J Mol Evol (2000) 50:232–237
DOI: 10.1007/s002399910026

JOURNAL OF MOLECULAR EVOLUTION

Characteristic Sequence Pattern in the 5- to 20-bp Upstream Region of Primate *Alu* Elements

Yoshimi Toda,[1,2] Rintaro Saito,[1,2] Masaru Tomita[1,3]

[1] Laboratory for Bioinformatics, Keio University, 5322 Endo, Fujisawa, 252-8520 Japan
[2] Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, 252-8520 Japan
[3] Department of Environmental Information, Keio University, 5322 Endo, Fujisawa, 252-8520 Japan

Received: 10 May 1999 / Accepted: 1 October 1999

Abstract. We conducted comprehensive sequence analysis of 5′ flanking regions of primate *Alu* elements. Information contents were computed and frequencies of 1024 pentanucleotides were measured to approximate the location of a characteristic sequence and to specify its pattern(s), which may be involved in the integration of *Alu* elements into their host genomes. A large number of samples was used, the wide region of the 5′ end of *Alu* elements was analyzed, and comparisons were made among different subfamilies. Through our analyses, "TTTTAAAAA" or "$(T)_m(A)_n$" can be stated as a candidate for the characteristic sequence pattern, which resides around the region 5 to 20 base pairs upstream of the 5′ end of *Alu* elements. This characteristic sequence pattern was more prominent in the sequences of younger *Alus*, which is a strong indication that the sequence pattern has a role at the time of *Alu* integration.

Key words: Retroposons — Retrotranscripts — *Alu* elements — SINEs — Integration target

Alu repetitive sequences constitute an *Alu* family of short interspersed nucleic elements (SINEs). They are a type of non-LTR retroposons abundantly found in genomes of various organisms and are commonly found in primate genomes (Weiner et al. 1986). Their copy number in a human genome is estimated to be several hundred thousand to a million, which accounts for approximately 5 to 10% of the entire human genome (Weiner et al. 1986; Okada 1991, 1994). An *Alu* element is about 300 base pairs (bp) long and consists of two similar subunits which are connected with an adenine-rich (A-rich) linker and followed by a 3′ poly(A) tail. This unit is usually flanked immediately by direct repeats at both the 5′ and the 3′ ends (Rinehart et al. 1980; Daniels and Deininger 1985).

Alu elements are subclassified into three groups, namely, Old, Middle, and Young. They are further classified into 12 subfamilies (Old—Jo, Jb; Middle—Sz, Sx, Sq, Sp, Sg, Sc; Young—Y, Ya1, Ya5, Yb8) in accordance with their putative time of proliferation and based on diagnostic positions in their nucleotide sequences (Batzer et al. 1996).

Despite extensive studies of retroposons, the mechanism of retroposition through which non-LTR retroposons multiple their copy numbers is still not yet fully understood and there is not a definite single model that describes the mechanism (Rogers 1985; Boeke and Stoye 1996). On the analogies of the proposed hypothetical model of L1 retroposition, a model of SINE retroposition can be assumed (Eickbush 1992; Luan et al. 1993; Boeke and Stoye 1996). According to the model (Fig. 1), transcription of a SINE that produces a SINE mRNA is followed by nicking of the target site in the minus strand of the host genome sequence. The 3′ end of mRNA is then anchored to the nicked site and reverse transcription takes place using the SINE mRNA as a template and the

Correspondence to: Masaru Tomita; *e-mail:* mt@sfc.keio.ac.jp

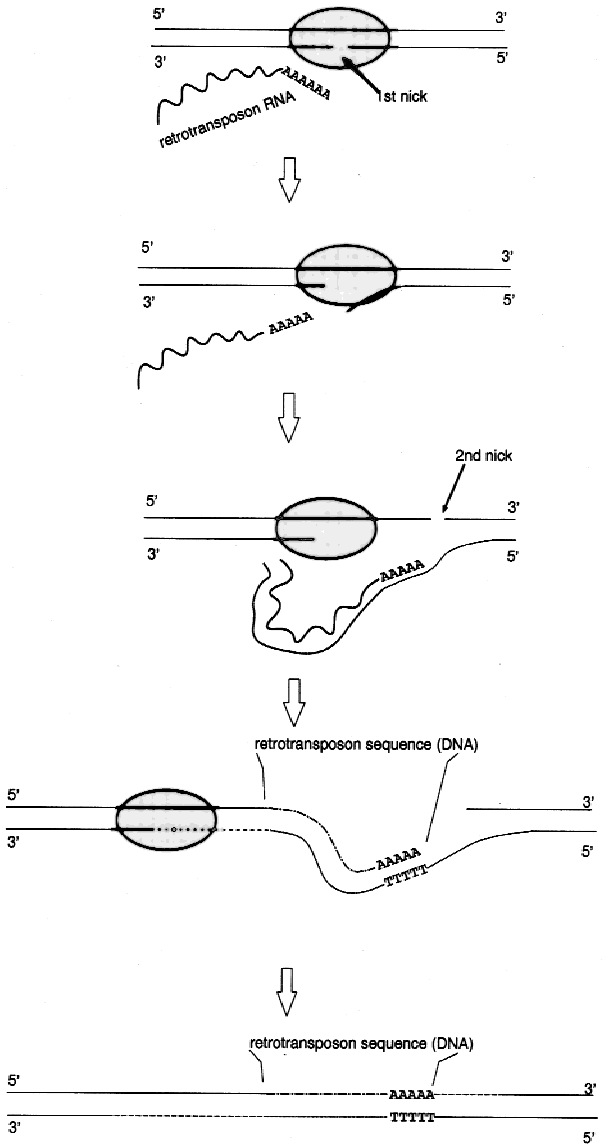

Fig. 1. A schematic model of *Alu* element integration. Modified after Luan et al. (1993) and Jurka (1997).

3′ end of the nicked site as a primer. The complement sequence is generated to form an RNA–DNA complex, and then a double-stranded DNA sequence of the SINE is generated, which is finally joined to the other nicked end to complete integration. The target-site duplication generates flanking direct repeats at both the 5′ and the 3′ ends of the SINE.

Site specificity of integration is known with some non-LTR retroposons such as the R2 element of *Bombyx mori* (R2Bm) (Luan et al. 1993). Their integration is associated with the 28SrRNA genes and some others are associated with certain tandem arrays. On the other hand, retrotranscripts such as SINEs had been considered non-site specific as reviewed by Boeke and Stoye (1996). However, the recent study by Jurka (1997) suggested the contrary. According to their study, there are possible consensus sequence patterns at integration sites of certain mammalian retroposons, which strongly suggests sequence-specific enzymatic involvement that mediates integration. For their analyses, 344 human *Alu* elements and 56 rodent ID elements that retain full length with identical flanking repeats at both the 5′ and the 3′ ends were carefully selected manually.

Table 1. The number of sample sequences in each subfamily used in the analyses

Subfamily	No. of samples
Old	
Jo	2,853
Jb	5,677
Middle	
Sz	4,279
Sq	2,728
Sp	2,011
Sx	4,527
Sg	2,462
Sc	1,514
Young	
Y	3,132
Ya1	105
Ya5	169
Yb8	206
Total	29,663

In our current study, to demonstrate the existence and the pattern of a characteristic sequence involved in *Alu* integration, comprehensive analyses of a wide 5′ upstream region of the *Alu* elements in 12 subfamilies were conducted in a fully automatic manner. “Information content” (Schneider et al. 1986) was computed and frequencies of all the possible pentanucleotide patterns were measured at each nucleotide position in the 5′ flanking region of *Alu* elements.

For an exhaustive search of any possible sequence pattern(s) associated with *Alu* elements, primate *Alu* elements longer than 250 bp were extracted from the NCBI GenBank database (release 103.0), along with their 5′ flanking sequences up to 500 bp long. The CENSOR program (Jurka et al. 1996) was used for extraction and subclassification of *Alu* elements. The 5′ flanking sequences of *Alu* elements up to 500 bp were extracted using our original program. Altogether, 29,663 *Alu* and their immediate upstream sequences were extracted and classified into 12 subfamilies. The number of analyzed *Alu* elements of each subfamily is shown in Table 1.

All the sample sequences were adjusted to the 5′ end of *Alu* elements when used for our analyses, while in Jurka (1997), the 5′ flanking repeats were adjusted to the left and the 3′ flanking repeats to the right to align the regions outside of the SINE sequences and the flanking repeats. These are direct repeats, which are oriented in the same direction and had been generated immediately flanking both the 5′ and the 3′ ends of the *Alu* elements as the result of the retroposon integration. Automatic detection of such flanking direct repeats is difficult be-

cause they are frequently absent, partial, or nonpaired. Thus, positions referred to in this report do not necessarily indicate precise locations in the analyzed region but include errors within ±5 bp.

Information content values were computed as indices in an exhaustive search to approximate the location of a possible characteristic sequence pattern(s) related to *Alu* integration. Information content is more useful than the χ^2 scale when searching for characteristic sites or sequence patterns that are new and unknown (Schneider et al. 1986). The χ^2 scale is used frequently in the sequence analyses in deciding consensus patterns after simple sequence alignment as also used by Jurka (1997) and Jurka and Klonowski (1996). However, as Schneider et al. (1986) discuss, it is not the most suitable scale to use to capture the general characteristics of the sequences to be analyzed in various ways.

The following formula was used to compute information content at each position l (Schneider et al. 1986).

$$\text{Information content}_{(l)} = \sum_{i=a,t,c,g} O_{(i,l)} \log_2 \frac{O_{(i,l)}}{E_{(i)}} \qquad (1)$$

where O_i is the observed frequency of nucleotide i at position l and E_i is the expected frequency of nucleotide i.

The observed frequency, $O_{(i,l)}$ is the actual frequency of the nucleotide i at position l. The expected frequency of the nucleotide i, $E_{(i)}$, is computed by taking the nucleotide composition of the whole analyzed sequences into consideration. By dividing $O_{(i,l)}$ by $E_{(i)}$, information content was standardized to avoid influence from nucleotide composition of analyzed sequences. A high information content at a certain position indicates a small variety in the nucleotide composition, i.e., a specific nucleotide appears at the position more often than expected.

To identify a specific sequence pattern(s) that may be connected to the existence of *Alu* elements, frequency values of all possible 1024 pentanucleotides were measured. The number of observed pentanucleotides at each position was divided by the number of analyzed sequences, and the resulted frequency was plotted for each position.

The existence of a characteristic nucleotide pattern was made apparent by plotting the information content values onto a graph. In all the analyzed 12 subfamilies, the information content value is high at the positions between −20 and −10 bp upstream of the 5′ end of *Alu*s. This suggests the existence of a characteristic sequence pattern(s) around this region. As shown in Fig. 2, the information content value is the highest with the youngest subfamilies. Average information content values of Ya1, Ya5, and Yb8 are indicated by "Young" in Fig. 2. This is because the sequence pattern(s) is(are) better conserved in the younger subfamilies than the older ones. A

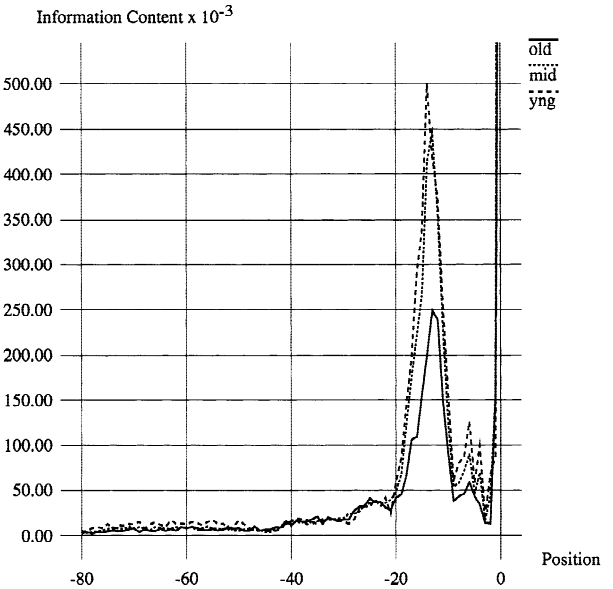

Fig. 2. Information content in the 5′ upstream region of *Alu* elements.

relatively short period of time has passed since the integration of the younger *Alu* elements into the host genomes. The subfamilies Jo and Jb, which are evolutionarily the oldest, show lower information content values compared to the younger ones. This is shown as "Old" in Fig. 2. These observations indicate that the region between −20 and −10- bp upstream of old *Alu* elements has accumulated more mutations in the course of time and lost its original sequence pattern(s). In contrast to the region between −20 and −10 bp upstream of *Alu* elements, the information content values in the region farther upstream of the −20 bp position are constantly low and show no differences among subfamilies, indicating that this region contains no characteristic sequence pattern.

Our observation that the region between 20 and 10 bp preceding the 5′ end of *Alu* elements shows high information content values is consistent with the results by Jurka (1997). Jurka computed χ^2 values of the suspected consensus patterns for each position of the 15-bp region immediately preceding the 5 end of flanking repeat sequences using a smaller data set of 344 human *Alu* and 56 rodent ID elements. The positions between −2 and +4 from the border of the 5′ adjacent and flanking sequences show significantly high χ^2 values. Allowing errors within ±5 bp in the referred location, the region between −20 and −10 bp upstream of the 5′ end of *Alu* elements where we observed significantly high information content values corresponds approximately to their −2 to +4 region. Consistent with their analysis, our result strongly suggests the existence of a certain consensus sequence(s) around this region. Our results, along with Jurka's, support the hypothesis that there is a characteristic sequence pattern involved in the mechanism of *Alu* integration.

Table 2. The most frequent pentanucleotide patterns observed at the positions between −20 and −11 from the 5′ end of *Alu* elements[a]

−20	aaaaa	0.0367239
	ttaaa	0.0167258
	tttaa	0.0157258
	taaaa	0.0133624
	tttta	0.0107263
	ttttt	0.0103627
	aaaag	0.0071792
	aaata	0.0070514
	agaaa	0.0065809
	aataa	0.0063857
−19	aaaaa	0.0431779
	ttaaa	0.0236342
	taaaa	0.0210890
	tttaa	0.0177257
	tttta	0.0136351
	ttttt	0.0109290
	agaaa	0.0090901
	aataa	0.0090901
	aaaga	0.0076357
	aaata	0.0073707
	aagaa	0.0066358
	aaaag	0.0064540
−18	aaaaa	0.0513590
	taaaa	0.0319971
	ttaaa	0.0309972
	tttaa	0.0246341
	aaaat	0.0117262
	agaaa	0.0111808
	aagaa	0.0111808
	aaata	0.0108383
	aaaga	0.0101809
	aaaag	0.0093628
	tttta	0.0091810
	aataa	0.0089992
−17	aaaaa	0.0683574
	taaaa	0.0435415
	ttaaa	0.0379965
	tttaa	0.0191801
	aaaat	0.0179075
	agaaa	0.0171803
	aagaa	0.0156349
	aaaga	0.0117262
	aataa	0.0109990
	aaaag	0.0103627
	aaata	0.0098173
	tttta	0.0076357
−16	aaaaa	0.0850832
	taaaa	0.0499955
	ttaaa	0.0318153
	aaaat	0.0284520
	agaaa	0.0215435
	aagaa	0.0182711
	aaaga	0.0149077
	tttaa	0.0144532
	aaaag	0.0137260
	aaata	0.0112717
	aataa	0.0109081
	tttta	0.0061813
−15	aaaaa	0.0888101
	taaaa	0.0411781
	aaaat	0.0382692
	aaaag	0.0246341
	aagaa	0.0214526
	ttaaa	0.0203618
	agaaa	0.0175439
	aaata	0.0174530
	aaaga	0.0124534
	aataa	0.0109990
	tttaa	0.0107263
−14	aaaaa	0.0785383
	aaaat	0.0367239
	aaaag	0.0272702
	aaata	0.0242705
	taaaa	0.0240887
	aaaga	0.0172712
	aagaa	0.0160894
	agaaa	0.0158167
	ttaaa	0.0133624
	aataa	0.0112717
	tttaa	0.0067353
−13	aaaaa	0.0632670
	aaaag	0.0235433
	aaata	0.0230888
	aaaat	0.0209072
	aaaga	0.0160894
	taaaa	0.0150895
	aagaa	0.0149986
	aataa	0.0116353
	agaaa	0.0098173
	ttaaa	0.0074539
	tttaa	0.0059086
	tttta	0.0048608
−12	aaaaa	0.0543538
	aaaat	0.0161789
	aagaa	0.0150882
	aaaag	0.0145428
	aaaga	0.0136339
	aataa	0.0122705
	taaaa	0.0116342
	aaata	0.0109980
	agaaa	0.0106344
	ttaaa	0.0078168
	tttaa	0.0063977
−11	aaaaa	0.0483548
	aagaa	0.0133612
	aaaga	0.0112707
	aaaat	0.0109980
	aaata	0.0109071
	aaaag	0.0102709
	taaaa	0.0100891
	agaaa	0.0098164
	ttaaa	0.0082256
	aataa	0.0079985

[a] The first column shows the position; the second column lists the patterns; the third column shows the frequency of listed patterns. Each frequency is the average of that of each pattern from three subfamilies, Jb, Sx, and Y.

To identify a nucleotide sequence pattern(s) causing the high information content values, frequencies of all of the 1024 pentanucleotides were measured at each nucleotide position preceding the 5′ end of *Alu* elements. Among the 1024 pentanucleotide patterns, AAAAA, TAAAA, TTAAA, and TTTAA are outstandingly the most frequent in the region between −20 and −10 bp upstream of *Alu* elements. The most frequent patterns at each position along with their frequencies are listed in

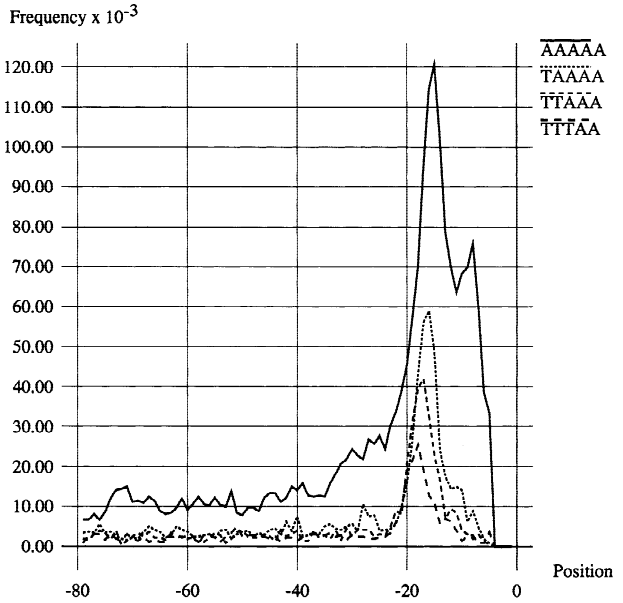

Fig. 3. Frequencies of four most frequent pentanucleotides in the young subfamilies.

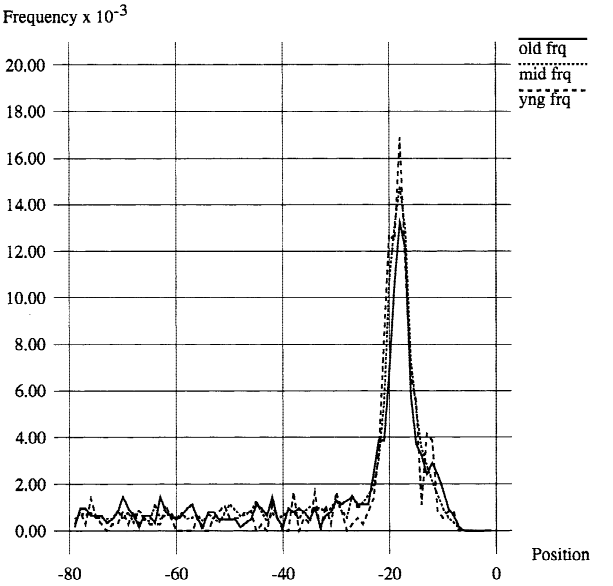

Fig. 4. Frequencies of the heptanucleotide TTTAAAA.

Table 2 and the frequency distribution of the four most frequent patterns is shown in Fig. 3.

Assembling the most frequent pentanucleotides, which are located at positions next to each other, we obtained a nonanucleotide pattern, TTTTAAAAA, which seems to be the characteristic sequence pattern in the region. Because of the limited number of sample sequences of some subfamilies, six subfamilies (Jb, Jo, Sz, Sx, Sq, Y) with sample sequences more than 2500 were analyzed for frequencies of the heptanucleotide, TTTAAAA, at each position and the results are shown in Fig. 4.[1] The highest frequency was observed in the region between −20 and −10 bp upstream of *Alu* elements with all analyzed subfamilies. The group of younger subfamilies showed higher frequencies around the region than that of older subfamilies (Fig. 4). The changes in frequency values among subfamilies imply that the observed sequence pattern is more conserved in the younger subfamilies than the older ones. This indicates that the analyzed characteristic pattern, TTTAAAA, already existed at the time of *Alu* integration, and as the result of accumulated mutations, the putative original pattern has degraded. Within the analyzed 500-bp region, the characteristic pattern was detected only in the limited area close to the SINE sequences. No significant pattern was detected in the region farther upstream of the area. This indicates that the region farther upstream is not directly involved in targeting the insertion site.

This heptanucleotide pattern can be represented as "$(T)_m(A)_n$" or "$(Y)_m(R)_n$", where *Y* represents a pyrimidine, *R* represents a purine, and *m* and *n* represent the number of nucleotides shown in parentheses. Based on the suggestion regarding bendability of a nucleotide strand given by Jurka et al. (1998), this pyrimidine–purine boundary, which resides upstream of a retroposon, may be a kinkable site. The purine oligonucleotide could be facing the histone complex, while the pyrimidine oligonucleotide could be facing outside. Thus, the boundary of the pyrimidine and purine tracts could be a bendable site that could be an easy target of a protein involved in the *Alu* integration. However, "$(T)_m(G)_n$", "$(C)_m(A)_n$", and "$(C)_m(G)_n$", which are the other possible pyrimidine–purine tracts, were not observed at a high frequency. The targeted integration site, thus, seems to be that (1) it has to contain a pyrimidine–purine tract and (2) it has to be recognized by a certain sequence specific endonuclease. "$(Y)_m(R)_n$" tract is a pattern recognized and nicked by the endonuclease encoded by the L1 element, and the purines are often A's (Tatout et al. 1998; Feng et al. 1996). With another supporting report by Tatout et al. (1998) in the case of plant SINE S1 retroposons, our results thus emphasize that *Alu* elements are not randomly integrated into host genomes. Also, their integration is likely to be dependent on the endonuclease encoded by the L1 element (Tatout et al. 1998; Feng et al. 1996).

In this study, comprehensive analyses of sequence patterns in the upstream regions of the 5′ end of *Alu* elements were conducted in a fully automatic manner on a large data set (1) by computing information contents and (2) by measuring the frequencies of pentanucleotide patterns. Our results show that the approximate location between −20 and −10 bp of the 5′ end of *Alu* elements is the only region within the analyzed 500-bp region that shows high information content values and high frequen-

[1] The nonanucleotide pattern, TTTTAAAAA, was also analyzed, but the results are not shown because of the limitation due to small sample size.

cies of certain pentanucleotides. This indicates that the region contains a characteristic sequence pattern whose primary candidate is "$(T)_m(A)_n$", which may be extended to TTTTAAAAA. The variance among evolutionarily different subgroups of *Alu*s is a strong indication that the sequence pattern existed at the time of *Alu* integration and has accumulated mutations in the course of time. Thus, as proposed by Luan et al. (1993) and supported by Jurka (1997), our results further emphasize that the integration of *Alu* elements into host genomes is sequence specific, suggesting the involvement of sequence-specific enzymes.

Acknowledgments. We thank Drs. Jerzy Jurka, Paul Klonowski, and Jolanta Walichiewicz at the Genetic Information Research Institute, Palo Alto, CA, for data on *Alu* elements. This work is supported in part by a Grant-in-Aid for Scientific Research on Priority Areas "Genome Science" from the Ministry of Education, Science, Sports and Culture, Japan.

References

Batzer MA, Deininger PL, Hellmann-Blumberg U, Jurka J, Labuda D, Rubin CM, Schmid CW, Zietkiewicz E, Zuckerkandl E (1996) Standardized nomenclature for *Alu* repeats. J Mol Evol 42:3–6

Boeke JD, Stoye JP (1996) Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In: Varmus H, Hughes S, Coffin J (eds) Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp 343–435

Daniels GR, Deininger PL (1985) Integration site preferences of the *Alu* family and similar repetitive DNA sequences. Nucleic Acids Res 13:8939–8954

Eickbush TH (1992) Transposing without ends: The non-LTR retrotransposable elements. New Biol 4:430–440

Feng Q, Moran JV, Kazazian HH Jr., Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905–916

Jurka J, Klonowski P (1996) Integration of retroposable elements in mammals: Selection of target sites. J Mol Evol 43:685–689

Jurka J, Klonowski P, Dagman V, Pelton P (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 20:119–121

Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc Natl Acad Sci USA 94:1872–1877

Jurka J, Klonowski P, Trifonov EN (1998) Mammalian retroposons integrate at kinkable DNA sites. J Biomol Struct Dyn 15:717–721

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (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

Okada N (1991) SINEs: Short interspersed repeated elements of the eukaryotic genomes. Trends Ecol Evol 6:358–361

Okada N (1994) [Retroposon as temporal landmarks of evolution]. Tanpakushitsu Kakusan Koso 39:2724–2735 (in Japanese)

Rinehart FP, Ritch TG, Deininger PL, Schmid CW (1981) Renaturation rate studies of a single family of interspersed repeated sequences in human deoxyribonucleic acid. Biochemistry 20:3003–3010

Rogers JH (1985) The origin and evolution of retroposons. Int Rev Cytol 93:187–279

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. J Mol Biol 188: 415–431

Tatout C, Lavie L, Deragon JM (1998) Similar target site selection occurs in integration of plant and mammalian retroposons. J Mol Evol 47:463–470

Weiner AM, Deininger PL, Efstratiadis A (1986) Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu Rev Biochem 55:631–661