MOLECULAR AND CELLULAR BIOLOGY, Jan. 1987, p. 231–236
0270-7306/87/010231-06$02.00/0

Vol. 7, No. 1

Nucleotide Sequence Analysis of a Cloned DNA Fragment from Human Cells Reveals Homology to Retrotransposons

ROLF M. FLÜGEL,[1]* BERND MAURER,[1] HELMUT BANNERT,[1] AXEL RETHWILM,[1] PAUL SCHNITZLER,[2] AND GHOLAMREZA DARAI[2]

Institute for Virus Research, German Cancer Research Center, Im Neuenheimer Feld 280,[1] and Institute of Medical Virology of the University, Im Neuenheimer Feld 324,[2] 6900 Heidelberg, Federal Republic of Germany

Received 19 March 1986/Accepted 4 June 1986

During molecular cloning of proviral DNA of human spumaretrovirus, various recombinant clones were established and analyzed. Blot hybridization revealed that one of the recombinant plasmids had the characteristic features of a member of the long interspersed repetitive sequences family. The DNA element was analyzed by restriction mapping and nucleotide sequencing. It showed a high degree of amino acid sequence homology of 54.3% when compared with the 5'-terminal part of the *pol* gene product of the murine retrotransposon LIMd. The 3' region of the cloned DNA element encodes proteins with an even higher degree of homology of 67.4% in comparison to the corresponding parts of a member of the primate *Kpn*I sequence family.

Members of the *Kpn*I family of long interspersed (LI) repetitive DNA sequences are repeated approximately 10^4 times per haploid genome (19). These DNA elements and related DNA sequence families have been identified in other mammals (4, 19). In some cases, rearranged and truncated forms of LI repetitive sequences have been isolated and characterized (14). In several cases, transpositions have been reported that affect cellular gene expression (3, 6, 12). Until recently there was no definite clue about the function, if any, of the LI repetitive sequence family members. The situation changed when the nucleotide sequences of full-length members of the repetitive DNA family revealed that they have the potential to encode proteins evolutionarily related to the retroviral gene product reverse transcriptase (15, 22). Thus, another link was established between transposable elements and retroviruses, thereby extending the concept of retrotransposons from lower eucaryotes to mammals (1, 2, 23).

Since retrotransposons reveal a broad spectrum of fascinating structural features, i.e., in one case long terminal repeats were reported to be inverted (5), nucleotide sequence analysis is the first step to gain insight into their function. Here, we report the nucleotide sequence analysis of a DNA element that was selected, isolated, and established during molecular cloning of proviral DNA of human spumaretrovirus (HSRV) that had been prepared from human embryonic lung fibroblast cells infected with HSRV.

MATERIALS AND METHODS

Cells and virus. Cells of human embryonic lung fibroblasts (HEL cells) were prepared and propagated as described previously (7). Virus (HSRV) was propagated on HEL cells as described previously (16).

Construction of recombinant plasmids. Native DNA from HSRV-infected HEL cells was extracted, deproteinized, and run on a 0.8% low-melting-point agarose gel. After staining with ethidium bromide, broad DNA bands were divided into five fractions (A to E) and isolated from agarose. Samples of the resulting DNA fractions were rerun and stained with ethidium bromide. Above a broad background, discrete and intense DNA bands became visible, particularly in fraction B, which moved at the approximate position of supercoiled human mitochondrial DNA or were of higher mobility. The *Hin*dIII DNA fragments of fraction B that corresponded to 4 to 6 kilobase pairs were isolated from the low-melting-point agarose gel, purified, and inserted into the *Hin*dIII sites of the pAT153 vector (13, 25). A total of 314 recombinant clones were analyzed. Subcloning of one recombinant clone, pHSRV-H-107, was performed with pUC18 and pUC19. In addition, the recombinant pHSRV-H-107 was amplified in *Escherichia coli* GM33-C119, a *dam* host.

DNA sequence analysis. Labeled DNA fragments were sequenced by the method of Maxam and Gilbert (17) as described previously (9). More than 90% of the sequence was determined from both strands or at least three times when the same strand was used.

Nucleic acid hybridization. DNAs were cleaved with different restriction endonucleases and were separated by agarose slab gel electrophoresis. The DNA fragments were transferred to nitrocellulose sheets and hybridized as described by Southern (21). Portions (0.5 μg) of individual DNA were labeled in vitro as described by Rigby et al. (18). Each sample (25 μl) contained 40 μCi of [α-^{32}P]dCTP (specific activity, 6,000 Ci/mmol; and [α-^{32}P]dATP (specific activity, 3,000 to 6,000 Ci/mmol).

Quantitation of homology. The *pol* and protease regions of the murine retrotransposon (15) and the primate LI sequence (19) were aligned with those of H-107 with the programs of Dayhoff (8).

RESULTS

Homology of *Hin*dIII DNA fragment H-107 to cellular repetitive DNA elements. To determine whether the recombinant clone pHSRV-H-107 is a member of a repetitive DNA sequence family, Southern blot hybridizations of H-107 DNA to DNA from uninfected and infected HEL cells were performed. HSRV-infected cells were used, since one of the recombinant clones had hybridized to cDNA. The results in Fig. 1 indicate positive and comparable hybridization signals of H-107 DNA to DNA bands of both uninfected and infected human HEL cells. The relative intensity of the

* Corresponding author.

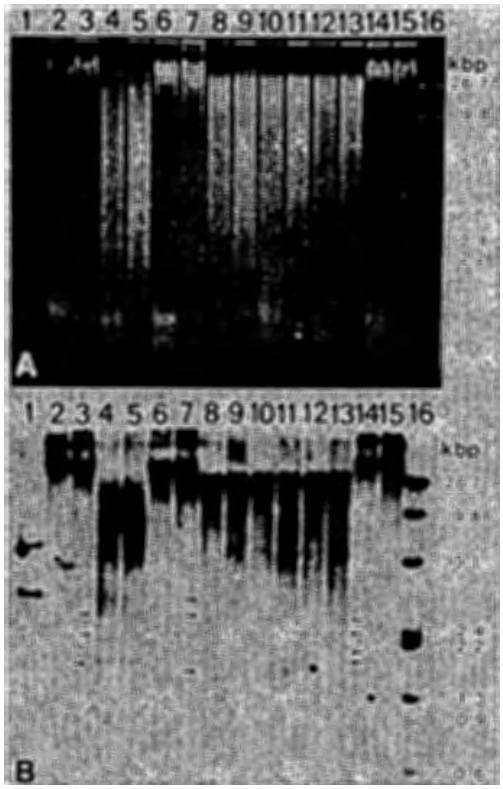

FIG. 1. Southern blot hybridization of recombinant pHSRV-H-107 to human DNAs from uninfected and HSRV-infected HEL cells. DNAs from uninfected cells (lanes 4, 6, 8, 10, 12, and 14) and HSRV-infected cells (lanes 5, 7, 9, 11, 13, and 16) were cleaved with restriction enzymes *Hin*dIII (lanes 4 and 5), *Cla*I (lanes 6 and 7), *Bam*HI-*Cla*I double digestion (lanes 8 and 9), *Bam*HI (lanes 10 and 11), *Bam*HI-*Sal*I double digestion (lanes 12 and 13), and *Sal*I (lanes 14 and 15) and separated electrophoretically on a 0.8% agarose gel. Undigested DNAs (lanes 2 and 3) and 0.06 pM of recombinant HSRV-H-107 DNA digested with *Hin*dIII (lane 1) were analyzed under the same conditions. Lambda DNA cleaved with the *Mlu*I enzyme (0.25 μg of unlabeled DNA and 0.01 μg of ^{32}P-labeled DNA) served as markers (lane 16) and as a control for electrophoretic transfer to nitrocellulose paper. A, Ethidium bromide staining; B, autoradiograph of the same gel after hybridization to ^{32}P-labeled recombinant pHSRV-H-107. Arrows mark positions of defined DNA bands hybridizing to H-107 DNA.

hybridization signals of individual *Hin*dIII DNA fragments indicate that the hybridizing DNA elements were present in multiple copies as expected for a member of a repetitive sequence family.

Nucleotide sequence analysis. Restriction maps of H-107 DNA are shown in Fig. 2. The strategy for determining the major part of the nucleotide sequence of recombinant clone H-107 by the procedure of Maxam and Gilbert (17) is also shown in Fig. 3. The 4,695-base-pair sequence was obtained by sequencing both strands and by sequencing individual fragments several times under different conditions. To minimize sequence errors, multiple-cut restriction enzymes were used to confirm the sequence of subfragments of H-107; in addition, the plasmid H-107 was grown in a *dam E. coli* host, and certain sequences were redetermined by making use of those cleavage sites (e.g., of the *Nde*I, *Cla*I, and *Bcl*I enzymes) that were methylated in the original plasmid H-107 that had been amplified in *E. coli* C600.

The resulting nucleotide sequence (Fig. 3) contains 64.4% A · T base pairs. There are two open reading frames that have retroviral analogs and that were used to orient the map of H-107. Open reading frames longer than 91 amino acid residues were not found in the opposite strand. The major open reading frame located closest to the 3′ terminus starts at nucleotide position 4150 and runs downstream for 142 codons to the *Hin*dIII site at the boundary between the insert of pHSRV-H-107 and pAT153. Further upstream at nucleotide position 3689, another open reading frame precedes the presumed *pol* gene overlapping it for 16 codons (Fig. 3). Unexpectedly, a homology of 54.3% was found when the sequence of the reverse transcriptase of the retrotransposon LIMd (15) was compared with the sequence of recombinant plasmid pHSRV-H-107 (Fig. 4). There are long runs of identical amino acid residues in the NH_2-terminal part of the *pol* gene, although they are derived from different hosts (mouse versus human). The degree of homology increased to 67.4% when this region of H-107 was aligned to a selected domain of the corresponding primate LI sequence. Figure 5 shows a comparison of amino acid sequences between two members of the LI family, namely, the LIMd and the primate LI, to the H-107 DNA element.

It is noteworthy that the region of high homology extends into the presumed retrovirus-like protease gene and abruptly stops eight amino acid residues upstream of the well-conserved domain DTFKAVC that compares to DTMKAFL in the LIMd and to DTFIAVC in the primate LI sequence (Fig. 4 and 5). This domain has been found to be conserved in other protease sequences and is assumed to be part of the catalytic center of the enzyme, because of its obvious similarity to the catalytic domains of cellular serine proteases (24).

DISCUSSION

The analysis of the primary structure of DNA element H-107 indicates that it has one region of strong homology to the predicted protease and to the NH_2-terminal part of reverse transcriptases. Since the degree of homology in amino acid sequences is higher (54.3%) to the corresponding sequence of the murine retrotransposon LIMd and even higher (67.4%) to a corresponding domain of the primate LI sequence (19, 20) than to any of the retroviral sequences, we conclude that H-107 DNA has to be part of human retrotransposon LIHs. It is intriguing that even cleavage sites for certain restriction enzymes are conserved in the primate LI and H-107 sequence, e.g., the *Kpn*I and *Hin*dIII sites at positions 4485 and 4689 (Fig. 3). The result of the positive blot hybridization of H-107 DNA to genomic DNA of human origin is consistent with the assumption that this DNA element is an essential part of a human retrotransposon.

Recently the 6.2 kilobase-pair DNA sequence of a full-length member, TβG41, of the *Kpn*I family of human DNA was reported by Hattori et al. (10). An alignment of the amino acid sequences of TβG41 to those of H-107 required us to assume not only multiple frameshifts but also the suppression of numerous termination codons within the TβG41 sequence. Nevertheless, a high degree of homology was again found to H-107 DNA, similar to that found for the primate LI sequence.

Although some transposable elements like 17.6 and Ty1 are flanked by retroviruslike long terminal repeats, the genes

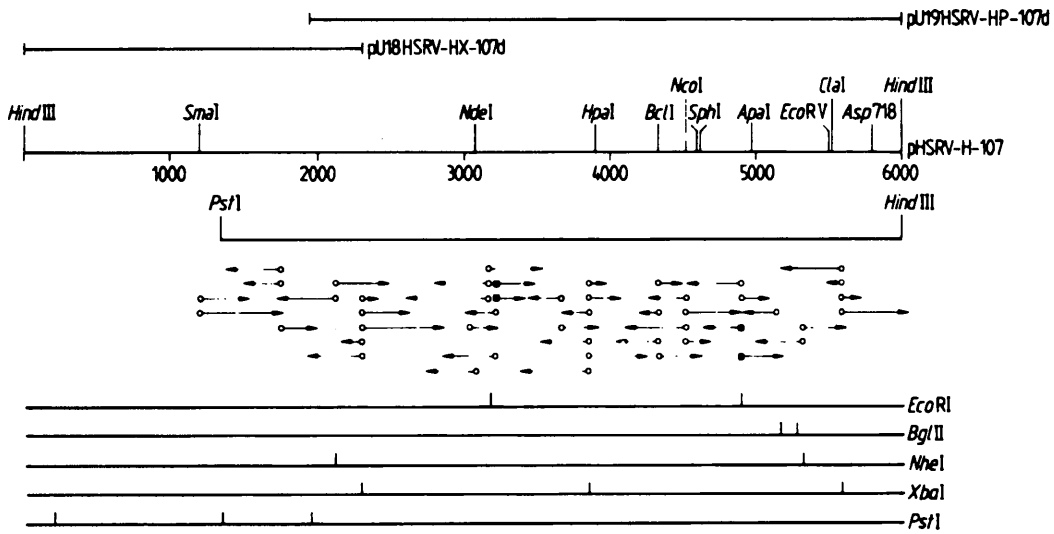

FIG. 2. Restriction maps of 6.0-kilobase-pair pHSRV-H-107 and two subclones and strategy for determining the nucleotide sequence of the *Pst*I-*Hin*dIII DNA fragment. The *pol* gene is on the right, and numbering indicates the distance in base pairs from the *Hin*dIII cleavage site. Recognition sites for restriction endonucleases that occur no more than twice in H-107 are indicated above the map (*Sph*I has two adjacent cleavage sites), whereas recognition sites for the enzymes *Eco*RI, *Bgl*II, *Nhe*I, *Xba*I, and *Pst*I are represented below the maps. DNA was digested with one of these enzymes or *Hpa*II, *Taq*I, and *Hinf*I and end labeled at its 5' termini (■) or 3' termini (○). The arrows below the map indicate the direction and extent of sequences determined for each fragment. The two top lines represent the location and size of two subclones, pU19HSRV-HP-107d and pU18HSRV-HX-H-H107d with respect to pHSRV-H-107. Some of the cleavage sites (*Nde*I, *Cla*I, *Bcl*I) were detectable only after growing H-107 in a *dam* strain of *E. coli*.

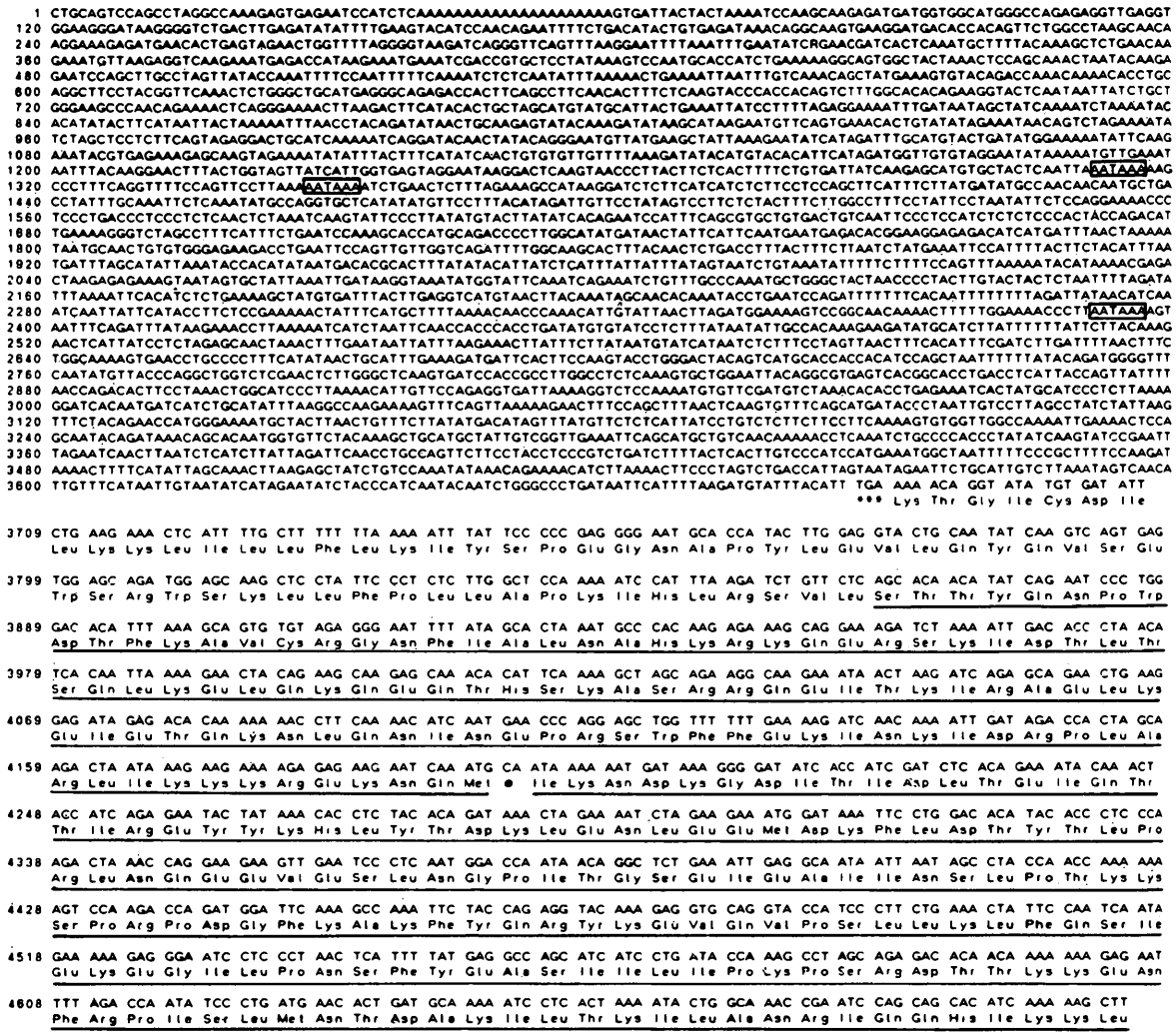

FIG. 3. DNA sequence of the 4,695-base-pair *Pst*I-*Hin*dIII DNA fragment of pHSRV-H-107. The amino acid sequences encoded by the DNA element are shown below the DNA sequences. The regions of strong homology to the murine retrotransposon LIMd (15) and to the primate LI sequence (19) are underlined. ***, Stop codon. Three poly(A) addition signal sequences in the 3'-untranslated region are boxed. ●, Postulated translation frameshift from the protease reading frame to the *pol* reading frame.

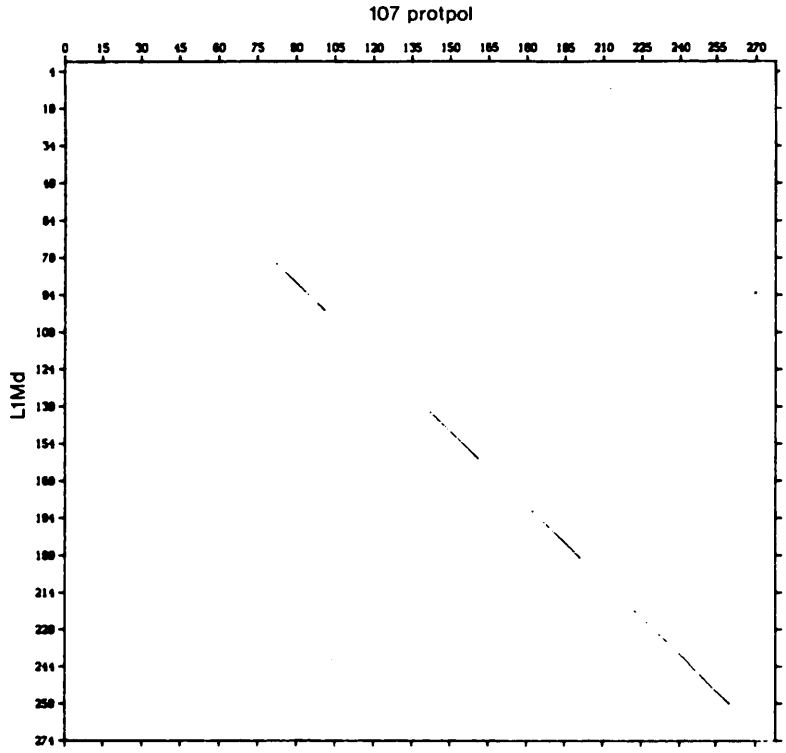

FIG. 4. Homology matrix comparison of the amino acid sequence encoded by the H-107 DNA element (abscissa) with those encoded by the murine retrotransposon LIMd (15). A computer program was used to generate diagonal lines indicating segments of 20 residues that show homology. Boundaries were set as follows: 3865 to 4695 for H-107 (Fig. 3), and 3715 to 4549 for LIMd (15).

of the murine retrotransposon LIMd are flanked at the 5′ end by multiple copies of a 208-base-pair direct tandem repeat and at the 3′ end by an adenine-rich sequence (15). A comparison of the 5′ sequences of H-107 DNA to those of LIMd does not reveal any obvious homology or similarity. There are some short direct repeats at the 5′ end of the sequenced part of H-107 and a tract of 20 adenine residues at position 40. The functional significance of these structures remains unknown.

The predicted protease gene of H-107 is in a different

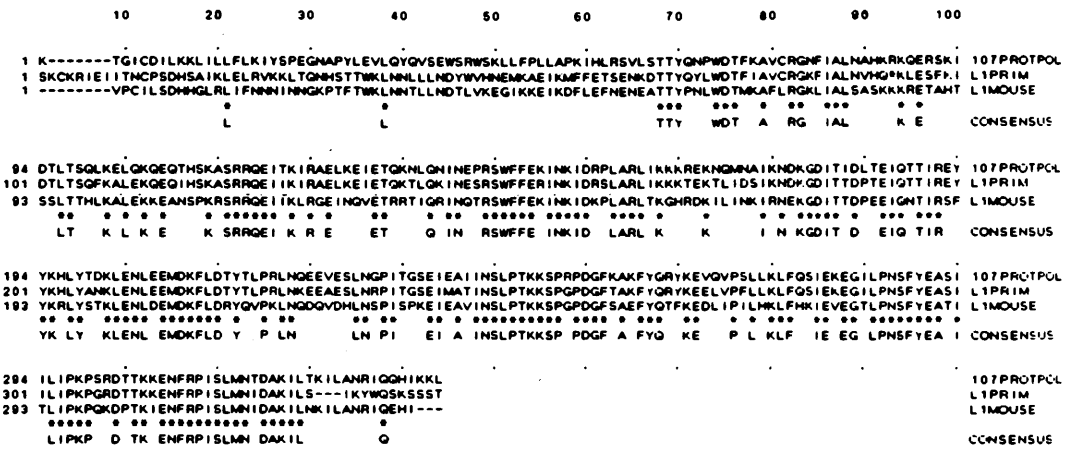

FIG. 5. Comparison of amino acid sequences between the protease and the amino-terminal part of reverse transcriptase of the murine retrotransposon LIMd (15) and of the primate LI sequence (19) to that of pHSRV-H-107. *, Identical amino acid residues. The single-letter code for abbreviating amino acids was used. For proper alignment a frameshift is postulated to occur in the primate LI protease region at position 75 and in the H-107 region at position 170, the boundaries being set to the beginning of the H-107 DNA sequence at the NH_2 terminus. The sequences are 3689 to 4695 for H-107, 2616 to 3645 for primate LI, and 3542 to 4549 for LIMd.

reading frame compared with that of the *pol* gene and overlaps the *pol* gene reading frame for 48 nucleotides. The position and nature of the translational block is a common feature among transposable elements and retroviruses and varies in different genetic elements (11, 26). It is interesting that aligning the H-107 protease region with the primate LI sequence requires a frameshift and thereby restores an aspartic acid and a threonine residue precisely at the site of the postulated catalytic center of serine proteases (24). It is furthermore remarkable that the analogous sequence DTKAVC encoded by the human sequence TβG41 (10) is identical to that of the H-107 sequence.

ACKNOWLEDGMENTS

We thank Howard Temin for critically reading the manuscript. We are grateful to Maxine Singer, National Institutes of Health, for kindly providing us the compiled sequences of the *Kpn*I family including unpublished data, and to M. Sprinzl, University of Bayreuth, for a tape of his compilation of tRNA sequences. We thank P. Loh for providing virus, John Taylor, Fox Chase Cancer Center, Philadelphia, for a gift of oligodeoxynucleotides as primers, and S. Suhai and J. Buchert for help in the computer-assisted programs. We thank H. zur Hausen for encouragement and continuing support.

This work is supported by the Deutsche Forschungsgemeinschaft (Re 647/1-1).

LITERATURE CITED

1. **Baltimore, D.** 1985. Retroviruses and retrotransposons: the role of reverse transcription in shaping the eukaryotic genome. Cell **40**:481–482.
2. **Boeke, J. D., D. J. Garfinkel, C. A. Styles, and G. R. Fink.** 1985. Ty elements transpose through an RNA intermediate. Cell **40**:491–500.
3. **Burton, F. H., D. D. Loeb, S. F. Chao, C. A. Hutchison III, and M. H. Edgell.** 1985. Transposition of a long member of the LI major interspersed DNA family into the mouse beta globin locus. Nucleic Acids Res. **13**:5071–5084.
4. **Burton, F. H., D. D. Loeb, C. F. Voliva, S. L. Martin, M. H. Edgell, and C. A. Hutchison III.** 1986. Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. J. Mol. Biol. **187**:291–304.
5. **Cappello, J., K. Hanselman, and H. F. Lodish.** 1985. Sequence of dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell **43**:105–115.
6. **Cooper, G. M., G. Goubin, A. Diamond, and P. Neiman.** 1986. Relationship of blym genes to repeated sequences. Nature (London) **320**:579–580.
7. **Daral, G., and K. Munk.** 1976. Neoplastic transformation of rat embryo cells with herpes simplex virus. Int. J. Cancer **18**:469–481.
8. **Dayhoff, M. O.** 1978. Survey of new data and computer methods of analysis, p. 1–8. *In* M. O. Dayhoff (ed.), Atlas of protein sequence and structure, vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, D.C.
9. **Flügel, R. M., H. Bannert, S. Suhai, and G. Darai.** 1985. The nucleotide sequence of the early region of the Tupaia adenovirus DNA corresponding to the oncogenic region E1b of human adenovirus 7. Gene **34**:73–80.
10. **Hattori, M., S. Hidaka, and Y. Sakaki.** 1985. Sequence analysis of a Kpn I family member near the 3′ end of human β-globin gene. Nucleic Acids Res. **13**:7813–7827.
11. **Jacks, T., and H. E. Varmus.** 1985. Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. Science **230**:1237–1242.
12. **Katzier, N., G. Rechavi, J. B. Cohen, T. Unger, F. Simoni, S. Segal, D. Cohen, and D. Givol.** 1985. "Retroposon" insertion into the cellular oncogene c-myc in canine transmissible venereal tumor. Proc. Natl. Acad. Sci. USA **82**:1054–1058.
13. **Koch, H.-G., H. Delius, B. Matz, R. M. Flügel, J. Clarke, and G. Darai.** 1977. Molecular cloning and physical mapping of the Tupaia herpesvirus genome. J. Virol. **55**:86–95.
14. **Lerman, M., R. E. Thayer, and M. F. Singer.** 1983. Kpn I family of long interspersed repeated DNA sequences in primates: polymorphism of family members and evidence for transcription. Proc. Natl. Acad. Sci. USA **80**:3966–3970.
15. **Loeb, D. D., R. W. Padgett, S. C. Hardies, W. Shehee, M. B. Comer, M. H. Edgell, and C. A. Hutchison III.** 1986. The sequence of a large LIMd element reveals a tandemly repeated 5′ end and several features found in retrotransposon. Mol. Cell. Biol. **6**:168–182.
16. **Loh, P. C., and F. S. Matsuura.** 1981. The RNA of the human syncytium-forming (foamy) virus. Arch. Virol. **68**:53–58.
17. **Maxam, A., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA **74**:560–564.
18. **Rigby, P. W. J., M. Dieckmann, C. Rhodes, and P. Berg.** 1977. Labeling deoxyribonucleic acid to high specificity in vitro by nick translation with DNA polymerase I. J. Mol. Biol. **114**: 237–256.
19. **Singer, M. F., and J. Skowronski.** 1985. Making sense out of LINES: long interspersed repeat sequences in mammalian genomes. Trends Biochem. Sci. **10**:119–122.
20. **Skowronski, J., and M. F. Singer.** 1985. Expression of a cytplasmid LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proc. Natl. Acad. Sci. USA **82**: 6050–6054.
21. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.
22. **Temin, H. M.** 1981. Origin of retroviruses from cellular moveable genetic elements. Cell **21**:599–600.
23. **Temin, H. M.** 1985. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons and retrotranscripts. Mol. Biol. Evol. **2**:455–468.
24. **Toh, H., R. Kibuno, M. Hayashida, T. Miyata, W. Kugimiya, S. Inouye, S. Yuki, and K. Saigo.** 1985. Close structural resemblance between putative polymerase of a Drosophila transposable genetic element 17.6 and pol gene product of Moloney murine leukaemia virus. EMBO J. **4**:1267–1272.
25. **Twigg, A., and D. Sheratt.** 1980. Trans-complementable copy-number mutants of plasmid Col E1. Nature (London) **283**:216–218.
26. **Varmus, H. E.** 1985. Reverse transcriptase rides again. Nature (London) **314**:583–584.