MOLECULAR AND CELLULAR BIOLOGY, Feb. 1986, p. 411–424
0270-7306/86/020411-14$02.00/0

Vol. 6, No. 2

Structure of the Highly Repeated, Long Interspersed DNA Family (LINE or L1Rn) of the Rat

ETTORE D'AMBROSIO, SUSAN D. WAITZKIN,† FRANK R. WITNEY,‡ ANNE SALEMME,§ AND ANTHONY V. FURANO*

Section on Genomic Structure and Function, Laboratory of Biochemical Pharmacology, National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases, Bethesda, Maryland 20892

Received 25 July 1985/Accepted 14 November 1985

We present the DNA sequence of a 6.7-kilobase member of the rat long interspersed repeated DNA family (LINE or L1Rn). This member (LINE 3) is flanked by a perfect 14-base-pair (bp) direct repeat and is a full-length, or close-to-full-length, member of this family. LINE 3 contains an approximately 100-bp A-rich right end, a number of long (>400-bp) open reading frames, and a ca. 200-bp G+C-rich (ca. 60%) cluster near each terminus. Comparison of the LINE 3 sequence with the sequence of about one-half of another member, which we also present, as well as restriction enzyme analysis of the genomic copies of this family, indicates that in length and overall structure LINE 3 is quite typical of the 40,000 or so other genomic members of this family which would account for as much as 10% of the rat genome. Therefore, the rat LINE family is relatively homogeneous, which contrasts with the heterogeneous LINE families in primates and mice. Transcripts corresponding to the entire LINE sequence are abundant in the nuclear RNA of rat liver. The characteristics of the rat LINE family are discussed with respect to the possible function and evolution of this family of DNA sequences.

Repeated DNA sequences are present in the genomes of all metazoans (6). A class of highly repeated DNA that has been studied in primate and mouse genomes (1, 10, 14, 18, 19, 31, 32, 37, 49) has been called long interspersed repeated DNA (referred to as LINES [52] or, more recently, the L1 family [59]). These families contain long members (several kilobases) that are repeated >20,000 times per genome and are responsible for the prominently stained electrophoretic bands that are seen when total genomic DNA is digested with the appropriate restriction endonuclease; the names of the endonucleases were originally used to denote these families.

Extensive studies on mouse and primate LINEs (see references 43 and 53 for recent reviews) revealed several major features that these families share. First, many cloned members have, at what has been called the right or 3′ end, a putative polyadenylation site, AATAAA, followed by an A-rich sequence (14, 54, 60). Second, although full-length members are 6 to 7 kilobases (kb) long (1, 14, 19), many cloned members are truncated and most often are missing a variable portion from their left end. Furthermore, there appear to be many more genomic copies of the right end than of the left end of certain cloned members of these families (14, 19, 59). As a consequence, the mouse and primate LINE families are quite heterogeneous. Third, both mouse (33) and primate (53) LINE families contain numerous open reading frames (ORFs) that evolved as if they are, or were, bona fide protein-coding sequences (33). Fourth, both primate and mouse LINE families are highly transcribed (22, 26, 29, 47) by RNA polymerase II (20, 50, 56), although there are conflicting results about the extent to which transcripts are poly(A)$^+$ (26, 50) and to what extent LINE transcription may be asymmetric (50, 56). It has been suggested that the truncated LINE members that end in A-rich 3′ ends are incomplete DNA copies (retrotranscripts) of poly(A)$^+$ LINE transcripts (14, 43, 55, 59, 60).

The rat also contains a highly repeated family of transcribed LINE sequences (65). Members of this family have been (or still are) undergoing transposition in the rat genome, since at least three single-copy loci are polymorphic owing to the presence or absence of members of this family (12, 28).

Here we report the DNA sequence of a full-length 6.7-kb member (LINE 3) of the rat LINE (L1Rn) family as well as that of about half of another member (LINE 4). From these results, as well as from those derived from the analysis of other cloned copies of this family and from one- and two-dimensional restriction enzyme analysis of total genomic DNA, we estimate that most of the 40,000 or so copies of this family are ≥6.5 kb long and that together they account for as much as 10% of the rat genome. Therefore, the LINE family of rats is much more homogeneous than the heterogeneous LINE families in mice and primates. LINE 3 is flanked by a 14-base-pair (bp) direct repeat, and it, as well as two other members, has near each end a 150- to 200-bp sequence that is about 60 to 65% G+C. The G+C-rich regions are not homologous to each other or to the consensus sequence of the long terminal repeats of mammalian retroviruses (that also are relatively G+C rich) (9), but contain sequences that are similar to type II DNA synthesis arrest sites of the simian virus 40 (SV40) and parvovirus genomes (61). The rest of the LINE sequence is somewhat more A+T rich (60 to 62%) than total rat DNA (58% A+T [51]), and since the distribution of A is quite asymmetric (one strand is 40% A), the LINE sequence is characterized by stretches of A's (or T's). LINE 3 contains a number of ORFs that begin with an initiation codon and extend for 400 bp or more. LINE 4

* Corresponding author.
† Present address: George Washington University Medical School, Washington, DC 20005.
‡ Present address: Bio-Rad Laboratories, Richmond, CA 94804.
§ Present address: Computer Center Branch, Division of Computer Research and Technology, National Institutes of Health, Bethesda, MD 20892.

contains a 1,941-bp ORF that corresponds to two somewhat shorter ORFs in LINE 3. Hybridization with unfractionated liver nuclear RNA shows that the entire LINE sequence is transcribed. These properties and other structural features of the rat LINE family are discussed with respect to the evolution and function of this family of repeated sequences.

MATERIALS AND METHODS

DNA sequence determination. LINE-containing clones of a λ Charon 4A library of rat DNA (kindly provided by Thomas D. Sargent, National Institute of Child Health and Human Development, Bethesda, Md.) were selected as previously described, by using the repeat DNA clone pR4A1 (65) (see Fig. 3). Two, called λ4A1-3 and λ4A1-4, were extensively characterized. The DNA sequence of the LINE-containing portion of both, as well as all but about 500 bp of the non-LINE portion of one (λ4A1-3) (see upper diagram in Fig. 3), was determined by using the chain termination procedure (44) on the appropriate subclones with the M13 vectors mp18 or mp19 (40). Where appropriate, *Exo*III deletion clones were made by the method described by Henikoff (21). Large portions of LINE segments B and C (see Fig. 3) could not be propagated in either orientation in either of the M13 vectors or their plasmid counterparts, the pUC vectors, although they were readily propagated in pBR322 (3). In the latter vector the DNA segments would not be downstream of the *Escherichia coli* (*LacZ*) promoter or translational start signals, and we reasoned that transcription or translation, or both, of B or C segment sequences were incompatible with vector replication or were lethal to *E. coli*. Since recombinants could not be recovered from the M13 vectors even when the transfectants were plated in the absence of the *lacZ* inducer isopropyl-β-D-thiogalactopyranoside (IPTG), we prepared deleted derivatives of M13mp18 and M13mp19 (referred to as Δ18 and Δ19, respectively), which lack the *lacZ* promoter and its translation initiation site. This was done by digestion of the appropriate replicative form of the M13 vectors with *Ava*II and *Eco*RI (for mp18) or *Ava*II and *Hin*dIII (for mp19). The ends were made flush by treatment with 2.5 U of the Klenow fragment of *E. coli* DNA polymerase I (New England BioLabs) and each deoxynucleoside triphosphate (at 50 μM) in a 10-μl reaction volume for 3 min at 37°C and ligated with the appropriate linker (P-L Biochemicals, Inc.) (*Eco*RI to produce Δ18 and *Hin*dIII to produce Δ19) by using T4 DNA ligase (New England BioLabs).

Since these vectors produce colorless plaques whether or not they contain an insert, they were treated, after digestion with the appropriate restriction endonuclease, with calf intestine phosphatase (Boehringer Mannheim Biochemicals) (0.06 U/800 ng of vector for 1 h at 37°C). The phosphatase was inactivated by heating for 3 h at 70°C, and the vectors were used directly in ligation reactions. So prepared, these vectors produced 0 to 10 colorless plaques per 20 ng of vector after incubation in a ligation reaction without added insert DNA. We encountered no difficulties in propagating, in either orientation, LINE segments as large as 2.5 kb, which is the largest we tried.

The dideoxy reactions were carried out with a reagent kit from New England BioLabs and Klenow enzyme from Bethesda Research Laboratories. Sequencing gels (0.20 mm thick) were polymerized onto glass plates (16). Computer analysis of the DNA sequence was carried out by using previously published programs (8, 63) as well as a program for detecting short regions of homology between long sequences (M. Kanehisa and D. Lipman, personal communication). The Protein Sequence Database of the Protein Identification Resource (release 5.0, May 1985, of the National Biomedical Research Foundation, Georgetown University Medical Center, Washington, D.C.) was searched for proteins homologous to those encoded by the LINE ORFs by using the program SRCHGP which implements the search algorithm described by Wilbur and Lipman (63). This program compares the query sequence with each sequence in the database for overall or "global" homology and returns as output the database sequences that produce the 40 highest alignment scores. Significant homology is readily apparent and is indicated by alignment scores that differ by many standard deviations from the mean score, which is also generated in the search. This mean score can be considered the alignment score that would be generated between the query sequence and one to which it is randomly related (see reference 63 for a detailed explanation of the scoring and the assessment of statistically significant alignments). The search parameters were set to those recommended (63) to give optimal sensitivity and speed. As a consequence, the search was carried out at near-maximal sensitivity (63).

Other techniques. Electrophoresis, blots, hybridizations, and preparation of radioactive hybridization probes were done as described previously (65). Two-dimensional gel electrophoresis (40a) was kindly carried out by Sylvia L. Bunting and Michael M. Seidman (National Cancer Institute, Bethesda, Md.). DNA dots (24) were prepared by using a Schleicher and Schuell Minifold I apparatus.

RESULTS

DNA sequence and overall structure. Figure 1 shows a summary of the sequencing reactions carried out on the LINE family member (LINE 3) and some of the flanking DNA in the λ4A1-3 clone as well as those performed on part of the LINE family member (LINE 4) and flanking DNA in the λ4A1-4 clone. Figure 2 shows the DNA sequences, and Fig. 3 shows a diagram of these sequences, illustrating their overall relationship to each other, as well as some of the structural features derived from the DNA sequence. Figure 3 also shows the region of LINE sequence that corresponds to three previously sequenced repeat DNA clones that we isolated from a library of reannealed rat repeat DNA sequences (65), including pR4A1, which was used to select the present clones. We have divided the rat LINE into four large segments, A, B, D, and C, from left to right. Written in this way, the rat LINE is oriented in the same way that LINE families in mice and primates are usually presented (54).

The LINE 3 member begins at or near the left-hand *Bam*HI site. DNA to the left of the *Bam*HI site does not give a detectable hybridization signal with total radioactive rat DNA (results not shown). Furthermore, this DNA contains a 14-bp sequence that is duplicated at the right end of LINE 3 (see heavy upward-pointing arrows in Fig. 3) and is not present in the corresponding position of LINE 4. Therefore, this 14-bp repeat probably represents a target site duplication that occurred when LINE 3 inserted into its present location. The DNA just beyond the left-hand *Bam*HI site is highly repeated, since it hybridizes quite strongly to total rat DNA (see below). The first 200 bp or so of DNA sequence is unusually G+C rich for rat DNA (65 versus 42% for total genomic DNA [51]). Very shortly, however, the base composition becomes somewhat more A+T rich (60 to 62%) than that of total rat DNA and is quite asymmetric with respect to A content, such that the strand shown in Fig. 2 is about 40% A and therefore contains numerous stretches of A. The composition and asymmetry are retained for the entire

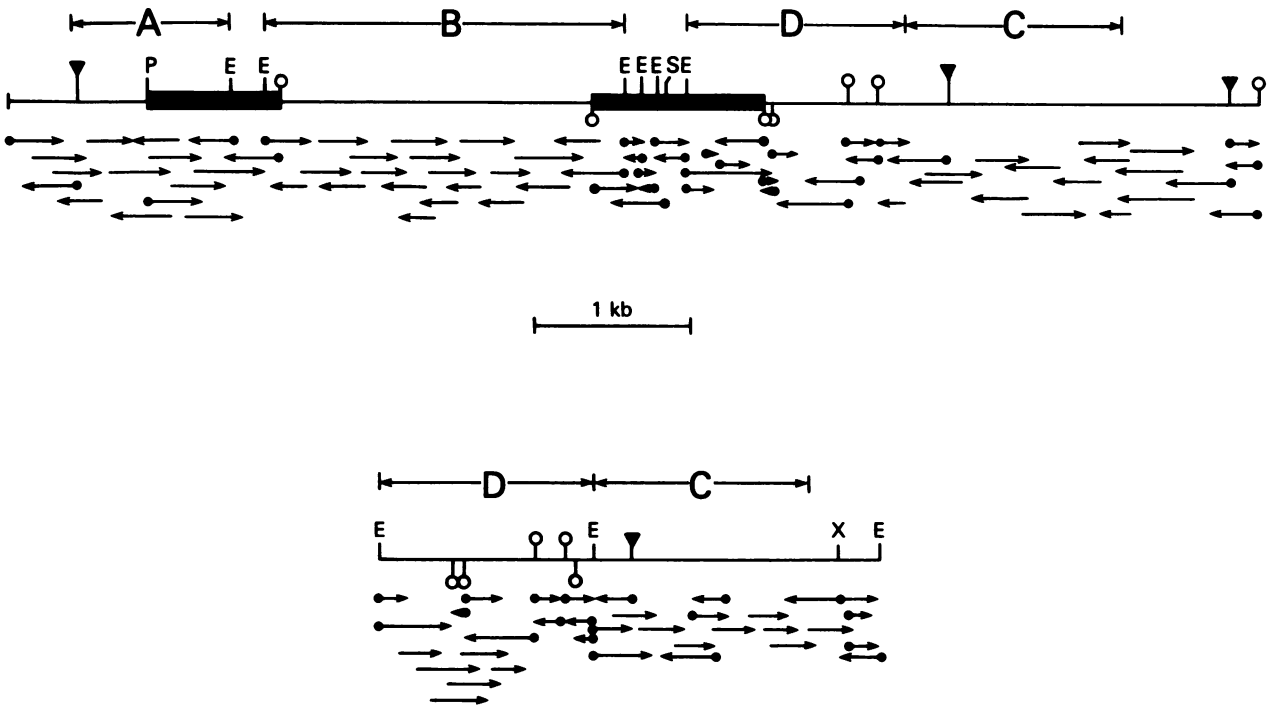


FIG. 1. Summary of the sequencing reactions performed on LINEs 3 and 4. The upper diagram depicts the LINE-containing portion (LINE 3) and some of its flanking DNA from the λ4A1-3 clone, and the lower diagram shows part of the LINE (LINE 4) and its flanking DNA from λ4A1-4. Arrows indicate the portion of sequence obtained with individual clones and point in the direction sequenced. Those without tails represent *Exo*III deletion clones, and those with tails (●) represent clones of the indicated restriction enzyme fragments: P, *Pst*I; E, *Eco*RI; S, *Sph*I; X, *Xba*I; ▼, *Bam*HI; ♀, *Hin*dIII; ♁, *Bgl*II. Arrows with tails that are not aligned with the above restriction sites represent clones of *Sau*3A fragments. The orientation of the *Hin*dIII fragment of segment D, LINE 4, was determined by comparison with the corresponding region of LINE 3. The *Eco*RI sites between the A and B segments are not shown. The heavy bars indicate the *Pst*I (P)-*Hin*dIII (♀) fragment that spans the A and B segments and the *Bgl*II (♁) fragment that spans the B and D segments (see Fig. 3). These fragments were cloned and sequenced to determine the number and orientation of the numerous small *Eco*RI fragments present between the A and B or B and D segments (see Fig. 2 and 3).

length of the repeat unit until the right end is reached, which is at or near the point where the sequences of LINEs 3 and 4 diverge (cross-hatched regions in Fig. 3). Just before this point the composition of both DNA sequences changes somewhat abruptly to a second G+C-rich region, which for 160 to 180 bp reaches 62% in both LINEs 3 and 4. About 85 bp after the G+C-rich region of each member, sequence homology between LINEs 3 and 4 ends, and after about another 45 bp the second member of the 14-bp direct repeat which flanks LINE 3 is reached. Therefore, the C segment of these LINE members is about 1.35 kb long (see below), and the overall length of the LINE 3 member is 6.7 kb.

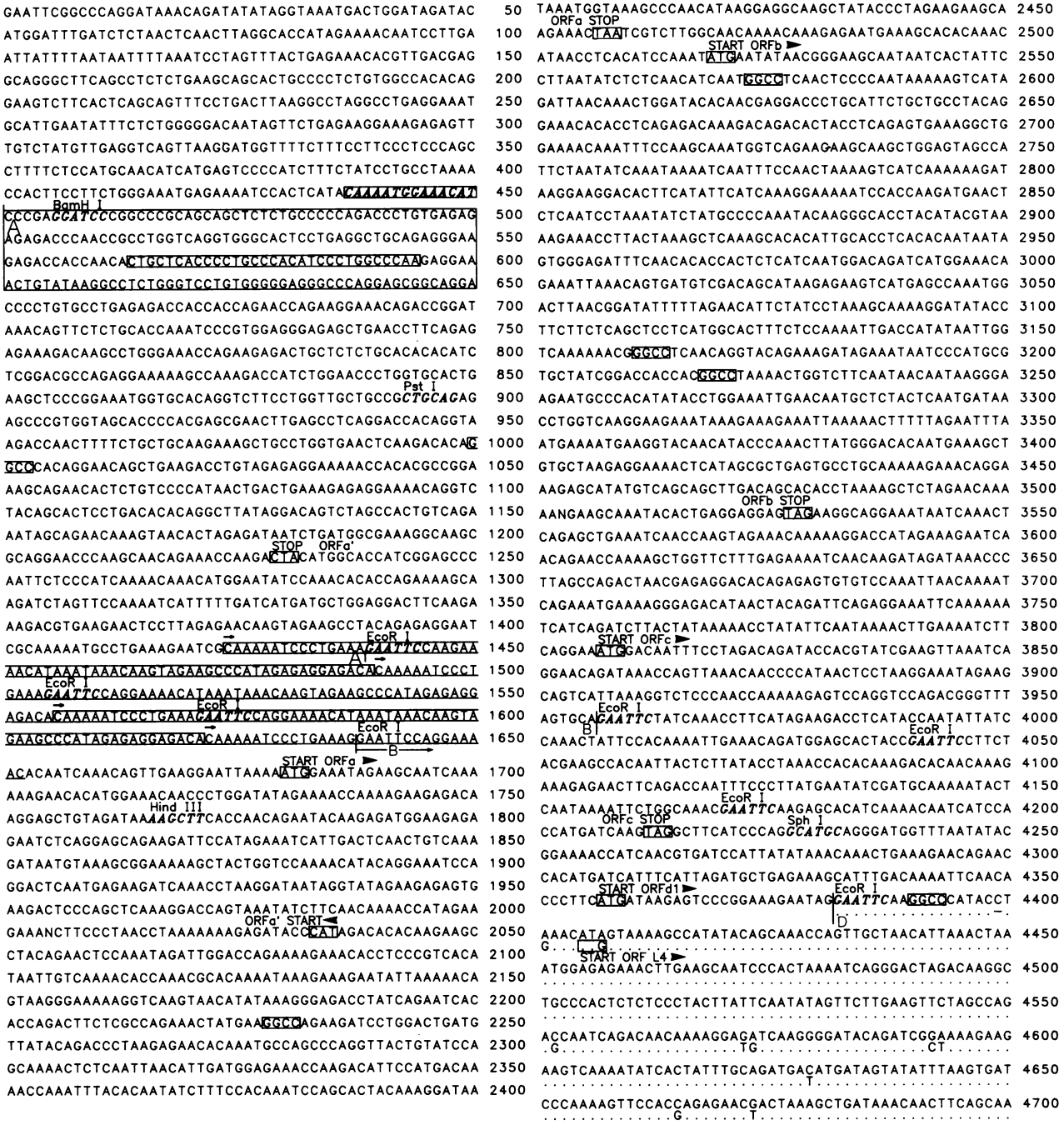
Although DNAs to the right of LINEs 3 and 4 are not homologous, both contain a >30-bp stretch of alternating purines and pyrimidines (A-T, LINE 3; G-T, LINE 4) which have the potential to form Z DNA (39).

The remaining 10 kb of DNA in the λ4A1-4 clone and 4 kb of DNA in the λ4A1-3 clone are not highly repeated, except for a portion of the most rightward *Eco*RI fragment of λ4A1-3 (see upper diagram in Fig. 3). This fragment, but not the ones labeled A30 and A44, hybridized to radioactive total rat DNA (results not shown). DNA sequence determination of the right-hand *Eco*RI fragment showed that it contained a single 180-bp member of a rat short interspersed family (or SINE [52]) that is very homologous to the mouse B2 family (27) (results not shown).

The ca. 200-bp G+C-rich regions that are near each end of the LINE 3 member do not contain palindromes (i.e., inverted repeats), are not homologous to each other, and are not homologous to the relatively G+C-rich (56%) consensus sequence of long terminal repeats of mammalian retroviruses (9). However, each contains sequences that closely resemble one version of type II DNA synthesis arrest sites identified in the SV40 and parvovirus genomes (61) (Fig. 2 legend). Some of the type II arrest sites, which, in contrast to type I sites, do not contain palindromic sequences, arrest DNA synthesis in vitro not only by the eucaryotic α polymerase but also by retroviral reverse transcriptase and the Klenow fragment of *E. coli* DNA polymerase I (61). While carrying out the DNA sequencing, we found that the right-hand G+C-rich regions of both LINEs 3 and 4, but none of the other DNA that we sequenced, contain a strong DNA synthesis arrest site for the Klenow polymerase (Fig. 2, vertical arrows). This was seen when the strand shown in Fig. 2, but not the opposite strand, was being copied (Fig. 2 legend). This behavior is typical of type II sites (61).

The remainder of the LINE sequences contain neither inverted nor direct repeats except for a tandem array of 3.5 copies of a short (65- to 70-bp) sequence between the A and B segments of LINE 3 (Fig. 2 and 3). This tandem array is traversed by one of the ORFs in LINE 3. However, before discussing this and certain other features of the LINE family, we show that the sequenced structures are quite representative of the genomic copies of the rat LINE family.

Relationship between cloned LINE members and genomic copies of the rat LINE family. The 6.7-kb long LINE 3



member contains two *Bam*HI sites 5.5 kb apart that delineate a fragment which contains the B (2.3-kb) and D (1.35-kb) segments. The B segment is flanked by *Eco*RI sites, and although LINE 3 lacks the *Eco*RI site that separates the D and C segments (owing to a single-base deletion), the D segment of seven of eight other cloned copies of rat LINE members (including LINE 4) is also flanked by *Eco*RI sites (Fig. 3) (results not shown).

To determine whether genomic LINE members contain similarly located *Bam*HI or *Eco*RI sites, total rat DNA was digested with either enzyme. The stained electrophoretic gel of the *Bam*HI digestion contained a prominent 5.5-kb band, and the *Eco*RI digestion contained prominent 2.3- and 1.35-kb bands (see below). The 2.3-kb B segment of LINE 3 hybridized almost exclusively to the 2.3-kb *Eco*RI band of genomic DNA, and the 1.35-kb D segment of LINE 4 hybridized almost exclusively to the genomic 1.35-kb *Eco*RI band; both segments hybridized strongly to the 5.5-kb *Bam*HI band as well as to other, mostly larger, fragments (results not shown).

We extended these results by resolving *Eco*RI-*Bam*HI double digests on two-dimensional gels. In these experiments the *Bam*HI fragments are first separated in one dimension and then digested in situ with *Eco*RI (40a). From the data shown in Fig. 3 and the results mentioned above, we expected that a significant number of the genomic copies of

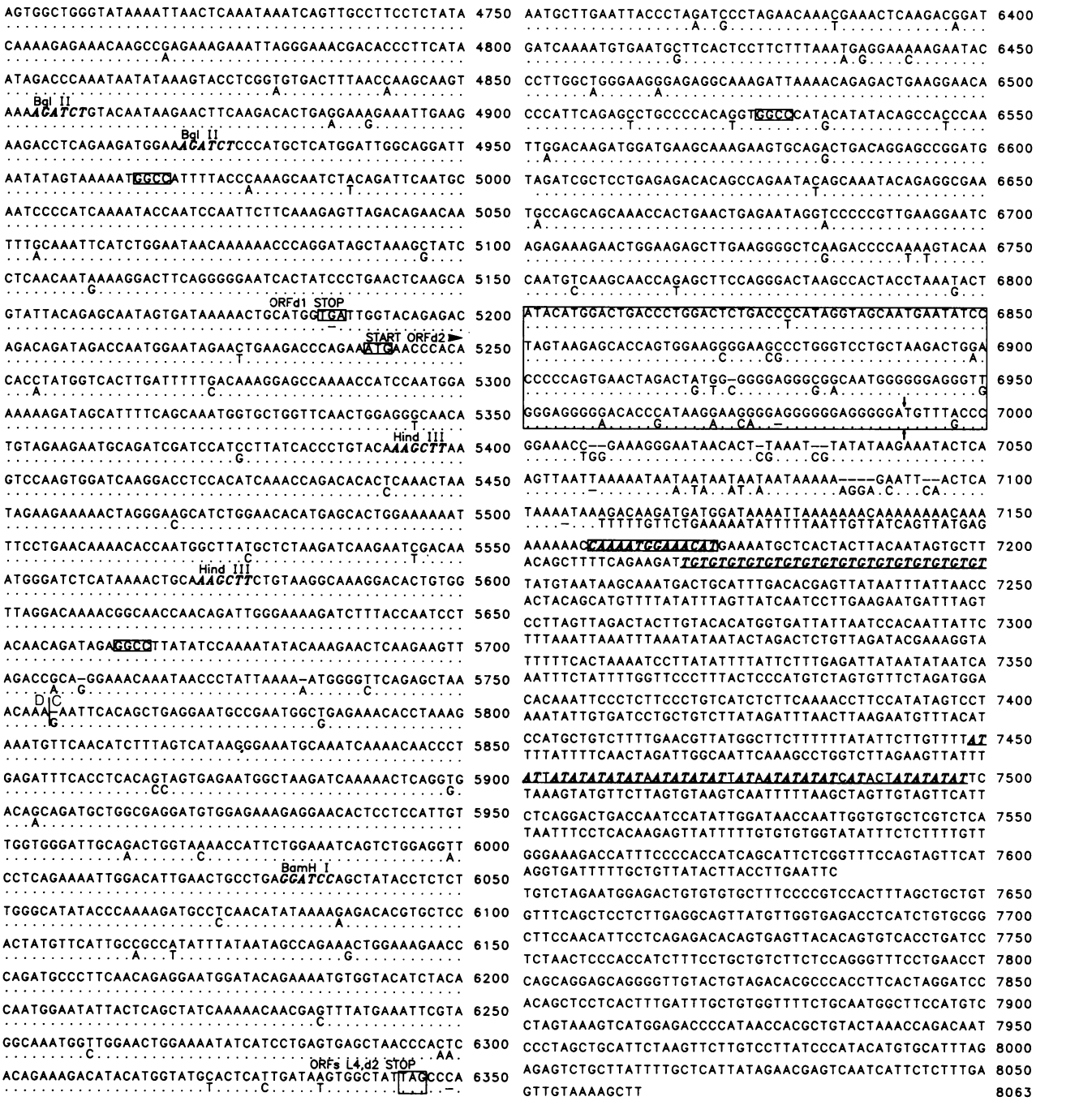


FIG. 2. DNA sequence of rat LINE family members. The top sequence is that of the LINE 3 member (see Fig. 3, middle diagram), and the bottom one is of the LINE 4 member (see Fig. 3, lower diagram), written as a ● where it is the same as LINE 3 or as a — or a letter where it is not. Some of the restriction endonuclease sites are indicated, and the relevant *Hae*III sites (see Fig. 5) are boxed. The boundaries of segments A, B, D, and C (see Fig. 3) are indicated by vertical lines. The direct repeats that flank LINE 3 are italicized and boxed, and members of the tandem array of 3.5 ca. 65-bp sequences in LINE 3 are alternately boxed and underlined and marked with arrows. The initiation and termination codons and sense of the ORFs are also indicated. The stretches of alternating purines and pyrimidines that are just beyond the right end of LINEs 3 or 4 are italicized and underlined, and the G+C-rich regions near the ends of LINEs 3 and 4 are demarcated in large boxes. The box within the left-hand G-C cluster of LINE 3 has the same motif as type II DNA arrest sites (see text) in SV40 or parvovirus DNA, and the vertical arrows within the right-hand G-C cluster of LINEs 3 and 4 show where DNA synthesis by the Klenow polymerase is arrested in vitro. This stop occurs only when the strand shown is being copied (see text), and the G-rich region to the left of the arrows has the same motif as certain type II DNA synthesis arrest sites in SV40 DNA (61). The DNA from positions 4920 to 6346 is 85% homologous, with no gaps, to the corresponding portion of the mouse LINE family (mouse sequence from reference 60). Position 6346, which is the last nucleotide of the ORFd2 (and ORF4) stop codon, corresponds to the last nucleotide of the mouse LINE ORF stop codon (33). From this position to their respective right ends, each of which is about 750 bp beyond their ORF stop codons, the mouse and rat LINE sequences abruptly diverge and show only 31% homology, or about that expected from chance alone. The rat D segment corresponds to the mouse MIF-1 sequence, and the C segment sequence between the right-hand *Bam*HI site to about 200 bp beyond the stop codon corresponds to the mouse *Bam* 5 sequence. The mouse and rat sequences are not homologous over these 200 bp (see above). The remainder of the rat LINE C segment corresponds to but is not homologous to the mouse R sequence (see references 14, 54, and 59 for alignment of the mouse sequences).

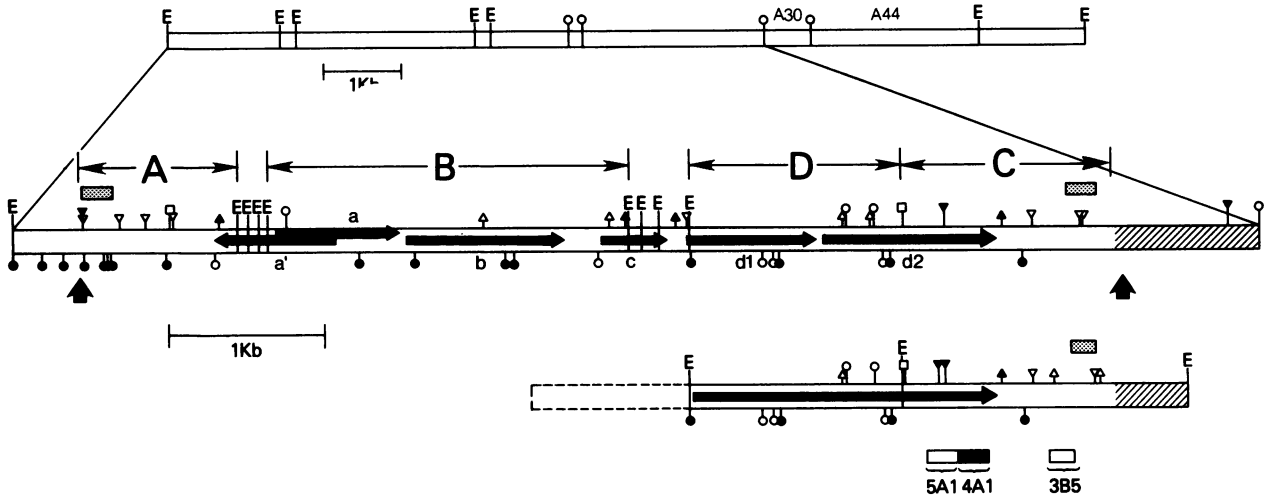


FIG. 3. Diagrammatic representation of rat LINE family members. The top diagram shows a partial restriction endonuclease map of the genomic clone, λ4A1-3, in which the LINE 3 member (middle diagram) resides. The terminal *Eco*RI sites (E) of the top diagram are those at the junction between rat and vector DNA. The lower diagram shows all of the LINE 4 member that is present in another genomic clone, λ4A1-4. The region indicated with dashed lines abuts vector DNA, and partial (60%) sequence determination of this region (data not shown) showed that it is homologous to the corresponding regions of LINE 3 that are displayed immediately above it (i.e., part of segment B and the *Eco*RI fragments between the B and D segments). The remainder of the LINE 4 member and all of the LINE 3 member (middle diagram) were completely sequenced (Fig. 1 and 2). The letters, A, B, D, and C, indicate the LINE segments, and the right end of the C segments is assumed to be at or near the point of divergence of the two sequences, which is indicated by the cross-hatched region (see text). In addition to *Eco*RI (E), the sites for some other retriction endonuclease sites are also shown: *Hin*dIII (▽), *Bam*HI (▼), *Pvu*II (▽), *Hae*III (▲), *Bgl*II (△), *Msp*I (▽), *Taq* (▽), and *Bcl*I (▽). The heavy solid arrows contained within the diagrams represent ORFs and are referred to by the lower-case letters immediately below or above them. The heavy up arrows under LINE 3 indicate the position of the direct repeat that flanks this member. The shaded bars above each terminal region of LINE 3 and above the right end of LINE 4 indicate the G+C-rich regions, and the open and filled boxes under LINE 4 that are labeled 5A1, 4A1, and 3B5 indicate the LINE sequence in previously described clones that were isolated from the repeated DNA fraction of rat DNA (65). (See text for additional details.)

the A, B, and D LINE segments should be released from the 5.5-kb *Bam*HI fragments by *Eco*RI digestion. The ethidium bromide-stained two-dimensional gel (Fig. 4, upper left-hand panel) shows that three prominent *Eco*RI fragments (B, 2.3 kb; D, 1.35 kb; A, 1 kb) are released from the 5.5-kb *Bam*HI band by *Eco*RI digestion. Each hybridizes strongly to total rat DNA and therefore contains highly repeated DNA (Fig. 4; panel tot).

We next hybridized a series of blots of double *Eco*RI-*Bam*HI digests of total rat DNA with clones of LINE segments A, B, D, and C, respectively. Since only one prominent, repeat DNA-containing band was present at the level of the 1-kb A fragment, we used a blot of a one-dimensional gel for hybridization to the A segment clone (Fig. 4A, lower right-hand corner). As the stained gel shows (Fig. 4A, left photograph), the B, D, and A bands are clearly visible in the double *Eco*RI-*Bam*HI digest (lane EB), and hybridization of the A segment clone is mainly to the 1-kb A band (lane EB, right photograph).

Panels B, D, and C show the hybridization of the clones of these respective LINE segments to blots of two-dimensional gels. The clones of the B or D segments hybridize mainly to the 2.3-kb (B) or 1.35-kb (D) *Eco*RI bands, respectively, derived from the 5.5-kb, or larger, *Bam*HI fragments. Figure 4D, as well as less exposed autoradiograms of it and the autoradiograms shown in panels B and tot, indicate that at least one-half of the 2.3-kb (B) and 1.35-kb (D) *Eco*RI fragments are derived from 5.5-kb *Bam*HI fragments. These results, along with those shown in panel A, indicate that at least half of the genomic copies of the A, B, and D LINE segments are present in 5.5-kb *Bam*HI fragments, as depicted in the diagram at the top of Fig. 4. This diagram was derived from our sequence data of LINE members 3 and 4 and from partial restriction enzyme analysis and hybridization studies of five other LINE-containing genomic clones (results not shown). Almost all of the other genomic copies of the A, B, and D LINE segments are in larger *Bam*HI fragments, which could be due to the absence of either of the *Bam*HI sites (diagram at the top of Fig. 4).

Figure 4C shows that the C segment clone hybridizes largely to a smear of different-sized *Eco*RI-*Bam*HI fragments that stopped abruptly at 1.1 kb. (The 0.6- and 0.7-kb *Eco*RI-*Bam*HI fragments which account for only a small percentage of the total hybrids are not accounted for by the structures diagrammed in Fig. 3.) The smear of hybridization is expected because most of the C segment DNA would be contained in *Eco*RI-*Bam*HI fragments that span the right-hand LINE–non-LINE DNA junctions (diagram at the top of Fig. 4). Since LINE members are interspersed in the genome (65), non-LINE restriction enzyme sites should be located randomly with respect to the ends of the various LINE members. Therefore, the restriction enzyme fragments that contain these C segment sequences should be random in size down to the distance between the *Bam*HI site in the C segment and the right end of the C segment (diagram at the top of Fig. 4).

The DNA sequences of LINEs 3 and 4 (Fig. 2) show that the C segments of these LINE members extend to the right about 1.1 kb beyond the *Bam*HI site (Fig. 2 and 3). Since this distance corresponds to the sharp demarcation at 1.1 kb of the C segment hybridization pattern (Fig. 4C), then most genomic LINE members extend at least 1.1 kb beyond the right-hand LINE *Bam*HI site or at least 1.35 kb beyond the *Eco*RI site that separates the D and C segments of most

LINE members (see above). Therefore, the minimum length of the genomic C segment *Eco*RI fragments should be 1.35 kb; this is verified by Fig. 5a, which shows an *Eco*RI digest of total rat DNA hybridized to a C segment clone.

The results in Fig. 4A can also be analyzed as above and indicate that the minimum length of the genomic A segment *Eco*RI fragments (lane E) is slightly larger than the 1-kb *Bam*HI-*Eco*RI double-digestion fragment (lane EB). This indicates that the genomic copies of this family extend somewhat beyond the left-hand *Bam*HI site. The discrete *Eco*RI fragment that is slightly longer than 1 kb (lane E) accounts for about 5% of the total hybridization of the A segment clone to this blot. Therefore, a small number of LINE members may actually contain an *Eco*RI site here, which would mean that some members may extend even further to the left than this. Therefore, LINE 3 may not be a true full-length member.

Regardless of this, all of the above results indicate that almost all of the genomic A, B, D, and C segments are organized in structures that are quite similar to the 6.7-kb LINE 3 member. Furthermore, as we show below, each of these segments is present to about the same extent in the rat genome.

The DNA sequences shown in Fig. 2, as well as that of the LINE at the insulin 1 locus in certain rats (28), indicated that the locations of *Hae*III sites in the D and C segments were highly conserved. Therefore, to further examine the relationship between the cloned and genomic LINE members, we digested total rat DNA with *Hae*III and hybridized blots of electrophoretic gels of the digest with clones of the C, D, or B LINE sequence. Figure 5d shows that *Hae*III produces a series of stained bands that clearly stand out over the background and give somewhat broad but well-defined densitometric peaks, the sizes of some of which are also given. A *Hae*III digest of LINE 3 would contain LINE fragments (to the nearest 0.05 kb), going from right to left, of 0.85 kb (mostly C segment), 0.7 and 0.55 kb (D segment), and 1.2, 0.6, 0.35, and 1.25 kb (B segment) (Fig. 5, top diagram); these correspond to most of the bands observed on the stained gels. Figure 5b shows that most of the hybridization of C, D, and B segment clones to total rat DNA are to *Hae*III fragments of the expected size.

The relationship between the genomic *Hae*III fragments detected by the above LINE segment clones and the position of *Hae*III sites in LINE 3 (and in LINE 4) is not fortuitous, because another randomly selected rat LINE clone (containing just the 5.5-kb *Bam*HI fragment) that was sequenced by M. B. Soares, E. Schon, and A. Efstratiadis (personal communication) contains almost exactly the same distribution of *Hae*III sites that is shown at the top of Fig. 5.

We also mapped the location of the *Eco*RI and *Bam*HI sites with respect to the *Hae*III site in genomic copies of the C segment by hybridization of the relevant double digests of total rat DNA with the C segment clone pR4A1 (Fig. 5c). These results confirm the conclusion that most of the LINE family members have an *Eco*RI site separating the D and C segments (see above), and also show that the right-hand *Bam*HI site within the C fragment is highly conserved.

Representation of LINE segments in nuclear RNA and genomic DNA. Figure 6 shows the hybridization of radioactive liver nuclear RNA to an *Eco*RI-*Bam*HI restriction enzyme digest of total rat DNA that was resolved in two dimensions. The B, D, and A bands hybridize well to nuclear RNA, as do several other discrete *Eco*RI-*Bam*HI fragments that do not correspond to prominent ethidium bromide-stained bands (compare Fig. 6 with Fig. 4). Nuclear RNA did not hybridize to the DNA that is to the left of the left-hand *Bam*HI site (Fig. 2 and 3) (results not shown). Since earlier results (65) showed that the C segment repeat clones pR4A1, pR5A1, and pR3B5 (Fig. 3) also hybridize strongly to unfractionated liver nuclear RNA (a result which we verified by using a clone containing the entire C segment; data not shown), the entire LINE is transcribed in liver. Preliminary results with total RNA from kidney, muscle, and a rat cell line indicate that the entire rat LINE is also transcribed in these cells (data not shown).

To determine the relative genomic copy number of each LINE segment, total rat DNA was hybridized to dots of M13 clones containing either of the complementary strands of the A, B, D, or C segments. The dots were also hybridized with M13 DNA. Except for the C segment, the amount of hybridization of total rat DNA to each clone is proportional to the length of the repeated DNA segment in the clone (for quantitation see Fig. 6, top diagram and legend). These results, and others not shown (Fig. 6 legend), indicate that whereas LINE segments A, B, and D are present to about the same extent in the genome, there is an excess (1.2- to 1.6-fold) of C segment sequences.

ORFs in the rat LINE sequence. LINE 3 contains six ORFs that are >400 bp long; each begins with an initiation codon and ends with a termination codon (Fig. 2 and 3). ORFa′ (807 bp) is oriented opposite to the other five: ORFa (780 bp), ORFb (1,011 bp), ORFc (408 bp), ORFd1 (831 bp), and ORFd2 (1,104 bp). LINE 4 also contains an ORF (ORF4; 1,941 bp), which corresponds to ORFd1 and ORFd2 and terminates at the same position as ORFd2 (Fig. 2). The LINE sequence upstream of the insulin 1 gene in certain rats (LINE I1) also contains an ORF (ORFI1) that initiates and terminates in exactly the same positions as ORFd2 (28; Soares et al., personal communication). The rat LINE C segment sequence published by Scarpulla (45) and the C segment of the LINE at the *Mlvi-2* locus (LINE-*Mlvi-2*) in some rats (12; A. V. Furano, C. C. Somerville, P. N. Tsichlis, and E. D'Ambrosio, manuscript in preparation) also contain ORFs. These C segment sequences begin at the *Eco*RI site that separates the D and C segments, and their ORFs, beginning with the first three nucleotides of the sequence, GAA, are in phase with ORF4 and terminate at the same position as ORF4 (and ORFd2 and ORFI1). Note that ORFd2 contains three single-base deletions in a 50-bp region (positions 5709 to 5756; Fig. 2). These deletions are not present in ORF4 or ORFI1, and if only one or two of these were present, ORFd2 would terminate within 100 bp of the remaining one(s).

We compared the LINE ORFs with other protein-encoding sequences in four ways. First, we found no significant homology between the putative proteins encoded by any of the LINE ORFs and the May 1985 contents of the Protein Sequence Database of the Protein Identification Resource. However, since the program we used was designed to detect global, or overall, homology between proteins (see Materials and Methods) (63), short, local regions of similarity would not have been revealed.

Second, we found that all of the ORFs differ significantly in their codon usage from that of other mammals (17) for most of the 18 amino acids for which a codon choice is possible. For all the ORFs except ORFa′, codon usage was in most cases correlated with the A content of the codons, which is not surprising, given the asymmetric distribution of A in these ORFs (see above). For most of the codons for which A content would not bias codon use, the usage of codons in the LINE ORFs could be correlated with the

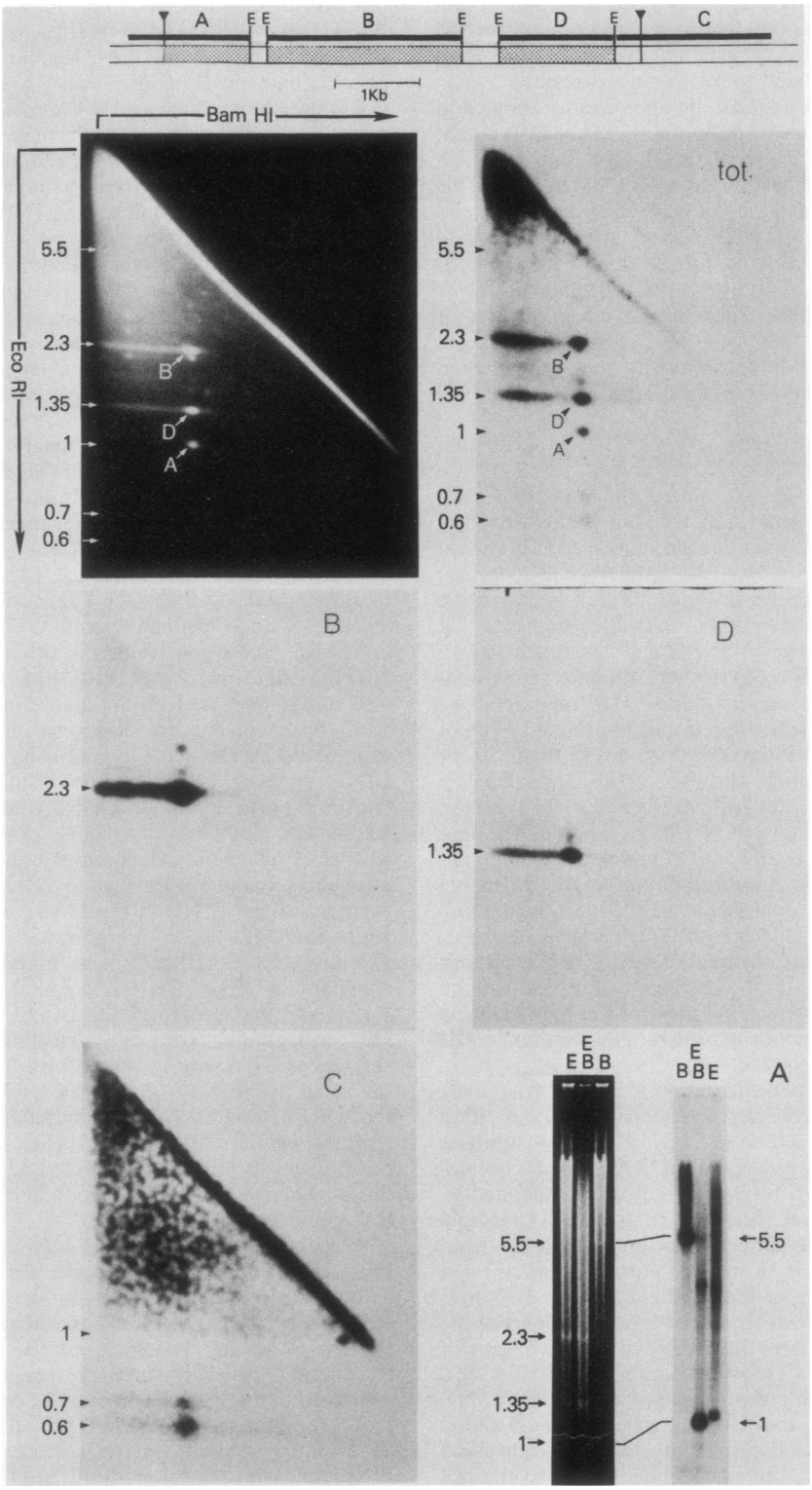


FIG. 4. Two-dimensional restriction enzyme analysis of genomic copies of rat LINE sequences. For all of the panels except panel A (see below), about 25 μg of rat DNA was digested with an excess of *Bam*HI at 37°C for 6 h. After phenol extraction and alcohol precipitation, the DNA was redigested with *Bam*HI under the same conditions. The digest was resolved by electrophoresis in a 1% gel of low-melting-point agarose and digested in situ with *Eco*RI, and the fragments were resolved in the second direction on a 1% agarose gel (40a). After staining, the fragments were blotted to nitrocellulose and then hybridized to total rat DNA (tot) or clones that contained LINE segments B, D, or C (see heavy solid bars of top diagram [E, *Eco*RI; ▼, *Bam*HI] and Fig. 2). After washing, the blots were exposed to X-ray film; the corresponding autoradiograms are shown in panels tot, B, D, and C, respectively. The mottled appearance of panel C is due to a blotting artifact related to the lot of nitrocellulose. The upper left-hand panel shows a photograph of the stained gels and the direction of

relatively low G+C content of the LINE sequence compared with that of most other coding sequences (15). Therefore, it seems that most of the differences between codon use of the LINE ORFs and that of other protein-encoding sequences are simply a reflection of the base composition of the ORF. This conclusion also generally applies to ORFa', which is quite T rich.

Third, we compared the number of base substitutions between ORF4 and its counterparts in other LINE sequences (Fig. 3; also see above) that do not lead to an amino acid change (silent) with those that do (replacement). Depending on the pair compared, 29 to 37% of the total base substitutions were silent. We excluded the amino acid differences due to the change in reading frame caused by the base deletions in ORFd2. The average for the five pairwise comparisons with ORF4 was 33% silent substitutions, a value that is in the lower range of those reported by Jukes (23) for 15 different sets of protein-encoding sequences, which ranged from 33 to 92% silent substitutions.

Finally, we applied to the LINE ORFs the algorithm TESTCODE that was devised from empirical observations of coding sequences by Fickett (15). The algorithm yields a probability of coding from 0 to 1.0 that is indifferent to codon choice but is correlated with the periodic distribution of nucleotides in about 95% of bona fide coding sequences and is lacking in noncoding sequences, and, to a lesser extent, in the G+C content of the sequence (15). The values that we obtained for the "probability of coding" were as follows: ORFa', 0.4; ORFa, 0.77; ORFb, 0.4; ORFc, 0.92; ORFd1, 0.77; ORFd2, 0.29; ORF4, 0.07; ORFI1, 0.04 (by using just the part of the sequence (the first 797 bp) determined by us [28]); the ORFs in the two C segment sequences (see above), 0.4 each. Only ORFc has a value that predicts coding. Each of the others has a value that corresponds to "no opinion" (0.4 to 0.77) or predicts noncoding (<0.4) (15). This distribution of TESTCODE scores indicates that, taken together, the LINE ORFs are atypical coding sequences. For example, with bona fide coding sequences, TESTCODE returns a value consistent with noncoding only 5 to 12% of the time and a value consistent with "no opinion" 18 to 23% of the time (15, 58).

As a test of our program and of whether the relatively high A content (about 40%) and somewhat low G+C (about 20% each of G and C) content of all of these ORFs (except ORFa', which is T rich) were responsible for their scores, we tested four coding sequences from the A+T-rich *E. coli* bacteriophage T4 which were not yet sequenced when Fickett derived his parameters. These sequences correspond to gene 41 (33% A, 16.5% C, 20.7% G), gene 61 (36.7% A, 15.3% C, 19.1% G), and T4 deoxyadenosine methylase gene (39.2% A, 12.6% C, and 15.1% G), and an ORF that corresponds to a recently proposed T4 gene, gene 69 (36.6% A, 14.5% C, 21.1% G) (30; B. Alberts, personal communication). The probabilities of coding for these sequences were 0.92, 0.92, 0.77, and 0.92, respectively: a distribution consistent with those earlier reported for bona fide coding sequences (15, 58).

DISCUSSION

The 6.7-kb member of the rat LINE family (rat LINE 3, L1Rn 3) has a ca. 200-bp G+C-rich (≥60%) region near each end, but is otherwise somewhat more A+T rich than the total rat DNA, contains a number of ORFs, and contains neither inverted nor direct repeats except for a tandem array of 3.5 copies of a ca. 65-bp sequence that is about 1 kb from the left end, which we shall also refer to as the 5' end. Comparison of the LINE 3 sequence with that of others we determined (LINE 4, Fig. 2; the LINE at the insulin 1 locus [LINE I1] [28]; the C segment of the LINE at the *Mlvi-2* locus [Furano et al., manuscript in preparation]) or with the sequence of a 5.5-kb *Bam*HI rat LINE fragment provided by Soares et al. (personal communication) and the C segment sequence published by Scarpulla (45) indicates that G+C-rich terminal regions, ORFs, and the tandem array of short repeats may be typical features of the rat LINE family. Furthermore, in each case an A-rich sequence of variable length and base composition is at the right-hand end of the C segment. However, only three of the five A-rich stretches are preceded by the putative polyadenylation signal, AATAAA. Members of the mouse and primate LINE families (14, 54, 60), as well as other mammalian interspersed repeated families (43, 48), also have A-rich sequences at their right ends.

Except for C segment sequences, total rat DNA contains about equal amounts of each LINE segment (Fig. 4 and 6), and most of the genomic copies of this family are quite similar in overall structure to LINEs 3 and 4 (Fig. 4 and 5). These results also show that the *Eco*RI sites to the left of the left-hand *Bam*HI site of most genomic members are randomly located with respect to this *Bam*HI site (Fig. 4A), whereas the *Eco*RI sites to the right of the right-hand *Bam*HI site are positioned ≥1.1 kb beyond this *Bam*HI site, which is 1.35 kb beyond the *Eco*RI site that separates the D and C segments in most members of this family (Fig. 3, 4C, and 5a). Therefore, most genomic copies of the rat LINE family are at least 6.7 kb long and extend from somewhere near the left-hand *Bam*HI site to ≥1.1 kb beyond the right-hand *Bam*HI site. Since the rat genome contains about 50,000 copies of the LINE C segment sequence that is in clone pR4A1 (Fig. 3) (65), and since as many as 80% of them are present in typical LINE structures (Fig. 5b and c; Fig. 6 and its legend), then as much as 10% of the rat genome ($0.8 \times 50{,}000 \times 6.7$ kb) has a DNA sequence that is generally similar to the one shown in Fig. 2.

The C segment, taken in its entirety, is 1.2 to 1.6 times more highly repeated than the rest of the LINE (Fig. 6 legend), and our earlier results (65) suggested that the 3B5 portion of the C segment may be as much as twofold more highly repeated than the 4A1 portion (Fig. 3). It is possible that the "extra" copies of C segment sequences are orga-

electrophoresis after each restriction enzyme digestion. The stained fragments labeled A, B, and D correspond in size to the expected *Eco*RI-*Bam*HI fragments that are shaded in the top diagram, which shows only the *Bam*HI sites and some of the *Eco*RI sites of the LINE (see Fig. 3 and text). The unshaded, smaller *Eco*RI or *Eco*RI-*Bam*HI fragments will have migrated off of these gels. The length in kilobases of the indicated fragments was determined by comparison to the migration of appropriately-sized marker fragments. For panel A (lower right-hand corner), ca. 5 μg of rat DNA was digested with *Eco*RI, *Bam*HI, or both, and the products were resolved in a 1% agarose gel which was stained and photographed (left photograph) and then blotted to nitrocellulose. After hybridization with a clone that contained LINE segment A (see heavy solid bar, top diagram), the blot was washed and exposed to X-ray film (right photograph). The size of the indicated fragments was determined as described above. The autoradiographs in panels B, D, and A were exposed for 40 h; the others were exposed for 7 days.

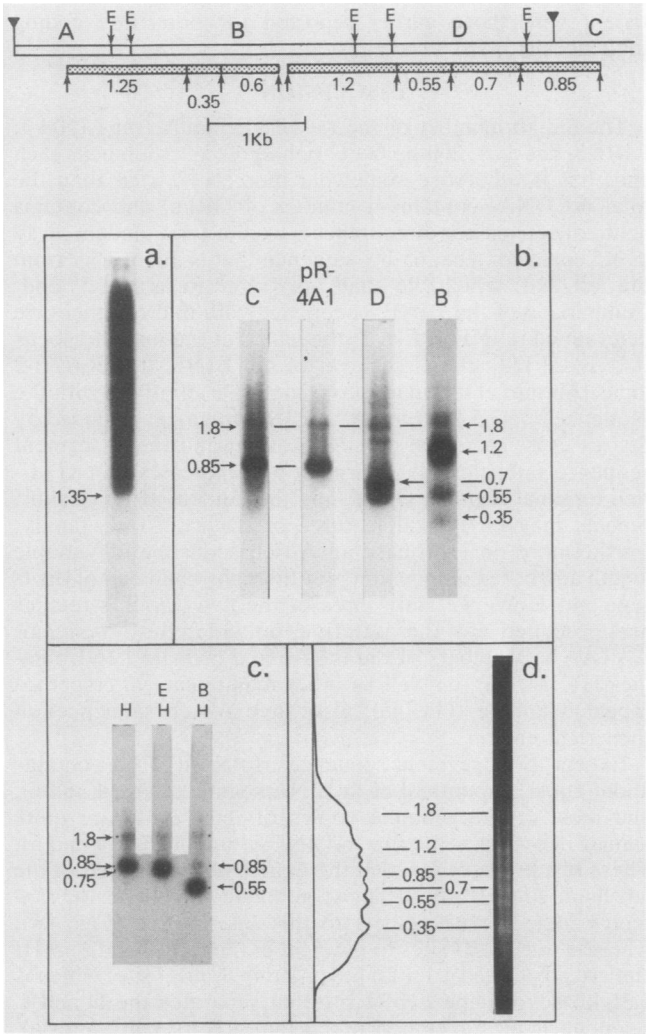


FIG. 5. Restriction enzyme analysis of genomic copies of rat LINE sequences. Rat DNA (5 to 10 μg) was digested with *Eco*RI (panel a), *Hae*III (all lanes of panel b and lane H of panel c), or *Hae*III and either *Eco*RI or *Bam*HI (lanes EH and BH, respectively, of panel c). The gels were blotted, and the blots were hybridized to a C segment clone (panel a and lane C of panel b) or to the D or B segment clones (lanes D and B of panel b). These were the same C, D, and B segment clones used for the experiment shown in Fig. 4. The pR4A1 lane of panel b and all the lanes of panel c were hybridized with the pR4A1 clone which contains about 200 bp of C segment sequence (Fig. 3). Panel d shows the photograph and densitometric trace of a stained gel of total rat DNA digested with *Hae*III. The sizes in kilobases of the indicated fragments were determined by comparison with the migration of appropriately sized marker DNAs and, for the blots shown in panels b and c, by rehybridizing these blots (after removing the first probe with alkali [65]) with a clone of rat satellite I DNA to detect the "ladder" of 370-bp monomer and its higher multiples that are generated by *Hae*III digestion of satellite I (13, 41). At least 70% of the stained 0.35-kb band in panel d is due to the *Hae*III fragments derived from the ca. 100,000-member rat satellite I family and will not hybridize to LINE sequences (13).

nized as the permuted clusters in which some copies of different, highly repeated transcribed rat DNA sequences reside (65). These sequences include those present in clones pR5A1, pR4A1, and pR3B5 (Fig. 3), and our recent sequence data on the LINE member inserted at the *Igh* locus of an Osborne-Mendel rat (12) support the permuted organization of C segment sequences. The *Igh* LINE is similar to LINE 3, but has an extra C segment that is contiguous to but oriented in the opposite direction as the normal C segment (results not shown).

Liver nuclear RNA contains transcripts of the entire LINE sequence (Fig. 4) (65), and in experiments not shown here we have detected abundant, relatively discrete-sized, >5-kb LINE transcripts among the heterogeneous-sized population of LINE transcripts in the total RNA of liver, kidney, muscle, and a rat cell line. We are now determining which DNA strand of the LINE family members is represented in the total RNA and in the RNA of the subcellular fractions from these various sources.

Long, relatively discrete transcripts (≥4.5 kb) of the human LINE family have also been detected (47, 55), and other studies on the transcription of mouse (20) and human (50, 56) LINE families indicate that they are transcribed by RNA polymerase II. LINE 3 contains potential polymerase II start sites (i.e., TATA-containing sequence [5]) in or near both G+C-rich clusters. Furthermore, both clusters resemble the ca. 200-bp G+C-rich regions that are thought to serve as polymerase II promoters for certain "housekeeping" genes (11, 35, 42, 66). However, the LINE G-C clusters differ from the putative housekeeping promoters in that the LINE G-C clusters show a typical mammalian DNA bias against the CpG sequence (34), which is capable of being methylated, whereas the housekeeping promoters do not. The LINE 3 sequence does not contain promoter sequences for RNA polymerase III. Although the members of the tandem array between the A and B segments (Fig. 2 and 3) are about the length of some enhancer sequences, they do not contain the enhancer core nucleotides that have been identified in certain enhancer elements (25, 62).

Both LINEs 3 and 4 contain ORFs. Therefore, in this way the rat LINE family is similar to the mouse and primate LINE families, which also contain ORFs (33, 53). Martin et al. (33) showed that the ORFs in mouse LINE sequences have evolved like bona fide protein-encoding sequences; i.e., silent base changes (no amino acid change) are more prevalent than replacement base changes. This is also true of ORF4 and its counterparts in the other rat LINES. However, owing to the numerous replacement changes, the putative proteins synthesized from these ORFs would form a family of related proteins which would be quite different from that encoded by ORF4. Although this does not argue for or against protein synthesis from these ORFs, the prediction of noncoding or no opinion by TESTCODE for all but one of the LINE ORFs (see Results) suggests that these ORFs, as a class, are atypical coding sequences (see Results) (15). In this way they resemble ORFs in certain stable RNAs (58).

Other workers have suggested that all of the primate or mouse LINE ORFs examined to date are pseudogenes (33, 53) or relics of the intact (longer) ORF(s) present in the few functional progenitor LINE family members from which most present-day nonfunctioning LINE members were derived (14, 43, 55, 59, 60). Two pairs of ORFs in LINE 3 can be fused into two longer ones by only one genetic change between each pair. Deletion of the T at position 3527 eliminates the stop codon of ORFb and puts it in phase with ORFc. Deletion of the G at position 5186 eliminates the stop codon of ORFd1 and puts it in phase with ORFd2, thereby generating an ORF that is somewhat longer than and has the same reading frame as ORF4. In addition to these, two deletions at different sites between ORFs a and b and likewise between ORFs c and d1 connect all of the ORFs

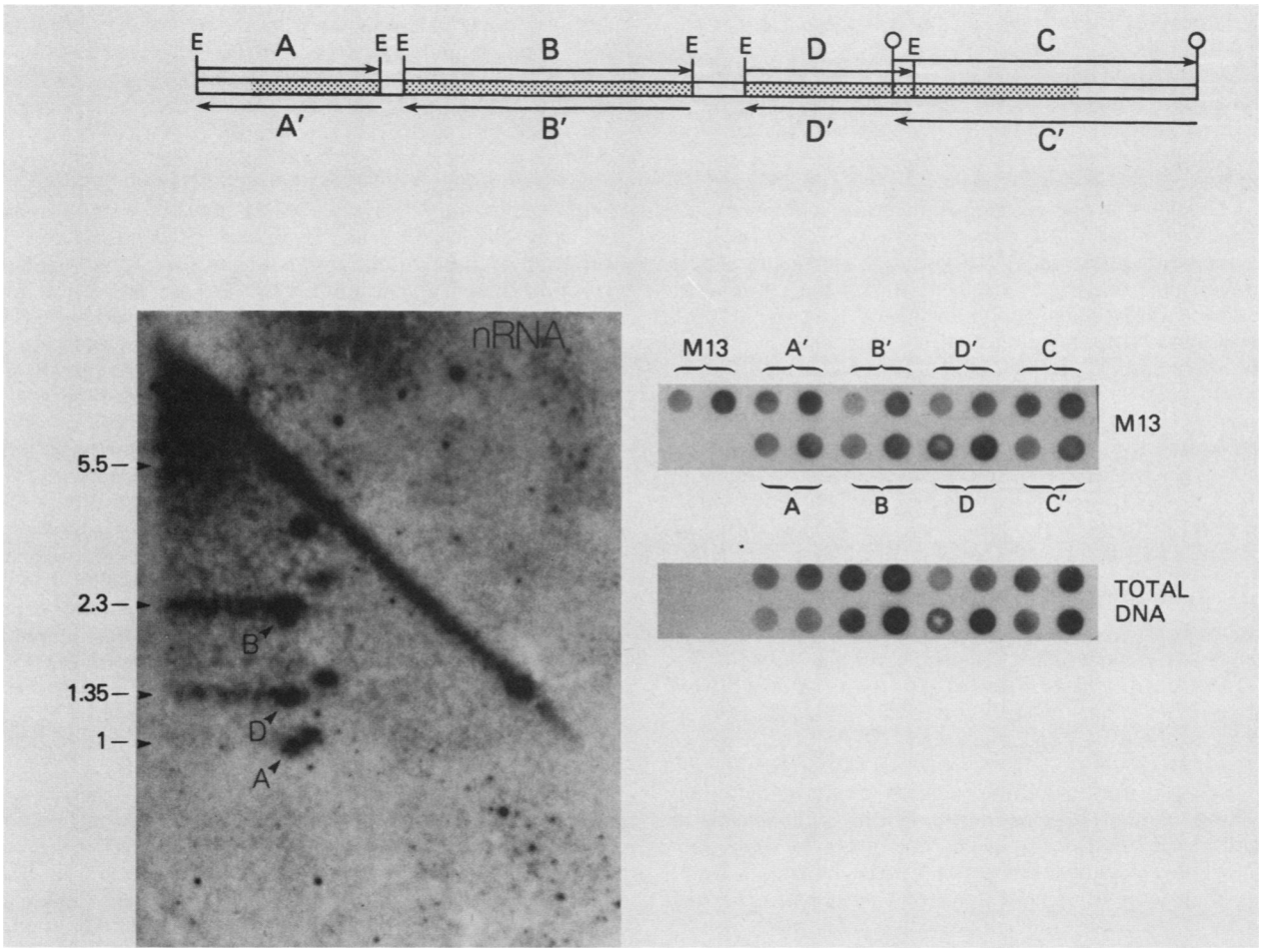


FIG. 6. Representation of LINE segments in nuclear RNA and in genomic DNA. The large photograph on the left shows a blot of a two-dimensional gel of an *Eco*RI-*Bam*HI double digest of total rat DNA (Fig. 4) that was hybridized to total liver nuclear RNA, prepared, and radiolabeled with [γ-^{32}P]ATP and polynucleotide kinase as described previously (65). The diagram at the top of the figure (E, *Eco*RI; ♀ . *Hin*dIII) shows LINE segment-containing M13 clones derived from the LINE-containing λ genomic clones (Fig. 3). The shaded regions indicate the portion corresponding to LINE segments A, B, D, and C, respectively. The A', B', D', and C' clones contain the complementary strand of the sequence shown in Fig. 2. Note that the C segment clones contain 0.17 kb of D segment and therefore contain 1.35 + 0.17, or 1.52, kb of LINE sequence. These clones were hybridized to radioactive total rat DNA and M13 vector DNA as follows. Two different amounts of each clone DNA, as well as M13 vector DNA, were applied as dots (24) to nitrocellulose filters. (Note that the placement of the C and C' dots is reversed with respect to the placement of the other dots.) Depending on the DNA, the lower amounts were 100 to 150 ng and the higher amounts were 200 to 300 ng. These amounts of DNA provide a minimum of either 7- to 10-fold or 14- to 20-fold over the total amount of complementary LINE segment sequence in the 200 ng of total rat DNA used as a hybridization probe (assuming that about 10% of total rat DNA is LINE DNA). However, the hybridization reaction was stopped when the LINE sequences in the probe would have attained a Cot (6) of about 1.7×10^{-3}. At this point less than 5% of the LINE sequences in the probe would have reannealed and less than 0.1 ng of LINE DNA was bound to any of the dots (as calculated from the counts per minute bound and the specific activity of the probe). Therefore, depending on the dot, only 0.002 to 0.008 of the filter-bound DNA contained hybrids when the reaction was stopped. We limited the extent of the hybridization reaction to minimize concatenation of the probe in solution and subsequent hybridization of the concatenates to the filter-bound DNA. Since the concatenates could contain networks of LINE and non-LINE repeated DNA families (18, 43, 65), this could lead to spurious results. Because the rate of the hybridization reactions with the various dots is determined both by the concentration of DNA in the probe and the amount of DNA on the filter, we controlled for this by hybridizing M13 DNA to a duplicate set of dots. After the dots were exposed to X-ray film, the counts per minute of total rat DNA that hybridized to each dot was determined in a liquid scintillation spectrometer and normalized to the counts per minute of M13 DNA that hybridized to the respective dots. Depending on the clone and the amount of DNA applied to the filter, 1,500 to 5,500 cpm of M13 DNA was bound to the various dots. The average values for the ratio of the counts per minute of rat DNA to the counts per minute of M13 DNA for A, B, D, and C were 0.17 (0.19, 0.14), 0.42 (0.45, 0.38), 0.24 (0.23, 0.24), and 0.3 (0.28, 0.31), respectively. (The values in parentheses are from each of the complementary strands of each cloned segment.) Therefore, the relative hybridization of rat DNA to the A, B, D, and C segment-containing clones was 1, 2.5, 1.4, and 1.8, respectively. After correcting these values for the relative length of LINE DNA in these clones, which is 1, 2.3, 1.35, and 1.52, respectively, we calculate the genomic ratio of the A, B, D, and C segments to be 1.0:1.1:1.0:1.2. These results were confirmed in experiments (not shown) in which the relative hybridization of radioactive total rat DNA to the LINE segments released by *Eco*RI digestion of the λ4A13 or λ4A14 genomic clones (Fig. 3) was determined by densitometry of the autoradiograms of the relevant blot hybridizations. In this case we estimated the genomic ratio of A, B, D, and C segments to be 1.0:1.0:1.3:1.6.

into a single 4.6-kb ORF. This indicates that the ORFs of LINE 3 could well have been derived from a long progenitor ORF. However, there is no way of knowing whether this occurred before, during, or after transposition of LINE 3 to its present site (see below). Furthermore, a priori, there is no reason to conclude that LINE 3 is now an inert member of the LINE family. For example, it could still be transcriptionally active.

The perfect 14-bp direct repeat that flanks LINE 3 probably represents a target site duplication caused by the transposition of the LINE 3 member to its present location. Other members of the rat LINE family as long as the LINE 3 member have undergone transposition in the rat genome, since the LINE-related polymorphisms at two of three different single-copy loci are due to the presence or absence of ≥6.5-kb members of this family (12). It has been proposed that the transposition and amplification of the mouse and primate LINE families occur largely by the insertion of DNA copies (retrotranscripts) of LINE transcripts into the genome (14, 43, 55, 59, 60). Important to this proposal is the idea that the A-rich end of various members of these families is the relic of the poly(A)$^+$ tail of a retrotranscribed LINE transcript, and, in fact, this is why the end of the LINE contiguous to the A-rich end is considered the 3′ (or right) end (14, 54, 60).

Although transposition and amplification of rat LINE sequences could occur by retrotranscription, the putative type II DNA synthesis arrest sites near each end of rat LINE members (Fig. 2 and legend) could also allow these processes to occur as a consequence of DNA replication. If, under certain conditions, DNA synthesis is arrested at these sites, then a LINE-containing replicative intermediate may exist long enough for the newly synthesized strands to hybridize to each other and be extruded from the replication intermediate by reannealing of the parental strands. Replication intermediates of SV40 can do this in vitro (67) and perhaps as well in vivo (4). Also, defined (but as yet unidentified) DNA arrest sites in rat DNA can stall replication forks proceeding in either direction from an integrated polyoma replicon in vivo (2).

As for LINE sequences, depending on numerous factors including the extent to which either arm of the replication fork was arrested, a partial or full-length LINE member as well as some flanking DNA could be extruded and detached from the replicative intermediate and eventually be reintegrated elsewhere in the genome. Integration into target sites that contain A-rich stretches could account, in part, for the variable-length A-rich right end of most LINE members. Our DNA sequence determination of the "empty" and LINE-containing target sties at the insulin 1 and *M1vi-2* loci (Furano et al., in preparation), as well as of the target site duplication that flanks LINE 3 (Fig. 2), shows that the A-rich right end of these members became contiguous with an A-rich stretch in the target site. Since A-rich stretches are simple sequences and are subject to various means of expansion (and contraction) (57), it is possible that the A-rich end of a LINE member is a composite of "true" LINE sequence (e.g., AATAAA) and target site sequence that has been modified with time. The A-rich tails of both LINEs 3 and 4 (Fig. 2) seem to have undergone expansion, at least in part, since each contains internal repeats. The tendency of various mammalian interspersed repeated sequences to insert into A-rich stretches has recently been noted by Rogers (43). Finally, the extra copies of C segment sequences may be related to more frequent or prolonged arrests at the putative right-hand DNA arrest site, which strongly arrests DNA synthesis in vitro (Fig. 2), than at the left-hand site. The proposed mechanism, which could also account for rapid amplification of the arrested sequences (4, 46), would be made more plausible by demonstrating that the potential DNA synthesis arrest sites are recognized in vivo.

The fact that most genomic copies of the rat LINE family are full length and generally similar to each other (Fig. 4 through 6) is in contrast to the heterogeneous state of the mouse or primate LINE families (7, 14, 19, 29, 36, 59). The heterogeneity of the mouse LINE family is quite evident when restriction enzyme fragments of mouse DNA are analyzed by the two-dimensional electrophoretic technique. Hybridization with the mouse LINE sequences that correspond to either the B or D segments of the rat LINE did not produce the simple hybridization patterns shown in Fig. 4B and D, but instead produced a very complex pattern. A significant amount of the total hybridizaton by either mouse probe was to a wide variety of fagments in addition to those expected from the structure of an archtypical member of the mouse LINE family (M. Seidman, S. L. Bunting, S.-M. Cheng, and A. C. Peacock, personal communication).

The heterogeneity of the mouse LINE family can be partly explained by the fact that many of its members are missing a variable portion from their 5′ (left) end (14, 59). This is also true of the primate family (19), and part of the evidence for this is that these genomes contain at least 6 times as many copies of the 3′ end as of the 5′ end of certain cloned members of these families (14, 19, 59). By contrast, the rat genome contains only about 1.2 to 1.6 times as many copies of the 3′ C segment as of the 5′ A segment of the rat LINE family (see above and Fig. 6 legend).

Another explanation for the heterogeneity of the mouse or primate LINE families compared with the rat LINE famliy is that several distinct but related ancestral LINE familes were amplified or maintained during the evolution of primates and mice, whereas in the rat there was only one. There are numerous examples of species-specific amplification of related but distinct repeated DNA sequences (7, 48, 52, 64), and the complete lack of homology between certain corresponding portions of the rat and mouse LINE families (see legend to Fig. 2) is consistent with the existence of different ancestral rodent LINE families. The presence of different but related LINE families in mice and primates could also account for some of the discrepancy between the genomic copy number of the 5′ and 3′ ends of certain cloned members of the mouse or primate LINE families (14, 19, 59), and the hybridization data presented by Grimaldi et al. (19) are consistent with the presence of more than one version of the LINE family in the African green monkey genome.

Whatever the explanation for the distinctive states of the LINE families in rats, mice, and primates, this difference is not the only one among the repeated DNA families of these animals. Perhaps as striking is the fact that satellite DNA sequences account for only 1 to 2% of the total rat genome but 10 to 20% of the total DNA in mice or primates (38, 41, 52). Understanding the biological significance of these rather large-scale differences among the DNA composition of these animals should help determine the role or effect of repeated DNA.

ACKNOWLEDGMENTS

We thank M. Seidman for communicating to us, prior to publication, his results on the two-dimensional restriction enzyme analysis of the mouse LINE family, B. Alberts for permission to analyze the T4 bacteriophage gene 41 and gene 61 DNA sequences, and A. Efstratiadis and his colleagues M. Soares and E. Schon for

communicating to us, prior to publication, their sequence of the 5.5-kb *Bam*HI fragment of a rat LINE family member.

LITERATURE CITED

1. **Adams, J. W., R. E. Kaufman, P. J. Kretschmer, M. Harrison, and A. W. Nienhuis.** 1980. A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. Nucleic Acids Res. **8**:6113–6128.
2. **Baran, N., A. Neer, and H. Manor.** 1983. "Onionskin" replication of integrated polyoma virus DNA and flanking sequences in polyoma-transformed rat cells: termination within a specific cellular DNA segment. Proc. Natl. Acad. Sci. USA **80**:105–109.
3. **Bolivar, F., R. L. Rodriguez, P. J. Greene, M. C. Betlach, H. L. Heyneker, H. W. Boyer, J. H. Crosa, and S. Falkow.** 1977. Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene **2**:95–113.
4. **Botchan, M., W. Topp, and J. Sambrook.** 1978. Studies on simian virus 40 excision from cellular chromosomes. Cold Spring Harbor Symp. Quant. Biol. **43**:709–719.
5. **Breathnach, R., and P. Chambon.** 1981. Organization and expression of eucaryotic split genes coding for proteins. Annu. Rev. Biochem. **50**:349–383.
6. **Britten, R. J., and D. E. Kohne.** 1968. Repeated sequences in DNA. Science **161**:529–540.
7. **Brown, S. D. M., and G. Dover.** 1981. Organization and evolutionary progress of a dispersed repetitive family of sequences of widely separated rodent genomes. J. Mol. Biol. **150**:441–466.
8. **Brutlag, D. L., J. Clayton, P. Friedland, and L. H. Kedes.** 1982. SEQ: a nucleotide sequence analysis and recombination system. Nucleic Acids Res. **10**:279–294.
9. **Chen, H. R., and W. C. Barker.** 1984. Nucleotide sequences of the retroviral long terminal repeats and their adjacent regions. Nucleic Acids Res. **12**:1767–1778.
10. **Cheng, S.-M., and C. L. Schildkraut.** 1980. A family of moderately repetitive sequences in mouse DNA. Nucleic Acids Res. **8**:4075–4090.
11. **Dush, M. K., J. M. Sikela, S. A. Kham, J. A. Tischfield, and P. J. Stambrook.** 1985. Nucleotide sequence and organization of the mouse adenine phosphoribosyltransferase gene: presence of a coding region common to animal and bacterial phosphoribosyltransferases that has a variable intron/exon arrangement. Proc. Natl. Acad. Sci. USA **82**:2731–2735.
12. **Economou-Pachnis, A., M. A. Lohse, A. V. Furano, and P. N. Tsichlis.** 1985. Insertion of long interpreted repeated elements (LINEs) at the *Igh* and *MLvi-2* loci of rats. Proc. Natl. Acad. Sci. USA **82**:2857–2861.
13. **Epstein, D. A., F. R. Witney, and A. V. Furano.** 1984. The spread of sequence variants in *Rattus* satellite DNAs. Nucleic Acids Res. **12**:973–988.
14. **Fanning, T. G.** 1983. Size and structure of the highly repetitive *Bam*HI element in mice. Nucleic Acids Res. **11**:5073–5091.
15. **Fickett, J. W.** 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. **10**:5303–5318.
16. **Garoff, H., and W. Ansorge.** 1981. Improvements of DNA sequencing gels. Anal. Biochem. **115**:450–457.
17. **Grantham, R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier.** 1981. Codon catalog usage is a genome strategy modulated for gene expression. Nucleic Acids Res. **9**:r43–r74.
18. **Grimaldi, G., and M. F. Singer.** 1983. Members of the *Kpn*I family of long interspersed repeated sequences join and interrupt α-satellite in the monkey genome. Nucleic Acids Res. **11**:321–338.
19. **Grimaldi, G., J. Skowronski, and M. F. Singer.** 1984. Defining the beginning and end of KpnI family segments. EMBO J. **3**:1753–1759.
20. **Heller, D., M. Jackson, and L. Leinwand.** 1984. Organization and expression of non-Alu family interspersed repetitive DNA sequences in the mouse genome. J. Mol. Biol. **173**:419–436.
21. **Henikoff, S.** 1984. Unidirectional digestion with exonuclease III creates targeted break points for DNA sequencing. Gene **28**:351–359.
22. **Jackson, M., D. Heller, and L. Leinwand.** 1985. Transcriptional measurements of mouse repeated DNA sequences. Nucleic Acids Res. **13**:3389–3403.
23. **Jukes, T. H.** 1980. Silent nucleotide substitutions and the molecular evolutionary clock. Science **210**:973–978.
24. **Kafatos, F. C., C. W. Jones, and A. Efstratiadis.** 1979. Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. Nucleic Acids Res. **7**:1541–1552.
25. **Khoury, G., and P. Gruss.** 1983. Enhancer elements. Cell **33**:313–314.
26. **Kole, L. B., S. R. Haynes, and W. R. Jelinek.** 1983. Discrete and heterogeneous high molecular weight RNAs complementary to a long dispersed repeat family (a possible transposon) of human DNA. J. Mol. Biol. **165**:257–286.
27. **Krayev, A. S., T. V. Markusheva, D. A. Kramerov, A. P. Ryskov, K. G. Skryabin, A. A. Bayev, and G. P. Georgiev.** 1982. Ubiquitous transposon-like repeats B1 and B2 of the mouse genome: B2 sequencing. Nucleic Acids Res. **10**:7461–7475.
28. **Lakshmikumaran, M. S., E. D'Ambrosio, L. A. Laimins, D. T. Lin, and A. V. Furano.** 1985. Long interspersed repeated DNA (LINE) causes polymorphism at the rat insulin 1 locus. Mol. Cell. Biol. **5**:2197–2203.
29. **Lerman, M. I., 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.
30. **Macdonald, P. M., and G. Mosig.** 1984. Regulation of a new bacteriophage T4 gene, 69, that spans an origin of DNA replication. EMBO J. **3**:2863–2871.
31. **Manuelidis, L.** 1980. Novel classes of mouse repeated DNAs. Nucleic Acids Res. **8**:3247–3258.
32. **Manuelidis, L.** 1982. Nucleotide sequence definintion of a major human repeated DNA, the *Hin*dIII 1.9 kb family. Nucleic Acids Res. **10**:3211–3219.
33. **Martin, S. L., C. F. Voliva, F. H. Burton, M. H. Edgell, and C. A. Hutchison III.** 1984. A large interspersed repeat found in mouse DNA contains a long open reading frame that evolves as if it encodes a protein. Proc. Natl. Acad. Sci. USA **81**:2308–2312.
34. **McClelland, M., and R. Ivarie.** 1982. Asymmetrical distribution of CpG in an "average" mammalian gene. Nucleic Acids Res. **10**:7865–7877.
35. **Melton, D. W., D. S. Konecki, J. Brennard, and C. T. Caskey.** 1984. Structure, expression, and mutation of the hypoxanthine phosphoribosyl transferase gene. Proc. Natl. Acad. Sci. USA **81**:2147–2151.
36. **Meunier-Rotival, M., and G. Bernardi.** 1984. The Bam repeats of the mouse genome belong in several super families the longest of which is over 9 kb in size. Nucleic Acids Res. **12**:1593–1608.
37. **Meunier-Rotival, M., P. Soriano, G. Cuny, F. Strauss, and G. Bernardi.** 1982. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc. Natl. Acad. Sci. USA **79**:355–359.
38. **Miklos, G. L. G., D. A. Willcocks, and P. R. Braverstock.** 1980. Restriction endonuclease and molecular analyses of three rat genomes with special reference to chromosome rearrangement and speciation problems. Chromosoma **76**:339–363.
39. **Nordheim, A., E. M. Lafer, L. J. Peck, J. C. Wang, B. D. Stollar, and A. Rich.** 1982. Negatively supercoiled plasmids contain left-handed Z-DNA segments as detected by specific antibody binding. Cell **31**:309–318.
40. **Norrander, J., T. Kempe, and J. Messing.** 1983. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. Gene **26**:101–106.

40a. **Peacock, A. C., S. L. Bunting, S. P. C. Cole, and M. Seidman.** 1985. Two-dimensional electrophoretic display of restriction fragments from genomic DNA. Anal. Biochem. **149**:177–182.

41. **Pech, M., T. Igo-Kemenes, and H. G. Zachau.** 1979. Nucleotide sequence of a highly repetitive component of rat DNA. Nucleic Acids Res. **7**:417–432.
42. **Reynolds, G. A., S. K. Basu, T. F. Osborne, D. J. Chin, G. Gil, M. S. Brown, J. L. Goldstein, and K. L. Luskey.** 1984. HMG

CoA reductase: a negatively regulated gene with unusual promoter and 5′ untranslated regions. Cell **38**:275–285.
43. **Rogers, J. H.** 1985. The origin and evolution of retroposons. Int. Rev. Cytol. **93**:187–279.
44. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
45. **Scarpulla, R. C.** 1985. Association of a truncated cytochrome C processed pseudogene with a similarly truncated member from a long interspersed repeat family of rat. Nucleic Acids Res. **13**:763–775.
46. **Schimke, R. T.** 1984. Gene amplification in cultured animal cells. Cell **37**:705–713.
47. **Schmeckpeper, B. J., A. F. Scott, and K. D. Smith.** 1984. Transcripts homologous to a long repeated DNA element in the human genome. J. Biol. Chem. **259**:1218–1225.
48. **Schmid, C. W., and W. R. Jelinek.** 1982. The Alu family of dispersed repetitive sequences. Science **216**:1065–1070.
49. **Shafit-Zagardo, B., F. L. Brown, J. J. Maio, and J. W. Adams.** 1982. *Kpn*I families of long, interspersed repetitive DNA's associated with the human β-globin gene cluster. Gene **20**:397–407.
50. **Shafit-Zagardo, B., F. L. Brown, P. J. Zavodny, and J. J. Maio.** 1983. Transcription of the *Kpn*I families of long interspersed DNAs in human cells. Nature (London) **304**:277–280.
51. **Shapiro, H. S.** 1976. Distribution of purines and pyrimidines in deoxyribonucleic acids, p. 241–311. *In* G. D. Fasman (ed.), Handbook of biochemistry and molecular biology: nucleic acids, vol. 2. CRC Press, Cleveland, Ohio.
52. **Singer, M. F.** 1982. Highly repeated sequences in mammalian genomes. Int. Rev. Cytol. **76**:67–112.
53. **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.
54. **Singer, M. F., R. E. Thayer, G. Grimaldi, M. J. Lerman, and T. G. Fanning.** 1983. Homology between the *Kpn*I primate and *Bam*HI (MIF-1) rodent families of long interspersed, repeated sequences. Nucleic Acids Res. **11**:5739–5745.
55. **Skowronski, J., and M. F. Singer.** 1985. Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proc. Natl. Acad. Sci. USA **82**:6050–6054.
56. **Sun, L., K. E. Paulson, C. W. Schmid, L. Kadyk, and L. Leinwand.** 1984. Non-Alu family repeats in human DNA and their transcriptional activity. Nucleic Acids Res. **12**:2669–2690.
57. **Tautz, D., and M. Renz.** 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. **12**:4127–4138.
58. **Tramontano, A., V. Scarlato, N. Barni, M. Cipollano, A. Franzè, M. F. Macchiato, and A. Cascino.** 1984. Statistical evaluation of the coding capacity of complementary DNA strands. Nucleic Acids Res. **12**:5049–5059.
59. **Voliva, C. F., C. L. Jahn, M. B. Comer, C. A. Hutchison III, and M. H. Edgell.** 1983. The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. Nucleic Acids Res. **11**:8847–8859.
60. **Voliva, C. F., S. L. Martin, C. A. Hutchison III, and M. H. Edgell.** 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. J. Mol. Biol. **178**:795–813.
61. **Weaver, D. T., and M. L. DePamphilis.** 1984. The role of palindromic and non-palindromic sequences in arresting DNA synthesis *in vitro* and *in vivo*. J. Mol. Biol. **180**:961–986.
62. **Weiher, H., M. König, and P. Gruss.** 1983. Multiple point mutations affecting the simian virus 40 enhancer. Science **219**:626–631.
63. **Wilbur, W., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA **80**:726–730.
64. **Witney, F. R., and A. V. Furano.** 1983. The independent evolution of two closely related satellite DNA elements in rats (*Rattus*). Nucleic Acids Res. **11**:291–304.
65. **Witney, F. R., and A. V. Furano.** 1984. Highly repeated DNA families in the rat. J. Biol. Chem. **259**:10481–10492.
66. **Yang, J. K., J. N. Masters, and G. Attardi.** 1984. Human dihydrofolate reductase gene organization. J. Mol. Biol. **176**:169–187.
67. **Zannis-Hadjopoulos, M., M. Persico, and R. G. Martin.** 1981. The remarkable instability of replication loops provides a general method for the isolation of origins of DNA replication. Cell **27**:155–163.