J Mol Evol (1987) 25:65–73

Journal of
Molecular Evolution
© Springer-Verlag New York Inc. 1987

The Sequence of the 3.3-Kilobase Repetitive Element from *Dipodomys ordii* Suggests a Mechanism for Its Amplification and Interspersion

Paul Keim and Karl G. Lark

Department of Biology, University of Utah, Salt Lake City, Utah 84112, USA

Summary. DNA from the kangaroo rat, *Dipodomys ordii,* contains a 3.3-kb, highly repeated sequence that is interspersed throughout the genome in small tandem clusters. One 3.3-kb unit has been cloned into pBR322 and the nucleotide sequence determined. The clone used was shown to be representative of the bulk of such sequences found in the genomic DNA. The sequence contains 10 homologous subunits each ca. 260 bp in length. Comparison of these to one another yielded a 258-bp consensus sequence containing a 35-bp terminal inverted repeat. Two unique stretches also occur. One of these contains a region that could serve as a promoter for RNA polymerase III; the other contains a sequence related to the ARS sequences of yeast. It is proposed that an ancestral sequence similar to the consensus sequence was amplified to 10 or more units, and that, subsequently, two other sequences were inserted. The properties of these insertions may have led to the dispersal of the sequence throughout the genome.

Key words: Repetitive DNA — DNA amplification — DNA interspersion — *Dipodomys ordii*

Introduction

Mammalian genomes contain highly repeated interspersed sequences (Singer 1982). Studies of such sequences have suggested mechanisms for their amplification and dispersal in the genome, such as unequal crossing-over, gene conversion, and possibly transposition (Dover 1982; Finnegan 1985). Recently it has been proposed that because many of these sequences contain control elements of RNA polymerase III, they may be amplified and dispersed via an RNA intermediate that is inserted into a chromosomal site (retroposon) and then copied by reverse transcriptase (Rogers 1985a,b). We describe here a repeated sequence from the kangaroo rat, *Dipodomys ordii,* which contains structural features suggesting a mechanism for its amplification.

The kangaroo rats (genus *Dipodomys*) have been shown to have a variable genome. There has been rapid chromosomal evolution (Stock 1974), and the satellite DNAs in *Dipodomys* species show variation involving the amplification or loss of GC-rich sequences. It is clear that massive genomic changes have occurred in the Dipodomii and that these molecular changes are correlated with speciation within the genus (Hatch et al. 1976).

The genome of *D. ordii* contains a highly repeated 3.3-kb sequence. Sequences hybridizing to this element are dispersed throughout the genome (Liu and Lark 1982). In a previous study we reported that portions of this sequence are homologous and can recombine when introduced into bacteria as cloned inserts contained in phage or plasmid vectors (Keim et al. 1984). We have now obtained the complete sequence of the 3.3-kb unit. The structure suggests that an ancestral 258-bp sequence was amplified into a 10-unit tandem repeat. Later, two other sequences were inserted into the tandem. We propose that these insertions possess special qualities that have resulted in the massive amplification and dispersal of the 3.3-kb sequence throughout the genome.

Offprint requests to: K.G. Lark

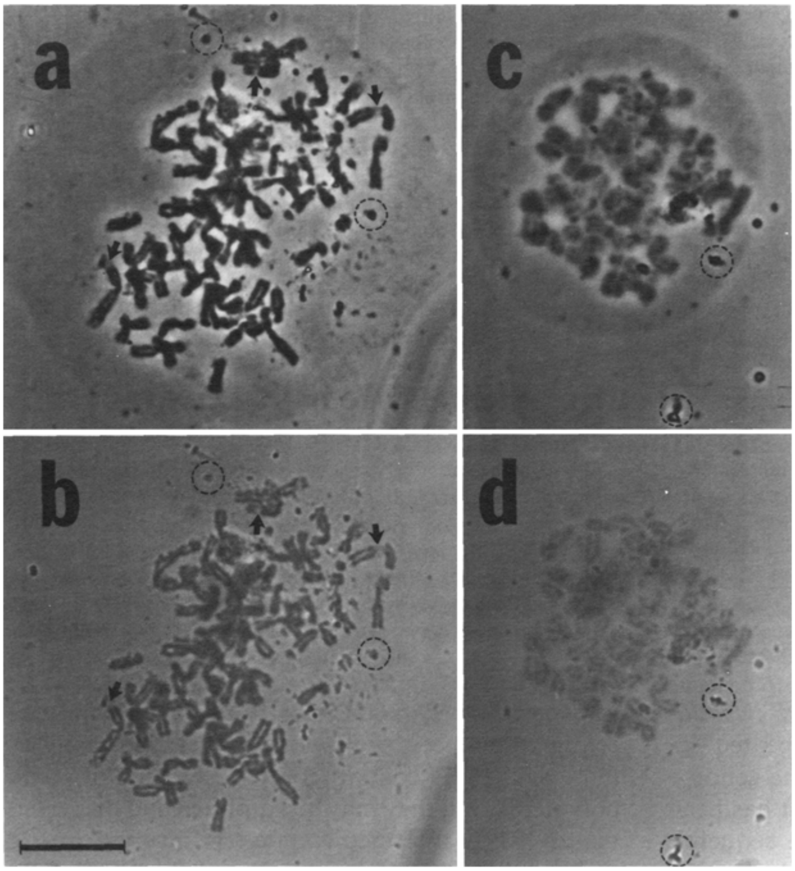

Fig. 1. In situ hybridization of the 3.3-kb repeated sequence to *D. ordii* metaphase chromosomes. Metaphase chromosomes were prepared from cultured *D. ordii* cells as described in Materials and Methods. A portion of the slide containing the metaphase spreads was hybridized with a biotinylated probe prepared from KR-2 DNA. The metaphase shown in **a** and **b** is taken from this portion of the slide. The entire slide was then treated with streptavidin-conjugated alkaline phosphatase, staining of the slide with 5-bromo-4-chloro-3-indolyl phosphate and nitro-blue tetrazolium was developed for 30 min, and the entire slide was then stained with orcein in order to visualize the chromosomes. The control metaphase in **c** and **d** was taken from a portion of the slide *not* hybridized to the probe. Dye was deposited on the hybridized regions of the chromosome, changing their color and defining them as darker regions in the photograph. Examples of nonhybridized regions are shown in **a** and **b** by the arrows; background clumps of dye have been indicated within the circled areas. These demonstrate that both preparations were treated with streptavidin-conjugated alkaline phosphatase. **a** and **c** Phase contrast; **b** and **d** Bright field. Bar = 20 µm

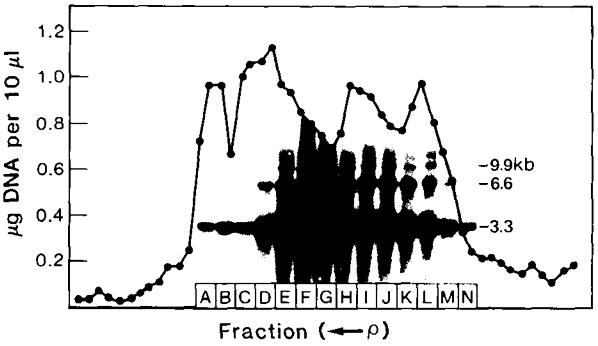

Fig. 2. Hoechst dye cesium chloride isopycnic equilibrium centrifugation of *D. ordii* genomic DNA. Genomic DNA from *D. ordii* liver was centrifuged to equilibrium in cesium chloride with Hoechst dye 33258. The gradient was fractionated and the DNA content of each fraction determined by ethidium bromide fluorescence (LePecq and Paoletti 1966). Two micrograms of DNA from the sets of fractions (A–N) was digested with *Bam*HI. These DNAs were separated by agarose gel electrophoresis, transferred to nitrocellulose, and probed by hybridization to radioactive KR-2 DNA. The autoradiograph is superimposed on the gradient profile. The densities of fractions A and N were 1.729 g/ml and 1.684 g/ml, respectively

Methods and Materials

Cloning. The KR-2 clone was isolated as described previously [the pBR322 clone described in Liu and Lark (1982)]. DNA was amplified in *Escherichia coli* strain 802 (Wood 1966) and purified by banding on CsCl (Liu and Lark 1982).

Dipodomys ordii Genomic DNA Isolation. Kangaroo rat genomic DNAs were isolated as described in Liu and Lark (1982). Briefly, livers were frozen with liquid nitrogen and ground with a mortar and pestle to a fine powder, which was mixed with 50 mM Tris (pH 8.0), 100 mM NaCl, 100 µg/ml proteinase K, and 0.5% sarkosyl, and digested at 65°C for 1 h. This crude preparation was phenol/chloroform-extracted three times, chloroform-extracted one time, and centrifuged to equilibrium on cesium chloride gradients in the presence of ethidium bromide. The ethidium bromide was removed with isopropanol, and the DNA exhaustively dialyzed against 10 mM Tris (pH 8.0) and 0.1 mM EDTA.

Hoechst Dye Cesium Chloride Isopycnic Centrifugation. Dipodomys ordii DNA was separated into different satellite fractions using Hoechst dye 33258 (0.4 µg/µg DNA), which binds to AT-rich DNAs and decreases their buoyant density. This technique was described by Keim (1986).

In Situ Hybridization. Cultured *D. ordii* cells, grown as previously described (Liu and Lark 1982), were used for preparing chromosome spreads. Chromosome preparation followed the protocol of T. Hori (Hori et al. 1985). The cells were blocked with colcemid for 1 h followed by a 10-min trypsin treatment. Liberated cells were collected by centrifugation and then swollen with a hypotonic KCl solution (0.075 M). Swollen cells were fixed with methanol/glacial acetic acid (3:1) and then spread on glass microscope slides. The DNA probe was created by nick translation (Rigby et al. 1977) of KR-2 DNA using biotin-11-deoxyuridine triphosphate (Bethesda Research Laboratories). The probe hybridization and visualization were done according to Engels (Engels et al. 1986) using streptavidin conjugated to alkaline phosphatase [Bethesda Research Laboratories' DNA detection system (cat. no. 8239)]. Orcein was used as a nonspecific DNA

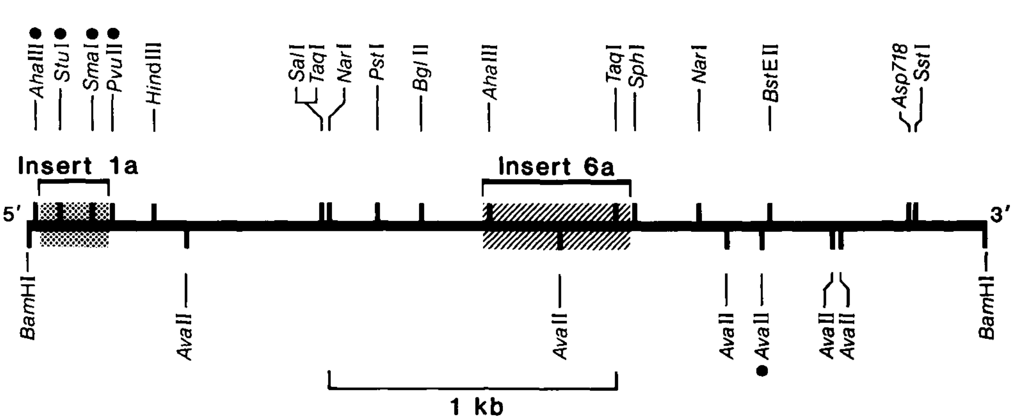

Fig. 3. Restriction map of the KR-2 clone. Restriction sites were determined by digestion of the KR-2 plasmid DNA with the indicated enzymes. All sites are in agreement with the sequence data. With the exception of the sites noted (●), the restriction sites indicated were used for end-labeling during sequence determinations. The positions of the proposed insertions 1a and 6a (see text) are indicated by shading

stain (Engels et al. 1986). Chromosomes were observed under oil immersion using a Zeiss photomicroscope.

DNA Restriction Analysis. Restriction enzymes were obtained from New England Biolabs, Bethesda Research Laboratories, and International Biotechnology, Inc., and were used according to the manufacturers' directions. DNA was resolved by agarose gel electrophoresis and transferred to a nitrocellulose membrane as described by Southern (1975). Molecular hybridization probes were radioactively labeled by nick translation (Rigby et al. 1977).

DNA Sequencing. All sequences were determined by the technique of Maxam and Gilbert (1980). Both the 5' and the 3' strands were sequenced using the restriction sites illustrated in Fig. 3. End labeling of DNA fragments was accomplished as follows: 5' with T4 polynucleotide kinase (Maniatis et al. 1982), 3' by polymerization using Klenow fragment (Maniatis et al. 1982), and by terminal transferase using ddATP alpha P32 (Amersham Corporation). Typically 300 bp of sequence adjacent to the labeled end were determined.

Computer Analysis. Homology comparisons were made using a program generously supplied by Dr. John Shepherd (Biocenter, University of Basel). The coding potential of the KR-2 sequence was determined by John Shepherd using a method described previously (Shepherd 1981). The secondary structure of the consensus sequence (Fig. 7) was prepared using a program generously provided by Dr. M. Zuker (Zuker and Stiegler 1981). A search of sequence homology within the Genbank library of sequences was carried out by the National Biomedical Research Foundation using their search program.

Results

More than 10^5 copies of the 3.3-kb sequence represented by KR-2 are found within the genome of *D. ordii* (Liu and Lark 1982). The copies must occur in tandem arrays, because enzymes that cut the sequence at a single restriction site (e.g., *Hin*dIII or *Bam*HI) generate the 3.3-kb fragment from animal genomic DNA (Liu and Lark 1982). This fragment appears, however, to be distributed throughout most of the genome, because in situ hybridization shows that the KR-2 clone hybridizes to all chromosomes and that hybridization is *not* confined to a particular region of each chromosome. Figure 1a and b presents the results of in situ hybridization using a biotinylated KR-2 probe and subsequent development with alkaline phosphatase. (A control, nonhybridized, metaphase is shown in Fig. 1c and d. The metaphase in Fig. 1c and d was taken from the same slide as that in Fig. 1a and b, but from a region that was not hybridized to the biotinylated probe. This region, however, was treated with streptavidin-conjugated alkaline phosphatase and orcein stain.) Hybridized regions appear as regions of greater contrast in both phase contrast (Fig. 1a) and in bright field (Fig. 1b) microscopy. Arrows indicate examples of regions to which the probe did not bind; such regions are rare.

The KR-2 sequence appears to be distributed in small clusters, because the 3.3-kb fragment can be generated by restriction enzyme digestion of all fractions of DNA separated according to their AT/GC content (Fig. 2). In this experiment, DNA was centrifuged to equilibrium in CsCl-containing Hoechst 33258 dye, which decreases the buoyant density of AT-rich DNA and thus enhances the separation of different fractions of genomic DNA. [The DNA of *D. ordii* contains a large proportion of GC-rich satellite DNA (Hatch et al. 1976).] Fractions were digested with *Bam*HI and the presence of the 3.3-kb fragment noted. (The presence of 6.6- and 9.9-kb fragments is due in part to incomplete digestion and in part to infrequent loss of *Bam*HI restriction sites; see below.) The sequence of one 3.3-kb fragment was obtained from the clone KR-2 [Liu and Lark (1982); referred to as KR-1 by Keim et al. (1984)]. Figure 3 presents a restriction map of this insert, including the sites used for obtaining the sequence.

Because the sequence was derived from a single clone, it was important to ascertain whether it was representative of the more than 10^5 copies present in the animal. A preliminary study (Keim et al.

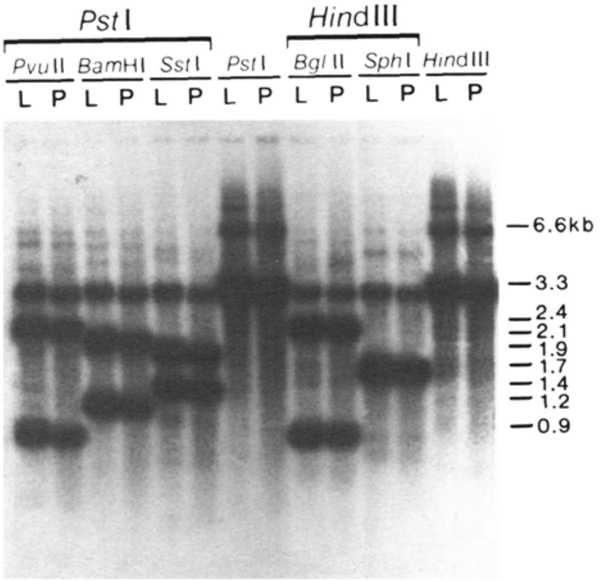

Fig. 4. Analysis of *D. ordii* genomic DNA with restriction enzymes. Genomic DNAs from two populations (L, Little Sahara; P, Pelican Lake) were digested with one or two restriction enzymes. These DNAs were separated by agarose gel electrophoresis, transferred to nitrocellulose, and probed by hybridization to radioactive KR-2 DNA. The sizes of fragments were determined by comparison to DNA standards (not shown) and correspond to the sizes predicated from the sequence of KR-2 (see Fig. 3)

1984) already had compared about 700 bp of the KR-2 sequence with the same region in two other clones. In addition, genomic DNA isolated as a 3.3-kb restriction fragment was sequenced directly (Maxam and Gilbert 1980) and compared to the clones. In all, 2400 bp of sequence were compared and differences were detected at 16 positions. On the average, one nucleotide change was observed per 150 nucleotides of sequence (0.7%). The KR-2 3.3-kb sequence contains a number of restriction sites, of which several (e.g., *Bam*HI and *Hin*dIII) had already been shown to be present in the majority of the 3.3-kb sequences found in the genome (Liu and Lark 1982). Figure 4 presents examples of restriction digests of *D. ordii* genomic DNA obtained with *Pst*I or *Hin*dIII in conjunction with several other enzymes. In each case, the data show that the fragments obtained are consistent with the sequence of KR-2 shown in Figs. 3 and 5. In addition to *Bam*HI, *Hin*dIII, *Pst*I, *Pvu*II, *Sph*I, *Bgl*II, and *Sst*I, other restriction sites that occur in KR-2 corresponded to sites found in the majority of the genomic DNA 3.3-kb fragments. These included *Aha*III, *Sal*I, *Stu*I, and *Taq*I. The *Asp*718 site found in KR-2 is polymorphic within the genomic DNA, and about 50% of the genomic sequences lack an *Asp*718 site. The remainder contain the site found in KR-2. KR-2 lacks a *Bst*EII site, which is found adjacent to the *Bam*HI site in the majority of genomic 3.3-kb sequences. With the exception of this *Bst*EII site, the correspondence of restriction sites found in KR-2 with restriction sites found in the genomic DNA is further evidence that the sequence of the KR-2 clone is representative of the 3.3-kb repeat as it occurs throughout the DNA of the animal.

The data in Fig. 4 also show that 3.3-kb fragments are present in all of the preparations digested with two enzymes. These fragments were present despite all precautions taken to ensure that the DNA had been completely digested. We believe that these fragments represent sequence variation in which the loss of a restriction site, corresponding to either one of the two enzymes, has occurred. (The loss of the *Bst*EII site from KR-2 is probably an example of such a change.) Such sequence variation is similar to variation observed in other repeated DNA, e.g., the α-satellite of the African green monkey (Thayer et al. 1981; Lee and Singer 1982) or others reviewed by Miklos (1985). Densitometer scans of autoradiographs that had been exposed for shorter times allowed us to estimate that 4.5% of the DNA was found in the 3.3-kb fragments, corresponding to a variation frequency of 0.37% per nucleotide (assuming that each of the six nucleotides in each of two sites has an equal probability of change).

The sequence of the 3.3-kb fragment from KR-2 is presented in Fig. 5. The main features of the sequence are the presence of 10 homologous subunits of ca. 260 bp each (indicated as [1] to [10] in Fig. 5) and two nonhomologous regions, believed to be insertions within subunits 1 and 6 (indicated as insertions 1a and 6a in Figs. 3 and 5). Inserts 1a and 6a showed no homology to each other or to any other part of the sequence (comparative data not shown).

The homologous regions were identified by computer and aligned (Fig. 6) after correcting for several small (1–5-bp) insertions or deletions. By comparing the subsequences to one another, a consensus sequence was established (Fig. 6). It contains a pair of almost perfect inverted terminal repeats (ca. 35 bp each), and the secondary structure of the consensus sequence [generated by computer using a folding program of Zuker and Stiegler (1981)] indicates that a stem of about 40 bp can be formed utilizing pairing of G with T. A search through the Genbank data base failed to reveal any sequences related by homology to the consensus sequence.

In Fig. 7 the consensus sequence is compared to the 3.3-kb sequence from Fig. 5 (using a homology comparison program provided by Dr. John Shepherd). It can be seen (Fig. 7A) that the 10 repeated subunits are separated into two groups by two unique regions (inserts 1a and 6a). [The presence of the inverted terminal repeats also can be detected (Fig. 7B).]

```
                                                               <---POLY A TAIL                    <--BOX B
1)                   10]|[1                         |INSERT 1a AAAACAAAACAGAACAAAACAAAACA---------CCGAACTTG---
GGATCCAGGTCACGGTCAAACAATAATGTTTTAAATCCCCGAAGTTCTGAGTGCTCTGTATACTTTTGTTTTGTCTTGTTTTGTTTTGTCGGTCGTGGGGCTTGAACTTT

                                                                                <---- BOX A
111)----------------------------------------------------------------------------CC--CAAATAA-------------------
AGGCCTGGGCGTTGTCCCTCAGCTCTTCAAAGTCAAGGCCAGCGCTCTACCTTGAGCTATAGCGCAACTTCCGGTTTTCTGGTGGTTTATTGGAGATAGGTGTCTCACCT

221)-----------------------------------------------------------------------------END INSERT-1a|
TCTCTCCTGCCCGGGCTTGCTTGCTTGCAACCGTGTTCCTCAGATTTCCGGCTCCTGAGTAGATAGGATTACAGCTGTGAGCCACCAGTGCCTGGCTCTATCTTACTCTT
331)
TGCATCCTGAGTTCCTAAACACTTGTGACCTCTCTCACAATGAAAAGAAAAAAATTTCACCTGTTATTTTAGACAGAATGAGGCACTTTCAATATAGCTAGGATGCAGAG

441)                                                                                               1]|[2
CAAAGCTTCTTGGGAAAGCTGGGTGTTGTGAATTTCTCTCTGGAGGGGTGTCTCTTCCTTGGGGAAGGTGTAGAGCCCTCCTTCCCGGAGGGCCTTAAGACTGTTCTAGG
551)
GGTCCCATATTCTGAGTACTCTATCTTATCGTTTGTACCTGCGCTCTCTAACACTGGTCACTTCTTCCAGAATTGGAAGAACAAAGTTTGGACGGTTTACTGTGGATAGT
661)
GCAAGGGTCATTAAGCCCTGCTAAGATTCAGAGCACAGTTGACTCGGACAGATCATGCTTTCTATTTCTACTAAGTGGCTTGTCTTGCTTTCGGGAAAAGTTGAGAGCAC

771)                    2]|[3
TCAGAACTTGCAGACCTTAATACAATTTTCTTTATGTTCTCCAAATTCTGAGTACTCTCTCTTGTTGTTTGCATCCTGTGCTCCCAAACCCTCCACAGCTCTCTCAAATT
881)
TGAAAGTAAAACGGATTGTCCACGGGATGGTGACACCAAGAGGTGCTTTGACTCATGTCAGGGTGCGCACAGGGGCTGCTAGTAACCGACTGATATTGTCTATTTCTTTT

991)                                                          3]|[4
CGGTGGTTGTTCTTTTCTTGGTCGACGGTATAGAGCATTCAGATCCTTGGCGCCATGAAACAATTTTGTTTTAAGTTCCCCAAATTCTCAGTGCTCTATCTTGTCGTTTT
1101)
CATCCTGAGCATCCTAACACTGGTCAGCCCACCCTGAACTGCAAGGGAAAAATTCTGACTTGTTACTGTGGACAGTGTAAGGGGCTTGAAGCCCTGCTAGGCTGCAGACC

1211)                                                                                     4]|[5
GTTGTCGCCTGTGATAGAACCATGCTGTCTATTTCTCTTTCCTGGCTTGTATCTCCTTTGAGGAATTGTAGAGCCTTCAGAACTTGGGCAATATTGTTTACCTCACCAAA
1321)
GGTCTGAGTGCTCTATGTTGGCGTTTACAACCTGAGATCTCAAAGACTTTTCAACCCTCCCAGTGGTAAGGAGAAAAACGGAACCTGTGATTGTAGACAGGATGAGACAC
1431)
TTTCTGTCTTGCTATTACGCAGACCAGAGCTTATCGGAAGATTGGTGCTGTCAATTTCTCTATAGAGGCGTGTCACTTCTATGGGGAAGGATTAAACCCTTCATTGCTTG

1541)      5]|[6                                      |INSERT 6A<---ARS HOMOLOGY---------->-------------------
GGGGCCTTAAAACTGTTCTAAAGTCCCCGTGTTCTGAGTGCTCCATCTTTTCTTGTTTTTAAATTTATTTATTTGTTCATTTATTTTTTTACAAAACAATGTACAGACG

1651)---------------------------------------------------------------------------------------------------------
GGTTCCAATTACATGTGCAAGATGATGAGTACATTTTTTATAAAGTATTGTAACGTCGTCTTTGTTTCCCATTCCCACCTTTTCCCATCCTCCTCCTTTCCTCCCCCACA

1761)---------------------------------------------------------------------------------------------------------
GTCCTCCCCTCTCCCATACCCCCACCACTCACAAACCAATGTTGTAAAGTTCGTTTTGGACATAACGAATTGTCCATCGTATTTTTGCTATGGGTCCACCTGTATCCCTA

1871)---------------------------------------------------------------------------------------------------------
TTGAAACGAATTGTTTTCCCCCTTCCCCTCCACAAATTATACCGACATGAAATACAGTGAAACAGTTGTTATAAAATGGTGGAGAGACTTACAAAGTAAAAAAGAAATGA

1981)---------------------------------------------------------------------------------------------------------
TCTTATGAATGTGTCATAAACAATCGCTTGAAAACAAATAAAAAAAACTTTCGAATACATATATTAGGGTTCAGTTCATTAACATCACCTTATCTGACTTTCTCTGCTTT
2091)
INSERT6 a|END
ACTCAAACTTTCGTTTGCATGCTGAACTCGCTAACACTGGTCAGTTCTCCCAGAATTGGAAGAAAAAATTGGACTGGTTCCTGTGGATGGACAGTATAAGTGGCTTTCAA
2201)
CCTAGCTATTATGCAGACCGGAGAAACCTGGGTTAGAAAGAAGCAGTCTACTTGTATTTAGAGGCTTACCTCTCCTTTGAGGAAGATGTAGATCATTGAGAAATTGACCC

2311)     6]|[7
CTGAAAAACAATTTTGTTTTAGGGCGCCGAATTTCTTTGTTGCTCTATCTAGTCGTTTGCATCCTGTGCTCTCAAACACTTTTCAGCTCTCCCAGAATTGCAAGGAATAT
2421)
AATTGGACCTGTTATTGTTGACAGTATGTGGAACAGGATGTAGATTAGGGCTTAGTGAGAAAGGTCGGCGCTGTCCATATATCTTTGGAGGCTTGTCTCTCCATTGGGGA

2531)                            7]|[8
AGGAGTAGAGCATCCTCACTTTGGGACCTTAAAACTGTTTTAAGGTCACCAATTTCGGAGTGCTCTCTAATGTCGTTTGCATCCTGAGCTGACTAACACAGATCAGCTCT
2641)
CCCAAAATTGGATGAAACAAATTGGACTGGGCTCTGTGGACAGTGTGAGTGGCATTAAGCCCTGATGGATGCAGACCAGAGTTGCCTGGGCGAGAAAAATGCAGTCTATT

2751)                                                              8]|[9
TCTATTTAGCGGCTTTTCTCTCCTTTGGGGAAGGTGTAGAGCACTCAAAACTTGGGGACCTTAAAACAATTGTTTTAAGGTCCCCAAGTTTTGAGTGCTCTTTCTTGTTA
2861)
TTTGCATCCTGTGCTCCCAAAACCTCCTCAGCTCTCCCAAATATTTCAAAGAAATCTTTTGACCTATGGCTGATGATACAGAGAGGGCCTGTGAACCAAGCCAGGATGCA

2971)                                                                                                    9]|[10
CATAGGAGCTACTAGAGACTGAAATATGCTGTCTATTTCTCTTTTGTGTCTTGTCTCTTCATTATGGAAGGTGGAGAGTATTCACAACTGAGGCGACCACAAACAGTTCT
3081)
GTTTTAAGGTACCCAAGTTCTCAGGGCTCTATCTTGTCGTTTTCATCCTGAGCTCCCTAACACTGCTGAGCAATCCCAGAATTGCAAGGAAAAAATTGGCACTGGTTAAT
3191)
GGGGACAGTGTGAGGAGGTTGAAGCCCTGCTAGGATGCAGACCGAAGATGCCTGGGATAGACCGATGCTGTCTAACTCTCCTTACTGCCCTGTATTACCGTTGGGGAATA
3301)
TGTAGAGCATTTGGATCC
```

Fig. 5. The nucleotide sequence of KR-2. The sequence of KR-2 is written 5′ to 3′ as shown in Fig. 3. The 10 258-bp related subsequences are delineated by] ▌ [. Inserts 1a and 6a are marked ---. The RNA pol III promoter sequence, poly A tail, and ARS homology are highlighted (▁). The direct terminal repeats associated with inserts 1a and 6a are noted GCTCT or CTT. In addition, the region homologous to a sequence flanking human asn-tRNA is denoted ▙▁▁▁▟

Insert 1a is flanked by direct repeats of five base pairs (GCTCT). One of these direct repeats is found in the consensus sequence, whereas the other is unique to subsequence [1] and has been placed within the insert (see Fig. 5). In addition, the complementary strand of insert 1a contains a promoter for RNA polymerase III (Ciliberto et al. 1983) adjacent to a poly A tail. This is noted in Fig. 4 (bp 63 to 201). A search through the Genbank data base indicated homologies only in the region containing the RNA pol III promoter and the poly A tail. In this region some similarity to the Alu family was found. In addition, a striking match (41 out of 46 bp) was found (bp 271 to 317) with a sequence that occurs as a flanking region of a minority class of human asparagine tRNA (Ma et al. 1984). The significance of this homology is not clear.

Insert 6a is flanked by a direct repeat of the tri-

```
TTTGTTTTAAGGTCCCCAART-TCTGAG-TGCTCTATCTTGTCGTTTGCATCCTGAGCTCCC?AACACTGGTCAGCTCTCCCAGAATTG?
[1]                           |insert 1a
AA........ATC...G..G........-...........ACTC.............T...TA......T..G.C.....T..C...GAA
[2]
--....C..G..GT....TA..........A.........A.......T.-....C......T...........CT...T.........G
[3]
-..TC....T.T..T....A..........A....C......T............T......A...C..CCA........T..A.T...A
[4]
...........T.......A.....C.....................T..........AT..T.............C.A...T...C..C
[5]
A......--.CC..A....AGG...............G...G.....A..A......A..T.A..G...TT...A.C.......TGG.-A
[6]                                     |insert 6a
--....C...A......GTG..............C.....-..........G....A...G.T............T.............G
[7]
.........G..CG..G..T.....TT.T..........A...............T....T.A......TT..................C
[8]
--............A....T....G..........C.AA....................GA.T.....A.4............A.....G
[9]
--.................G...T...........T......TA...........T......A..AC..CC............A.TA.TT
[10]
.C...........A.....G.....C...G.................T..............T.......C.G...AA...........C


AAGAAAAAAA-TTGGACCT--GGTTAT?GT----GACAGT?TGAGGGGCTTTAAGCC-??GCTAGGATGCAGACCRGAGCTGCCTGGGAY
[1]
.............TC......-.....TT.A........AA.....CA....C.ATATA-.............G.AA....T.T.....A
[2]
.....C...GT......--.G.T...CT..G.....T...GCA....T.A.......CT-....A...T....G.AC..T..A..C...C
[3]
...T....CGGA.T.T..A..C.GG..G..........CCAA....T.....G.CT.A-T.TC...G...GC..AG.G.....TA.TA.C
[4]
...GG........CTGA....T....CT..G.........G.A......-..G.AG.CCT......C........GTT.TC.....T..T
[5]
.G..G.-....CG.A......-..G..T..A........GA....ACA....CT.T..TT....TT.C.......A.....-TA.C...A
[6]
......-..........TG...T.CC.GTGGATG......A.A..T......C.A...TA....TT.........G...AAA......TT
[7]
...G..T.T.A..........T.AT.GT............A..T..AA.AGG.T.-T.AGAT....GCTT..T---...AAAGG.C..C-
[8]
.T....C..........TG....C.C.GTG..........G....T...A........CT.-AT...........A...T........CG
[9]
C.A.G...TCT..T.....AT..C.GAT.A....T....AGA.G.CCTG-.G..C.A.AGC.-.........C.TAG....A.TA.A..C
[10]
...G...........CA...G.T.AA.G.G..........G....A.G..G....CTGCT-..............GA..A.........T


AGA??GATGCTGTCTATTTCTCTTTAG?GGCTTGTCTCTCCTTTGGGGAAGG-TGTAGAGCATTCAGAAC-TTGGGGRCCTTAAAACAAT
[1]
..CTG.G..T...GA........C.G.A..GG.......T.C...................CC..CTTC..CG.A..G......G..---
[2]
...T-C.....T.........AC.A..T.........-.G....C.....AAG.TG......C...........CA.A......T.....
[3]
C..CT...AT...........T..CG.T..-....TCT.TTC...TCG.C.G.A..............T.C....C.-..A.G.......
[4]
...ACC...................CCT.......A.........A....T-.........C..............----------....
[5]
-..TT.G.......A........A...A...G....A..T..A..........AT..A.C.C....TTG........G.........---
[6]
...AA..A..A.....C..G.A.....A.....AC..........A.....A.......T....G....A....ACCC.TGA........
[7]
--------......C..A.A.....G.A.............A...........A........-..CTC.....T...A.........---
[8]
...AAA....A..........A.....C.....T............................C...A..........A............
[9]
T..AAT...................T.T.T.........T.A..AT.........G....T.....C.....GA..CGA.CAC.....G.
[10]
...CC...........AC....C...CT.C.C...A.TA..C........TA............TG..T..CA..TCA.GG.C.......
```

Fig. 6. Consensus sequence of the 258-bp subunits. A consensus sequence of the 258-bp subunits was determined by comparison of the 10 individual subsequences [1–10]. The consensus is shown above the individual sequences. Dots (.) indicate agreement with the consensus; disagreements are noted as A, C, G, or T. (–) indicates a deleted nucleotide. The position of inserts 1a and 6a are indicated (▌), as are the terminal inverted repeats (▬)

nucleotide CTT, one copy of which is found in the consensus sequence while the other is unique to subsequence six (again included in insert 6a in Fig. 5). In addition, it contains a region homologous to autonomously replicating sequences (ARS) of yeast and *Drosophila* (see Fig. 5). Within this region, an 11-bp sequence occurs (Fig. 5, bp 1619 to 1630 inclusive) in which 10 of the 11 bases match the consensus sequence for an essential element common to most ARS sequences from yeast (Broach et al. 1982). A search of the Genbank data base revealed homologies with ARS regions of yeast and *Drosophila* extending beyond the 11-bp ARS consensus described above. The significance of the features of inserts 1a and 6a will be discussed.

Recently, we have begun to characerize DNA from other species of *Dipodomys*. Using different portions of the KR-2 clone as probes to examine the DNA of *Dipodomys merriami*, we have been unable to detect homology to insert 6a or the 258-bp repeated subunits. However, we have found evidence for repeated interspersed DNA homologous to insert 1a (Keim, unpublished data).

Discussion

The kangaroo rat, *D. ordii*, contains a highly repeated (more than 10^5 copies per genome) 3.3-kb sequence (Liu and Lark 1982). The 3.3-kb repeated sequence of *D. ordii* appears to be interspersed throughout the genome in small clusters. Thus, it is distributed over most of the chromosomal material, as judged by in situ hybridization (Fig. 1); *D. ordii* DNA of widely differing densities contains the 3.3-kb sequence (Fig. 2); and partial digestion with re-

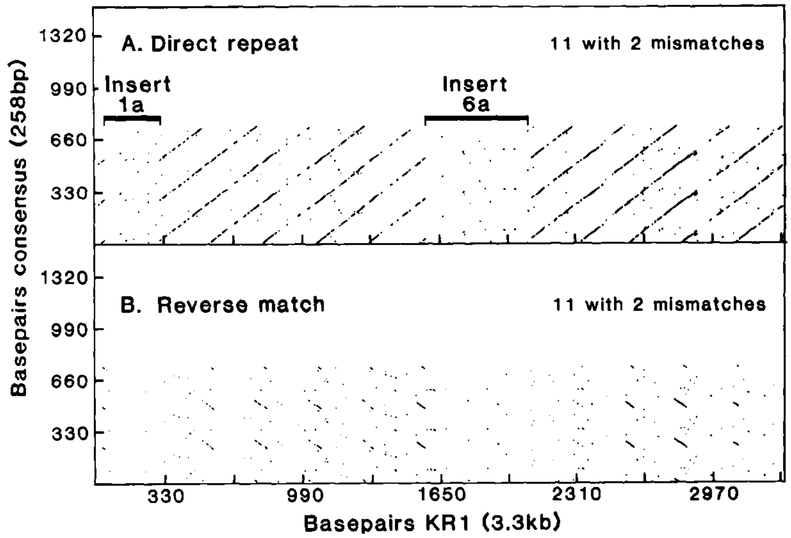

Fig. 7. Homology between the 258-bp consensus sequence and the KR-2 sequence. The KR-2 sequence was compared directly to the 258-bp consensus sequence (**A**) and as a reverse match to the 258-bp sequence (**B**). The consensus sequence was run as a loop and is therefore repeated. The diagonal lines with positive slope in **A** are indicative of direct repeats, while the diagonals with negative slope in **B** indicate inverted matching repeats. Regions in **A** devoid of homology contain the proposed insertions 1 and 6a (see text)

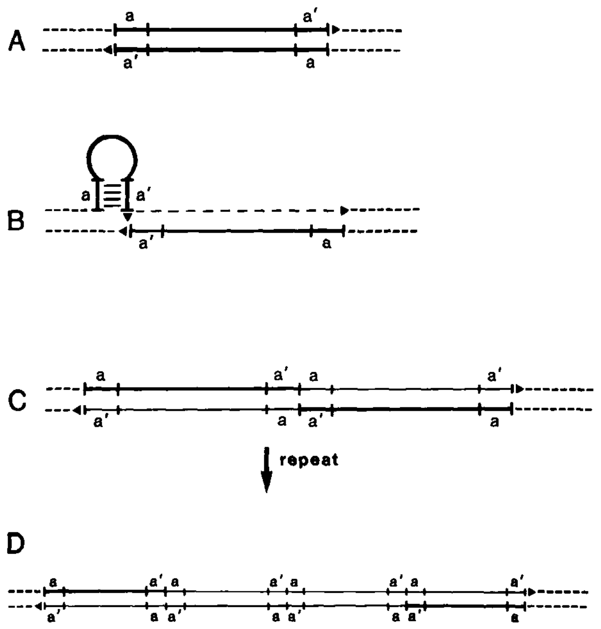

Fig. 8. A model of replicative duplication of the consensus DNA sequence. The 258-bp consensus sequence contains a set of terminal inverted repeats, which suggests this mechanism of replicative duplication. The sequence (**A**) can form a stem-loop (**B**) by pairing its inverted repeats. The stem acts as the primer for further synthesis using the appropriate strand as template (**B**). The stem-loop itself is copied, completing the duplication (**C**). This process could be repeated using a "dimer" stem-loop (**D**), which would duplicate a pair of 258-bp sequences

striction endonucleases liberates single copies, some dimers and trimers, and only small amounts of higher multimers (Liu and Lark 1982).

The structure of the KR-2 clone appears to be representative of the majority of this class of repeated sequences, based on a limited sequencing of genomic DNA and of two other clones (Keim et al. 1984) and on the arrangement in the genomic DNA of the restriction endonuclease sites found in the KR-2 clone. An RNY analysis (Shepherd 1981) of the sequence (carried out by Dr. John Shepherd) indicated a lack of any preferred reading frame, similar to the absence of preferred reading frames in ribosomal RNA genes. This suggests that no translatable information is encoded in the sequence.

The KR-2 sequence has the following interesting characteristics. It contains 10 subsequences ([1–10]) that are strongly related by sequence homology and two subsequences (inserts 1a and 6a) that are unique. A comparision of the 10 similar subsequences yields a consensus sequence that contains terminal inverted repeat sequences 35 bp in length (Figs. 6 and 7B).

These data suggest that the 3.3-kb sequence arose by tandem amplification of some ancestral sequence similar to the consensus sequence. After amplification, two sequences were inserted (inserts 1a and 6a in Fig. 4) and subsequently the 3.3-kb unit was spread throughout the genome, reaching a copy number of more than 10^5 copies per genome. We would like to speculate on how this process occurred.

Tandem Amplification of the Ancestral Sequence

When the individual subsequences ([1–10]) (in Fig. 6) are compared with the consensus sequence, the odd sequences [1, 3, 5, 7, 9] have features in common that differentiate them from the even [2, 4, 6, 8, 10] (see Table 1). The odd series have less homology to the consensus (59–79 mismatches) than do the even (39–56 mismatches). In 67% of the cases in which the same substitution is found in two or more subsequences, the substitution occurs only in the odd or only in the even series. In only 33% of the mismatches are substitutions in an even also found in an odd series. (In general, base substitutions are more frequently transversions than transitions.) Using common substitutions it is possible

Table 1. Relationships between the 258-bp consensus sequence and the 10 individual repeated subunits found within the 3.3-kb sequence from *D. ordii*

	Subsequence									
	Even					Odd				
	[8]	[4]	[0]	[6]	[2]	[7]	[5]	[1]	[9]	[3]
No. of errors	39	44	55	56	56	59	65	66	71	79
No. of errors										
Common with										
Odd	14	16	27	24	25	31	45	54	39	36
Even	27	32	32	35	35	29	12	22	22	21
Relatives										
1st	[6]	[10]	[4]	[2]	[4]	[5]	[1]	[5]	[3]	[9]
2nd	[2]	[2]	[7]	[8]	[6]	[10]	[7]	[3]	[1]	[1]

The 10 subunits are designated in brackets as in Figs. 5 and 6. Errors denote differences between individual subsequences and the consensus sequence. Common errors are identical base substitutions that occur in different individual subsequences at the same position. (The numbers of identical substitutions found in individual subsequences and all other odd or even subsequences are noted.) Finally, these common substitutions have been used to determine nearest relatives on the assumption that closest relatives will be those that share the largest number of identical substitutions

to identify those subsequences that are most closely related. When this is done (Table 1), it is found that the closest relatives of evens are evens and of odds are odds. In addition to the relationships summarized in Table 1, a 5-bp deletion occurs at the junction of subsequences [1] and [2], [5] and [6], and [7] and [8] (Fig. 6), suggesting that these pairs of subsequences are related by descent.

These data suggest that the separation into even and odd subsequences occurred at an early stage and that this separation was maintained. We suggest that an ancestral sequence similar to the consensus sequence was duplicated by the mechanism shown in Fig. 8. During DNA replication, the inverse repeats at the two ends of the sequence on one strand paired, looping out the sequence in such a way that its complement could be used as a template to extend the 3′ end of the looped-out strand. This would result in duplication of the sequence in one strand, and subsequent copying of the loop (repair) would result in a completion of the duplication on both strands. One half of this duplex underwent fewer changes than the other (possibly because it was maintained under selective pressure, while the other was free to undergo change). We believe that these two copies of the original sequence were, respectively, the progenitors of the even and odd series of subunits characterized in Table 1.

The observed pattern of subsequence relationships is not consistent with unequal crossing-over beginning with the duplex described in Fig. 8C. However, these relationships can be explained by an additional cycle of replicative duplication, as proposed in Fig. 8D, followed by replicative duplication and/or unequal crossing-over. Amplification may have continued to reach a total length of more than 10 of these subunits.

Spread of the 3.3-kb Unit through the Genome of D. ordii

We believe that the 10 subunits were spread recently throughout the genome as the result of acquiring two insertions, 1a and 6a. We have already noted that insert 6a contains the sequence of one of the essential elements associated with autonomously replicating sequences from yeast. [In addition, it bears a homology (20 bp with three mismatches) with a sequence from African green monkey cells believed to contain a replicative origin (Zannis-Hadjopoulos et al. 1985).] The formation of circular autonomous replicating unit may have occurred as a result of acquiring an origin of replication, carried within insert 6a, and excision of a ca. 3-kb element comprised of this insert and the 10 tandem subunits ([1–10]). An active origin of replication would have allowed this sequence to be maintained and to spread by recombination, resulting in its insertion into sites of homology located throughout the genome. Such homology was provided, we believe, by acquisition of insert 1a creating a 3.3-kb sequence similar to the one we have studied. Insert 1a contains an RNA polymerase III promoter together with a poly A tail of the type (CAx)y. Using the retroposon mechanism described by others (see review by Rogers 1985a), insertion 1a could have spread through the genome. Wherever this sequence was located, the 3.3-kb plasmid could then insert by homologous recombination. In addition, multiple nested insertions could occur in one position, creating tandem repeats of the 3.3-kb unit. Our recent finding that *D. merriami* contains interspersed repeated DNA sequences homologous to insert 1a supports the idea that the spread of insert 1a may have occurred prior to the dispersal of the 3.3 kb throughout the genome.

There is about 0.7% variation between nucleotides observed in sequence analysis of genomic DNA and three different clones of the 3.3-kb repeat (Keim et al. 1984). Loss of restriction sites for *Pst*I, *PVu*II, *Bam*HI, *Hin*dIII, *Sst*I, *Bgl*II, or *Sph*I (in Fig. 4, the relative proportion of 3.3-kb fragments that remain uncut by one of the two enzymes) suggests a frequency of base change of about 0.4%. If nucleotide changes have accumulated at about 0.35%/million years, as suggested by Helm-Bychowski and Wilson (1986), then the 3.3-kb repeats may have dispersed through the genome 1–2 million years ago. The fossil evidence indicates that the Dipodomii underwent speciation 1–2 million years ago, or in early Pleistocene (Stock 1974). Thus, the amplification and interspersion of the 3.3-kb sequence may be correlated with the emergence of *D. ordii* as a species.

Acknowledgments. We thank Dr. John Shepherd for computer programs and the RNY analysis of the sequence. We especially thank him for stimulating discussions and critique. We thank Dr. T. Hori for helpful advice in preparing chromosome spreads for in situ hybridization. This work was supported primarily by NAID-AI-10056 to K.G.L. and in part by grants ACS-IRG IN-154 to P.K. and BSRG-507-RR07092 to K.G.L.

References

Broach JL, Li Y-Y, Feldman J, Jagmaram M, Abraham J, Nasmyth KA, Hicks JB (1982) Localization and sequence analysis of yeast origins of DNA replication. Cold Spring Harbor Symp Quant Biol 47:1165–1173

Ciliberto G, Raugei G, Costanzo F, Dente L, Cortese R (1983) Common and interchangeable elements in the promotons of genes transcribed by RNA polymerase III. Cell 32:725–733

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Engels WR, Preston CR, Thompson P, Eggleston WB (1986) *In situ* hybridization to drosophila salivary chromosomes with biotinylated DNA probes and alkaline phosphatase. BRL Focus 8:6–8

Finnegan DI (1985) Transposable elements in eukaryotes. Int Rev Cytol 93:281–326

Hatch TT, Dodner AJ, Mazrimas JA, Moore DH (1976) Satellite DNA and karyotypic variation in kangaroo rats (genus *Dipodomys*). Chromosoma (Berl) 58:155–168

Helm-Bychowski KM, Wilson AC (1986) Rate of nuclear DNA evolution in pheasant-like birds: evidence from restriction maps. Proc Natl Acad Sci USA 83:688–692

Hori T, Ayusawa D, Shimizu K, Kayama H, Seno T (1985) Assignment of human gene encoding thymidylate synthase to chromosome 18 using interspecific cell hybrids between thymidylate synthase-negative mouse mutant cells and human diploid fibroblasts. Somatic Cell Mol Genet 11:227–283

Keim P (1986) Isolation of plant nuclei. In: Packer L, Douce R (eds) Plant membranes. Methods in enzymology (in press)

Keim P, Thliveris AT, Meenen EA, Lark KG (1984) Rearrangements of an immunoglobin-like sequence mediated by a specific prokaryotic recombination system. In: Cantor H, Chess L, Sercarz E (eds) Regulation of the immune system. UCLA Symposia on Molecular and Cellular Biology, New Series 18:511–526

Lee TNH, Singer MF (1982) Structural organization of alpha-satellite DNA in a single monkey chromosome. J Mol Biol 161:323–342

LePecq J, Paoletti C (1966) A new fluorometric method for RNA and DNA determination. Anal Biochem 17:100–107

Liu L-S, Lark KG (1982) The red function of phage L mediates the alteration of an interspersed repeated DNA sequence from the kangaroo rat *Dipodomys ordii*. Mol Gen Genetics 188: 27–36

Ma DP, Lund E, Dahlberg JE, Roe BA (1984) Nucleotide sequence of two regions of the human genome containing Asn-tRNA genes. Gene 28:257–262

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

Maxam AM, Gilbert W (1980) Sequencing end-labeled DNA with base specific cleavage. Methods Enzymol 65:499–560

Miklos GLG (1985) Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes. In: MacIntyre RJ (ed) Molecular evolutionary genetics. Plenum Press, New York, pp. 241–321

Rigby PWJ, Diekmann M, Rhodes C, Berg P (1977) Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J Mol Biol 113:237–251

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

Rogers JH (1985b) Origins of repeated DNA. Nature 317:765–766

Shepherd JCW (1981) Method to determine the ready frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc Natl Acad Sci USA 78:1596–1600

Singer MF (1982) Highly repeated sequences in mammalian genomes. Int Rev Cytol 76:67–112

Southern E (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98: 503–517

Stock AD (1974) Chromosome isolation in the genus *Dipodomys* and its taxonomic and phylogenetic implications. J Mammal 55:505–528

Thayer RE, Singer MF, McCutchan TF (1981) Sequence relationshps between single repeat units of highly reiterated African green monkey DNA. Nucleic Acids Res 9:169–181

Wood WB (1966) Host specificity of DNA produced by *Escherichia coli*: bacterial mutations affected the restriction and modification of DNA. J Mol Biol 16:118–133

Zannis-Hadjopoulos MG, Kaufmann SS, Wong RL, Lechner E, Karawya J, Hesse J, Martin RG (1985) Properties of some monkey DNA sequences obtained by a procedure that enriches for DNA replication origins. Mol Cell Biol 5:1621–1629

Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxilary information. Nucleic Acids Res 9:131–148

Received June 13, 1986/Revised and accepted November 20, 1986