Short Interspersed Repeats in Rabbit DNA Can Provide Functional Polyadenylation Signals[1]

Dan E. Krane and Ross C. Hardison

Department of Molecular and Cell Biology, The Pennsylvania State University

Analysis of 37 short repetitive elements (SINEs) in rabbit DNA that are known as C repeats has revealed three that contribute functional polyadenylation signals to genes into which they have been inserted. Similar roles have been attributed to particular individual SINEs in rodents and primates before, suggesting that these roles may be common to SINEs in all mammalian orders. Although most SINEs appear to have little influence on the genome individually, the observation that three of 36 rabbit C repeats provide functional sequences suggests a mechanism for the maintenance of SINEs within mammalian genomes.

Introduction

Interspersed repetitive elements in mammals can be classified by length as either SINEs (short), which are typically <500 bp, or LINEs (long), whose full-length copies range in size from 6 to 8 kb (Singer 1982). Within each length class, families can be identified on the basis of sequence similarity. Virtually all mammalian SINEs share several structural similarities such as internal RNA polymerase III promoters, oligo (A)-rich tracts at their 3′ ends, and an absence of open reading frames (Weiner et al. 1986). However, sequence comparisons show that SINEs present in high copy number in one taxonomic order have different origins than do the highly repeated SINEs in other orders (Schmid and Shen 1985). In the proposed model for the retrotransposition of SINEs (Rogers 1985; Weiner et al. 1986), the RNAs transcribed from these repeats are reverse transcribed and reinserted into the genome at staggered breaks, as evidenced by resulting flanking direct repeats. SINEs and LINEs appear to be currently propagating, as is shown by recent insertions of repeats (e.g., see King et al. 1986; Kazazian et al. 1988). Rabbit (*Oryctolagus cuniculus*) C repeats are typical of SINEs, and >200,000 copies of this element are found interspersed throughout the rabbit genome (Cheng et al. 1984; Hardison and Printz 1985). With an average length of 350 bp, they account for ⩾5% of the total DNA in rabbits.

The role that repetitive DNA elements play in the evolution and function of mammalian genomes has been a matter of debate since their discovery (reviewed in Rogers 1985; Weiner et al. 1986). They can act negatively by inactivating genes when they insert into coding (Kazazian et al. 1988) or regulatory regions. Frequently repeats insert into regions of little or no function, and hence the event is likely to be neutral (Sawada et al. 1985; Margot et al. 1989). It has also been argued that these elements are selfish DNA whose self-propagation provides no benefit to their hosts (Doolittle and Sapienza 1980; Orgel and Crick 1980).

1. Key words: C repeats, short interspersed repeats, SINEs, genome evolution, polyadenylation.

Address for correspondence and reprints: Dr. Ross C. Hardison, Department of Molecular and Cell Biology, The Pennsylvania State University, University Park, Pennsylvania 16802.

Mol. Biol. Evol. 7(1):1–8. 1990.

0737-4038/90/0701-0001$02.00

In contrast, a few individual members of these diverse repeat families appear to make positive contributions to the evolution of mammalian genomes. For example, Alu repeats are part of the coding regions for the human (*Homo sapiens*) genes for decay-accelerating factor and for a B-cell growth factor (Caras et al. 1987; Sharma et al. 1987). Also, the short B2 repeats can provide functional polyadenylation signals in at least three rodent genes (Kress et al. 1984; Ryskov et al. 1984; Rothkopf et al. 1986). We present evidence that rabbit short interspersed repeats known as C repeats can also contribute functional polyadenylation signals. This observation implies that this positive effect is probably common to this class of repeats in many mammalian orders.

Material and Methods

The consensus sequence was generated by alignment of 36 C repeats (to be reported elsewhere) from the rabbit genome. Along with those reported earlier (Hardison and Printz 1985), the sequences include 15 from the β-globin gene cluster in rabbits (Margot et al. 1989), 11 from the rabbit α-globin gene cluster (Cheng et al. 1988; and sequences to reported elsewhere), and 10 obtained by computer-assisted searches of GenBank.

A multiple alignment of these sequences was made by first generating two-way alignments by using the Genetic Computer Group's Gap program (Devereux et al. 1984), which uses the alignment algorithm of Needleman and Wunsch (1970). These two-way alignments were compared and aligned by inspection, and gaps were inserted to increase the overall similarity. A consensus sequence was generated on the basis of this multiple alignment and was used to search the GenBank data base with either the WordSearch (Devereux et al. 1984) or FASTA (Pearson and Lipman 1988) computer programs. Several previously unrecognized repeats were found in this way and were added to the growing multiple alignment. The multiple alignments of the three repeats with the consensus sequence shown in figure 1 were generated by making pairwise alignments which were in turn combined by inspection. Percent similarities between sequences were determined with the Gap program, which accepts matches with ambiguous nucleotides as half-matches, by using a gap penalty of 4.5 and a gap length penalty of 0.30/nucleotide missing in the gap.

Results

Analysis of 36 C repeat sequences has yielded an improved consensus sequence for this family of repeats. The last reported consensus was based on the sequence of six C repeats (Hardison and Printz 1985) and was heavily biased toward what now appears to be a subfamily of the overall group of repeats (to be reported elsewhere). The variability within this larger data set is much greater (an average of 25% dissimilar to the common consensus) than that reported for Alu repeats (Schmid and Shen 1985; Hwu et al. 1986) and for other SINEs (Rogers 1985), and, as a result, the consensus sequence reported in figure 1 contains several ambiguous positions.

As this consensus was being generated, it was used to search the GenBank data base to find other C repeats. In most cases, these sequences had not been previously recognized as members of the C repeat family. Three of the repeats identified in this way contribute functional polyadenylation signals to actively transcribed genes (Okino et al. 1985; Rebiere et al. 1987; Boggaram et al. 1988).

The first of these cases is the gene for isozyme 4 of cytochrome P-450. The rabbit genes for the homologous isozymes 4 and 6 of cytochrome P-450 differ radically in

```
                10        20        30        40        50        60        70        80        90
ConC          GGGGYNRGYRYTGTGGCDCAGTRGGTTAAGCCKCYRCYTGCRRYRCYG-GCATCCCATATBGGMGTSCYYGGTTCDAGTCCCRGCTGCTCCWCTTC      95
MHC         TA..CCTGA.CGA....-------------G...T.TG.T.A.AGTG.CT-..........G.AC---AGT.....GCA...TG........A....   1667
PS Apo    AAGA....CCA.CGC....T.GT..CA...A.....G.CA.C...AATG.TAT..C.......T.G.AC---ACC.....G..A...GA.......AT...   2726

             100       110       120       130       140       150       160       170       180        190
ConC      YRATCCAGCTCYCTGCTAATGYGCCTGGGARAGCAGYRGARGATGGCCCAAGTGCTTGGGYCCCTGCCACCCACRTGGGAGACCHG-GAWGAAGCTCCTG   194
MHC       AG......T..C......G..-........GA.....TG..G.........G.........C....---..A...A.A.......CA-..A....A.T...   1762
PS Apo    TG.........C.........T........A.....AGA.T....A..G...........C.....A-..T.T.A..........TGG..AA...G.....   2825
iso 4                                                                                         .A.CA-..T...   1616

              200       210       220       230       240       250        260       270               280
ConC      GCTCCTGGCTTYRGMCYGGCNCAGYYCYGGCYRTTGYRGCCATTTGGGGAGTGAACCAGYRG-ATGGAAGAYCTYTCTCTCTSTCT---------CTCTC   284
MHC       .G..T......TG.AGCT..T..ATT.T...CA...TG....CA..............ATG.-........C..C.......G...----------...C.   1852
PS Apo    ..C........TG.C.T.A.C.GAGC.CA..TA...CG.....C...........G...CA.-..A.....C..T.G.....CC..--------------   2910
iso 4     ..------...CT.C.CT..C.G.CC.T...TG..A-GT.................-..CAAG........T..C.......C...CTCTCTCTTT.GC.   1708

              290       300       310       320       330       340       350
ConC      TCTCTCTCTSTSTAACTCTGCCTTTCAAATAAATAAAWAAAAAAAAWWAAAAAAAAWAAAAAAAAAAAAAAAA                            355
MHC       ...T..CACC.G.......A...C..............T....T..AAT...GT.TTTTTTTT                                      1914
PS Apo    -----....C.G.......A...C..............C.G.T...CAC.TCTTTA..G..T..........AAAAA                        2982
iso 4     ...GCT...C.G..........................T                                                              1746
```

FIG. 1.—Alignment of three C repeat sequences with a consensus sequence. ConC = C repeat consensus sequence based on 36 members of this family of repeats; MHC = 3′ untranslated region of a gene from the MHC of rabbits (Rebiere et al. 1987); PS Apo = 3′ untranslated region of the gene for the major apoprotein of rabbit pulmonary surfactant (Boggaram et al. 1988); iso 4 = 3′ untranslated region of the gene for rabbit cytochrome P-450 isozyme 4 (Okino et al. 1985). Periods indicate matches with the consensus nucleotide at that position, while nucleotides are shown for positions that either are different from the consensus or are represented as ambiguous positions in ConC (R = purine; Y = pyrimidine; K = G or T; W = A or T; S = G or C; M = A or C; B = not A; D = not C; V = not T; H = not G; N = A, G, C, or T). A dash (–) indicates a gap inserted to improve the alignment. The first polyadenylation signal in the consensus sequence is shown in boldface. Numbers in the right-hand margin are the positions in individual GenBank files. The ConC sequence is numbered above the lines.

their 3′ untranslated regions, including the use of different polyadenylation signals (Okino et al. 1985). Both genes are expressed at high levels when induced by 2,3,7,8-tetrachlorodibenzo-*p*-dioxin, and the sizes of their mRNAs differ from homologues in the mouse (*Mus musculus*) because of extensive divergence 3′ to their termination codon. Our analysis of the sequence (figs. 1 and 2*a*) shows that a C repeat has inserted into the 3′ untranslated region of the isozyme 4 gene to provide the sequence AATAAA that is used as a polyadenylation signal (Proudfoot and Brownlee 1976). The nucleotides at the 5′ end of the isozyme 4 sequence in figure 1 are the first that match the consensus, indicating that 180 nucleotides of this repeat either have been deleted or were not transposed when the remaining portion of the repeat was inserted.

In the second case, the 3′ untranslated region of the rabbit gene for the major apoprotein of pulmonary surfactant contains two different polyadenylation signals (Boggaram et al. 1988). Comparison of the cDNA sequences to the consensus C repeat sequence (fig. 1) shows that a C repeat provides the second polyadenylation signal (fig. 2*a*). Densitometric analysis of Northern blots shows that this signal is used 20% of the time to make the longer of two mRNAs from that gene (Boggaram et al. 1988).

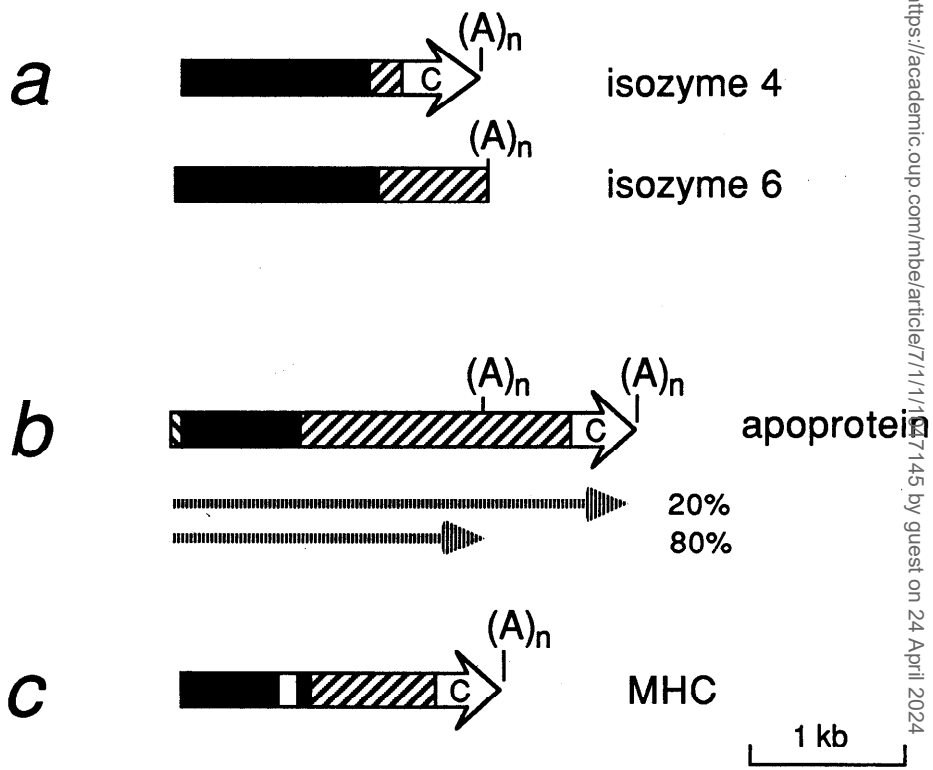

FIG. 2.—Diagrams illustrating the positions of C repeats that provide polyadenylation signals within cDNA clones. Protein coding regions are represented by solid black boxes (■); 5′ and 3′ UT regions are represented by hatched boxes (▨); C repeats are represented by arrowheads (⟹); and unspliced introns are represented by open boxes (□). The positions of polyadenylation signals are marked by $(A)_n$. *a*, cDNAs for rabbit cytochrome P-450 isozymes 4 and 6 (Okino et al. 1985). *b*, cDNA for the major apoprotein of pulmonary surfactant in rabbits, with its two polyadenylation signals. Hashed lines marked beneath the drawing show the lengths of the transcripts, along with their relative abundance (Boggaram et al. 1988). *c*, cDNA of the partially processed transcript for a gene from the MHC of rabbits (Rebiere et al. 1987).

In the third case, the polyadenylation signal reported for a gene from the major histocompatibility complex (MHC) of rabbits (Rebiere et al. 1987) is also generated by the insertion of a C repeat (figs. 1 and 2*a*). It was previously noted that a repetitive element was present in the 3′ untranslated region of this gene (Rebiere et al. 1987), and a comparison with the consensus shows that it is a C repeat (fig. 1). The C repeat is at the 3′ end of the cDNA, and it contains overlapping AATAAA sequences, indicating that this C repeat contributes the polyadenylation signal used in the processing of the message. Although this mRNA from a major histocompatibility gene is unusual in that its fourth intron has not been spliced out and in that part of its predicted 3′ untranslated region is similar to the last two introns and exons of several homologous genes in humans and rabbits, it is transcribed from a single-copy gene and in a tissue-specific manner, indicating that the mRNA may be functional (Rebiere et al. 1987).

For the three cases described above, the sequences assigned as polyadenylation signals perfectly match the consensus C repeat (fig. 1) and are positioned at the 3′ end of the cDNA clones (fig. 2*a*). Additionally, the measured length of mRNA from both the pulmonary surfactant (Boggaram et al. 1988) and the cytochrome P-450 isozyme 4 (Okino et al. 1985) genes indicates that the polyadenylation signals in C repeats are the ones used in the processing of the RNAs. It is striking that both the consensus and several individual C repeats contain a tandem and sometimes overlapping series of four conventional polyadenylation signals, AATAAA (fig. 1), suggesting that these are not randomly occurring sequences in the A-rich tract but may be conserved. Other, less well-characterized and apparently more variable sequences 3′ to the polyadenylation site are also involved in polyadenylation (Gil and Proudfoot 1984; Sadofsky et al. 1985), and presumably these other sequences are downstream from the C repeat insertion sites in these three cases.

Discussion

It is not uncommon for SINEs to be found within the introns and the 3′ flanking regions of the primary transcripts of genes, but these repeats are removed during the processing that produces mature mRNA (Jelinek and Schmid 1982). Still, some SINEs are retained in mature mRNA (frequently in the 3′ untranslated region) even after their processing has been completed. Ryskov et al. (1984) showed that when B2 repeats are present in mature poly(A^+) RNA of mouse liver, they are at the 3′ end of the transcript and are always in the orientation which would allow their conserved polyadenylation signals to be used in processing. Specific examples of this are B2 repeats that provide polyadenylation signals to a mouse class I histocompatibility gene (Kress et al. 1984) and to a rat glutathione S-transferase gene (Rothkopf et al. 1986). Members of the human Alu family of repeats also have well-conserved polyadenylation signals at their 3′ ends (Jelenik and Schmid 1982). Human Alu repeats have also been found in the 3′ untranslated regions of some genes, such as the human lysozyme gene (Chung et al. 1988), but that repeat is in the wrong orientation for its polyadenylation signal to be used, as is the B1 repeat in a mouse androgen-regulated gene (King et al. 1986). Repetitive elements other than SINEs also have the ability to provide importable polyadenylation signals. The 2.3-kb-long retrovirus-like human retrotransposon called THE 1 comprises the bulk of the 3′ untranslated region of a human calmodulin-related gene and provides a functional polyadenylation signal (Deka et al. 1988). The role of three rabbit C repeats in polyadenylation discussed here appears to be the same as that of the rodent and human repeats, implying that polyadenylation may be one function provided by some individual SINEs in all mammals.

Another function provided by a small subset of SINEs is to encode portions of polypeptides. Human Alu repeats account for approximately 32 codons in the 3′ portion of the genes for decay-accelerating factor and for a B-cell growth factor (Caras et al. 1987; Sharma et al. 1987). Also, the CCAAT box of the θ-globin gene in higher primates is part of an Alu repeat sequence (Kim et al. 1989). Thus the known examples of functional sequences provided by SINEs include promoter, RNA-processing, and protein-coding sequences.

Transposable elements can provide a selective advantage to bacteria in competitive growth conditions (Hartl et al. 1983), and the independent propagation of repetitive elements in many different mammalian orders (Hardison and Printz 1985; Rogers 1985; Sawada et al. 1985; Weiner et al. 1986; Margot et al. 1989) suggests that they may provide some selective advantage in higher organisms as well. However, the molecular mechanisms for this advantage are not clear. The results presented here show that some individual copies of repetitive elements can contribute sequences useful to the organism in which they occur. Although most copies of repetitive elements examined to date appear to have little effect on the expression of neighboring genes, in fact a notable portion of C repeats (three of 36) so far analyzed contribute functional sequences to the rabbit genome. Along with other more general positive effects that the repeats might contribute, including healing chromosome breaks (Voliva et al. 1984) and interruptions of gene conversion (Schimenti and Duncan 1984), the results reported here suggest a mechanism by which some SINEs can provide a selective advantage to the organism in which they occur.

Sequence Availability

These sequences have been deposited in GenBank under accession number M28239.

Acknowledgments

We thank Andrew G. Clark for helpful discussions and the Biocomputing Center at the Penn State Biotechnology Institute for use of their facilities. This work was supported by Public Health Service grant DK27635 and by RCDA DK01589 to R.C.H.

LITERATURE CITED

BOGGARAM, V., K. QING, and C. R. MENDELSON. 1988. The major apoprotein of rabbit pulmonary surfactant. J. Biol. Chem. **263**:2939–2947.

CARAS, I. W., M. A. DAVITZ, L. RHEE, G. WEDDELL, D. W. MARTIN, and V. NUSSENZWEIG. 1987. Cloning of decay-accelerating factor suggests novel use of splicing to generate two proteins. Nature **325**:545–549.

CHENG, J.-F., D. E. KRANE, and R. C. HARDISON. 1988. Nucleotide sequence and expression of rabbit globin genes ζ1, ζ2 and ζ3. J. Biol. Chem. **263**:9981–9993.

CHENG, J.-F., R. PRINTZ, T. CALLAGHAN, D. SHUEY, and R. C. HARDISON. 1984. The rabbit C family of short interspersed repeats: nucleotide sequence determination and transcriptional analysis. J. Mol. Biol. **176**:1–20.

CHUNG, L. P., K. SATISH, and S. GORDON. 1988. Cloning the human lysozyme cDNA: inverted Alu repeat in the mRNA and in situ hybridization for macrophages and Paneth cells. Proc. Natl. Acad. Sci. USA **85**:6227–6231.

DEKA, N., E. WONG, A. G. MATERA, R. KRAFT, L. A. LEINWAND, and C. W. SCHMID. 1988. Repetitive nucleotide sequence insertions into a novel calmodulin-related gene and its processed pseudogene. Gene **71**:123–134.

DEVEREUX, J., P. HAEBERLI, and O. SMITHIES. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.

DOOLITTLE, W. F., and C. SAPIENZA. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature **284**:601–603.

GIL, A., and N. J. PROUDFOOT. 1984. A sequence downstream of AAUAAA is required for rabbit β-globin mRNA 3′-end formation. Nature **312**:473–474.

HARDISON, R. C., and R. PRINTZ. 1985. Variability within the rabbit C repeats and sequences shared with other SINEs. Nucleic Acids Res. **13**:1073–1088.

HARTL, D. L., D. E. DYKHUIZEN, R. D. MILLER, L. GREEN, and J. DE FRAMOND. 1983. Transposable element IS50 improves growth rate of *E. coli* cells without transposition. Cell **35**:503–510.

HWU, H. R., J. W. ROBERTS, E. H. DAVIDSON, and R. J. BRITTEN. 1986. Insertion and/or deletion of many repeated DNA sequences in humans and higher ape evolution. Proc. Natl. Acad. Sci. USA **83**:3875–3879.

JELINEK, W. R., and C. W. SCHMID. 1982. Repetitive sequences in eukaryotic DNA and their expression. Annu. Rev. Biochem. **51**:813–844.

KAZAZIAN, H. H., C. WONG, H. YOUSSOUFIAN, A. F. SCOTT, D. G. PHILLIPS, and S. E. ANTONARAKIS. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature **332**:164–166.

KIM, J.-H., C.-Y. YU, A. BAILEY, R. HARDISON, and C.-K. J. SHEN. 1989. Unique sequence organization and erythroid cell-specific nuclear factor-binding of mammalian θ1 globin promoters. Nucleic Acids Res. **17**:5687–5700.

KING, D., L. N. SNIDER, and J. B. LINGREL. 1986. Polymorphism in an androgen-regulated mouse gene is the result of the insertion of a B1 repetitive element into the transcription unit. Mol. Cell. Biol. **6**:209–217.

KRESS, M., Y. BARRA, J. G. SEIDMAN, G. KHOURY, and G. JAY. 1984. Functional insertion of an Alu type 2 (B2 SINE) repetitive sequence in murine class I genes. Science **226**:974–977.

MARGOT, J. B., G. W. DEMERS, and R. C. HARDISON. 1989. Complete nucleotide sequence of the rabbit β-like globin gene cluster: analysis of intergenic sequences and comparison with the human β-like globin gene cluster. J. Mol. Biol. **205**:15–40.

NEEDLEMAN, S. B., and C. D. WUNSCH. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**:443–453.

OKINO, S. T., L. C. QUATTROCH, H. J. BARNES, S. OSANTO, K. J. GRIFFIN, E. F. JOHNSON, and R. H. TUKEY. 1985. Cloning and characterization of cDNAs encoding 2,3,7,8-tetrachlorodibenzo-*p*-dioxin-inducible rabbit mRNAs for cytochrome P-450 isozymes 4 and 6. Proc. Natl. Acad. Sci. USA **82**:5310–5314.

ORGEL, L. E., and F. H. C. CRICK. 1980. Selfish DNA: the ultimate parasite. Nature **284**:604–607.

PEARSON, W. R., and D. J. LIPMAN. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85**:2444–2448.

PROUDFOOT, N. J., and G. BROWNLEE. 1976. 3′ Non-coding region sequences in eukaryotic messenger RNA. Nature **263**:211–214.

REBIERE, M., P. N. MARCHE, and T. J. KINDT. 1987. A rabbit class I major histocompatibility complex gene with a T cell-specific expression pattern. J. Immunol. **139**:2066–2074.

ROGERS, J. H. 1985. The origin and evolution of retroposons. Int. Rev. Cytol. **93**:187–279.

ROTHKOPF, G. S., C. A. TELAKOWSKI-HOPKINS, R. L. STATISH, and C. B. PICKETT. 1986. Multiplicity of glutathione S-transferase genes in the rat and association with a type 2 Alu repetitive element. Biochemistry **25**:993–1002.

RYSKOV, A. P., P. L. IVANOV, D. A. KRAMEROV, and G. P. GEORGIEV. 1984. Universal orientation and 3′-terminal localization of repetitive sequences of the B2 family in mRNA. Biochemistry **18**:74–83.

SADOFSKY, M., S. CONNELLY, J. L. MANLEY, and J. C. ALWINE. 1985. Identification of a

sequence element on the 3′ side of AAUAAA which is necessary for simian virus 40 late mRNA 3′-end processing. Mol. Cell. Biol. **5**:2713–2719.

SAWADA, I., C. WILLARD, C.-K. J. SHEN, B. CHAPMAN, A. WILSON, and C. W. SCHMID. 1985. Evolution of Alu family repeats since the divergence of human and chimpanzee. J. Mol. Evol. **22**:316–329.

SCHIMENTI, J. C., and C. H. DUNCAN. 1984. Ruminant globin gene structures suggest an evolutionary role for Alu-type repeats. Nucleic Acids Res. **12**:1641–1655.

SCHMID, C. W., and C.-K. J. SHEN. 1985. Pp. 323–358 *in* R. J. MACINTYRE, ed. Molecular evolutionary genetics. Plenum, New York.

SHARMA, S., S. METHA, J. MORGAN, and A. MAIZEL. 1987. Molecular cloning and expression of a human B-cell growth factor gene in *Escherichia coli*. Science **235**:1489–1492.

SINGER, M. F. 1982. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell **28**:433–434.

VOLIVA, C. F., S. J. MARTIN, C. A. HUTCHISON, and M. H. EDGELL. 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. J. Mol. Biol. **178**: 795–813.

WEINER, A. M., P. L. DEININGER, and A. EFSTRATIADIS. 1986. Nonviral retroposons: genes, pseudogenes and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. **55**:631–661.

ALAN M. WEINER, reviewing editor

Received June 9, 1989; revision received September 8, 1989

Accepted September 8, 1989