The EMBO Journal vol.3 no.8 pp.1753 – 1759, 1984

Defining the beginning and end of *Kpn*I family segments

Giovanna Grimaldi[1], Jacek Skowronski and Maxine F.Singer

Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205, USA
[1]Present address: Istituto Internazionale di Genetica e Biofisica, Via Guglielmo Marconi 10, 80125 Napoli, Italy

Communicated by G.Bernardi

Comparison of the sequences at the ends of several newly cloned and full length members of the monkey *Kpn*I family with one another and with previously described monkey and human segments defines the nucleotide sequence at the two termini. No terminal repeats either direct or inverted are noted within full length family members which may or may not be immediately flanked by direct repeats. At the 3′ terminus, several family members have polyadenylation signals followed by a d(A)-rich stretch. The genomic frequency of segments within the full length element increases markedly from the 5′ to the 3′ terminus, consistent with the cloning of various truncated family members. One such truncated version joined to a low copy number DNA segment is inserted in monkey α-satellite where the combination appears to have been amplified in conjunction with the satellite itself.

Key words: repeated sequences/*Kpn*I family/moveable elements

Introduction

Several papers published over the past few years have reported the existence of families of long sequences that are interspersed and highly repeated in the genomes of both primates (*Kpn*I family), (Adams *et al.*, 1980; Manuelidis and Biro, 1982; Shafit-Zagardo *et al.*, 1982a,1982b; Grimaldi and Singer, 1983) and rodents (MIF-I or *Bam*HI family) (Soriano *et al.*, 1983; Meunier-Rotival and Bernardi, 1984; Voliva *et al.*, 1983; Fanning, 1983). The families have been collectively termed LINES (Singer, 1982a,1982b) and *Kpn*I and *Bam*HI sequences were shown to be homologous at least in regions where primary sequence data are available (Singer *et al.*, 1983). The longest known *Bam*HI (Fanning, 1983) and *Kpn*I (Adams *et al.*, 1980) family members are estimated to be ~ 7 kbp and 6.4 kbp in length, respectively. However, structural analysis of cloned family sequences revealed that the LINE families include a complex assortment of members of variable length that share some but not all sequences present in longer repeated units (Adams *et al.*, 1980; Thayer and Singer, 1983; Lerman *et al.*, 1983; Grimaldi and Singer, 1983; Potter, 1984; Miyake *et al.*, 1983; Voliva *et al.*, 1983). Moreover, common sequences within some of these shorter line family members are not necessarily co-linear but may be either reordered or inverted relative to one another (Thayer and Singer, 1983; Lerman *et al.*, 1983; Potter, 1984; Gebhard and Zachau, 1983). Similarly an inversion occurs in a human cDNA clone containing *Kpn*I sequences (DiGiovanni *et al.*, 1983).

We have been interested in defining the two ends of the longest *Kpn*I family members and in estimating the frequency at which different segments from within these members occur in primate genomes. One end (the right or 3′ end as usually written, see Figure 1) of several already described human and monkey family members is roughly defined by the end of the homology between them; these include several short members (Thayer and Singer, 1983; Lerman *et al.*, 1983; Potter, 1984) and several units of undetermined length (DiGiovanni *et al.*, 1983; Potter and Jones, 1983). In two cases the common 3′ end abuts known but unrelated DNA sequences, namely, satellite DNA (Thayer and Singer, 1983; Potter and Jones, 1983). However, none of these *Kpn*I sequences were derived from a known full length member and moreover there is significant variation between the apparent 3′ termini. Similarly there is ambiguity concerning the left end. The border of homology between several characterized sequences has been noted (Miyake *et al.*, 1983; Nienhuis *et al.*, personal communication; Potter, 1984) but only one of the sequences derives from a member that is over 6 kbp long, *Kpn*-TβG41 (Miyake *et al.*, 1983). Also, only in the case of *Kpn*-TβG41 have the two ends of a single long member been described.

We report here the isolation and characterization of several newly cloned *Kpn*I family members that are over 6 kbp long and co-linear throughout. Together with prior work, the reported data establish, to a high level of confidence, the nature of the two ends of the long family members. Also, using small subcloned regions from within long family members as probes, we demonstrate that the abundance of *Kpn*I family sequences in the monkey genome increases markedly from the left to the right end of the long members.

Previously we described three different cloned segments in which monkey α-satellite is joined to a *Kpn*I unit (λ Caα6, 7 and 11). The three *Kpn*I units are co-linear and one of them, which is flanked on both sides by α-satellite, was estimated to be at least 6 kbp long (Grimaldi and Singer, 1983). We presented evidence indicating that at least two of these units might be truncated ~1 kbp at the 5′ end. To establish the relationship of these three units to complete *Kpn*I sequences, we have compared them with the newly isolated full length *Kpn*I members.

Results

Isolation and analysis of KpnI family members

One aliquot of a monkey genomic library (AGM-*Eco*RI library II, see Grimaldi and Singer, 1983) was screened with probes representing the left and central segments of a long *Kpn*I family member (pCaα6.10.1 and pCaα6.5, respectively). Another aliquot was screened with probes representing the central and right segments (pCaα6.5, and *Bam*HI/*Eco*RI, respectively (see Figure 1 and Materials and methods for a description of the probes). Three phage were selected from the first aliquot (λF1, λF2, λE2) and one from the second (λB3). Each of the four phage was then mapped by

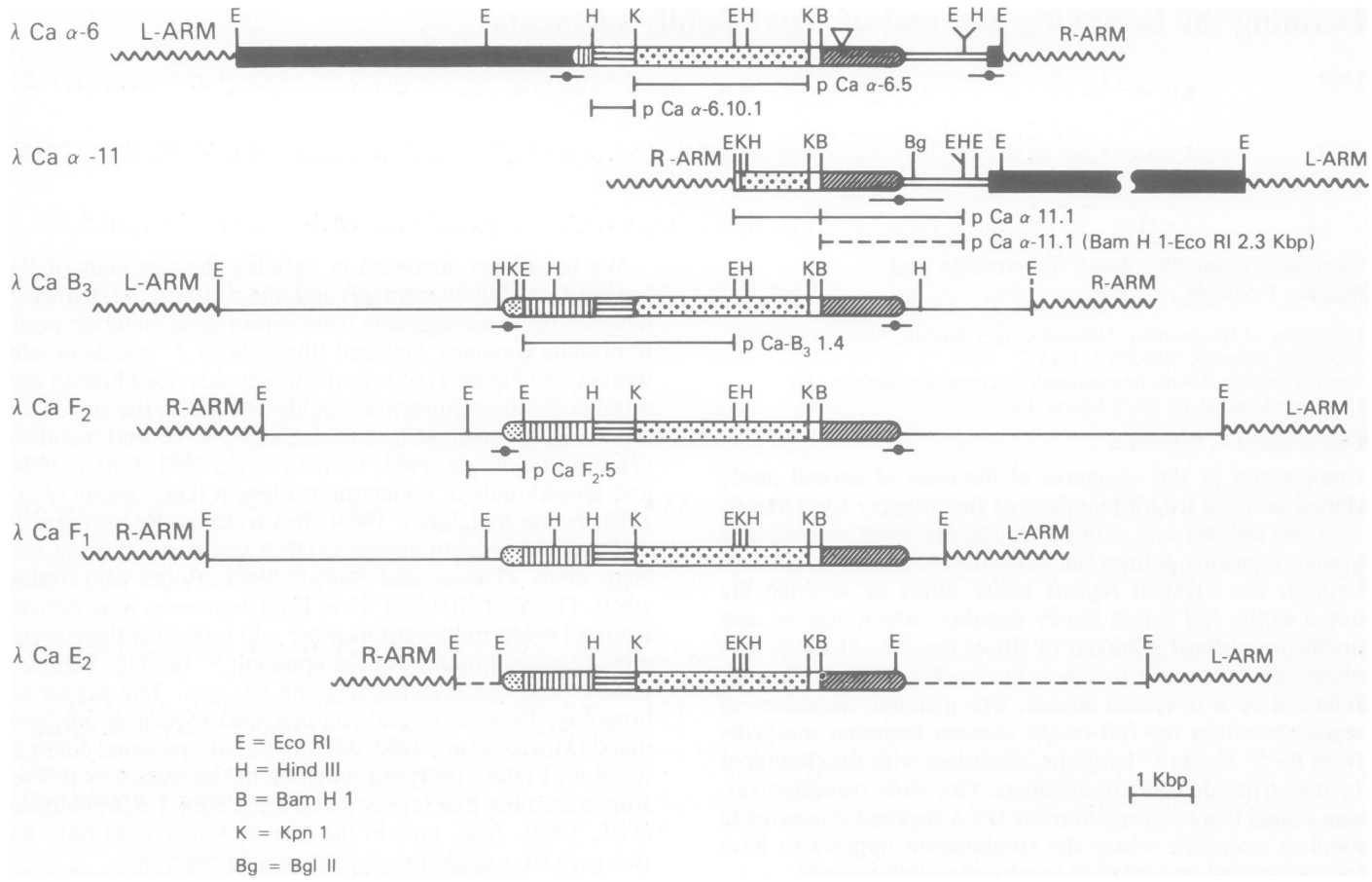

Fig. 1. Structure of *Kpn*I family members. Six phage isolated from monkey libraries are presented in this figure; each contains a *Kpn*I family member. λCaα6 and λCaα11 were previously described (Grimaldi and Singer, 1983). λCaF_2, λCaF_1 and λCaE_2 were selected using pCaα-6.5 and pCaα-6.10.1 as probes (see text). λCaB_3 was selected using the *Bam/Eco* probe derived from pCaα11.1. The figure represents a summary of the structural data available for the *Kpn*I members (open area) cloned in each phage. Symbols: ▨ anneals to *Bam/Eco* probe; ▩ anneals to pCaα6.5; ▤ anneals to pCaα6.10.1; ▥ anneals to pB3-1.4; ▩ anneals to pF2-.5; ■ anneals to pCa1004 (α-satellite); ∇ sequence homologous to *Alu* family member; —●— indicates where primary sequence data are available; ⊢—⊣ and ⊢- - -⊣ indicate source and extent of various subclones; ═ unrelated sequence linked to *Kpn*I members in λCaα6 and λCaα11; ∿∿ phage arms. The length of the region flanking the *Kpn*I segment in λCaE2 (----) is uncertain.

Southern analysis with all three probes. The order of the restriction fragments hybridizing to the three probes was similar in all the phage (see Figure 1) and in agreement with the general structure of long *Kpn*I members previously determined (Grimaldi and Singer, 1983; Shafit-Zagardo *et al.*, 1982b; Manuelidis and Biro, 1982).

The new phage were further characterized by heteroduplex analysis to establish the total length of common sequence present in each. The pairs analyzed were λF1 – λF2, λF1 – λE2, λF2 – λE2 (the orientation of λB3 in the phage precluded its use). In each case the duplex region measured between 6.1 and 6.5 kbp (Table I); no evidence for interruptions was found along the heteroduplex lengths. The position of the heteroduplex region and the lengths of the single-stranded loops between the heteroduplex regions and the λ-arms was in agreement with the restriction endonuclease and hybridization data in each pair of phage analyzed.

The four new phage as well as λCaα6 in which the *Kpn*I family segment is flanked by α-satellite were also analyzed by Southern hybridization to two additional probes, pB3-1.4 and pF2.5, derived from the left hand sides of two of the new phage (Figure 1). All five phage hybridized with pB3-1.4. However, while the length of the hybridizing region in all the new phage was the same (and thus equivalent to the length of the probe itself) the total length of the hybridizing region in λCaα6 was at least 500 bp shorter than the probe. Moreover, while all four new phage hybridized with pF2.5, no hybridization was observed to λCaα6. These results, which are summarized in Figure 1, are consistent with the earlier conclusion that about 1 kbp of a long *Kpn*I sequence is missing from the left end of λCaα6 (see below).

Table I. Lengths of heteroduplex regions formed between pairs of *Kpn*I family members in λ-phage

Pair of phage	Length (kbp)
λF2 – λE2	6.14 ± 0.36
λF2 – λF1	6.57 ± 0.47
λF1 – λE2	6.58 ± 0.16

These experiments indicate: (i) the lengths of *Kpn*I sequences in λB3, λF2, λF1 and λE2 correspond closely to the estimated length of the longest *Kpn*I member described by Adams *et al.* (1980) and (ii) the homology between the newly isolated *Kpn*I sequences extends further at the left end than does their homology with the *Kpn*I sequence present in λCaα6. The data altogether suggest that the newly isolated *Kpn*I members closely resemble *Kpn*-Tβ G41, the ~6.4-kbp *Kpn*I unit downstream of the human β-globin gene (Adams *et al.*, 1980; Miyake *et al.*, 1983; Shafit-Zagardo *et al.*, 1982b).

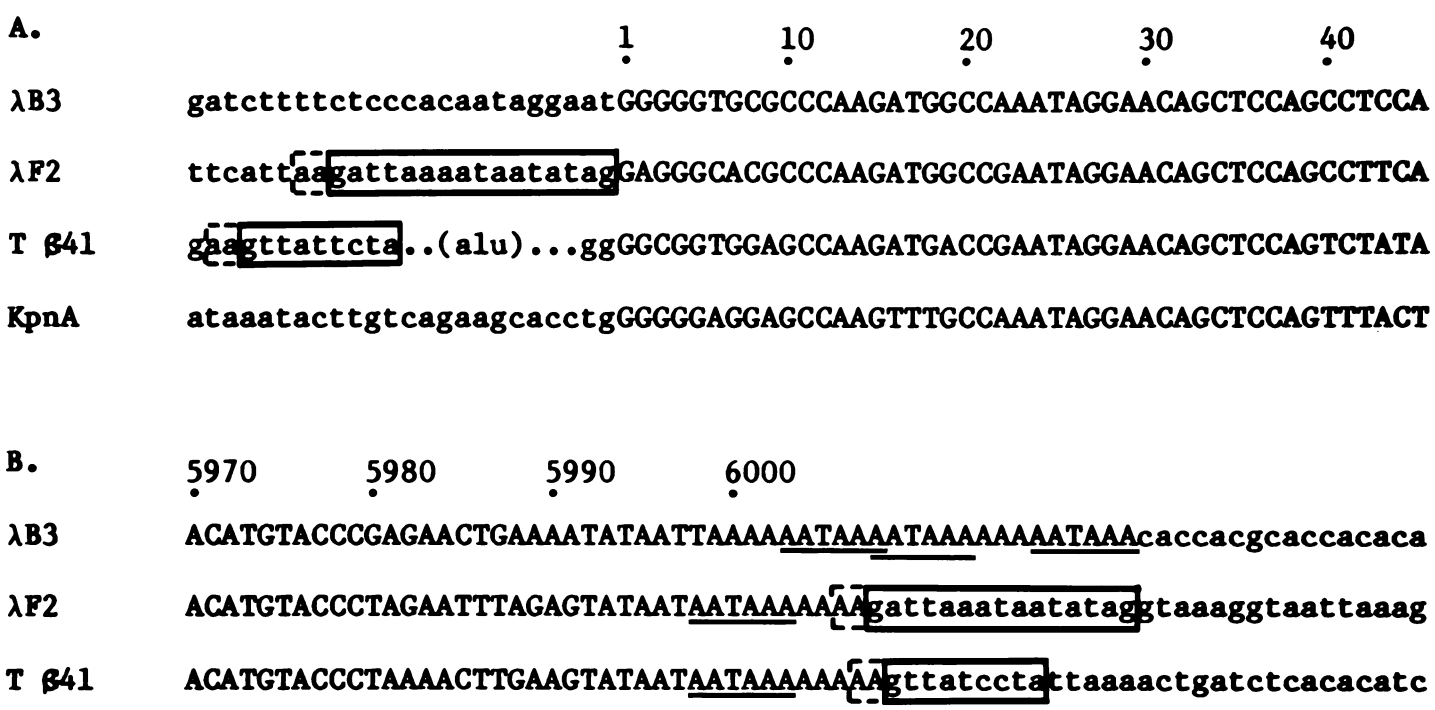

Fig. 2. Alignment of common sequences at the 5′ and 3′ ends of long *Kpn*I segments. The proposed *Kpn*I sequences are in capital letters and flanking segments are in lower case letters. Direct flanking repeats are boxed; dotted boxes indicate base pairs that cannot be unambiguously assigned to the repeats. The sequences of λB3 and λF2 were determined in this work; *Kpn*-TβG41 (Miyake *et al.*, 1983) and *Kpn*A (Potter, 1984) were described previously. **(A)** 5′ end. Base pair 1 is the first common base in all the sequences. The *Alu* family sequence that immediately precedes the *Kpn*I segment in TβG41 is not written out. **(B)** 3′ end. Polyadenylation signals (AATAAA) are underlined. The numbering is based on the sequence of a full length *Kpn*I family as compiled and deduced in this laboratory both from published sequences and our own unpublished data.

Determination of the 5′ and 3′ ends of long KpnI family sequences

We determined the nucleotide sequence at the 5′ (left) and 3′ (right) ends of the two long *Kpn*I sequences in λB3 and λF2 (Figure 1). In Figure 2A, we aligned the 5′ end sequences with the 5′ end of *Kpn*A, an internally rearranged unit flanked on both sides by human α-satellite (Potter, 1984) and of *Kpn*-Tβ G41 (Miyake *et al.*, 1983; Nienhuis *et al.*, personal communication). The homology among the four members starts at the same nucleotide position (position number 1 in Figure 2A) namely, the one that directly abuts α-satellite in *Kpn*A. In Figure 2B, we aligned the newly determined 3′ end sequences with the 3′ end of the *Kpn*-TβG41. The homology among the three sequences ends within a d(A)-rich stretch. In the case of *Kpn*-λF2 this is followed by an 18-bp sequence that is repeated perfectly at the 5′ end boundary (Figure 2). In λB3 no flanking direct repeats were observed (Figure 2).

Determination of the boundaries of the KpnI family sequences joined to α-satellite

We have already pointed out that the *Kpn*I unit in λCaα6 is missing ~1 kbp that commonly occurs at the left-hand end of the longest family members. Nevertheless, the total length of the segment interrupting α-satellite in λCaα6 is >6 kbp and previously published results demonstrate that throughout most of this length the restriction endonuclease sites and sequence arrangement are typical of the bulk of the genomic *Kpn*I elements. The segment adjoining α-satellite in λCaα6 is co-linear with that in λCaα11 (Figure 1) except that the latter is short at the left end where it is joined to the phage arm. However, although the distance from the common *Bam*HI site to the common 3′ terminus is only ~1200 bp in sequenced family members (Lerman *et al.*, 1983; DiGiovanni *et al.*, 1983), the distance from the *Bam*HI site to the end of the non-α-satellite segments in λCaα6 and 11 is >2 kbp. (A small part of the extra length in λCaα6 is accounted for by an *Alu* sequence.) Thus, it seemed likely that these segments contain additional sequences at the far right end that are not typical of *Kpn*I family members.

We determined the nucleotide sequences at the boundaries with α-satellite of the >6 kbp segment in λCaα6. They are aligned in Figure 3 with the α-satellite consensus sequence (Rosenberg *et al.*, 1978). At the place at the left side of *Kpn*-λCaα6 (Figure 3A) where homology to α-satellite degenerates, the sequence can be aligned starting at nucleotide number 982 from the 5′ terminus of the *Kpn*A sequence determined by Potter (1984); the following 70 bp, which is all that we analyzed, are homologous to the *Kpn*A sequence. The exact left-hand boundary between the truncated *Kpn*I sequence and the α-satellite sequence cannot be unambiguously defined. The λCaα6 sequence at this junction (boxed in Figure 3A) matches both the α-satellite consensus (positions 157−171) and the *Kpn*A sequence (position 982−996). The alignment of the left end of *Kpn*I-λCaα6 with *Kpn*A and α-satellite unambiguously indicates that the *Kpn*I-λCaα6 segment lacks over 900 bp characteristic of the 5′ end of the longest family members.

At the right end of the 6 kbp unit in λCaα6, the homology with α-satellite starts at positions 15 or 17 of the satellite consensus sequence (Figure 3B). This shows that a minimum of 18 bp of α-satellite sequence flanking the inserted DNA unit is deleted. The sequence joining the satellite segment (italics, Figure 3B) is not homologous to any portion of the *Kpn*I family sequence or to α-satellite; the lack of homology extends for more than 1 kbp to the left of the rightwards α-satellite sequence (data not shown). The junction between this

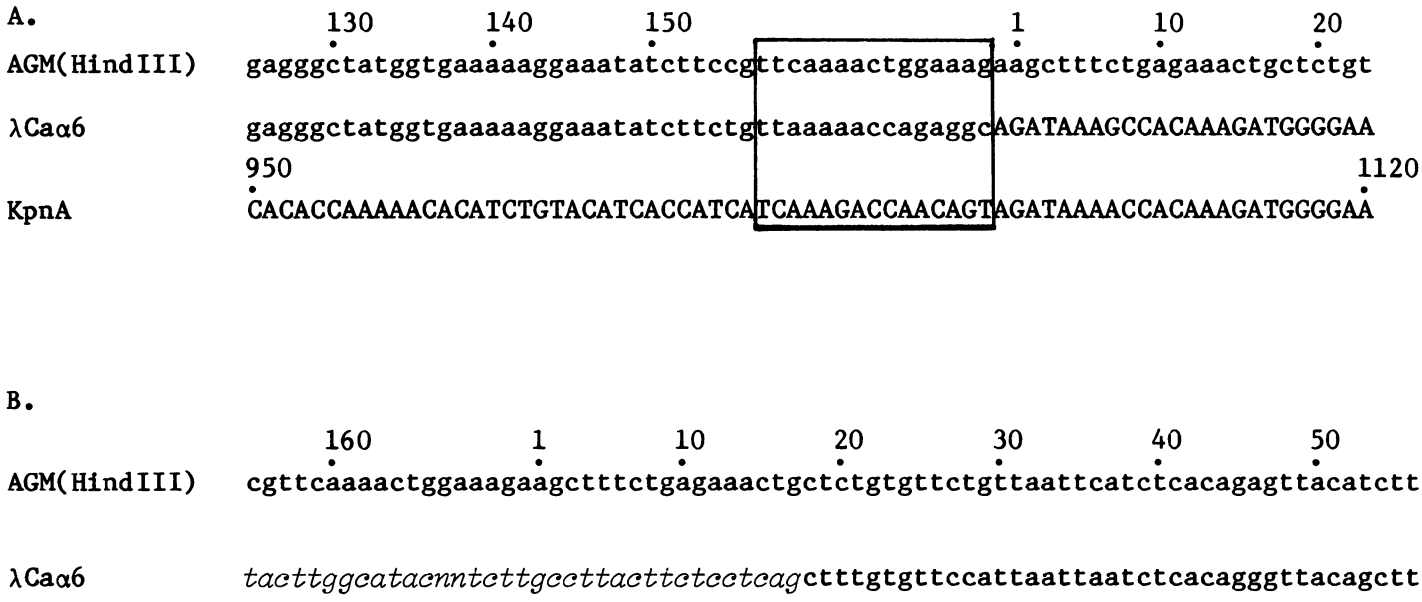

Fig. 3. Sequences at the borders between α-satellite and non-satellite segments in λCaα6. Sequences identified as *Kpn*I family by comparison with known sequences are in capital letters; α-satellite sequences are in lower case letters; uncharacterized sequences are in italics. For comparison, the α-satellite consensus sequence (Rosenberg *et al.*, 1978) and the relevant sequence of *Kpn*A (Potter, 1984) are given using the original numbering (note that in the sequence of *Kpn*A, residue 115 corresponds to the 5′ end of the *Kpn*I element, Potter, 1984). **(A)** The border at the left side of λCaα6 (see Figure 1). A 15-bp region of some homology between α-satellite and *Kpn*A is boxed; because of this, the precise junction between *Kpn*I and α-satellite cannot be defined. **(B)** The border at the right side of λCaα6 (see Figure 1). The nn in λCaα6 represents a gap of ~5 bp in the determined sequence; within the gap is a *Hpa*II site that was utilized in the sequencing strategy.

sequence and the right hand terminus of *Kpn*I was determined. Here the *Kpn*I sequence ends in a segment homologous to the 3′ end of the longest members including a d(A)-rich stretch (not shown). We conclude that the 6 kbp unit interrupting α-satellite includes, from left to right, a *Kpn*I segment that lacks >900 bp of sequences from the far left but then goes through the entire *Kpn*I sequence followed by an unrelated sequence.

We next confirmed that the sequence downstream from the 3′ end of the *Kpn*I family members in λCaα11 and λCaα7 is the same as that in λCaα6 (Figure 4). Two subclones of the pertinent regions of λCaα6 and λCaα11 were prepared, pCaα6.12.1 and pCaα11.11 (see Figure 4). pCaα6.12.1 hybridizes with the expected *Eco*RI fragments: one in λCaα6 and two in λCaα11 (Figure 4A). The one in λCaα6 and one of the two in λCaα11 contain the junction with α-satellite (see Figure 1). pCaα11.11 also hybridized with the expected *Hind*III and *Eco*RI fragments in the two phage (not shown). Both probes also hybridized to λCaα7 but neither one to λB3 or λF2 (not shown). The latter observation confirms the earlier suggestion that the sequences close to the right hand junction with α-satellite in λCaα6, 7 and 11 are not part of the *Kpn*I family (shown as a thin open bar in Figure 1). However, that sequence is reiterated in combination with the *Kpn*I family within α-satellite as indicated by the isolation of the three distinct cloned segments λCaα6, 7 and 11. We investigated the frequency of the combined arrangement in the genome. pCaα6.12.1 hybridizes to three abundant genomic fragments in *Eco*RI digests of monkey DNA; these are the same sizes as the annealing fragments in λCaα6 and λCaα11. Similarly, pCaα11.11 hybridizes to genomic *Hind*III and *Eco*RI fragments that are the same size as those generated from λCaα6 and λCaα11 (not shown). Thus the bulk of these genomic sequences appear to occur between *Kpn*I and α-satellite sequences as they do in the phage. Quantitative analysis by dot blot hybridization indicates that there are of the order of $1-2 \times 10^2$ copies per haploid genome (data not shown). Genomic blots and quantitative dot blots indicate that the sequences of the two subclones occur at least 10 times less frequently in the human genome and within a different context. The most likely explanation is that a low copy number segment was inserted into α-satellite next to a *Kpn*I family member (either as a single segment or in two separate events) and then the whole unit was amplified in the course of ongoing rearrangement of the satellite itself.

The genomic frequency of different regions within long KpnI family members varies

Two *Kpn*I family members that are much shorter than 6 kbp were recently studied in this laboratory. Both *Kpn*I-RET which is 829 bp (Thayer and Singer, 1983) and *Kpn*I-LS1 which is 2206 bp (Lerman *et al.*, 1983, and R.E.Thayer, unpublished) contain sequences from the 3′ end of long *Kpn*I units and terminate in the common A-rich stretch seen in Figure 2B. These findings suggest that the sequences from the 3′ end of long *Kpn*I units might be more frequent in the genome than those from the 5′ end. Therefore we investigated the relative genomic frequency of seven different and non-overlapping regions from within long *Kpn*I family members using the quantitative dot-blot hybridization method and appropriate subcloned probes (Figure 5). Altogether the probes cover >90% of a full length unit.

We prepared six replicate sets of nitrocellulose filters each containing $1-2$ μg of each of the seven linearized recombinant subclones. Three were annealed with either ^{32}P-labeled λB3, λF2 or λCaα11, the phage from which the subclones were made. The relative amount of radioactivity that annealed to each of the subclones with the three different phage

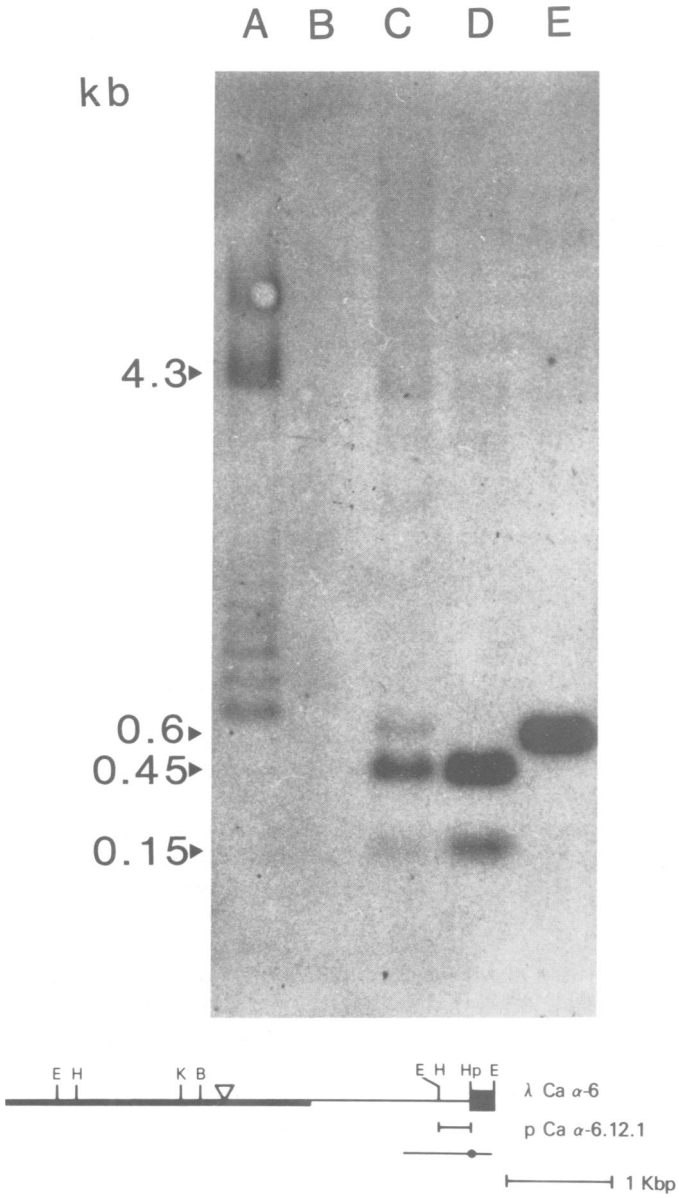

Fig. 4. Annealing of pCaα6.12.1 to restriction endonuclease digests of λCaα6, λCaα11, monkey and human DNA. **Bottom:** map of the right hand portion of insert in λCaα6. **Top:** DNA samples were digested with restriction endonucleases, electrophoresed on 1% agarose gels, transferred to nitrocellulose and annealed with ^{32}P-labeled pCaα6.12.1. **Lane A:** 10 μg monkey DNA, partial *Hind*III digest; **lane B:** 10 μg HeLa cell DNA, *Hind*III; **lane C:** 10 μg monkey DNA, *Eco*RI; **lane D:** 50 ng λCaα11, *Eco*RI; **lane E:** λCaα6, *Eco*RI. The sizes of the bands in λCaα6 and λCaα11 are indicated to the left.

DNAs varied <10% and there was no indication of any bias due to divergence between homologous sequences on the three phage. Since each of the subcloned segments is present only once in each of the phage, we could use these control data to normalize the experimental data (see below) to conditions of equal abundance of each subcloned region. The remaining four sets were annealed with ^{32}P-labeled total monkey DNA, each at a different hybridization stringency, and the amount of radioactivity annealed to each dot was normalized as just described. Under the conditions used, the fraction of ^{32}P-labeled sequence which annealed to its filter-bound complement was <20% of its initial amount, as determined in a model experiment at the most permissive temperature. There were no significant differences between the normalized results over the range of hybridization conditions although the extent of hybridization varied 7-fold under the different conditions. The mean result is presented in Figure 5. The data indicate that the relative abundance increases slowly over ~3.5 kbp starting at the far left end (5′) end of the long unit; the increase in frequency is no >2-fold. About 2.5 kbp or less from the 3′ end, the relative abundance begins to rise, increasing as the right end is approached. Sequences at the 3′ end are 4–5 times more abundant in the genome than those of the 5′ end. This finding is consistent with the fact that all except one (see Miyake *et al.*, 1983) characterized *Kpn*I family members align with the 3′ end of the longest units and suggests that family members that are truncated by varying extents at the 5′ end occur about five times more frequently than do full length units.

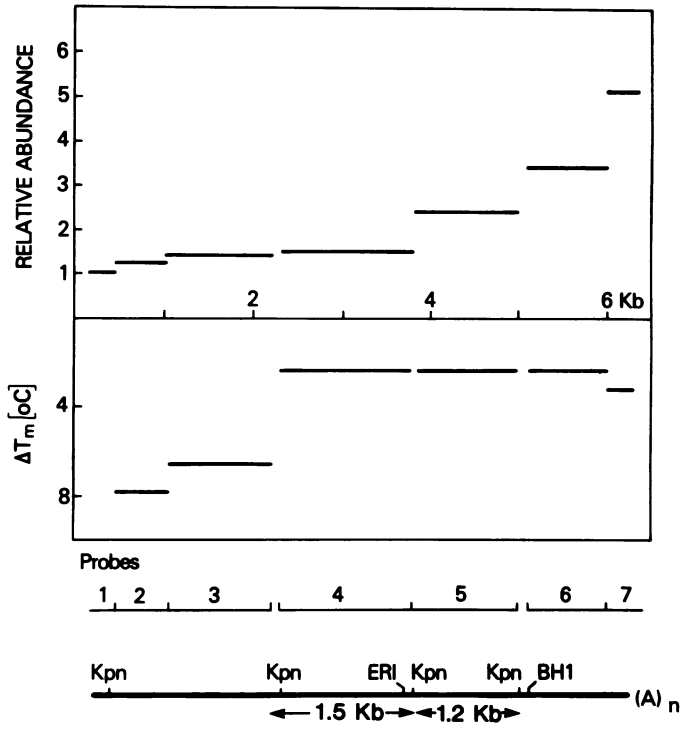

Fig. 5. Relative copy number in the monkey genome of different regions from within *Kpn*I family members. The procedures are described in the text and in the section on Materials and methods. The abscissa represents the ~6 kbp of a full *Kpn*I element. One ordinate gives the relative copy number of each probe. The second ordinate shows the difference in melting temperature (△Tm) measured for the probe annealed to itself compared with the probe annealed to genomic DNA; the duplexes were formed at 47°C (see Materials and methods). The probes were all subcloned as follows (see Figure 1): **1**, an *Eco*RI fragment from λF2 extending from low copy number sequence to the left to the *Eco*RI site ~200 bp from the 5′ border of the *Kpn*I unit in pEMBL9 (Dente *et al.*, 1983); **2**, an ~0.6-kbp *Eco*RI/*Hind*III fragment from λB3 in pUC8 (Vieira and Messing, 1982); **3**, an ~1-kbp *Hind*III/*Cla*I fragment from λB3 in pUC8; **4**, an ~1.5-kbp *Kpn*I fragment from λCaα7 in pUC8c2 (a derivative of pUC8 containing a *Kpn*I site kindly provided by G.F.Crouse, Frederick Maryland Research Facility, National Cancer Institute); **5**, an ~1.2-kbp *Kpn*I fragment from λCaα7 in pUC8c2; **6**, an ~1-kbp *Bam*HI/*Xho*I fragment from λCaα11 in pBR322; **7**, an *Xho*I/*Eco*RI fragment from λCaα11 extending from ~200 bp inside the 3′ end of the *Kpn*I unit to the next *Eco*RI site (note that the copy number of the non-*Kpn*I segment in probe 7 is at most 200 (see text) in pBR322. Probes 5 and 6 were kindly provided by R.E.Thayer. The map at the bottom is provided to assist in the localization of the probes and depicts some common restriction endonuclease sites in *Kpn*I segments.

We also determined the stability of the hybrids formed

(Figure 5). The less abundant sequences within 2.5 kbp of the 5′ end are associated with a significantly higher sequence divergence than are those closer to the 3′ end. The actual copy number of subcloned segments 2 and 5 was also determined (see Materials and methods). The average of two determinations was $4.0 \pm 0.4 \times 10^3$ and $8.6 \pm 1.4 \times 10^3$ for segments 2 and 5, respectively.

Discussion

Over the past few years several groups have investigated both the genomic organization of the *Kpn*I family members and (or) the structure of individual cloned units. However, the available nucleotide sequences did not allow definition of common 5′ and 3′ ends. The data suggested that at the 3′ end, the homology between different members degenerates in a d(A)-rich stretch of variable length and sequence. This is true both of genomic segments and cDNA clones. However, many of the sequences are derived from variously truncated and rearranged elements. We wanted to determine if similar 3′ ends occur in the longest and apparently complete *Kpn*I elements. We also wanted to define the 5′ end of complete family members particularly to see if any sequences are repeated at the two extremes of these elements.

We determined the sequence at the 5′ and 3′ ends of two *Kpn*I family members (*Kpn*-λB3 and *Kpn*-λF2) that are >6 kbp long as determined by hybridization to different subcloned *Kpn*I probes and heteroduplex analysis. The alignment of the two 5′ ends (Figure 2) shows that the homology between the two sequences starts at a G-rich region. This region of λF2 and λB3 matches the start of two other *Kpn*I family members whose nucleotide sequences were reported while our work was in progress (*Kpn*A, Potter, 1984; *Kpn*-TβG41, Miyake *et al.*, 1983). Comparing only two sequences, Miyake *et al.* (1983) placed the 5′ end 6 bp further downstream from the first common nucleotide at which the homology between the four elements (Figure 2A) starts. The alignment of these four sequences indicates that the 5′ end is sharply defined by the residue numbered 1 (Figure 2A). Unpublished work by A.Nienhuis and his colleagues (personal communication) on two additional family members confirms this starting point.

At the 3′ end, the homology between *Kpn*-λB3, *Kpn*-λF2 and *Kpn*-TβG41 terminates in a d(A)-rich stretch analogous to that in the truncated *Kpn*I family members previously analyzed. However, *Kpn*-λF2 and *Kpn*-TβG41 are more similar to one another than to the others. In both, the number of A residues is relatively small and the sequences match perfectly in this region; the homology terminates two or four A residues (see below) downstream from identically positioned polyadenylation sites. This region is followed in λF2 by 16–18 bp that are perfectly repeated at the 5′ end boundary (see Figure 2, boxed sequences). The presence of this direct repeat immediately flanking the 5′ and 3′ end of *Kpn*-λF2 suggests that a target site duplication occurred coincident with an insertion event. The position of this repeat at the 3′ end allows us to define the last possible nucleotide of the *Kpn*-λF2 sequence. A similar structure is recognizable in *Kpn*-TβG41 where 9–11 bp in the d(A)-rich region are repeated >6 kbp upstream immediately preceding the beginning of the *Alu* sequence joined to the 5′ end of this *Kpn*I element (Miyake *et al.*, 1983). Miyake and co-workers interpreted this observation as the duplication of a target sequence upon insertion of an *Alu-Kpn*I composite unit. These features support the definition of *Kpn*-λF2 and *Kpn*-TβG41 as members of a subset of *Kpn*I sequences that are ~6 kbp in length and co-linear throughout, share 5′ and 3′ ends, and have identically positioned polyadenylation signals followed by a few d(A) residues. However, this subset does not appear to include all *Kpn*I members that are ~6 kbp. *Kpn*-λB3 shares the 5′ end sequence but differs at the 3′ end and contains three polyadenylation signals at different positions within a very long d(A) stretch. We have not observed repeated sequences flanking *Kpn*-λB3 but cannot eliminate the possibility that a longer segment including *Kpn*-λB3 is bounded by flanking repeats, as is the case with *Kpn*-TβG41. However, immediately after the 3′ end of *Kpn*-λB3 there is a 50-bp long stretch of alternating purines and pyrimidines composed mainly of alternating Cs and As. Flanking target site duplications have been reported surrounding the truncated family member *Kpn*I-RET (Thayer and Singer, 1983). Notably, the size of the duplications varies from one family member to another.

Comparison of the sequences that define the two ends of long family members, regardless of subset, indicates that neither direct nor indirect terminal repeats occur within the *Kpn*I segment itself. The fact that some family members are associated with flanking direct repeats that may be target site duplications suggests that some *Kpn*I family members may be mobile. If so, then the absence of terminal repeats suggests that the *Kpn*I family is more like the F-family of moveable elements in *Drosophila* than other transposable elements. Like *Kpn*I family members, F-elements have polyadenylation signals at their 3′ ends and some F elements are truncated at the 5′ end (DiNocera *et al.*, 1983). The suggestion that *Drosophila* F-elements resemble processed pseudogenes except that no discrete gene product is known (DiNocera *et al.*, 1983) can be extended to *Kpn*I units.

Truncated *Kpn*I family members often lack sequences from the left end of the 6-kbp unit. The three segments joined to α-satellite in the cloned segments in λCaα6, 7 and 11 lack >900 bp of left end sequence. Moreover they are joined at the 3′ end to an unrelated segment that is repeated ~150 times in the monkey genome. The overwhelming majority of these repeats are in a context similar to that in λCaα6, 7 and 11, namely, bordered by *Kpn*I sequence at one end and α-satellite at another. No comparable abundance of the non-*Kpn*I segment occurs in human DNA. It is likely that one copy of the combination of truncated *Kpn*I segment with the unrelated sequence found its way into α-satellite and then was amplified in association with amplification of the satellite itself. The situation illustrates a novel way in which genomic segments may be amplified but it seems unlikely that these segments are of any inherent significance.

The data presented here indicate that sequences associated with the far left end of *Kpn*I units occur <4000 times in the monkey genome while those at the far right end occur five times more frequently or ~20 000 times. Thus, there are a maximum of 4000 full length units and probably fewer since at least one cloned unit is missing 3′ end sequences (Miyake *et al.*, 1983). Our estimates agree with the data of Adams *et al.* (1980) who measured the copy number in human DNA with a probe representing the 5′ half of a full unit. The gradient of copy number is unique for repeated sequence families with the exception of the rodent repeat family called *Bam*HI, MIF1 or L1 (Voliva *et al.*, 1983; Bennett and Hastie, 1984) which is homologous at least in part to the *Kpn*I family (Singer *et al.*, 1983). Also, the direction of the gradient in these two LINE families is unusual; except for the F-family,

other repeated sequence families that have truncated members, most notably the U1-, U2- and U3-RNA gene families (Van Arsdell and Weiner, 1984), tend to lack 3′ end sequences, not 5′.

Another notable feature of the *Kpn*I family is the gradient of increasing homology among family members from the 5′ to the 3′ end. Comparison of several available sequences from both ends confirms the difference indicated by the $\triangle T$ms. Whether these changes reflect different extents of sequence homogenization or some selective pressure imposed by an (unknown) function remains to be clarified.

Materials and methods

All cloning, sequencing and hybridization experiments were carried out as previously described unless indicated otherwise (Grimaldi and Singer, 1983; Thayer and Singer, 1983). The three phage λCaα6, 7 and 11 were isolated from a monkey genome library and were previously characterized as were the subclones pCaα6.10.1 and pCaα11.1 and the *Bam/Eco* fragment derived from pCaα11.1 (Grimaldi and Singer, 1983) (see also Figure 1). pCaα1004 is a clone containing monkey α-satellite sequence (Thayer *et al.*, 1981). Various other subcloned probes used in the present experiments are described in Figures 1 and 5.

The relative genomic abundance of various segments from within *Kpn*I family members was measured by the quantitative dot-blot procedure (Kafatos *et al.*, 1979). Approximately 1 μg samples of linearized recombinant plasmids were immobilized on nitrocellulose using the Schleicher-Schull dot-blot manifold (Schleicher and Schull, Inc., Keene, New Hampshire). The conditions for pre-annealing and annealing were: 0.6 M sodium chloride, 0.06 M sodium citrate (4 x SSC), 40% formamide, 0.5% SDS, 0.2% each of bovine serum albumin, Ficoll and polyvinylpyrrolidone, 0.1% sodium pyrophosphate and 200 μg/ml carrier *E. coli* DNA at 42°, 47°, or 65°C for 18 h. After hybridization, filters were washed extensively under the same salt and temperature conditions as the original annealing and the hybrids were melted in 2 x SSC, 40% formamide. The Cerenkov radiation was determined for individual dots after washing and subsequent melting (at 5°C intervals) directly in the washing medium. For copy number measurements six 2-fold dilutions of a known amount of monkey DNA, λB3 DNA and λF2 DNA were immobilized on the same nitrocellulose sheet and hybridized in the conditions given above (42°C) with an excess of ^{32}P-labeled clone probe.

References

Adams,J.W., Kaufman,R.E., Kretschmer,P.J., Harrison,M. and Nienhuis, A.W. (1980) *Nucleic Acids Res.*, **8**, 6113-6128.

Bennett,K.L. and Hastie,N.D. (1984) *EMBO J.*, **3**, 467-472.

Dente,L., Cesareni,G. and Cortese,R. (1983) *Nucleic Acids Res.*, **11**, 1645-1655.

DiGiovanni,L., Haynes,S.R., Misra,R. and Jelinek,W.R. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 6533-6537.

DiNocera,P.O., Digan,M.E. and Dawid,I.B. (1983) *J. Mol. Biol.*, **168**, 715-727.

Fanning,T.G. (1983) *Nucleic Acids Res.*, **11**, 5073-5091.

Gebhard,W. and Zachau,H.G. (1983) *J. Mol. Biol.*,, **170**, 255-270.

Grimaldi,G. and Singer,M.F. (1983) *Nucleic Acids Res.*, **11**, 321-338.

Kafatos,F.C., Johns,C.W. and Efstratiadis,A. (1979) *Nucleic Acids Res.*, **7**, 1541-1552.

Lerman,M.I., Thayer,R.E. and Singer,M.F. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3966-3970.

Manuelidis,L. and Biro,P. (1982) *Nucleic Acids Res.*, **10**, 3221-3239.

Meunier-Rotival,M. and Bernardi,G. (1984) *Nucleic Acids Res.*, **12**, 1593-1608.

Miyake,T., Migita,K. and Sakaki,Y. (1983) *Nucleic Acids Res.*, **11**, 6837-6846.

Potter,S.S. and Jones,R.S. (1983) *Nucleic Acids Res.*, **11**, 3137-3153.

Potter,S.S. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 1012-1016.

Rosenberg,H., Singer,M. and Rosenberg,M. (1978) *Science (Wash.)*, **200**, 394-402.

Shafit-Zagardo,B., Maio,J.J. and Brown,F.L. (1982a) *Nucleic Acids Res.*, **10**, 3175-3193.

Shafit-Zagardo,B., Brown,F.L., Maio,J.J. and Adams,J.W. (1982b) *Gene*, **20**, 397-407.

Singer,M.F. (1982a) *Cell*, **28**, 433-434.

Singer,M.F. (1982b) *Int. Rev. Cytol.*, **76**, 67-112.

Singer,M.F., Thayer,R.E., Grimaldi,G., Lerman,M.I. and Fanning,T.G. (1983) *Nucleic Acids Res.*, **11**, 5739-5745.

Soriano,P., Meunier-Rotival,M. and Bernardi,G. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 1816-1820.

Thayer,R.E., Singer,M.F. and McCutchan,T.F. (1981) *Nucleic Acids Res.*, **9**, 169-181.

Thayer,R.E. and Singer,M.F. (1983) *Mol. Cell. Biol.*, **3**, 967-973.

Van Arsdell,S.W. and Weiner,A.M. (1984) *Nucleic Acids Res.*, **12**, 1463-1471.

Vieira,J. and Messsing,J. (1982) *Gene*, **19**, 259-268.

Voliva,C.F., Jahn,C.L., Comer,M.B., Hutchison,C.A. III and Edgell,M.H. (1983) *Nucleic Acids Res.*, **11**, 8847-8859.

Received on 22 May 1984