Volume 13 Number 4 1985 Nucleic Acids Research

Variability within the rabbit C repeats and sequences shared with other SINES

Ross C.Hardison and Richard Printz

Department of Biochemistry, Microbiology, Molecular and Cell Biology, 206 Althouse Laboratory, Pennsylvania State University, University Park, PA 16802, USA

Received 21 December 1984; Accepted 24 January 1985

ABSTRACT

The C family of short, interspersed repeats (SINES) is highly repeated in the rabbit genome, and most members have a structure suggestive of a model for their dispersal via reinsertion of a double-stranded copy of an RNA polymerase III transcribed RNA. We have determined the nucleotide sequence of additional members of the repeat family and have compiled them to obtain an improved consensus sequence. This compilation shows that although most regions of the repeat are well conserved, two regions show high variability. Some individual repeats are truncated, and one truncated repeat retains the characteristic structures of a retroposon. The consensus sequence for C repeats does not match the sequence of any other sequenced mammalian SINE over large regions, but short imperfect matches to several primate and rodent SINES are observed. A sequence similar to the 27 nucleotide consensus sequence TCCCAGCAACCACATGGGAGGCAGAGA was found in all mammalian SINES examined. The 3' portion of this sequence matches a DNA segment found at the replication origins of papovaviruses.

INTRODUCTION

Reannealed repetitive DNA from many species can be fractionated into two size ranges: long repeats (> 2000 bp) and short repeats (about 300 bp; ref. 1). These repeats tend to be interspersed with single copy DNA (2). Recent structural analysis of the long interspersed repeats, or LINES (3), identifies at least two distinct classes. One class of LINES resembles transposable elements and retroviral proviruses in that they contain long terminal repeats (either inverted or direct) and they are flanked by short direct repeats that differ between members of the repeat family. Examples of transposon-like LINES are the yeast Ty1 elements (4), the *Drosophila copia* (5) and FB (6) repeats, and the mouse intracisternal A particles (7). A second type of LINE is exemplified by the *Kpn* element of primates (8,9) and the L1Md repeat of mouse (10,11). These repeats, which are homologous to each other (12,13), are about 6 kb long, lack a terminal redundancy, have an A-rich tract at the 3' end (8-11,14,15), are flanked by short direct repeats (16), and are

transcribed by RNA polymerase II (17). Many of these properties are also found in Drosophila F repeats (18). Individual members of the non-transposon class of LINES are frequently truncated at the 5' end (11), and occassionally they vary from each other by permutation of blocks of common sequences (9). The A-rich 3' tract and flanking short direct repeats are characteristics of elements that have been proposed to amplify and move by reverse transcription of an RNA product followed by reinsertion into the genome (19,20). Such elements have been termed retroposons (21).

The short interspersed repeats (SINES, ref. 3) are also comprised of at least two general classes. One class resembles prokaryotic insertion sequences and retroviral LTRs in that they contain terminal inverted repeats and occassionally flank longer elements (as IS elements flank some transposons and LTRs flank retroviral proviruses). A good example of this class of SINE is the δ element of yeast which can either flank the long Ty 1 repeat or exist alone in the genome (4). Other yeast SINES that contain terminal inverted repeats are σ and τ, which are related to δ and are usually found 5' to tRNA genes (22). The genomes of artiodactyls contain about 10^5 copies of a SINE that has terminal inverted repeats and does not end in a 3' tract of A-rich sequence (23). These repeats have been identified in goats (24) and cows (23,25). A distinctly different class of SINEs is exemplified by the primate Alu (or rodent type 1) repeat. This class of SINE is transcribable by RNA polymerase III, has a 3' A-rich tract, and is flanked by short direct repeats (26). They do not exhibit the terminal redundacy of the IS-like SINEs. Several different short repeats fall into this class, including rodent B1 (26,27), rodent B2 (28,29), and rabbit C (30). They have been proposed to propagate by reverse transcription of the polymerase III transcript, followed by reinsertion into the genome (20). In this sense they are short retroposons which differ from the long retroposons described above in their length and in the class of polymerase that transcribes them. Thus in each size range of repeats, one can distinguish one class related to transposons and insertion sequences and another class (retroposons) that lack terminal redundancy and end in an A-rich tract. All classes are flanked by short direct repeats that presumably form by repair after reinsertion of the element at a staggered break in a chromosome.

Previous analysis of the C repeat family (30) showed that there are about 170,000 copies per haploid genome, with an average size of about 300 bp per repeat. Transcripts containing C repeat sequences are heterogeneous in size and are primarily confined to the nucleus. One of the C repeat members

sequenced is considerably different from two other members (only 64% similarity). Although the structural and transcriptional properties of rabbit C repeats are similar to the primate Alu SINES, the C repeat sequence is not homologous to the Alu repeat sequence.

In this paper, we present the nucleotide sequence of additional C repeat members. The new compilation shows that the variability among C repeats is largely localized to two regions within the repeat, and on the basis of this variability one can identify at least two subfamilies of C repeats. Using the refined consensus sequence, we have expanded the sequence comparisons searching for related sequences in mammalian SINES. C repeats are not as closely related to other repeats as are primate Alu and rodent type 1 SINES (26) or the mouse, hamster and rat type 2 SINES (29). However, short regions of imperfect matches were observed to many mammalian SINES, suggesting that extensive sequence exchanges may have occurred among SINES during their evolution. Every mammalian SINE examined contained a sequence similar to a 27 nucleotide consensus: TCCCAGCAACCACATGGGAGGCAGAGA. The 3' portion
C T T
of this sequence matches with a sequence found at the papovavirus origin of replication.

MATERIALS AND METHODS

1) Recombinant DNAs

The C repeats sequenced in this report are located in the rabbit β-globin gene cluster (32,33) and are contained on the plasmids pE3.4 (5' to gene β4) and pE6.3 (3' to gene ψβ2). The subclones were derived from the recombinant bacteriophage λRβ'G8 and λRβG2 (32). Maps of the repeats in these clones are presented by Shen and Maniatis (34) and Cheng et al. (30).

2) DNA Sequencing

The C repeat DNA in plasmids pE3.4 and pE6.3 was sequenced by the Maxam and Gilbert method (35,36).

3) Sequence Comparison

DNA sequences were compared using Zweig's (37) dot-plot program on an IBM XT. The criterion selected was 12 nucleotides matching in a window of 14 nucleotides, which gave a minimum of background while still detecting matches. The sequences were compared at a variety of criteria and in both orientations.

RESULTS

1) Consensus Sequence of C Repeats

Two additional C repeat members from the rabbit β-globin gene family were sequenced, one located 5' to the embryonic globin gene β4 (pE3.4) and another located between pseudogene ψβ2 and fetal-adult gene β1 (pE6.3). These data are compiled in Figure 1, along with the sequences of four C repeats previously sequenced in this laboratory (30). The rabbit uteroglobin gene contains a repetitive element in the large intron (38). The uteroglobin intron sequence from nucleotides 1822 to 1540 matches with a C repeat; this is also listed in Figure 1. The orientation of the uteroglobin C repeat is opposite to the direction of transcription of uteroglobin. Thus Figure 1 contains the sequences of five full-length and two truncated C repeats. The sequences were aligned by inspection, and gaps were introduced to improve the alignment. A consensus sequence was derived and is presented on the top line of Figure 1. Nucleotide positions that are not present in all members of the sequenced set but which do occur in at least two members are listed in lower case. Because inserts in some members are included in the consensus, the length of the consensus sequence (352 bp before the A rich 3' tract) is longer than the average size of C repeats (310 bp).

The 5' end of the C repeat consensus sequence shown in Figure 1 begins 14 nucleotides before box A of the RNA polymerase III internal control region. Our previous report (30) described the 5' ends as variable, but the compilation of additional C repeat sequences allows us to localize the 5' end more precisely. The 5' end given in Figure 1 begins six nucleotides before the consensus previously reported (30). All intact C repeats, as well as the C repeat in pE3.4, are flanked by direct repeats (underlined in Figure 1) that differ between each individual repeat.

The C repeats analyzed so far can be divided into at least two subfamilies based on sequence similarities. The repeats in clones pEB1.3, pEB2.0 and pE6. 3 are more similar to each other than to the repeats in clones pE1.65 and in the uteroglobin gene. Likewise, the pE1.65 and uteroglobin C repeats are more similar to each other than to the other members. This can be seen in Figure 1, especially in nucleotides 55 through 64, where the pE1.65 and uteroglobin C repeats contain an almost identical insert, and in nucleotides 82 through 98, where both show an apparent deletion. Examination of more individual C repeats may reveal even greater heterogeneity. The more closely related C repeats are not closely linked in the genome; pEB1.3, pEB2.0 and pE6.3 were each obtained from a different genomic clone. Conversely,

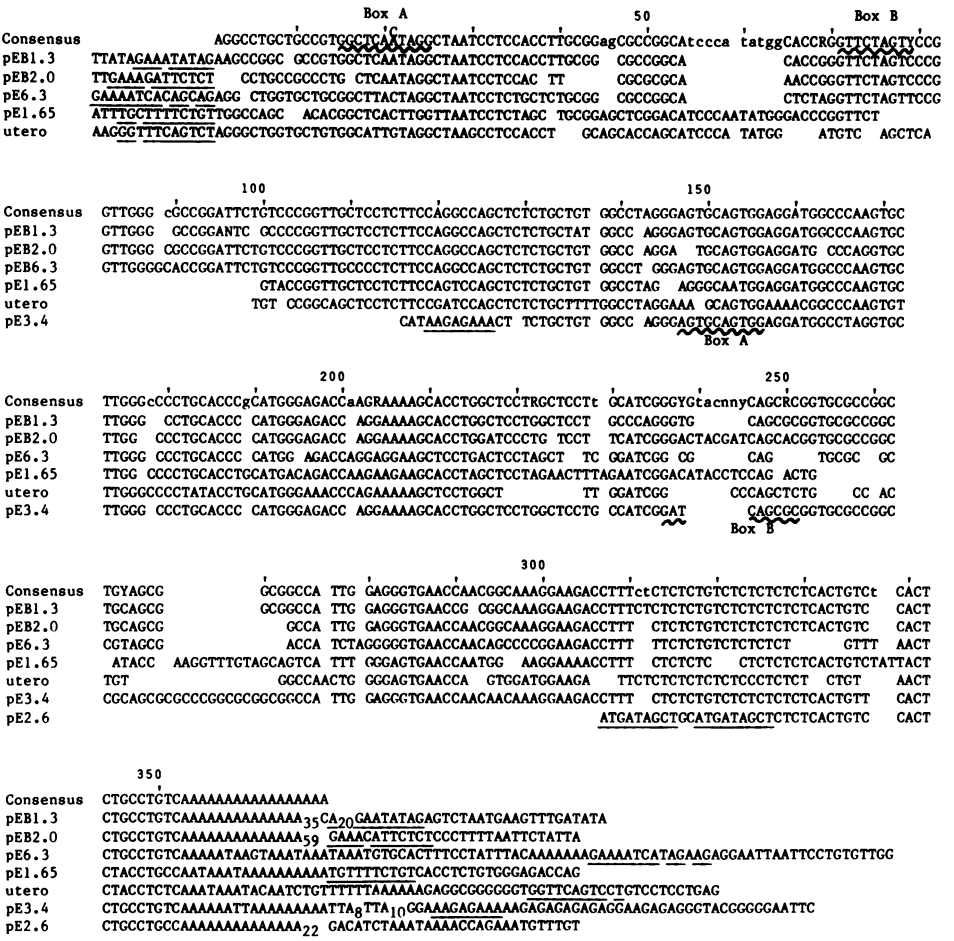


Figure 1. Nucleotide sequences of seven rabbit C repeats. The repeats are labelled by the plasmids containing them (30,32) or as "utero" for the C repeat located in the large intron of the uteroglobin gene (38). A consensus sequence is derived on the top line. The numbering begins with the first nucleotide of the consensus sequence. Sequences repeated before and after the repeat are underlined. Segments that match box A and box B of the RNA polymerase III internal control region (39) are marked with wavy underlining. Spaces are gaps introduced to improve the alignment.

closely linked repeat members are no more similar than are randomly chosen C repeats. Three of the repeats in Figure 1 are linked within the rabbit β-like globin gene family in the arrangement 5'-pE3.4C-β4-C-C-β3-pE1.65C-C-ψβ2-pE6.3C-β1-3', where sequenced C repeats are designated by the subclone

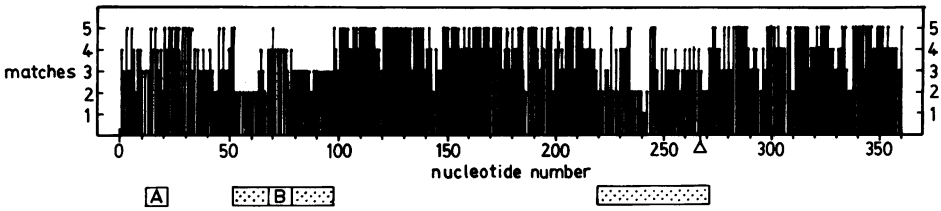


Figure 2. Variability within the C repeats. The number of times the consensus nucleotide appears at each position in the five intact C repeats is plotted against the position number. The most highly variable regions are indicated by the stippled boxes, and the box A and box B RNA polymerase III internal control sequences are marked as A and B. The inserts that occur between positions 269 and 270 of the consensus are indicated by the triangle.

containing them. The intact C repeats from this globin gene family (pE1.65 and pE6.3) are in two different repeat subfamilies, and the truncated C repeat (pE 3.4) has a novel internal duplication between nucleotides 255 and 270.

2) Variability Within C Repeats

While aligning the C repeat sequences in Figure 1, we noticed that much of the sequence variation occurred in two regions, approximately from nucleotides 50 through 100 and from nucleotides 220 through 270. This localized variability is presented graphically in Figure 2, which is a plot of the frequency of occurrence of the consensus nucleotide at each position. Only the five full-length repeats were included in this analysis. The consensus nucleotide occurs least frequently between nucleotides 55 and 98 and between nucleotides 222 and 272. More limited variability is scattered throughout the repeat, although the region from nucleotides 100 to 180 has remained the most constant. The positions of the bipartite RNA polymerase III internal control regions (box A and box B, ref. 39) are indicated in Figure 2. Box B occurs in the first variable region, and it is more highly conserved than the surrounding sequences. The first variable region can be used to divide the sequenced C repeats into two subfamilies, as noted in the previous section.

3) Truncation of C Repeats

Two of the C repeats listed in Figure 1 are shorter than the full-length repeat. Both the pE3.4 and pE2.6 repeats are shortened from the 5' end, and both 5' ends are in segments of repeating $(CT)_n$. The repeat member in pE2.6 is very short, with only 24 bp of C repeat sequence remaining before the A-rich 3' tract. It is not flanked by direct repeats and it could have formed

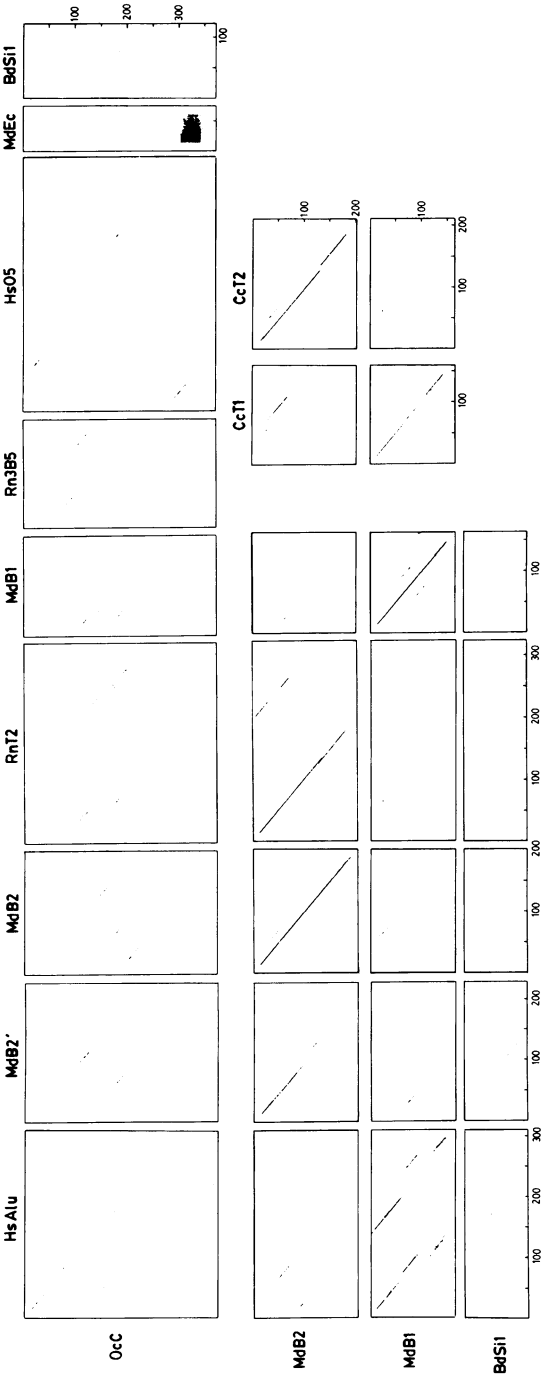
by a deletion between the $(CT)_n$ tract in the repeat and another $(CT)_n$ tract upstream. In contrast, the repeat member in pE3.4 is longer and is flanked by direct repeats. The flanking direct repeats suggest that the pE3.4 C repeat may have inserted into the genome in its shortened form. This shortened C repeat may be functional for transposition, since it contains sequences that match box A (nucleotides 147 through 156) and box B (nucleotides 236 through 251) of the RNA polymerase III internal control region (wavy underlining in Figure 1). Although this latter segment is within the second variable region of the C repeats, several of the individual repeats form a partial match with box B in the segment from nucleotides 231 to 251.

4) Relationship between Rabbit C Repeats and Other SINES

A dot-plot comparison of the C repeat consensus derived in Figure 1 with the sequences of other mammalian SINES did not reveal any extensive matches throughout the length of the repeat (Figure 3, top row of panels). Only short patches of similarity were found, and these frequently did not match corresponding positions within the repeats. The EC (evolutionarily conserved, ref. 44) repeat from mouse, which is mostly a simple repeating dinucleotide $(CT)_n$, matches the $(CT)_n$ stretch of the C repeat from positions 307 through 338. Since these simple repeats can match in several different frames, the dot-plot produces a block of dots rather than a simple diagonal.

A fairly high criterion (match 12 out of 14 nucleotides) was selected for the comparisons in Figure 3 in order to minimize the background, but this stringency does allow the detection of homology among SINES in different species. Human Alu is an imperfect dimer of an element related to the mouse B1 sequence (26), and this shows as the pair of diagonals in panel 1 of the third row of Figure 3. Likewise, the close similarity among the type 2 repeats of mouse (MdB2), rat (RnT2) and hamster (CcT2) is shown as the series of long diagonals in the second row of panels in Figure 3. Thus the C repeat of rabbits is not related to any SINE examined to the extent that the mouse, hamster and rat type 2 repeats or the rodent and primate type 1 (or Alu) repeats are related.

The sequences from other mammalian SINES that match with the C repeat consensus are aligned in Figure 4, and the extent of similarity is listed in Table 1. The matching segments usually involve different regions of the repeats. Some of the longer matches are with the human Alu repeat (C repeat positions 3-52 and 120-142), the rat 3B5 repeat (C positions 82-107), the mouse B2 repeat (C positions 145-171 and 205-239), the rat type 2 repeat (C positions 133-191), the human 05 repeat (C positions 292-324), and the mouse


BdSi1
MdEc
HsO5
Rn3B5
MdB1
RnT2
MdB2
MdB2'
HsAlu
CcT2
CcT1
OcC
MdB2
MdB1
BdSi1

EC repeat (C positions 307-338). These matching sequences, especially the shorter ones, could represent chance matches between random sequences, but some could reflect short regions of true homology (i.e. descent from a common ancestral sequence). It is difficult to satisfactorily discriminate between these possibilities, either statistically or functionally, so all observed matches are listed in Figure 4 and Table I. These data clearly show that none of the mammalian SINES are homologous throughout their length to the C repeat, but one cannot rule out the possibility that some short segments of the C repeat are derived from sequences now found in other mammalian SINES.

5) A Conserved Sequence in Many Mammalian SINES

Figure 4 shows that nucleotides 179 to 202 of the C repeat matches with DNA segments from both type 1 (Alu) and type 2 repeats. These matches are outside the previously noted homology between hamster type 1 and type 2 (29). The dot plots in Figure 3 also show that this region of the C repeat matches at least partially with all the sequences compared except mouse EC. After deriving a preliminary consensus, we searched other SINES for this sequence and in all instances we found a partial match, although the position varied in every repeat family. An alignment of this segment of the C repeat with the other repeat sequences examined is shown in Figure 5, and a consensus sequence is derived from the alignment. The sequences on lines 2-5 are from homologous regions of type 2 repeats, and the sequences on lines 6-8 are from homologous regions of type 1 repeats, so these lines do not represent comparisons between independent elements. This figure is a comparison among

Figure 3. A dot-plot comparison of mammalian SINE sequences. Matching sequences are shown as descending diagonals of dots; these patterns were produced using a program by Zweig (37), searching for 12 matches in a window of 14 nucleotides. The light grids occur at intervals of 50 nucleotides. The repeats are abbreviated as the initials of the genus and species following by the name of the repeat. The names and sources of the SINE sequences are: rabbit C = OcC=Oryctolagus cuniculus C (figure 1), human Alu = HsAlu = Homo sapiens Alu (26), mouse B2 variant = MdB2′ = Mus domesticus B2' (40), mouse B2 = MdB2 = Mus domesticus B2 (28), rat type 2 = rat dre 1 (41) = RnT2 = Rattus norvegius T2 (ref. 42; we included the second intact element and an adjacent modified element in the cluster of three repeats in the rat growth hormone gene intron), mouse B1 = MdB1 = Mus domesticus B1 (27), rat 3B5 = Rn3B5 = Rattus norvegius 3B5 (43), human 05 = Hs05 = Homo sapiens 05 (31), mouse EC = MdEC = Mus domesticus EC (44), artiodactyl short, abundant repeat = BdSi1 = Bovis domesticus SINE 1 (ref. 23; this repeat has been found in both cows and goats), hamster type 1 = CcT1 = Cricetus cricetus T1 (45), and hamster type 2 = CcT2 = Cricetus cricetus T2 (46). This nomenclature is based on suggestions of Voliva et al. (11).

Table 1. Similarity of SINE Segments

Segment in C repeats (nts)	SINE Compared	Segment in SINE compared (nts)	Number of matches -number of gaps	ave. length of segment	% Similarity
3-53	HsAlu	1-51	28	50	56
22-41	Hs05	72-91	12	20	60
82-107	Rn3B5	36-61	18	26	69
109-125	MdB2	100-116	13	17	76
110-121	RnT2	39-50	11	12	92
113-124	Rn3B5	141-152	11	12	92
120-142	HsAlu	228-250	15	23	65
133-191	RnT2	224-272	31	54	57
145-171	MdB2	125-151	18	27	67
168-197	Rn3B5	75-103	17	29.5	58
179-202	MdB2	68-90	15	23.5	64
179-202	MdB2'	65-87	11	23.5	47
179-202	MdB1	28-47	14	22	64
179-202	HsAlu	30-50	15	22.5	67
179-202	RnT2	68-90	15	23.5	64
205-239	MdB2	27-60	20	34	59
276-289	Rn5A1	106-119	11	14	78
292-324	Hs05	20-52	21	33	64
307-338	MdEC	1-32	28	32	88

Figure 4. Alignment of segments of mammalian SINES with the C repeat sequence. The rabbit C repeat consensus sequence is given in the first of each set of lines (OcC), and sequences shown to match in the Figure 3 dot-plots are aligned underneath. The position of each segment in its repeat element are given by the numbers before and after the segment. Nomenclature and sources of sequence data for the repeats are given in the legend to Figure 3.

					% Match	
					OcC	Consensus
OcC	178	CCCCTGCACCCGCATGGGAGACCAAGRAAA	207		-	78
MdB2	67	TCCCAGCAACCACAT GGTGGCTCACAACC	95	type	60	87
CcT2	64	TCCCAGCAACCACAT GGTGGCTCACAACC	92	2	60	87
RnT2	67	TCCCAGCAACCACAT GGTGGCTCACAACC	95		60	87
MdB2	64	TCCCAGTATACACAT GGCACCTCGAAACT	92		45	60
MdB1	27	TCCCAGCACTC GGGAGGCAGAGGCAG	52	type	64	80
CcT1	28	TCCCAGCACTC AGGAGGCAGAGGCAG	53	1	60	76
HsAlu	29	TCCCAGCACTT TGGGAGGCCGAGGTGG	55		63	75
HsAlu	87	AGCCTGG CCAACAT GGTGACCCCGTCTC	116		49	49
		AA				
BdSi1	45	TCCCAGGGACGG GGGAGCCTGGGGCTG	73		45	57
		GT				
Rn3B5	85	TCCCTGGAAGCTC TGGTTGCTGGCATTGT	113		42	53
Rn5A1	54	TCCCACCAACA AT GGAGG AGTGTTCC	29		48	68
Hs05	282	TCCCA CAATT TGGGAGATACAATTCA	318		36	46
		CGTGGGAATTC				
Rn4A1	39	GCCCTTCAACTGAAATGGATACAGAAATGT	70		54	54
		GG				
Consensus		TCCCAGCAACCACATGGGAGGCAGAGATCC			78	-
		C T T ARG				
		TT				
BK ori		GGAGGCAGAGGCGG				
SV40 ori		AGAGGCCGAGGCGG				

Figure 5. Alignment of SINE sequences in a highly conserved region. The numbers before and after the sequences refer to the nucleotide position in the repeat element. A consensus sequence is derived at the bottom and compared with sequences found at the origin of replication of the papovaviruses BK (47) and SV40 (48). The percentage of nucleotides that match the C repeat (OcC) or the derived consensus for this segment is given in the columns at the right of the figure. The % match calculation was limited to the region corresponding to 178-204 of OcC; the last 3 nucleotides do not form a good consensus among these SINES. The sequences for Rn4A1 and Rn5A1 are from Whitney and Furano (43).

9 independent sets of sequences, and substantial matches are found among all of them. The bovine IS-like repeat, BdSi1 (23), is a distinctly different class of SINE from the retroposon type 1, type 2, and C repeats, but it matches the consensus sequence for a 57% similarity. The rat repeats 3B5, 5A1, and 4A1 (43) are not obviously related to any other repeated sequences, but they also match the consensus sequence (53%, 68%, and 54% similarity, respectively). The rat 5A1 sequence is the complement of the published sequence (43). These similarities are detected even though each rat repeat has not been completely sequenced and it is not yet known whether or not the repeats are long or short.

Ten of the last 14 nucleotides of the consensus in Figure 5 match with a sequence found at the origin of replication of papovaviruses (47,48). The last 4 nucleotides in Figure 5 do not match the replication origin, and the

last 3 nucleotides are not highly conserved among SINES. However, the primate Alu (nucleotides 29-55) and rodent type 1 repeats match the replication origin sequence for 13 out of 14 nucleotides.

DISCUSSION

Structural variations among members of the rabbit C repeat family are of two types: 1) base substitutions and small insertions or deletions, which tend to be localized in two regions of the C repeat, and 2) large scale truncations, which have only been observed to involve the 5' ends. The fact that two regions of the C repeat (nucleotides 55-98 and 222-272) have accumulated more mutations than has the rest of the repeat element suggests that these regions are not required for the"activity" of C repeats, and conversely that the remainder of the repeat has some "activity." This "activity" could be limited to propagation of the C repeat (i.e. parasitic or selfish DNA) or it could include some host function as well. If the C repeats were simply "junk" DNA (inactive and functionless), then mutations should accumulate at random positions within the repeat, not in localized areas. Thus the nonrandom location of the bulk of the sequence changes argues for some activity of C repeats, albeit not necessarily a function advantageous to the whole cell. In this context it is interesting to note that the primate Alu repeat located 5' to the $\psi\alpha$-globin gene has diverged less than the surrounding DNA in a comparison between the human and chimpanzee sequences (49). In this case as well, the SINE is not behaving like inactive DNA.

The two examples of truncated C repeats are both shortened from the 5' end and begin in a stretch of repeating $(CT)_n$. Truncation from the 5' end is also seen in the long interspersed repeats such as primate Kpn (15) and mouse L1Md (11). In terms of the retroposon model for propagating repeats, these truncated members could arise by insertion in the genome of an incomplete reverse transcript. The repeat in clone pE3.4 fits this model well because it still has flanking direct repeats (Figure 1) presumed to form by repair after insertion at a staggered break.

Comparison of the C repeat consensus sequence with the sequences of other mammalian SINES failed to reveal any obvious homology. The search included all the common short repeats in humans (31), mouse (44) and rats (43). This absence of homology confirms that C repeats are not Alu-like (30), nor are they homologs of the rodent type 2 repeats or other known short repeats. Other repeats, however, are found in common between different species. For

example, the type 2 repeat is found at a high copy number in all rodents examined -mouse (B2, ref. 28), Chinese hamster (type 2, ref. 46) and rat (rat dre 1 or RnT2, ref. 41,42). Also, the type 1 or Alu-like sequence is found in two rodents (mouse B1, ref. 27, and Chinese hamster type 1, ref. 45) and primates (Alu repeat, ref. 26), but no Alu-hybridizing DNA was detected in rabbit genomic DNA (30). The absence of a type 1 repeat in rabbits suggests either that the type 1 repeat entered (or actively progagated in) the primate and rodent genomes separately or that it was lost from the rabbit genome. These events apparently occurred after the divergence of the lagomorph lineage from other mammals. The type 1 repeats and the C repeats are not stable, long-standing repetitive elements of the mammalian genome. The type 1 repeat is the only SINE so far found in species from more than one mammalian order.

Short segments of C repeat DNA do match with sequences from other repetitive elements. For example, the mouse EC repeat, which is a simple repeat of the dinucleotide CT, matches with the C repeat from nucleotides 307-338. Other matches involve parts of human Alu, mouse B2, human 05, and other repeats, but the matching sequences are usually from different regions within each repeat. A scrambled arrangement of sequences has been described previously for a comparison between mouse B2 and B1 repeats (28) and a rat type 2 repeat (clone 1B12, ref. 43). Also, the primate Kpn repeats contain permuted clusters of sequences (9). Earlier reassociation kinetics studies suggested that mouse (50) and Syrian hamster (51) DNA contain permuted clusters of repeats. The patches of matching sequences seen in the SINE comparisons in Figure 3 could result from an exchange of short DNA segments (30 to 50 bp) between different repeat members by recombination. After the sequences were acquired, they could become scrambled within the SINE, just as segments of the long interspersed repeats have been permuted within the repeat.

One segment of DNA, corresponding to C repeat nucleotides 178-204, has been observed in all SINES examined. This sequence is usually not part of a longer sequence that matches between two repeats, which suggests that it may be an independent element. Some clues as to a potential function for this conserved sequence may be gleaned from comparing it to sequences shown to be required for SINE transcription. The conserved sequence is found in two positions in the human Alu repeat. One position (nucleotides 29-52) is 3' to the RNA polymerase III promoter box A and is partially included in the segment required for enhancing transcriptional activity of the repeat (52).

Perhaps the conserved sequence plays a common role in transcription of many SINES. However, the conserved sequence is not located adjacent to an RNA polymerase III box A in all SINES.

The match between the 3' portion of the SINE conserved sequence and the DNA segment at the replication origin of some mammalian viruses suggests that the SINES may be involved in DNA replication. Such a sequence match and proposal has been made previously for primate Alu repeats (53) and the artiodactyl BdSi1 repeat (23). Recently it was demonstrated that Alu repeats could serve as replication origins in an in vitro, T-antigen dependent DNA synthesis assay (54). It would be informative to assay all the known SINES in this system to determined which contained operational origins in vitro and if the conserved sequence was required for this activity. If so, it is possible that members of the different classes of SINES (IS-like and retroposons) could serve as replication origins in different mammals. Of course, more passive functions for SINES are still possible, such as interrupting gene correction processes between duplication units (55).

ACKNOWLEDGEMENTS

We thank C.-K. J. Shen for helpful comments on the manuscript. This work was supported by PHS grants AM27635 and AM31961 from the NIH.

REFERENCES

1. Eden, F. C., Graham, D.E., Davidson, E. H. and Britten, R. J. (1977) Nucleic Acids Res. 4, 1553-1567.
2. Davidson, E. H., Hough, B. R., Amenson, C. S. and Britten, R. J. (1973) J. Mol. Biol. 77, 1-23.
3. Singer, M. F. (1982) Cell 28, 433-434.
4. Cameron, J. R., Loh, E. Y. and Davis, R. W. (1979) Cell 16, 739-751.
5. Spradling, A. C. and Rubin, G. M. (1981) Ann. Rev. Genet. 15, 219-264.
6. Potter, S. S. (1982) Nature (London) 297, 201-204.
7. Leuders, K. K. and Kuff, E. L. (1980) Proc. Natl. Acad. Sci. USA 77, 3571-3575.
8. Adams, J. W., Kaufman, R. E., Kretschmer, P. J. Harrison, M. and Nienhuis, A. W. (1980) Nucleic Acids Res. 8, 6113-6128.
9. Lerman, M. I., Thayer, R. E. and Singer. M. F. (1983) Proc. Natl. Acad. Sci. USA 80, 3966-3970.
10. Fanning, T. G. (1983) Nucleic Acids Res. 11, 5073-5091.
11. Voliva, C. F., Jahn, C. L., Comer, M. B., Edgell, M. H. and Hutchison, C. A. III (1983) Nucleic Acids Res. 11, 8847-8859.
12. Singer, M. F., Thayer, R. E., Grimaldi, G., Lerman, M. I. and Fanning, T. G. (1983) Nucleic Acids Res. 11, 5739-5745.
13. Burton, F. H., Voliva, C. F., Edgell, M. H. and Hutchison, C. A. III (1983) DNA 2, 82.
14. Gebhard, W., Meitinger, T., Hochtl, J. and Zachau, H. G. (1982) J. Mol. Biol. 157, 453-471.

15. DiGiovanni, L., Haynes, S. R., Misra, R. and Jelinek, W. R. (1983) Proc. Natl. Acad. Sci. USA 80, 6533-6537.
16. Miyake, T., Migata, K. and Sakaki, Y. (1983) Nucleic Acids Res. 11, 6837-6846.
17. Shafit-Zagardo, B., Brown, F. L., Zavodny, P. J. and Maio, J. J. (1983) Nature (London) 304, 277-280.
18. DiNocera, P. P., Digan, M. E. and Dawid, I. B. (1983) J. Mol. Biol. 168, 715-727.
19. Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T. and Gesteland, R. F. (1981) Cell 26, 11-17.
20. Jagadeeswaran, P., Forget, B. G. and Weissman, S. M. (1981) Cell 26, 141-142.
21. Rogers, J. (1983) Nature (London) 306, 113-114.
22. Genbauffe, F. S., Chisholm, G. E. and Cooper, T. G. (1984) J. Biol. Chem. 259, 10518-10525.
23. Watanabe, Y., Tsukada, T., Notake, M., Nakanishi, S. and Numa, S. (1982) Nucleic Acids Res. 10, 1459-1469.
24. Schon, E. A., Cleary, M. L., Haynes, J. R. and Lingrel, J. B. (1981) Cell 27, 359-369.
25. Schimenti, J. C. and Duncan, C. H. (1984) Nucleic Acids Res. 12, 1641-1655.
26. Schmid, C. W. and Jelinek, W. R. (1982) Science 216, 1065-1070.
27. Krayev, A. S., Kramerov, D. A., Skryabin, K. G., Ryskov, A. P., Bayer, A. A. and Georgiev, G. P. (1980) Nucleic Acids Res. 8, 1201-1215.
28. Krayev, A. S., Markusheva, T. V., Kramerov, D. A., Ryskov, A. P., Skryabin, K. G., Bayev, A. A. and Georgiev, G. P. (1982) Nucleic Acids Res. 10, 7461-7475.
29. Jelinek, W. R. and Schmid, C. W. (1982) Ann. Rev. Biochem. 51, 813-844.
30. Cheng, J.-F., Printz, R., Callaghan, T., Shuey, D. and Hardison, R. (1984) J. Mol. Biol. 176, 1-20.
31. Sun, L., Paulson, K. E., Schmid, C. W., Kadyk, L. and Leinwand, L. (1984) Nucleic Acids Res. 12, 2669-2690.
32. Lacy, E., Hardison, R., Quon, D. and Maniatis, T. (1979) Cell 18, 1273-1283.
33. Hardison, R., Butler, E. III, Lacy, E., Maniatis, T., Rosenthal, N. and Efstratiadis, A. (1979) Cell 18, 1285-1297.
34. Shen, C-K. J. and Maniatis, T. (1980) Cell 19, 379-391.
35. Maxam, A. M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560-564.
36. Maxam, A. and Gilbert W. (1980) Methods in Enzymology 65, 499-560.
37. Zweig, S. E. (1984) Nucleic Acids Res. 12, 767-776.
38. Suske, G., Wenz, M., Cato, A. C. B. and Beato, M. (1983) Nucleic Acids Res. 11, 2257-2271.
39. Ciliberto, G., Castagnoli, L. and Cortesi, R. (1983) Curr. Top. Dev. Biol. 18, 59-88.
40. Kalb, V. G., Glasser, S., King, D. and Lingrel, J. B. (1983) Nucleic Acids Res. 11, 2177-2184.
41. Lemischka, I. and Sharp, P. (1982) Nature (London) 300, 330-335.
42. Barta, A., Richards, R. I., Baxter, J. D. and Shine, J. (1981) Proc. Natl. Acad. Sci. USA 78, 4867-4871.
43. Witney, F. R. and Furano, A. V. (1984) J. Biol. Chem. 259, 10481-10492.
44. Bennett, K., Hill, R., Pietras, D., Woodworth-Gutai, Kane-Haas, C., Houston, J., Heath, J. and Hastie, N. (1984) Mol. Cell Biol. 4, 1561-1571.

45. Haynes, S., Toomey, T., Leinwand, L. and Jelinek, W. (1981) Mol. Cell Biol. 1, 573-583.
46. Haynes, S. R. and Jelinek, W. R. (1981) Proc. Natl. Acad. Sci. USA 78, 6130-6134.
47. Yang, R. C. A. and Wu, R. (1979) Science 206, 456-462.
48. Reddy, V., Thimmappaya, B., Dhar, R., Subramaniam, K., Zain, S., Pan, J., Celma, M. and Weissman, S. (1978) Science 200, 494-502.
49. Sawada, I., Beal, M. P., Shen, C.-K. J., Chapman, B., Wilson, A. and Schmid, C. (1983) Nucleic Acids Res. 11, 8087-8101.
50. Cech, T. R. and Hearst, J. E. (1976) J. Mol. Biol. 100, 227-256.
51. Moyzis, R. K., Bonnet, J., Li, D. W. and Ts'o, P. O. P. (1981) J. Mol. Biol. 153, 871-896.
52. Perez-Stable, C., Ayres, T. M. and Shen, C.-K. J. (1984) Proc. Natl. Acad. Sci. USA 81, 5291-5295.
53. Jelinek, W., Toomey, T., Leinwand, L., Duncan, C., Biro, P., Choudary, P., Weissman, S., Rubin, C., Houck, C., Deininger, P. and Schmid, C. (1980) Proc. Natl. Acad. Sci. USA 77, 1398-1402.
54. Ariga, H. (1984) Mol. Cell. Biol. 4, 1476-1482.
55. Hess, J. F., Fox, M., Schmid, C. and Shen, C.-K. J. (1983) Proc. Natl. Acad. Sci. USA 80, 5970-5974.