

# Length distribution of long interspersed nuclear elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection

Adam Pavlíček<sup>a</sup>, Jan Pačes<sup>a,b</sup>, Radek Zíka<sup>a,b</sup>, Jiří Hejnar<sup>a,\*</sup>

<sup>a</sup>*Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Flemingovo nám. 2, Prague 6, CZ-16637, Czech Republic*

<sup>b</sup>*Center for Integrated Genomics, Flemingovo nám. 2, Prague 6, CZ-16637, Czech Republic*

Received 25 June 2002; received in revised form 20 August 2002; accepted 23 September 2002

## Abstract

Deciphering the human genome includes reliable identification and structural characterization of individual retrotransposon elements. The most active group of autonomous transposable elements, the long interspersed nuclear elements (LINE), transpose themselves as well as other RNAs, including those of human endogenous retroviruses (HERV). During this transposition, however, the LINE-encoded reverse transcriptase (RT) often abortively dissociates from the RNA template, leaving a prematurely terminated, 5' truncated copy. We have analyzed the length distributions of LINEs and of processed pseudogenes derived from HERV-W. As expected, we have found that the majority of 5' truncated LINEs and HERV-W processed pseudogenes show a prevalence of very short elements terminated close to the 3' end. On the other hand, the number of complete elements is far above the expectation. The characteristic distribution in both cases indicates two important conclusions: (i) dissociation of LINE RT from the template cannot be fully explained by low processivity of RT modelled as a stochastic, Poisson-type process. (ii) Currently cited numbers of pseudogenes within the human genome are underestimated, since a large percentage of pseudogenes are terminated in the 3' untranslated region and remain undetectable in translated homology searches of protein databases against the human genome. © 2002 Elsevier Science B.V. All rights reserved.

**Keywords:** Alu; L1; Target-primed reverse transcription

## 1. Introduction

Long interspersed nuclear elements (LINE, LINE1 or L1), retrotransposons lacking long terminal repeats (LTR) and containing polyA tails, are the most active autonomous transposable elements in the human genome (Kazazian and Moran, 1998; Ostertag and Kazazian, 2001a). They are estimated to be present there in more than 500,000 copies, comprising 17% of the genome (Smit, 1996, 1999; IHGSC, 2001). However, due to frequent mutations, only 30–60 LINEs per haploid genome remain active and transpose along the genome (Sassaman et al., 1997; Kazazian, 1999; Ostertag and Kazazian, 2001a).

LINEs encode two proteins, one of which, ORF2 (second

open reading frame), has both endonuclease and reverse transcriptase (RT) activity (Mathias et al., 1991; Feng et al., 1996; Malik et al., 1999), and they are thought to integrate by a coupled reverse transcription/integration process called target-primed reverse transcription (TPRT; Luan et al., 1993). During TPRT (reviewed in Ostertag and Kazazian, 2001a), the endonuclease activity cleaves one strand of the DNA at its target site, producing a free 3'-hydroxyl at the DNA nick. After the retrotransposon RNA anneals at the break, the RT activity uses this RNA as a template and the 3'-hydroxyl as a primer for reverse transcription. The remaining steps of TPRT include the cleavage of the second DNA strand, integration of the cDNA, and completion of DNA synthesis. Upon completion of TPRT, a copy of the original retrotransposon is integrated in a new genomic location flanked by target site duplications. The 5' ends of most LINEs in the genome are either truncated or both inverted and truncated (Voliva et al., 1983; Smit, 1999; Boissinot et al., 2000; IHGSC, 2001; Ostertag and Kazazian, 2001b). Truncations have been hypothesized to occur because of a

*Abbreviations:* HERV, human endogenous retrovirus; LINE, long interspersed nuclear elements; RT, reverse transcriptase; TPRT, target-primed reverse transcription; UTR, untranslated region.

\* Corresponding author. Tel.: +420-2-2018-3443; fax: +420-2-2431-0955.

*E-mail address:* hejnar@img.cas.cz (J. Hejnar).

low processivity of the LINE reverse transcriptase or because of activity of cellular RNase H (see [Ostertag and Kazazian, 2001a](#) and refs. therein). If the RT dissociates from the RNA template before completion of the reverse transcription, the resulting insertion becomes truncated at the 5' end.

Similar 5' truncations have been described for processed pseudogenes. They were first defined as pseudogenes structurally collinear with their parental gene messenger RNA (mRNA) lacking promoters, introns, and, in general, without protein-coding capacity due to mutations and frequent stop codons. Their mRNA-derived structure, polyA tails at the 3' end and the presence of direct repeats of variable (5–15 bp) length suggested that their formation required RT, and these pseudogenes were termed processed pseudogenes ([Vanin, 1985](#); [Weiner et al., 1986](#); [Weiner, 2002](#)). Recently, *in vitro* experiments demonstrated the creation of reporter gene copies by the LINE machinery with all hallmarks of processed pseudogenes ([Esnault et al., 2000](#); [Wei et al., 2001](#)).

Retrotransposed repetitive elements like LINES or processed pseudogenes of human endogenous retroviruses (HERV) have particular advantages for studies of properties of pseudogenes. Since the mRNA structure is the same (LINES) or part of the DNA consensus sequences (HERVs), we can easily identify these sequences using standard tools for repeat detection. Given the sensitivity of RepeatMasker (Smit and Green RepeatMasker at <http://repeatmasker.genome.washington.edu>), it is possible to identify even very small fragments that could be overlooked by widely used fast heuristic searches. In this work, we analyzed the length distributions of LINES and processed pseudogenes of HERVs detected by a similarity search to full-length mRNA. Using this approach, we confirmed that the majority of LINES and processed pseudogenes are 5' truncated. We show that the 5' truncated copies of the youngest LINE family L1HS (Ta subset) display a power-law-like length distribution (except for the proportion of full-size copies), which changes for older families, probably due to post-integration processes. In addition, the number of complete elements in young LINE families is far above the expectation for random terminations. These data are at odds with the assumption of low processivity of RT randomly dissociating from an RNA template. Moreover, these results suggest that cited numbers of retro-transpositions of cellular mRNAs within the human genome, currently identified by the homology to protein coding sequences, are by far underestimated because the short pseudogenes, 5' truncated inside the 3' untranslated region (3'UTR), are neglected.

## 2. Materials and methods

### 2.1. Identification and length distribution of LINES in the human genome

The RepeatMasker program (Smit and Green Repeat-

Masker at <http://repeatmasker.genome.washington.edu>) with standard settings and Repbase Update libraries version 6.3 ([http://www.girinst.org/Repbase\\_Update.html](http://www.girinst.org/Repbase_Update.html); [Smit, 1996](#); [Jurka, 1998, 2000](#)) were used to identify LINES in the GoldenPath assembly (August 6 2001 Human Genome Working Draft at <http://genome.ucsc.edu/>). LINE fragments annotated by RepeatMasker as parts of one element were joined together as one element. For the size calculations, we only used elements with intact 3' ends, arbitrarily circumscribed as elements terminated at least 20 bp from the 3' end of the family consensus, to define the allowed uncertainty in the RepeatMasker detection. In contrast, elements terminated less than 20 bp from the 3' end of the family consensus are termed 3' truncated elements and were not included in this analysis. Insertions of other repeats into LINES were excluded from the length calculations.

### 2.2. Length distribution of HERV-W processed pseudogenes

We used the dataset of HERV-W (HERV17) pseudogenes obtained in our previous work ([Pavlíček et al., 2002](#)) from our HERV database (<http://herv.img.cas.cz>; [Pačes et al., 2002](#)). For 107 5' truncated elements of HERV-W with an intact 3' end, defined again as elements terminated at least 20 bp from the 3' end of the HERV-W mRNA consensus, we calculated the length after exclusion of insertions; in addition, 46 complete, untruncated pseudogenes (both 5' and 3' within 20 bp from the mRNA termini) were plotted separately.

## 3. Results

### 3.1. Length distribution of LINES in the human genome

We identified 557,011 independent LINES in the GoldenPath assembly of the human genome. The particular advantage of the RepeatMasker detection is that the program annotates fragments as parts of one element, irrespective of fragment orientations; thus, even the frequently found LINES with 5' terminal insertions ([Ostertag and Kazazian, 2001b](#)) are treated as one element. About 594 elements (0.108%) contain multiple inversions (up to fourfold inversions). For a few elements we detected unusual inversions of the terminal 3' sequence. For the length distribution we only took into account elements with intact 3' ends, after removal of insertions of other elements. LINES with terminal gaps over 20 bp were defined as 3' truncated and were excluded from the length calculation. This exclusion involved 317,516 (mainly old) LINES, or 57% (in majority they were both 5' and 3' truncated).

In the next step, we calculated the length distribution for selected LINE families. [Fig. 1](#) shows the length distribution of the youngest human-specific LINE family L1HS ([Smit et al., 1995](#)) or Ta, corresponding to subsets of Ta-0 and Ta-1 subfamilies ([Boissinot et al., 2000](#)). A power-law fit (log

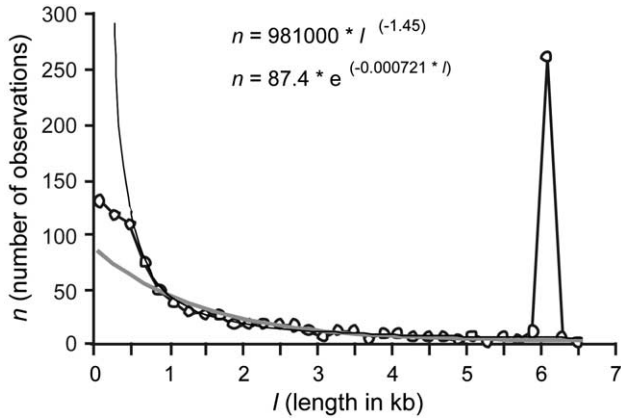


Fig. 1. Length distribution of the LINE L1HS family. Absolute numbers of L1HS elements within the 200 bp intervals are depicted in the plot. All 3' truncated copies and copies with insertions were excluded (see Section 2). The power-law (black) and exponential (gray) fit functions (see Section 3) are shown. All elements longer than 6 kbp were excluded from both fits.

of number  $n$  versus log of length  $l$  in bp) yielded  $n = 981,000 l^{-1.45}$ ; the power is  $-1.45 \pm 0.06$  (Jolicoeur error estimate, 5% level). The corresponding exponential fit function is also shown,  $n = 87.4 \exp(-0.000721 l)$ . For both fits, we excluded elements longer than 6 kb, since they contain full-length LINE copies. We also excluded the first two length intervals (points), because very short fragments of LINES are less likely to be detected by RepeatMasker and can be underrepresented: the threshold criterion, the Smith-Waterman score, is positively correlated with the length of the alignment. The length distribution of young L1HS is similar to a power-law distribution rather than to an exponential.

Fig. 2 shows a comparison of the length distributions between several LINE families. The power-law distribution of L1HS gradually changes for older families L1PA2,

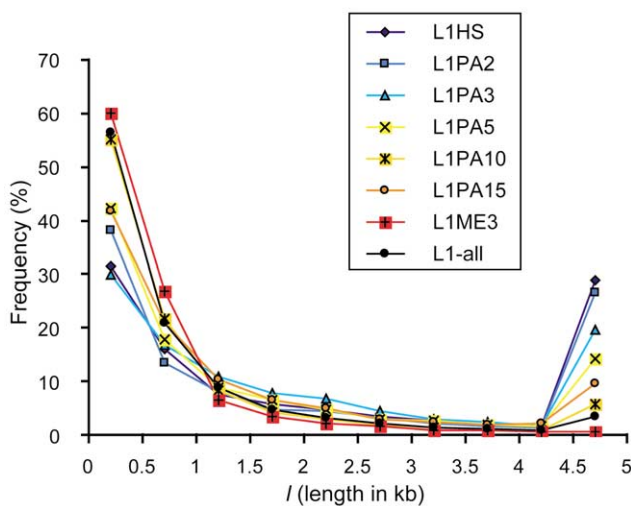


Fig. 2. Comparison of length distributions between several LINE families. Relative frequencies of the length distribution for several LINE families; 500 bp intervals were used. All 3' truncated copies and insertions were excluded (see Section 2). The general length profile of all LINE copies with an intact 3' terminus is shown (black curve).

L1PA3, and L1PA5, whose distributions became more exponential-like (data not shown). The proportion of 3' truncated and therefore excluded elements was 5.6% (59/1053) for L1HS, 3.4% (138/4065) for L1PA2, 4.0% (351/8726) for L1PA3, 4.2% (392/9295) for L1PA5, 8.2% (395/4807) for L1PA10, 10.6% (591/5556) for L1PA15 and 45.9% (2010/4377) for the L1ME3 family.

The percentage of the potential full-length LINES decreases with the age of a given family. The human-specific L1HS family (Smit et al., 1995; Boissinot et al., 2001) contains 28.7% full-length elements, in good agreement with estimated 35% for recent LINE insertions (Boissinot et al., 2000, 2001), whereas this proportion drops to less than 0.3% for the old L1ME3 family. As noted by Boissinot et al. (2001), chromosome Y has more complete LINE1 elements; in agreement, we detected 161/513 (31.4%) of full-sized Y-linked elements from L1HS, L1PA2–5 families. For all LINE elements, the average percentage of elements over 5 kbp is 3.2%.

### 3.2. Length distribution of HERV-W processed pseudogenes

We used our dataset of HERV-W pseudogenes in the human genome (Pavlíček et al., 2002) to calculate the lengths of these pseudogenes. Forty-six complete and 107 5' but not 3' truncated pseudogenes were separately included in the calculation. One pseudogene displays 5' inversion, similar to 5' inversions frequent in LINES (Ostertag and Kazazian, 2001b), another one contains 3' inversion rarely found in LINES (see above). Thirteen percent (23/176) of HERV-W pseudogenes were truncated at the 3' terminus of HERV-W mRNA, and were excluded from our analysis. All insertions were removed from HERV-W sequences. Fig. 3 shows the length distribution of the resulting selection of HERV-W processed pseudogenes. The power-law fit

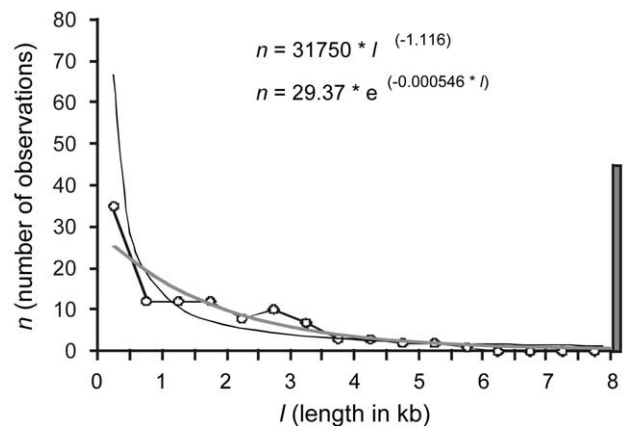


Fig. 3. Length distribution of HERV-W processed pseudogenes. Absolute numbers of HERV-W processed pseudogenes within 500 bp intervals are depicted in the plot. Data on HERV-W processed pseudogenes are from Pavlíček et al. (2002). All 3' truncated copies and insertions were excluded (see Section 2). Power-law (black) and exponential (gray) fit functions are shown. For both fits, we excluded all full-length elements (gray column).

function is  $n = 31,750 l^{-1.116}$ . The exponential fit function is also shown,  $n = 29.37 \exp(-0.000546 l)$ .

## 4. Discussion

### 4.1. Length distribution of LINEs and HERV-W processed pseudogenes

In the present work, we analyzed the length distribution of several LINE families and processed pseudogenes derived from endogenous retroviruses (Figs. 1–3). As expected (IHGSC, 2001; Boissinot et al., 2001), the majority of elements are very short, and very few elements range between 2 and 5 kbp. Over 5 kbp, there is a peak of potentially full-length elements, representing successfully terminated reverse transcripts. If the premature termination of TPRT were a Poisson-type process without memory, i.e. with equal probability of abortive RT dissociation at any nucleotide of reversely transcribed mRNA, then the length sizes should follow an exponential distribution. Instead, the length distribution of the youngest family L1HS (Ta) is rather of a power-law character, indicating a different process than random termination of TPRT (Fig. 1). Indeed, (1) the number of very short elements is clearly above the exponential expectation, i.e. the probability of termination is higher for early steps of integration compared to the random model; and (2) the number of full-length elements is far above the expectation. From the exponential fit we calculated the expected number of full-length elements (element longer than 6 kb) to be less than ten. However, 260 elements over 6 kb were observed; in other words, 26 times more than expected according to the exponential fit.

Given the purifying selection of LINEs that can change the length distribution (Boissinot et al., 2001), we first concentrated on the youngest family L1HS (Ta). The length distribution of 5'-truncated copies of this family has a power-law-like character (Fig. 1). The power-law fit decreases with the age of LINEs; the second best hit was found for L1PA2, followed by L1PA3 and L1PA5 families, where the distribution became more like a (truncated) exponential, but again, the peak of complete elements is far above expectations. For older LINEs, due to the degradation, the peak of full-length copies disappeared and the distribution became monotonously decreasing (Fig. 2).

The degradation of complete copies and other changes in the distribution are very likely results of multiple post-integration processes, including recombination, deletion and selection. It is well known that some, particularly full-length autosomal LINE insertions, are under negative selection (Boissinot et al., 2001). LINEs can have some beneficial function in X-chromosome inactivation (Bailey et al., 2000; however, see Chureau et al., 2002). LINEs in introns are negatively selected if they are in the same orientation as the gene (Smit et al., 1995; Smit, 1999). The LINE size also depends on the composition of flanking

sequences, e.g. LINEs from GC-rich regions tend to be 2–2.5 times shorter than LINEs in GC-poor regions (Smit, 1999). All these phenomena can change the original size distribution.

The rate of gradual shortening in time can be considered by comparing the proportions of 3' truncated elements in different LINE families. Whereas young families contain only about 5% of 3' truncated LINE elements, the proportion of 3' truncated copies reaches 50–60% for older families. In fact, the contribution of postintegration truncations (i.e. after TPRT and integration) could be roughly estimated from these differences. Also, very recently, endonuclease-independent LINE insertions through double-stranded break DNA repair were reported, predominantly truncated at 3' ends (Morrish et al., 2002); nevertheless, they were very likely rare in human evolution, since only a small fraction of young elements is truncated at the 3' end (see above).

The length distribution of HERV-W processed pseudogenes is similar to the distribution of LINEs. Again, there is a high peak of complete elements. However, contrary to LINEs that, in general, lack splicing (Ostertag and Kazazian, 2001a), HERV-Ws have a complex splicing pattern and the majority of their pseudogenes correspond to these splice variants, so it is difficult to calculate the expectation for full-length copies. The full size of the shortest splice variants is only 600–1300 bp, although the full size of HERV-W mRNA is about 9500 bp (Pavlíček et al., 2002).

Alus, another type of non-autonomous LINE-dependent repeats (see for example Ostertag and Kazazian, 2001a; Weiner, 2002), often escape the 5' truncation thanks to their very small size (about 280 bp at the consensus). For the young AluYa5 family, we found just 8.3% (300/3609) of 5' truncated copies compared to 3.7% (133/3609) of 3' truncated copies.

As can be seen in Figs. 1 and 3, the distributions of the L1HS family and of processed pseudogenes of HERV-W are not consistent with the random termination model. The first difference is the overrepresentation of short elements. It is possible that early invasion of the internal primer onto the LINE RNA before reverse transcription that has begun at the polyT primer or shortly thereafter (Ostertag and Kazazian, 2001b) can lead to premature inversions and, hence, to early truncations, and can reshape the size distribution. Nevertheless, excluding 5' inverted copies does not significantly change the size distribution (not shown). Frequent transduction of 3' sequences during LINE transposition (Pickeral et al., 2000; Goodier et al., 2000; IHGSC, 2001) can also shift the length distribution, especially in the small sizes, since the RT first integrates the transduced 3' sequence and after this, if not already terminated, proceeds further to the LINE sequence. However, the unexpectedly high number of complete elements (also found for processed pseudogenes of

ribosomal proteins, Zhang et al., 2002) definitely does not fit with the model of low-processive RT.

#### 4.2. Implications for the processed pseudogene detection

Currently, the translated homology search of protein databases against the human genome is the method of choice for detection of processed pseudogenes (Dunham et al., 1999; Goncalves et al., 2000; Venter et al., 2001; Harrison et al., 2002). This approach can obviously detect only pseudogenes homologous to protein coding sequences. However, eukaryotic mRNAs also contain untranslated regions, often with important *cis*-acting sequences regulating the mRNA stability, cytoplasmic transport, translational efficiency, etc. (Makalowski and Boguski, 1998; Pesole et al., 2001). Taking 1028 bp as a mean length of human 3'UTRs (Pesole et al., 2001), we theoretically overlook 47% (468/994), 72.3% (173,103/239,495) or 28% (49/176) of processed pseudogenes shorter than 1028 bp, provided that their distribution is similar to L1HS, all LINEs or HERV-W processed pseudogenes, respectively.

Considering the factor of 28% for HERV-W as a theoretical limit, this is clearly an underestimation of the real situation, since the commonly used criterion is a long uninterrupted homology to the closest matching protein (Goncalves et al., 2000; Venter et al., 2001; Harrison et al., 2002). However, the proportion of LINEs and HERV-W pseudogenes longer than 3 kb is low (Figs. 1–3), comprising mostly full-length elements, i.e. only 26% (46/176) of all HERV-W processed pseudogenes. It is, therefore, possible that only about 25–35% of processed pseudogenes are detected by conventional methods, in majority full-length. Indeed, 95% of processed pseudogenes of ribosomal proteins were found complete (Zhang et al., 2002), again indicating some other process than the low processivity of RT in the 5' truncations. Moreover, this can still be an underestimate, since the HERV-W mRNA is relatively short in comparison to the mean human genes (9.5 versus 27 kbp; IHGSC, 2001) and has a proportionally higher probability to be complete, i.e. untruncated. Also, the criterion of long continuous homology with the protein sequence can disregard some splice variants if large parts of coding sequences are spliced out.

Taken together, it is possible that as much as 65–75% of processed pseudogenes remain undetected by conventional methods, because they do not take into account the length distribution of processed pseudogenes. Homology searches using UTR databases (Pesole et al., 2002) can provide a better estimate of the number of processed pseudogenes.

## 5. Conclusion

In contrast to the expected randomness of TPRT termination, the length distribution of the young LINE L1HS (Ta) family (and partially also processed pseudo-

genes) is more similar to a power-law rather than to an exponential distribution. The unexpectedly high proportions of short elements and particularly a surprisingly high number of complete copies again indicate a different process than the low processivity of RT in the generating of 5' truncated copies. Since the probability of termination is highest immediately at the beginning of the TPRT, whereas after proceeding past the first 2–3 kb the integration is likely to terminate successfully, the mechanism of the 5' truncation might tentatively be linked to early steps in the TPRT reaction. Alternatively, premature termination can be a mixture of two or more independent processes. We also demonstrate that transposon-derived pseudogenes represent, thanks to the standard methodology for their detection, a powerful tool for the studies of processed pseudogene properties.

In conclusion, based on our results, the current estimate of 22,000–33,000 processed pseudogenes in the human genome (Goncalves et al., 2000) should be increased by a factor of at least three, taking into account all retro-transpositions of cellular mRNAs, not only those with a long homology to proteins (whether processed pseudogenes without homology to protein coding regions can be called pseudogenes is merely a matter of terminology). An open question remains: how is it possible to have such a high number of retroposed pseudogenes in spite of the strong *cis*-preference of the LINE proteins to mobilize its own RNA (Sassaman et al., 1997; Esnault et al., 2000; Wei et al., 2001)?

## Acknowledgements

We are very grateful to Oliver Clay for his help with distribution fitting and comments on the manuscript. We also thank Zhaolei Zhang for providing his unpublished results. This work was supported by grant No. 204/01/0632 of the Grant Agency of the Czech Republic to J.H.

## References

- Bailey, J.A., Carrel, L., Chakravarti, A., Eichler, E.E., 2000. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc. Natl. Acad. Sci. USA* 97, 6634–6639.
- Boissinot, S., Chevret, P., 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., Furano, A.V., 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol. Biol. Evol.* 18, 926–935.
- Chureau, C., Prissette, M., Bourdet, A., Barbe, V., Cattolico, L., Jones, L., Eggen, A., Avner, P., Duret, L., 2002. Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Res.* 12, 894–908.
- Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E.,

- Bruskiewich, R., Beare, D.M., Clamp, M., Smink, J.L., et al., 1999. The DNA sequence of human chromosome 22. *Nature* 402, 489–495.
- Esnault, C., Maestre, J., Heidmann, T., 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* 24, 363–367.
- Feng, Q., Moran, J.V., Kazazian, H.H. Jr, Boeke, J.D., 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87, 905–916.
- Goncalves, I., Duret, L., Mouchiroud, D., 2000. Nature and structure of human genes that generate retropseudogenes. *Genome Res.* 10, 672–678.
- Goodier, J.L., Ostertag, E.M., Kazazian, H.H. Jr, 2000. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* 9, 653–657.
- Harrison, P.M., Hegyi, H., Balasubramanian, S., Luscombe, N.M., Bertone, P., Echols, N., Johnson, T., Gerstein, M., 2002. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* 12, 272–280.
- IHGSC (International Human Genome Sequencing Consortium), 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Jurka, J., 1998. Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.* 8, 333–337.
- Jurka, J., 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* 16, 418–420.
- Kazazian, H.H. Jr, 1999. An estimated frequency of endogenous insertional mutations in humans. *Nat. Genet.* 22, 130.
- Kazazian, H.H. Jr, Moran, J.V., 1998. The impact of L1 retrotransposons on the human genome. *Nat. Genet.* 19, 19–24.
- Luan, D.D., Korman, M.H., Jakubczak, J.L., 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.
- Makalowski, W., Boguski, M.S., 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA* 95, 9407–9412.
- Malik, H.S., Burke, W.D., Eickbush, T.H., 1999. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* 16, 793–805.
- Mathias, S.L., Scott, A.F., Kazazian, H.H. Jr, Boeke, J.D., Gabriel, A., 1991. Reverse transcriptase encoded by a human transposable element. *Science* 254, 1808–1810.
- Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., Moran, J.V., 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat. Genet.* 31, 159–165.
- Ostertag, E.M., Kazazian, H.H. Jr, 2001a. Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* 35, 501–538.
- Ostertag, E.M., Kazazian, H.H. Jr, 2001b. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* 11, 2059–2065.
- Pačes, J., Pavlíček, A., Pačes, V., 2002. HERVd: database of human endogenous retroviruses. *Nucleic Acids Res.* 30, 205–206.
- Pavlíček, A., Pačes, J., Elleder, D., Hejnar, J., 2002. Processed pseudogenes of human endogenous retroviruses generated by LINES: their integration, stability, and distribution. *Genome Res.* 12, 391–399.
- Pesole, G., Mignone, F., Gissi, C., Grillo, G., Licciulli, F., Liuni, S., 2001. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* 276, 73–81.
- Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Mignone, F., Gissi, C., Saccone, C., 2002. UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.* 30, 335–340.
- Pickeral, O.K., Makalowski, W., Boguski, M.S., Boeke, J.D., 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* 10, 411–415.
- Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., Kazazian, H.H. Jr, 1997. Many human L1 elements are capable of retrotransposition. *Nat. Genet.* 16, 37–43.
- Smit, A.F., 1996. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* 6, 743–748.
- Smit, A.F., 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9, 657–663.
- Smit, A.F., Toth, G., Riggs, A.D., Jurka, J., 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* 246, 401–417.
- Vanin, E.F., 1985. Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.* 19, 253–272.
- Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al., 2001. The sequence of the human genome. *Science* 291, 1304–1351.
- Voliva, C.F., Jahn, C.L., Comer, M.B., Hutchison, C.A., Edgell, M.H., 1983. The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. *Nucleic Acids Res.* 11, 8847–8850.
- Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H.H., Boeke, J.D., Moran, J.V., 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol. Cell. Biol.* 21, 1429–1439.
- Weiner, A.M., 2002. SINEs and LINES: the art of biting the hand that feeds you. *Curr. Opin. Cell. Biol.* 14, 343–350.
- Weiner, A.M., Deininger, P.L., Efstradiatis, A., 1986. Non-viral retrotransposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* 55, 631–661.
- Zhang, Z., Harrison, P.M., Gerstein, M., 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* 12, 1466–1482.