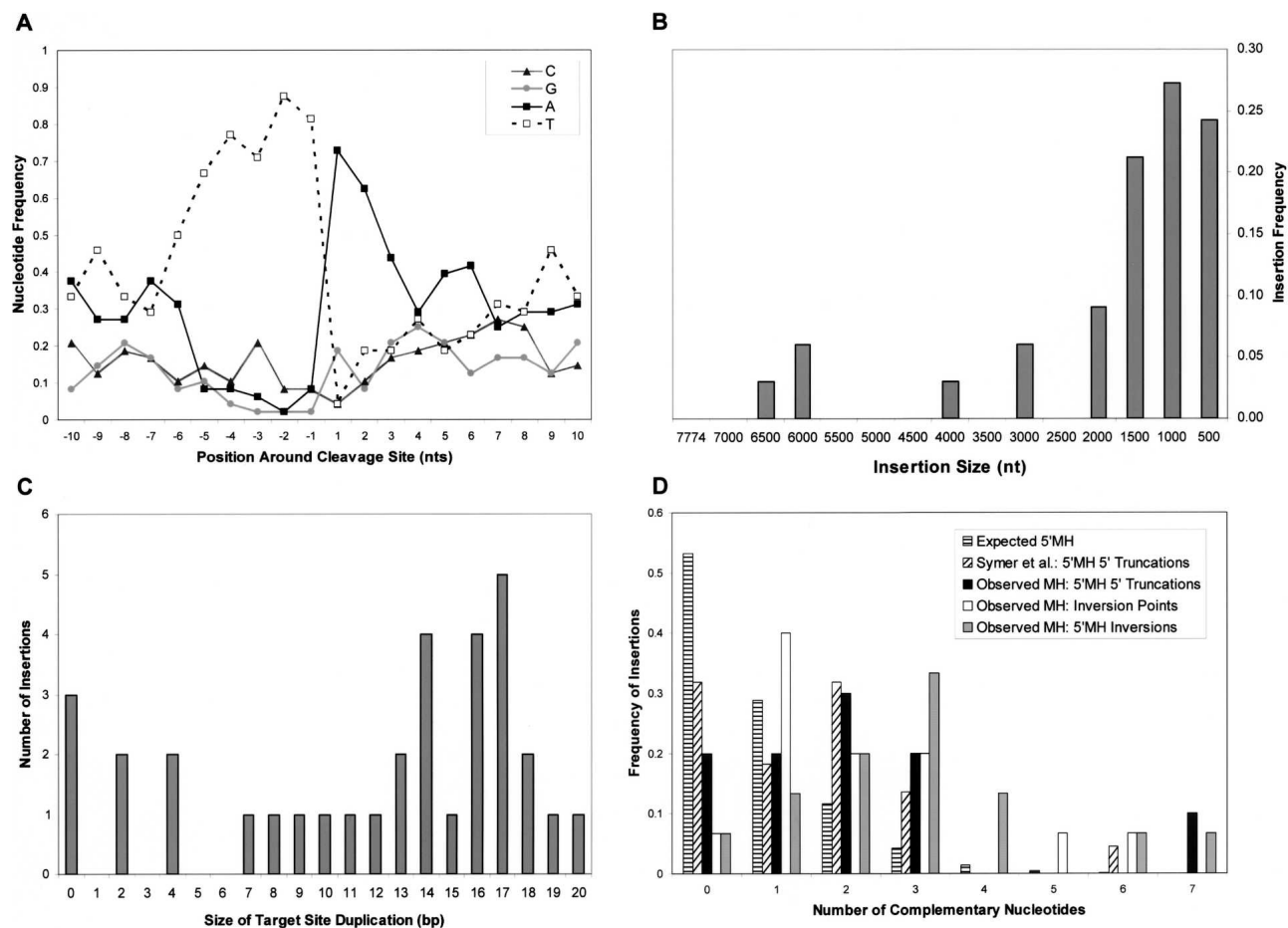
acterized L1 insertions in transgenic mice—two in Ostertag et al. (2002), one in Prak et al. (2003), and one in D.V. Babushok, E.M. Ostertag, and H.H. Kazazian Jr., (unpubl.)—two contained an inversion.

De novo L1 insertions were flanked by homopolymeric poly(A) tails of ~50–210 nt (mean  $103 \pm 38$  nt), in agreement with the long (mean 60–90 nt) poly(A)s observed in disease-causing L1 Ta elements (Chen et al. 2005) and in de novo cultured cell insertions (Gilbert et al. 2002; Symer et al. 2002). All but three de novo insertions in our study were flanked by target-site duplications (TSDs) of 2 to 20 nt (mean 13 nt) (Fig. 4C). Three insertions (~9%) contained no TSDs. Our results mirror the distribution of TSDs for L1-mediated disease-causing insertions (Chen et al. 2005), and for genomic L1s and *Alus* (Jurka 1997; Myers et al. 2002; Szak et al. 2002). We did not detect unusually long (>30 nt) TSDs that are uncharacteristic of genomic L1s, but were observed in 7%–27% of de novo insertions in cultured cells (Moran et al. 1996; Gilbert et al. 2002, 2005; Symer et al. 2002). Our 5' end recovery procedure allowed long-range amplification of full-length insertions containing up to ~1 kb of TSD. Because the majority of insertions are 5' truncated, and because long TSDs are typically under 1 kb (median 115 bp, range 34–6319 bp, only one of 28 over 1 kb) (Symer et al. 2002; Gilbert et al. 2005), it is unlikely that TSD length significantly impacted our insertion recovery (Supplemental Discussion 1).

In cultured human cells 16%–21.6% (Gilbert et al. 2002, 2005; Symer et al. 2002) of L1 integrants were accompanied by deletions of host DNA of 1 to >71,000 bps. In contrast, we did

**Table 1.** L1 Integration with respect to genomic landmarks

Genomic feature	Frequency of insertions (%)	Frequency in genome (%)	P-value
Gene annotations			
RefSeq	14.6	26.8	0.040
"known gene"	14.6	26.5	0.044
MGC	12.5	18.3	0.197
Genscan	66.7	66.6	0.562
Repeats	33.3	39.8	0.210
LINE	21.6	18.9	0.379
SINE	5.9	10.3	0.210
LTR	5.9	14.0	0.070
CpG Islands			
± 1 kb	0.0	1.5	0.390
± 25 kb	14.6	20.4	0.204



**Figure 4.** Characteristic features of L1 retrotransposition in de novo integrants. (A) Average frequencies of each base in the 10 nt surrounding the insertion sites of de novo integrants. (B) Insertion length distribution of 33 fully characterized integrants. Lengths of the inverted inserts were calculated by adding the lengths of inverted and direct segments. (C) The distribution of TSD lengths among 33 fully characterized inserts. (D) Observed microhomology (MH) frequencies at 5' boundaries of noninverted 5'-truncated insertions (black), and at the inversion points (white) and 5' boundaries (gray) of insertions containing inversions. Observed 5'MH frequencies were in agreement with those in cultured cells (Symer et al. 2002) (diagonal stripes), and were higher than expected for random ligation (Roth et al. 1985) using both the random nucleotide distribution (data not shown) and the nucleotide distribution in the 30-bp downstream of the insertion site (horizontal stripes).

not detect any deletions of host DNA at L1 integration sites. It is possible that large deletions were present among the 18 insertions that we could not fully characterize. However, we would have detected deletions up to 87 bp, because all inner-nested oligos used for recovery of 5' ends were at least 87 bp 5' to the putative insertion site (median 232 nt, range 87–570 nt). Extending our primer sets farther into the flanking host DNA (median 381 nt, range 158–759 nt) was similarly unsuccessful. Small target-site deletions ( $\leq 87$  nt [and  $\leq 158$  nts]) were found to accompany 8%–10.8% (Gilbert et al. 2002, 2005; Symer et al. 2002) of insertions in cultured cells. The lack of any deletions among the 33 characterized inserts permits us to estimate that frequency of small deletions in vivo is  $< 8.7\%$  ( $P = 0.050$ ), in agreement with the 1.2%–2.2% deletion frequency for genomic L1s (Vincent et al. 2003; Han et al. 2005; Ho et al. 2005) (Supplemental Discussion 2).

#### 5' boundary and inversion point microhomologies are common in de novo insertions in mice

Consistent with the high frequencies of microhomologies (MHs) between the 5' end of L1 RNA and flanking host DNA in cultured

cells (Symer et al. 2002; Gilbert et al. 2005) and in the human and mouse genomes (Martin et al. 2005; Zingler et al. 2005), a majority of noninverted 5' truncated de novo L1 integrants (excluding elements with extra 5' nt) contained MHs of at least 1 bp at their 5' boundaries (8/10, 80%) (Fig. 4D). As predicted by the twin priming hypothesis of inversion formation (Ostertag and Kazazian Jr. 2001b), there was a high frequency of MHs at both the 5' boundaries (14/15, 93.3%), and at the junctions of inverted and direct segments (inversion point MHs) (14/15, 93.3%) in inverted inserts.

#### Extra nucleotide formation at 5' ends of L1 integrants

The majority (7/11, 63.6%) of insertions with no MH at the 5' boundary contained extra nucleotide stretches of 2–47 bp between the flanking host DNA and the 5'-truncated L1 element. While most of these unexpected sequences were short (2–7 nt), making it difficult to ascertain their source, there were two instances of longer additions (26 and 47 nt) that allowed us to deduce their origin. In the first case, an addition of 26 nt oc-

**Table 2. Structural features of 33 characterized insertions**

Structural feature	Number observed
Full-length (~6 kb)	3
5'-truncated	30
Inverted	17
w/deletion	14
w/duplication	1
precise inversion	2
dual inversion	2
Extra 5' nucleotides	7
5' boundary microhomologies	22
noninverted 5' truncated L1s	8/10
inverted L1s	14/15
Inversion point microhomologies	14/15
Target site features	
deletion	0
precise integration	3
TSD ≤ 8 nt	6
TSD > 8 nt	24

curred in a 238-bp 5'-truncated integrant that inserted at a canonical 5'-TTTT'A-3' endonuclease cleavage site, terminated in a 90-nt poly(A), and was flanked by 13-bp TSDs. A closer examination of the unknown nucleotides revealed that they were a product of two short successive twin priming reactions by the RT, producing a "dual inverted fragment" (Fig. 5A,B). In all likelihood, the first twin priming reaction was initiated using a 1-bp MH of the L1 RNA and the 3'-most T of the 5' target-site duplication, creating the first inverted segment of 18 nt. It appears that at that point, L1 RT underwent a template-jump (or a template-slip), likely facilitated by a 1-bp MH between the 3'-most base of newly synthesized cDNA and L1 RNA. Eight base pairs from a nearby, partially overlapping site on L1 RNA were added to create the second inverted segment. Resolution of the structure did not involve any MH, and resulted in a final insertion of 264 bp. L1 RT template jumping during a twin priming reaction likely explains the addition of 7 nt in another integrant (Supplemental Discussion 3). While to the best of our knowledge, RT template jumping during twin priming has not been reported previously, our results are consistent with the conceptually similar TPRT initiation, which occasionally creates 3' extra nucleotides due to the repeated attempts of RT to fully engage the RNA template (Luan and Eickbush 1995; Eickbush et al. 2000; Cost et al. 2002).

The longer 47 extra-nucleotide addition was likely created by a series of three successive template jumps by L1 RT in the course of an otherwise unremarkable TPRT reaction (Fig. 5A,C). It appears that after the element inserted at a typical 5'-GTTT'A-3' endonuclease cleavage site and its 368 bp were reverse transcribed, the L1 RT jumped templates from the end of L1 RNA onto the upstream single-stranded region of host DNA. This template jump was likely facilitated by a 2-bp MH between the terminal nucleotides of the newly synthesized L1 cDNA and a single-stranded host DNA region within the 5' target site. This 1<sup>st</sup> template jump resulted in the addition of 11 bp complementary to the 5'-flanking host DNA, and was immediately followed by the 2<sup>nd</sup> template jump (or template-slip) of the L1 RT. The 2<sup>nd</sup> jump was likely facilitated by a 3-bp MH between the nascent cDNA end and the same single-stranded region in the 5' TSD, and added another 26 bp complementary to the host target site. A single, possibly untemplated, base was added, and the 3<sup>rd</sup> template jump occurred, copying seven more nucleotides. Another

two bases were added, and the complete 415-bp insertion was integrated.

## Discussion

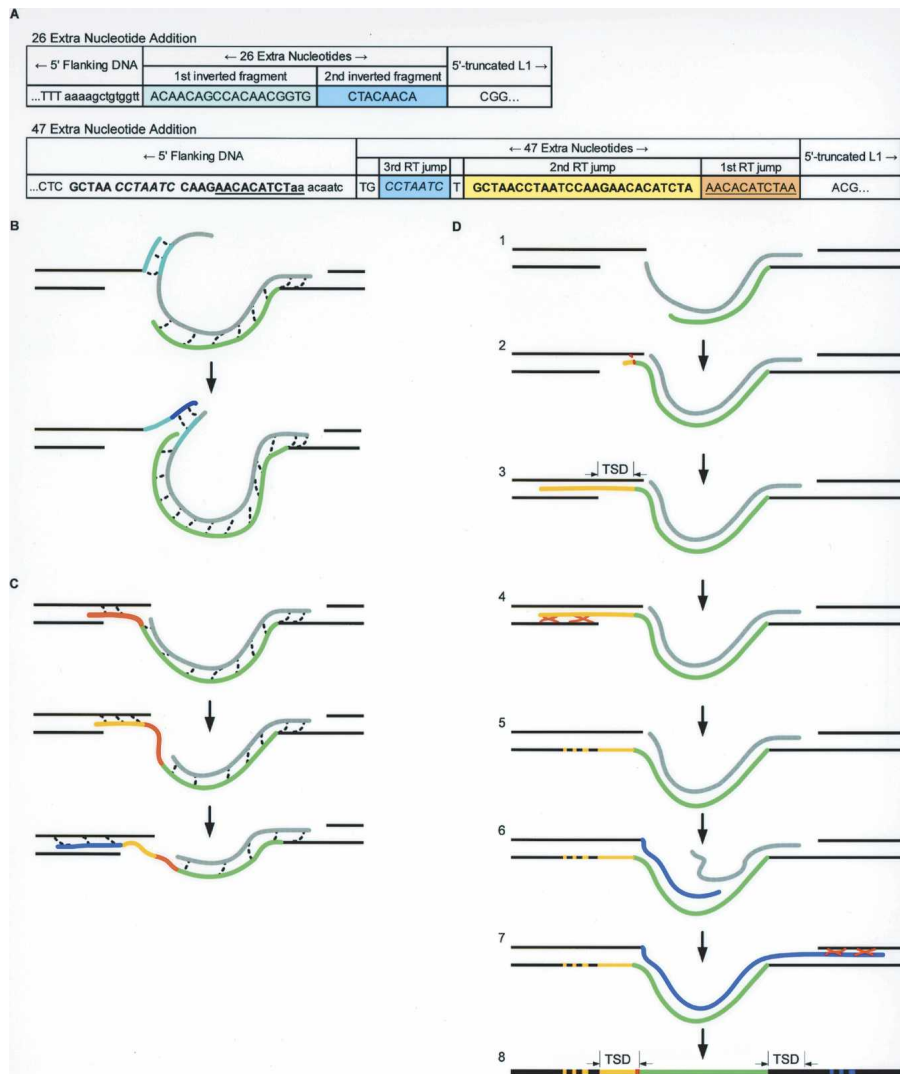
This work extends studies of L1 in cultured cells (Gilbert et al. 2002, 2005; Symer et al. 2002) by providing a comprehensive analysis of de novo retrotransposition events in vivo.

### Somatic mosaicism in mouse models of L1 retrotransposition

Traditionally, L1 retrotransposition was thought to be restricted to the germ line by DNA methylation (Ostertag and Kazazian Jr. 2001a; Bourc'his and Bestor 2004), with this regulation being impaired in cultured cells (Moran et al. 1996) and some human cancers (Miki et al. 1992). Somatic retrotransposition was recently noted during early embryogenesis in a pPolIII/5'UTR/L1<sub>RP</sub>/EGFP transgenic founder (possibly from an episomal location) (Prak et al. 2003), and in neuronal precursor cells of 5'UTR/L1<sub>RP</sub>/EGFP transgenic mice (Muotri et al. 2005). We also cannot rule out low-level somatic retrotransposition in our previous pPolIII mice (Ostertag et al. 2002), as rare somatic events may have been missed by genotyping PCR. However, because unambiguous Mendelian inheritance of detected insertions in pPolIII mice was lost when pHSP70-2, a strong meiotic promoter (Dix et al. 1996), was substituted for pPolIII, a strong ubiquitously active promoter (Ostertag et al. 2002), we conclude that this substitution shifted the bulk of retrotransposition from gametogenesis to post-fertilization. Somatic retrotransposition was seen in F<sub>2</sub> offspring carrying transgenes integrated in different locations, suggesting that the shift in retrotransposition timing is position independent. Instead, it is likely caused by a lag between the meiotic activation of pHSP70-2 and the retrotransposition event. This hypothesis is supported by several somatic insertions inherited in the absence of transgene; they likely resulted from L1 transcript production during parental meiosis, followed by germ-line transmission of the L1 RNPs and retrotransposition in early development. Transgene-independent inheritance of insertions argues against retrotransposition in adult somatic tissues, although we cannot rule out late somatic retrotransposition for all insertions. More studies are needed to determine whether somatic retrotransposition of a human L1 in transgenic mice accurately reflects the ability of native LINES to retrotranspose in somatic cells of their normal hosts, or whether it is a result of impaired silencing of human L1s in mouse cells. Notably, differences in retrotransposition timing observed for pPolIII (Ostertag et al. 2002) and pHSP70-2 could be harnessed for germ-line and somatic random mutagenesis applications. A similar strategy was used successfully with a DNA transposon (Dupuy et al. 2001, 2005; Collier et al. 2005).

### New L1 retrotransposition events modestly favor intergenic regions

The preference for insertion into intergenic regions is consistent with the underrepresentation of endogenous L1s in genes (13%–17% gene hit rates) (Medstrand et al. 2002; Szak et al. 2002; Boissinot et al. 2004), but is discordant with higher gene hit rates in cultured cells (~50%) (Symer et al. 2002). Differences in gene hit rates between our study and that of Symer et al. (2002) can be explained by the use of different gene annotations and retrotransposition assays. Importantly, Symer et al. (2002) used a modified *neo* cassette (Freeman et al. 1994) to confer G418 resis-



**Figure 5.** (A) 5' junctions of insertions containing 26 and 47 extra 5' nucleotides. Their likely formation is depicted in B and C, respectively. TSD bases are in lower case. (Flanking host DNA) Solid black lines; (L1 RNA) gray line; (nascent cDNA strand) light-green line; (conceptual base pairing) dotted black lines. (B) A schematic of the "dual inversion" mechanism that created the 26 extra nucleotides between the flanking host DNA and the 238-bp 5'-truncated insert. Extra nucleotides are a product of two successive twin priming reactions (in A, highlighted in light and dark blue). Likely, the first twin priming reaction was initiated using a 1-bp MH, synthesizing the first inverted fragment of 18 bp (light blue). Then, the RT underwent a template jump, creating the second inverted fragment of 8 bp (dark blue). (C) Successive L1 RT template jumps created an insertion of 47 nt in a 368-bp de novo integrant. The 47-nt addition contains three overlapping regions of homology to the host DNA immediately 5' to the insertion site (in A, underlined orange, bold and yellow, and italicized and blue). When L1 RT reached the end of the RNA template at the end of a TPRT reaction, it likely template jumped from L1 RNA to the host DNA strand. Nascent cDNA was elongated by 11 bases (orange) copying a portion of host DNA. A second RT jump occurred, copying the same region of host DNA for 26 nt (yellow). One base was added (data not shown), and a third RT template jump occurred extending for seven more nucleotides (blue). Two bases were added (data not shown), and the structure was resolved, creating an addition of 47 nt. (D) Model for L1 integration. (Flanking host DNA) Solid black lines; (L1 RNA) gray line; (nascent cDNA strand) green line; (nascent 2nd strand) blue line; (homologous recombinational repair) orange crossed lines; (recombination products [no sequence changes expected]) stippled black/yellow and black/blue lines; (5' MH-guided base pairing) dotted red line; (TSD) target site duplication. See Discussion for detailed explanation.

tance on L1 integrants. In contrast, our assay does not depend on marker expression. In fact, some of observed preference for intergenic regions may be due to selective loss of cells containing L1 insertions into essential genes. While it is highly unlikely that

any heterozygous somatic mutation would result in animal lethality, somatic mosaicism creates a possibility for a selective growth disadvantage in the affected cells. The potential for selective cell loss is likely greater in our study due to the gene-trapping action of our retrotransposition cassette. Nevertheless, the tremendous sensitivity of TAIL-PCR probably compensates for most selection effects. If this is correct, the small bias against integration into genes may reflect possible differences in the distribution of L1 endonuclease target sites or potential interactions with transcriptional machinery or chromatin modifications. High-throughput studies of insertion sites of unbiased de novo integrants are needed to confirm our observations and will be particularly important if L1 is to be used in gene therapy and random mutagenesis.

#### L1 RT template jumping creates extra 5' nucleotides and 5'-end microhomologies

A highly informative 47-nt addition (Fig. 5A,C) demonstrated the ability of L1 RT to undergo template jumps from the L1 RNA onto the immediately upstream, single-stranded regions of host DNA. This behavior is consistent with the efficient DNA-dependent DNA polymerase activity of L1 RT (Mathias et al. 1991) and with the well-characterized ability of the RT of R2 to readily undergo template jumps from the 5' terminal end of the template RNA to one of the terminal nucleotides at the 3' end of another RNA or single-stranded DNA template (Bibillo and Eickbush 2002b, 2004). Analogous to our results, multiple successive jumps of R2 RT can occur (Bibillo and Eickbush 2002b). RT template jumping was proposed to facilitate attachment of the 5' end of R2 to the host DNA during integration (George et al. 1996; Burke et al. 1999; Eickbush et al. 2000; Bibillo and Eickbush 2002b), and our data suggest that template jumping may also be a feature of L1 retrotransposition. A similar mechanism was postulated to create chimeric L1 insertions (e.g., U6-L1) (Hayward et al. 1997; Buzdin et al. 2002, 2003; Gilbert et al. 2005).

Template jumping by the R2 RT is facilitated by the occasional addition by the RT of a few untemplated nucleotides to the 3' end of the donor's cDNA, one or more of which may be able to anneal at the 3' end of the acceptor template (Bibillo and Eickbush 2004). An analogous property of L1 RT may be reflected in the frequent

5'MHs observed in our noninverted 5'-truncated integrants, as well as in those found in the human and mouse genomes and in cultured cells (Symer et al. 2002; Gilbert et al. 2005; Martin et al. 2005; Zingler et al. 2005). Likely, as seen in the 26 and 47 extra nucleotide additions in our study, a template jump by the L1 RT can be facilitated by pairing of a few bases between the terminal end of the nascent cDNA and bases on the acceptor template, producing an apparent 5'MH in the final insertion. Such annealing could be possible if the terminal bases of the nascent cDNA became exposed for binding either due to the transient melting of a few 5' base pairs in the cDNA/RNA duplex, or due to the continued degradation of the RNA molecule within the duplex. This mechanism provides an alternative to previous models of 5'MH creation, which invoked MHs in 5' truncation and priming of the second-strand synthesis (Voliva et al. 1984; Symer et al. 2002; Martin et al. 2005; Zingler et al. 2005) and guidance of second-strand cleavage (Symer et al. 2002).

### Model of 5'-end attachment and L1 integration

Based on the characteristics of L1 insertions described here and elsewhere, we propose an L1 integration model that accounts for all common features of L1 retrotransposition (Fig. 5D). First, a host DNA strand is cleaved by the L1 EN, reverse transcription of the L1 RNA is initiated in a standard TPRT reaction, and is followed by the downstream cleavage of the second strand (Fig. 5D, 1). Upon reaching the end of L1 RNA, the L1 RT attempts template jumping from the 5' end of the RNA template onto the upstream overhang of host DNA (Fig. 5D, 2). This template jump may be facilitated by annealing of one or more nucleotides at the 3' end of the nascent cDNA with those in the overhang of host DNA, or by the addition of several untemplated bases (Bibillo and Eickbush 2004). The former will result in an apparent 5' MH, the latter will create unexplained nucleotides at the element's 5' boundary. After the jump, L1 RT likely continues copying the host DNA region, adding a stretch of DNA complementary to the host's top strand to the L1 cDNA (Fig. 5D, 3). Depending on the length of added homologous DNA, the nascent strand can be joined to the host's bottom strand by simple nick ligation or by the host's homologous recombinational repair machinery (for review, see Thompson and Schild 2001) (Fig. 5D, 4 and 5). Homologous recombinational repair is thought to facilitate R2 5'-end integration (Fujimoto et al. 2004) and was reported for L1s integrating into pre-existing older L1 elements, resulting in hybrid or toggled insertions (Gilbert et al. 2002; Symer et al. 2002). Unlike this special case of toggled insertions, where host DNA had prior homology to the retrotransposon allowing for detectable sequence changes after recombination, a more general case will involve recombination between identical stretches of DNA, and no sequence changes will be produced. Recombination-mediated resolutions were also proposed to mediate the creation of large L1-mediated deletions (Gilbert et al. 2005; Han et al. 2005). After the attachment of the 5' end of L1 cDNA to the host's bottom strand, a second molecule of L1 RT likely completes the synthesis of the second strand (Fig. 5D, 6), either displacing RNA from the DNA:RNA duplex during the reaction, a property observed for R2 RT with RNA:RNA duplexes *in vitro* (Bibillo and Eickbush 2002a), or relying on RNA degradation by host enzymes. Finally, the nick is repaired by simple ligation or by the host's homologous recombination machinery (Fig. 5D, 7), creating a typical L1 insertion (Fig. 5D, 8).

One consequence of this model is that TSDs flanking the

insertion may be shorter than would have been predicted by the true second-strand cleavage site. This is because template jumping by the L1 RT may occur to internal regions of the host DNA overhang, as seen in our data and described for R2 RT. Such internal jumps will produce a shorter TSD, implying a smaller distance between the two cleavage sites and contributing to observed variability in TSD lengths.

Another implication is the likely mechanism of 5' truncations, previously hypothesized to result from MH-mediated initiation of second-strand synthesis (Voliva et al. 1984), poor processivity of the L1 RT, or degradation of L1 RNA by cellular RNase H (Ostertag and Kazazian Jr. 2001a). Biochemical studies of the R2 RT demonstrate that template jumps readily occur from the terminal 5' nucleotide of the R2 RNA, but are extremely inefficient from its internal positions (Bibillo and Eickbush 2004). If L1 RT behaves in a similar manner, its template jumps must be occurring from the RNA's 5' terminal nucleotide, implying prior 5'-end degradation of ~70% of L1 transcripts. The bimodal distribution of insertion lengths may thus reflect two different populations of L1 RNA in the host cell—the majority, which are quickly degraded by cellular RNases and give rise to the 5'-truncated L1 insertions, and the minority, which evade degradation and produce FL insertions. Altered susceptibility to degradation may be caused by the protective action of ORF1 protein or of the m<sup>7</sup>G-cap, loss of which may allow for the rapid degradation of the transcript's 5' end. While decapping of most cellular mRNAs requires prior shortening of the poly(A) tail (for review, see Collier and Parker 2004), we and others observed no difference between poly(A) lengths of 5'-truncated and FL integrants. This may imply that shortened L1 poly(A) tails are extended by the RT during TPRT initiation (Luan and Eickbush 1995; Eickbush et al. 2000; Cost et al. 2002) or are subject to additional polyadenylation after re-entry of L1 RNA into the nucleus.

A lower 5'MH frequency in FL elements (Zingler et al. 2005) may be explained by an altered position of the template jump by the L1 RT when the donor RNA template is full length. Biochemical studies of R2 demonstrated that m<sup>7</sup>G-capping of the donor RNA template reduced template jumps to internal locations on the acceptor template's 3' end (Bibillo and Eickbush 2004), with most jumps occurring to the 3'-most terminal nucleotide. Analogous behavior of the L1 RT will result in less 5'MH formation for the likely m<sup>7</sup>G-capped FL transcripts, and, together with a reduced probability of a random match between the G-rich L1 5' end and an AT-rich target site, will explain the lower 5'MH frequencies in FL insertions.

Inversions, characteristic of L1 integration, can be explained by a special case of the model, in which second strand synthesis is initiated prior to the completion of the first strand. Premature initiation will give a twin priming reaction, in which the host's top strand overhang is used as a primer and an upstream portion of L1 RNA—as a template for the inverted fragment; this priming in turn creates MHs commonly seen at the 5' boundary of inverted integrants (Ostertag and Kazazian Jr. 2001b). An upstream twin priming reaction will prevent the first RT from reaching the 5' end of the template RNA, effectively eliminating its template jumping ability. As the twin priming RT reaches the terminal nucleotide of the RNA template, it will attempt template jumping onto the single-stranded overhang formed by the direct segment of L1 cDNA. This situation is identical to the standard L1 integration (Fig. 5D, 2), with the L1 RT now jumping to the terminal or an internal nucleotide of the protruding cDNA end, often facilitated by one or more complementary bases. This, in

turn, creates the MHs commonly seen at L1 inversion points (Ostertag and Kazazian Jr. 2001b).

After the attachment of the 5' end of L1 cDNA, synthesis of the second strand is probably completed by the second RT molecule. In cases where first-strand attachment was initiated with an RT jumping to an internal nucleotide of the host DNA overhang, a host 3'→5' exonuclease activity would likely be needed to trim the short protruding 3' end of the host DNA prior to initiation of second-strand synthesis. Alternatively, L1 ORF2 may itself possess 3' exonuclease activity, analogous to the 3' exonuclease activity of a related apurinic/apyrimidinic endonuclease (APE) (Chou and Cheng 2002). Perhaps this 3' exonuclease activity is related to the second nuclease activity of L1 ORF2 protein observed in vitro (Cost et al. 2002). A 3' exonuclease activity was also observed for the sequence-specific R2 EN (Bibillo and Eickbush 2002b).

Clearly, much future biochemical work is needed to provide direct evidence for this model of L1 5'-end integration.

## Methods

Transgenic mice, PCR primers, genotyping PCRs, Southern blotting, statistical analysis, and analysis of microhomologies are described in the Supplemental material.

### Transgene construction

The markerless (mless) retrotransposition cassette was constructed by flanking the human  $\gamma$ -globin intron, obtained from the EGFP cassette (Ostertag et al. 2000) by splice acceptors from the human *BCL2* gene, obtained from the pRET.L<sub>L</sub> vector (Ishida and Leder 1999), a kind gift of Dr. P. Leder. Mless was cloned into the pJCC5-L1<sub>LRE3</sub> (Ostertag et al. 2000; Brouha et al. 2002) to create pBS-L1<sub>LRE3</sub>-mless. The L1<sub>LRE3</sub>-mless was then cloned into pCEP-Hsp70-2, a vector similar to pCEP4 (Invitrogen), but containing the mouse heat-shock protein 70-2 promoter (pHSP70-2, cloned from the mouse genome) (Dix et al. 1996) instead of the CMV promoter to create the transgene used in this study. Detailed cloning strategy is available upon request.

### Characterization of insertions and transgenes

Genomic locations of L1<sub>LRE3</sub>-mless insertions and transgenes were identified by TAIL-PCR essentially as described (Liu and Whittier 1995; Liu et al. 1995). Long specific primers were positioned in the SV40 poly(A)-signal sequence, allowing 3' flank amplification. A third reamplification reaction was added, and all three reamplification reactions were carried out as TAIL cycles. The products of TAIL-PCR reactions were resolved on a 1% agarose gel and visualized by EtBr staining. Single bands were gel-purified with GeneClean Spin Kit (Q-BIOgene), and either sequenced directly using AD primers (UPenn DNA Sequencing Facility) or cloned using TopoTA Cloning Kit for Sequencing (Invitrogen) and then sequenced using M13Forward and/or M13Reverse standard primers.

To characterize the 5' ends, multiple series of nested PCRs were performed for each insertion using nested genomic primers positioned between 87 and ~1500 bp upstream of the insertion location, and several sets of nested primers complementary to the insertion sequence. Individual PCRs were carried out according to the manufacturers' protocols. To assay shorter insertions, PCRs were carried out with Taq DNA Polymerase (Promega) using 160 ng of genomic DNA in a 20- $\mu$ l reaction volume and 3–3.5-min extension times. To characterize longer insertions, high-processivity Phusion Polymerase (MJResearch) was used with

3–5-min extension times. The PCR products from the primary reactions were diluted 1:100 and used in the nested PCRs. The products were gel-purified with Qiaquick Gel Extraction Kit (Qia- gen), cloned using TopoTA Cloning Kit for Sequencing (Invitrogen) (Taq PCR products) or Zero Blunt Topo Kit (Invitrogen) (Phusion PCR products), and sequenced using M13Forward and/or M13Reverse standard primers.

### Informatics

Genomic locations of insertions were found using UCSC Genome Browser (<http://genome.ucsc.edu/>) (Kent et al. 2002; Karolchik et al. 2003), NCBI BLAST server, or Celera Discovery System by searching for the 3' flanks obtained by TAIL-PCR. For inserts with unique genomic coordinates (48/51), neighboring genomic structures were surveyed against the mm6 assembly of the mouse genome in the UCSC Table Browser (Karolchik et al. 2004) using custom Perl scripts (available upon request). For genome composition calculations, nonoverlapping sequences belonging to a genomic feature were added for each chromosome using UCSC Table Browser annotation. These were weighted 2, 2, 3/2, and 1/2 (autosomes, Un\_random, X, and Y chromosomes, respectively) and added to obtain the total content for a genomic feature; genome proportion was obtained by dividing the total content by mouse mm6 genome size (autosomes, Un\_random, X, and Y chromosomes weighted 2, 2, 3/2, and 1/2, respectively). Insert structures were analyzed by comparing the filled and empty site sequences, and by comparing sequences of new and expected FL L1-mless integrants.

## Acknowledgments

We thank Dr. Robert F. Whittier for helpful suggestions on optimizing TAIL-PCR, Melissa Vetter for technical assistance, the Transgenic and Sequencing Cores of the University of Pennsylvania for generation of transgenic mouse lines and for DNA sequencing services, and members of the Kazazian lab, especially Maria del Carmen Seleme, for critical discussions, and John L. Goodier for critical review of the manuscript. D.V.B. was supported by the predoctoral Genetics Training Grant, H.H.K. and E.M.O. were supported by grants from the US National Institutes of Health.

## References

- Bibillo, A. and Eickbush, T.H. 2002a. High processivity of the reverse transcriptase from a non-long terminal repeat retrotransposon. *J. Biol. Chem.* **277**: 34836–34845.
- . 2002b. The reverse transcriptase of the R2 non-LTR retrotransposon: Continuous synthesis of cDNA on non-continuous RNA templates. *J. Mol. Biol.* **316**: 459–473.
- . 2004. End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J. Biol. Chem.* **279**: 14945–14953.
- Boissinot, S., Chevret, P., and Furano, A.V. 2000. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.* **17**: 915–928.
- Boissinot, S., Entezam, A., and Furano, A.V. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol. Biol. Evol.* **18**: 926–935.
- Boissinot, S., Entezam, A., Young, L., Munson, P.J., and Furano, A.V. 2004. The insertional history of an active family of L1 retrotransposons in humans. *Genome Res.* **14**: 1221–1231.
- Bourc'his, D. and Bestor, T.H. 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**: 96–99.
- Brouha, B., Meischl, C., Ostertag, E., de Boer, M., Zhang, Y., Neijens, H., Roos, D., and Kazazian Jr., H.H. 2002. Evidence consistent with human L1 retrotransposition in maternal meiosis I. *Am. J. Hum. Genet.* **71**: 327–336.
- Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H.,

- Moran, J.V., and Kazazian Jr., H.H. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci.* **100**: 5280–5285.
- Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**: 78–94.
- Burke, W.D., Malik, H.S., Jones, J.P., and Eickbush, T.H. 1999. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol. Biol. Evol.* **16**: 502–511.
- Bushman, F., Lewinski, M., Ciuffi, A., Barr, S., Leipzig, J., Hannenhalli, S., and Hoffmann, C. 2005. Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* **3**: 848–858.
- Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y., and Sverdlov, E. 2002. A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of I1. *Genomics* **80**: 402–406.
- Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T., and Sverdlov, E. 2003. The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. *Nucleic Acids Res.* **31**: 4385–4390.
- Chen, J.M., Stenson, P.D., Cooper, D.N., and Ferec, C. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum. Genet.* **117**: 411–427.
- Chou, K.M. and Cheng, Y.C. 2002. An exonucleolytic activity of human apurinic/aprimidinic endonuclease on 3' mispaired DNA. *Nature* **415**: 655–659.
- Christensen, S.M. and Eickbush, T.H. 2005. R2 target-primed reverse transcription: Ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol. Cell. Biol.* **25**: 6617–6628.
- Coller, J. and Parker, R. 2004. Eukaryotic mRNA decapping. *Annu. Rev. Biochem.* **73**: 861–890.
- Collier, L.S., Carlson, C.M., Ravimohan, S., Dupuy, A.J., and Largaespada, D.A. 2005. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**: 272–276.
- Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **21**: 5899–5910.
- Dix, D.J., Rosario-Herrle, M., Gotoh, H., Mori, C., Goulding, E.H., Barrett, C.V., and Eddy, E.M. 1996. Developmentally regulated expression of Hsp70-2 and a Hsp70-2/lacZ transgene during spermatogenesis. *Dev. Biol.* **174**: 310–321.
- Dupuy, A.J., Fritz, S., and Largaespada, D.A. 2001. Transposition and gene disruption in the male germline of the mouse. *Genesis* **30**: 82–88.
- Dupuy, A.J., Akagi, K., Largaespada, D.A., Copeland, N.G., and Jenkins, N.A. 2005. Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature* **436**: 221–226.
- Eickbush, D.G., Luan, D.D., and Eickbush, T.H. 2000. Integration of Bombyx mori R2 sequences into the 28S ribosomal RNA genes of *Drosophila melanogaster*. *Mol. Cell. Biol.* **20**: 213–223.
- Feng, Q., Moran, J.V., Kazazian Jr., H.H., and Boeke, J.D. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905–916.
- Freeman, J.D., Goodchild, N.L., and Mager, D.L. 1994. A modified indicator gene for selection of retrotransposition events in mammalian cells. *Biotechniques* **17**: 47–52.
- Fujimoto, H., Hirukawa, Y., Tani, H., Matsuura, Y., Hashido, K., Tsuchida, K., Takada, N., Kobayashi, M., and Maekawa, H. 2004. Integration of the 5' end of the retrotransposon, R2Bm, can be complemented by homologous recombination. *Nucleic Acids Res.* **32**: 1555–1565.
- George, J.A., Burke, W.D., and Eickbush, T.H. 1996. Analysis of the 5' junctions of R2 insertions with the 28S gene: Implications for non-LTR retrotransposition. *Genetics* **142**: 853–863.
- Gilbert, N., Lutz-Prigge, S., and Moran, J.V. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**: 315–325.
- Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol. Cell. Biol.* **25**: 7780–7795.
- 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.
- Han, K., Sen, S.K., Wang, J., Callinan, P.A., Lee, J., Cordaux, R., Liang, P., and Batzer, M.A. 2005. Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res.* **33**: 4040–4052.
- Hayward, B.E., Zavanelli, M., and Furano, A.V. 1997. Recombination creates novel L1 (LINE-1) elements in *Rattus norvegicus*. *Genetics* **146**: 641–654.
- Ho, H.J., Ray, D.A., Salem, A.H., Myers, J.S., and Batzer, M.A. 2005. Straightening out the LINEs: LINE-1 orthologous loci. *Genomics* **85**: 201–207.
- Hohjoh, H. and Singer, M.F. 1996. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J.* **15**: 630–639.
- . 1997. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO J.* **16**: 6034–6043.
- Ishida, Y. and Leder, P. 1999. RET: A poly A-trap retrovirus vector for reversible disruption and expression monitoring of genes in living cells. *Nucleic Acids Res.* **27**: e35.
- Jurka, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retrotransposons. *Proc. Natl. Acad. Sci.* **94**: 1872–1877.
- Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**: 51–54.
- Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**: D493–D496.
- Kazazian Jr., H.H. 2004. Mobile elements: Drivers of genome evolution. *Science* **303**: 1626–1632.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC. *Genome Res.* **12**: 996–1006.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- Liu, Y.G. and Whittier, R.F. 1995. Thermal asymmetric interlaced PCR: Automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. *Genomics* **25**: 674–681.
- Liu, Y.G., Mitsukawa, N., Oosumi, T., and Whittier, R.F. 1995. Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J.* **8**: 457–463.
- 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.
- Luan, D.D., Korman, M.H., Jakubczak, J.L., and Eickbush, T.H. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell* **72**: 595–605.
- Martin, S.L. and Bushman, F.D. 2001. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol. Cell. Biol.* **21**: 467–475.
- Martin, S.L., Li, W.L., Furano, A.V., and Boissinot, S. 2005. The structures of mouse and human L1 elements reflect their insertion mechanism. *Cytogenet. Genome Res.* **110**: 223–228.
- Mathias, S.L., Scott, A.F., Kazazian Jr., H.H., Boeke, J.D., and Gabriel, A. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254**: 1808–1810.
- Medstrand, P., van de Lagemaat, L.N., and Mager, D.L. 2002. Retroelement distributions in the human genome: Variations associated with age and proximity to genes. *Genome Res.* **12**: 1483–1495.
- Miki, Y., Nishishio, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., and Nakamura, Y. 1992. Disruption of the APC gene by a retrotranspositional insertion of L1 sequence in a colon cancer. *Cancer Res.* **52**: 643–645.
- 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.
- Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., and Moran, J.V. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat. Genet.* **31**: 159–165.
- Muotri, A.R., Chu, V.T., Marchetto, M.C., Deng, W., Moran, J.V., and Gage, F.H. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**: 903–910.
- Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V., et al. 2002. A comprehensive analysis of recently integrated human Ta L1 elements. *Am. J. Hum. Genet.* **71**: 312–326.
- Ostertag, E.M. and Kazazian Jr., H.H. 2001a. Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35**: 501–538.
- . 2001b. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**: 2059–2065.

- Ostertag, E.M., Luning Prak, E.T., DeBerardinis, R.J., Moran, J.V., and Kazazian Jr., H.H. 2000. Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic Acids Res.* **28**: 1418–1423.
- Ostertag, E.M., DeBerardinis, R.J., Goodier, J.L., Zhang, Y., Yang, N., Gerton, G.L., and Kazazian Jr., H.H. 2002. A mouse model of human L1 retrotransposition. *Nat. Genet.* **32**: 655–660.
- Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. 2001. Genomic characterization of recent human LINE-1 insertions: Evidence supporting random insertion. *Genome Res.* **11**: 2050–2058.
- Pavlicek, A., Paces, J., Zika, R., and Hejnar, J. 2002. Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: Implications for retrotransposition and pseudogene detection. *Gene* **300**: 189–194.
- Prak, E.T., Dodson, A.W., Farkash, E.A., and Kazazian Jr., H.H. 2003. Tracking an embryonic L1 retrotransposition event. *Proc. Natl. Acad. Sci.* **100**: 1832–1837.
- Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2005. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**: D501–D504.
- Rogic, S., Mackworth, A.K., and Ouellette, F.B. 2001. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11**: 817–832.
- Roth, D.B., Porter, T.N., and Wilson, J.H. 1985. Mechanisms of nonhomologous recombination in mammalian cells. *Mol. Cell. Biol.* **5**: 2599–2607.
- Salem, A.H., Myers, J.S., Otieno, A.C., Watkins, W.S., Jorde, L.B., and Batzer, M.A. 2003. LINE-1 pre-Ta elements in the human genome. *J. Mol. Biol.* **326**: 1127–1146.
- Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F., et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. Sci.* **99**: 16899–16903.
- Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. 2002. Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* **110**: 327–338.
- Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol.* **3**: research0052.
- Thompson, L.H. and Schild, D. 2001. Homologous recombinational repair of DNA ensures mammalian chromosome stability. *Mutat. Res.* **477**: 131–153.
- Vincent, B.J., Myers, J.S., Ho, H.J., Kilroy, G.E., Walker, J.A., Watkins, W.S., Jorde, L.B., and Batzer, M.A. 2003. Following the LINEs: An analysis of primate genomic variation at human-specific LINE-1 insertion sites. *Mol. Biol. Evol.* **20**: 1338–1348.
- Voliva, C.F., Martin, S.L., Hutchison III, C.A., and Edgell, M.H. 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. *J. Mol. Biol.* **178**: 795–813.
- Yant, S.R., Wu, X., Huang, Y., Garrison, B., Burgess, S.M., and Kay, M.A. 2005. High-resolution genome-wide mapping of transposon integration in mammals. *Mol. Cell. Biol.* **25**: 2085–2094.
- Zingler, N., Willhoeft, U., Brose, H.P., Schoder, V., Jahns, T., Hanschmann, K.M., Morrish, T.A., Lower, J., and Schumann, G.G. 2005. Analysis of 5' junctions of human LINE-1 and *Alu* retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res.* **15**: 780–789.

Received August 17, 2005; accepted in revised form November 9, 2005.