Non-Alu family interspersed repeats in human DNA and their transcriptional activity

Lily Sun, K.Eric Paulson[1], Carl W.Schmid*, Lisa Kadyk[2] and Leslie Leinwand[2]

Departments of Chemistry and [1]Biochemistry and Biophysics, University of California, Davis, CA 95616, and [2]Department of Microbiology and Immunology, Albert Einstein College of Medicine, Bronx, NY 10461, USA

Received 18 January 1984; Accepted 29 February 1984

Abstract

Randomly selected human genomic clones have been surveyed for the presence of non-Alu family interspersed repeats. Four such families of repeats have been isolated and characterized with respect to repetition frequency, interspersion, base sequence, sequence divergence, *in vitro* RNA polymerase III transcription, elongation of transcripts in isolated nuclei, and *in vivo* transcription. The two most abundant of the four families of repeats correspond to previously reported families of repeats, namely the Kpn I family and poly (CA). We conclude that most of the highly repetitive (> 50,000 copies) human interspersed repeats have already been identified. Two lower abundance repeats families are also described here.

The abundance with which each of these families is represented in nuclear RNA qualitatively corresponds to their genomic reiteration frequencies. Further, the complementary strands of each repeat family are approximately symmetrically transcribed. The abundance of these repeats in cytoplasmic RNA is qualitatively less than in nuclear RNA. The bulk of the *in vivo* transcriptional activity of these repeats thus appears to be non-specific read through from other promoters.

INTRODUCTION

The genomes of most eukaryotes contain repetitive DNA sequences which are interspersed with single copy sequences. In many eukaryotes such as sea urchin and human, these interspersed repeats are short, typically about 300 bp in length (1,2). An important distinction between the short interspersed repeats of these divergent organisms, is that in sea urchin DNA, no single family dominates the many different families of sequences whereas in human DNA a single relatively abundant family of sequences, the Alu family, comprises the majority of the 300 bp interspersed repeats (1). The existence of this major family of interspersed repeats does not preclude the presence of numerous additional families of interspersed repeats in human DNA, and as reviewed below, examples of non-Alu family interspersed repeats in human DNA have been previously reported. To compare genomic organization between sea urchin and human, it is interesting to ask whether these non-Alu family human interspersed repeats belong to a few or many different families of sequences.

The structure and biochemical properties of human Alu family members provide clues to the mechanism by which they are dispersed. Individual members of the family share an almost precise 5' end but terminate in an A rich 3' end of variable sequence and length (1,3,4). The Alu family members are flanked by short direct repeats which we are certain were duplicated upon insertion of an Alu family member at an unoccupied genomic site (5-8). Alu family members contain *in vitro* RNA polymerase III promoters and are related in sequence to 7 S RNA (1,9,17,18). Thus there is a very close analogy between Alu family members and pseudogenes for a number of different small RNAs (3,4,10,17). As yet it is not known whether some Alu repeats serve as genes.

These pseudogenes can best be thought of as retrogenes in which the RNA gene product was converted into a complementary DNA sequence which was reinserted at a new chromosomal location. This explanation accounts for the precise 5' end of these pseudogenes, their variable A rich 3' end, their resemblance to the corresponding transcription product and the duplicated chromosomal insertion sites (1,3,4,10,17). There are other examples of retrogenes which in addition to these features have been subjected to RNA splicing prior to chromosomal reinsertion (reviewed in 10 and 11). These processed messages might be thought of as "interspersed repeats" which belong to relatively low copy number multigene families. Conceivably most families of interspersed repeats are dispersed by a similar mechanism involving genomic reinsertion of a complementary DNA sequence. Alternatively, the dispersion products of known gene gamilies are not typical of randomly selected interspersed repeats. The issue is whether randomly selected interspersed repeats are generally retrogenes.

To provide insight into both of these issues the primary structure, transcriptional activity, genomic distribution and copy number of a few examples of randomly selected interspersed repeats have been determined.

MATERIALS AND METHODS

A. Recombinant DNAs

The human genomic clones surveyed for repetitive sequences were isolated from a λ Charon 4A genomic library which was constructed by Dr. Arthur Banks. Subclones in either pBR 322 or pBR 327 were constructed and grown by standard methods (36). M13 subclones for either sequencing or for use as hybridization probes were constructed by insertion into the universal cloning sites of M13 strains mp 8 or 9 (37).

B. DNA Sequencing

Except for limited Maxam Gilbert (38) sequencing of portions of a Kpn I family member, all sequencing was done by Messing's (37) modification of Sanger's dideoxy method.

C. RNA Techniques

Nuclear and cytoplasmic RNAs were prepared from Hela cells (a gift of Drs. P. Yau and M. Bradbury) as described by Weiner (18). In preliminary blot hybridization experiments these RNAs were found to contain a DNA contaminant which was removed by sedimentation equilibrium in buoyant CsCl.

Transcription by human lymphoblastoid nuclei in the presence of various α amanitin concentrations was performed and assayed as described (39). *In vitro* transcription by RNA polymerase III (a gift from Mr. C. Perez-Stable and Dr. J. Shen) was performed as described by Wu (40).

D. Filter Hybridization Methods

DNA was blotted from agarose gels onto nitrocellulose by Smith and Summer's (41) modification of Southern's procedure. Denatured RNA electrophoresed on 1.2% agarose gels in 2.2M formaldehyde was transferred to nitrocellulose as described by Lehrach et al. (42). Hybridization to these filters was performed by standard methods (36). The hybridization on Northern blots was abolished by preincubating the blots in alkali, thus demonstrating the absence of DNA contamination in the RNA preparations.

Melting temperatures of hybridized repeats were estimated by thermal elution in 1 x SSC of P^{32} labelled probe DNA which had been hybridized to DNA (10 μg) spotted onto nitrocellulose filters. Spot Cot analysis using both scintillation counting and radioautography was used to estimate repetition frequencies of repetitive sequences (41). For this purpose serial dilutions of total human DNA and appropriate plasmid cloned DNAs were immobilized as spots on nitrocellulose. The nitrocellulose strips were hybridized to radiolabelled repetitive sequence probes. The radiolabelled probes were prepared from M13 subclones of repetitive elements by use of the universal primers (see A above). For determination of the "O" copy number both human DNA and a pBR clone of a 2.6 kb Eco Rl fragment containing the entire O_4 repeat depicted in Figure 1 were annealed with a radiolabelled M13 subclone containing a 500 bp Alu I - Eco Rl restriction fragment of the O_5 repeat. The Alu cloning site is positioned 170 bp 5' to the O_5 sequence depicted in Figure 1; the Eco Rl cloning site is position 329 of the O_5 sequence depicted in Figure 1. As an illustrative sample calculation of a particular experiment: 1 ng of a 5.8 kb plasmid O_4 subclone hybridized as much repetitive label as 63 ng of human DNA. This suggests that there is approximately one copy of this repeat per every 365 kb (63 x 5.8 kb) of human DNA or alternatively 6,900 copies (2.5×10^9 bp/3.65×10^5 bp) in the human genome. The copy number of poly (CA) was similarly estimated by annealing a pBR subclone containing a 2.9 kb Hind III restriction fragment with a radiolabeled M13 subclone of the 2.9 kb Hind III fragment. This M13 subclone contains a 346 bp Alu insert which includes the poly $(AC)_{17}$ tract reported in the text. This same pBR subclone contains an additional repetitive element, called K. The copy number of the K element was estimated by use of a radiolabelled M13 subclone containing a 123 bp Alu restriction fragment. The 5' end of this fragment is at position 135 (...CCCCAGCTTCCC...); the 3' end is at position 258 (...AAGACAGCTGAGG...) of the sequence reported in the text.

Colony screening was also employed to estimate the repetition frequency of various repeat families. For this purpose several thousand recombinant λ phages from the human library were immobilized onto nitrocellulose and annealed to radiolabeled cloned repetitive sequences. The repetition frequency "f" of a given family is estimated from the fraction of phages "F" which hybridize to the repetitive sequence according to the equation $f = 1.25 \times 10^{+5} \times F$. This equation assumes that the human genome (2.5×10^6 kb) can be represented by 1.25×10^5 λ recombinants each of which has a 20 kb insert size (i.e. $1.25 \times 10^5 = 2.5 \times 10^6/20$).

RESULTS

1) Isolation of non-Alu Interspersed Repeats

Ten randomly selected clones were probed in parallel with a cloned Alu family member and total human DNA. Non-Alu family repeats are identified as those restriction fragments which anneal to human DNA but not to the Alu family member. In principle, the number of such non-Alu family repeats (at least six in this work) and the amount of human DNA surveyed (about 200 kb = 10 clones x 20 kb per clone) provides an estimate of the number of non-Alu family interspersed repeats. In practice, this value is a serious underestimate of the number of non-Alu family repeats for at least two reasons. First, this method is insensitive to low copy number repeats. Although the literature suggests that the presence of a 100 fold cloned repeat is detectable by this approach (19), our own experience reported here is that we were barely able to detect a 1000 fold repeat. Second, a non-Alu family repeat which is closely linked to either an Alu family member or another non-Alu repeat would not be detected except by high resolution restriction mapping.

By the differential hybridization of a cloned Alu family member and total human repeats to these ten clones, four non-Alu family repeats have been detected. The previously undescribed repeats are called "O" and "K". The poly (CA) and Kpn I families correspond to previously reported families of repeats.

2) Repetition frequency and interspersion of the non-Alu repeats

The repetition frequency of the four non-Alu family repeats has been determined by hybridization, particularly spot Cot analysis (Table I). The repetition frequency of the Kpn I family member, previously determined by a hydroxylapatite Cot analysis, Table I (12), is corrected for sequence mismatching and length effects as has been done in careful estimates of the repetition of the Alu family (13).

Library screening demonstrates that each family is interspersed throughout the genome at a frequency corresponding to its repetition frequency (Table I). The good agreement between the repetition frequencies estimated by hybridization and by library screening is strong evidence that each family is broadly distributed throughout the genome. The alternative, which is disproven by this comparison, is that a given repeat family might be tightly clustered in certain regions of the genome.

3A) The O family: structure

The DNA base sequences of two O family members are compared in Figure 1. The homology between these two sequences ends abruptly at the 5' side. The 3' end of O-4 continues past that of O-5 as a long internally repeated A rich sequence. Both of these features are typical of Alu repeats (1,17). A recognizable direct repeat, agtaatc, occurs precisely at the 3' non-homology boundary of clone O-5. An imperfect eleven nucleotide direct repeat can be recognized as flanking the 5' end of O-4 and its 3' A rich region. As reviewed in the Introduction, these features are hallmarks of retrogenes (1,3,4,10,17). One O clone (O_5) contains a 57 bp insertion element which is also flanked by short direct repeats (Figure 1). This could be an unprocessed intervening sequence which is present in the parent RNA molecule or a parasitization of an interspersed repeat by a second family of interspersed repeats. The following S1- bot hybridization experiment was employed to determine the relative abundance of the two O

Table 1. Repetitive DNA Families in the Human Genome.

Repetitive	Length (ℓ)	Frequency (f)		% Mismatching		%
DNA family	(bp)	Spot cot[a]	Colony screening[b]	Tm[c]	Sequence[d]	Total Genome[e]
poly (CA)	∿ 40	130,000	100,000 (Ref. 29)	f'	-	0.2
K family	∿ 200	1,050	1,400	11	-	∿ 0.01
O family	373-435	4,500	2,000 - 37,000[g]	10	10	0.07 - 0.08
Kpn I family	670-6,400	50,000 (Ref. 12)	40,000	10 (Ref. 12)	11	1.4 - 12.8
Alu family	300	$0.5 \times - 1 \times 10^6$ (Ref. 13)		19 (Ref. 13)	20 (Ref. 16)	6 - 12

a. Unless otherwise indicated the repetition frequency (f) was estimated by spot hybridization as described in Methods.

b. The recombinant λ phages from the human DNA library were screened using cloned repetitive DNA probes. The frequency "f" is estimated from the fraction of the phages which hybridize to the repetitive sequence as described in Methods.

c. The Tm's of the renatured duplex formed between the probe and the parent clone DNA and between the probe and total sheared human DNA were determined. Each 1°C depression in Tm corresponds to 1% base-pair mismatching (15).

d. Nucleotide sequences of at least two clones from a repetitive DNA family were compared. Each point mutation, insertion or deletion regardless of size is taken as one mismatch.

e. The % total genome was calculated using the formula $\ell f \times 100/2.5 \times 10^9$ using the spot Cot f.

f'. The 10°C depression in Tm can be attributed to the short length of the poly (CA) repeat.

g. The lower estimate was obtained by counting only those recombinant phages that anneal strongly to the probe. The higher estimate includes phage that anneal weakly to the probe.

```
                                                              *    **                 * *           *  100
0–4   aaaaaagtagataaatgacgaatgaTGTATTAGTCTGTTTTCACACTGCTGATAAAAACATAA      CTGAGACTGTGGAGAAAAAGAGGTTTAATTG
0–5   tagcaatgaatttgctcatagtaacTGTATTAGTCTGTTTTCACACTGCTGATAAAGACATTTTTCTTCCTGAGACTG GAAAAAAAAGAGGTTCAATTG

            * *      *              *             *                                                    200
0–4   GACTTACATTTCCACATGGCTGGGGAGGCCACAGAATCATAGCGAGAGGTGAAAGGCACTTCTTACATGGTGGTGGCAAGAGG
0–5   GACTTATAGTTCCACTTGGCTGGGGAGGCCTCAGAATCATAGCGGGAGGTGAAAGGCACTTCTTACATGGTGGTGGCAAGAGGAAAATGAGGAAGAAGCA

                                               *       *               ***     **   **           *  *  300
0–4                                           CCTATTCACGATCATGAGAATAGCATGGGAAAGATCAGTTCCCATGATTCAATTGCTTCC
0–5   AAAGGGGAAACCCCTGATGAACCCATCAGATCTCATGAGACTTATTCACTATCATGAGAATAGCACAAGAAAGGCCAGGCCCCATGATTCAGTTACCTCC

             *           ***     *             *                  *                  *  *              400
0–4   CCCTGGGTCCCTCCCACAACATGTGG AGTTCTGGGAGATATAATTCAAGTTGAGATTTGAATGGGGACACAGCCAAACCATATCAAATGAATATGCTAG
0–5   CCCTGGGCCCCTCCCACAATTCGTGGAATTCTGGGAGATACAATTCAAGTTGAGATTTGGATGGGGACACAGCCAAACTATGTCagtaatcacggtcga

                                                                                                       500
0–4   AAATGAGGAAAACAAAAATCAAAAGGTACAAAAAAGGAAAAAAAAAAAAAAAAAGAACAGAGGatgcataatgaagagatttaatta
0–5   gactatctgcaatttagtgaatcaagcctgtggtttacggcaactcagagggcttttgtgcatgtgcaaggtgggttgggcagcgc
```

Figure 1. pBR subclone O_4 was derived from a randomly selected λ genomic clone, O_5 was obtained by screening the human library with O_4. Differences are noted by asterisks. Each repeat is flanked by imperfect short direct repeats as indicated by arrows. The extra DNA in O_5 (positions 187 - 240) is also flanked by a five base direct repeat as indicated by arrows.

variants that are shown in Figure 1.

Renatured human DNA was treated with S1 nuclease to release duplex repetitive sequences (14). By gel electrophoresis, the resulting duplex DNA shows a smear of lengths and a prominent 300 bp band which is assigned to the Alu family (14). O family clones anneal to three discrete bands in this S1 digest (Figure 2): a major band of 420 bp, and faint bands of 370 and 270 bp against a heterogeneous background of hybridization. The heterogeneous background might result from either length heterogeneity of O repeats or from the activity of S1 nuclease against mispaired duplex molecules. For example, the most stringent S1 digestion (lane C, Figure 2) almost eliminates the major 420 bp band. Since the sizes of two of these bands (420 bp and 370 bp) correspond to the two variant sizes (411 bp and 360 bp) reported in Figure 1, each may be represented as multiple genomic copies but the 420 bp variants are more numerous.

The two sequenced O clones also differ by a number of point mutations (Figure 1). Is this divergence typical of O family members or in using the first clone as a probe to isolate the second have we perhaps selected for closely related members of the family? The extent of divergence of repetitive sequences can be estimated from the depression in their thermal stability relative to well paired duplex DNA, according to the rule that a 1% base pair mismatching in duplex DNA depresses its melting temperature by 1° (15). There is good agreement between the divergence observed between sequences of the O clones and the value estimated from the melting tempera-

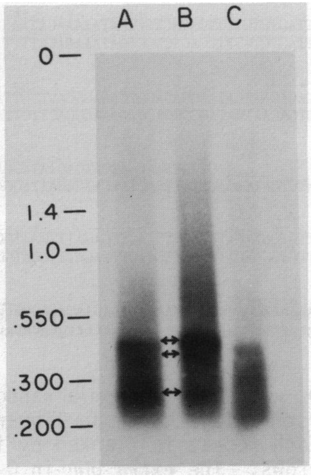


Figure 2. Length variants of the O family of repeats. Total human DNA (12 μg) renatured to a Cot of 68 was digested with the indicated amount of S-1 nuclease for the indicated length of time (14). The duplex repetitive sequence fragments were separated by electrophoresis on a 1.5% agarose gel and transferred to a nitrocellose filter which was hybridized with a ^{32}P-labelled M13 subclone of the O_5 repeat. Size markers are indicated in kilobase pairs (Kbp): (a) 24 units of S-1, 120 min, (b) 14 units of S-1, 30 min and (c) 45 units of S-1, 120 min.

ture of an O clone renatured to total human DNA, Table I. The divergence observed for the two sequenced members of the O family is typical of the divergence of the family as a whole. The pattern of divergence is also of interest. 25% of the point mutations in the two O family repeats are transversions (8 transversions, 24 transitions, Figure 1). A similar bias in point mutations (28% transversions) has been observed for Alu family members (16).

3B) The O family: transcriptional activity

The view advanced in the Introduction is that some dispersed repeats are reverse transcriptions of RNAs (3,4,10,17). The transcriptional activity of O family repeats has been examined by Northern blot analysis, *in vitro* transcription by RNA polymerase III, and transcription by isolated nuclei.

Neither the O_5 subclone of the O family nor the other three non-Alu repeat families reported below are transcribed *in vitro* by RNA polymerase III (Figure 3) and unpublished). A positive control is the transcriptional activity of the Alu family member (lane 8) which results in a long transcript identical to that observed in reference 47. Other bands as exhibited

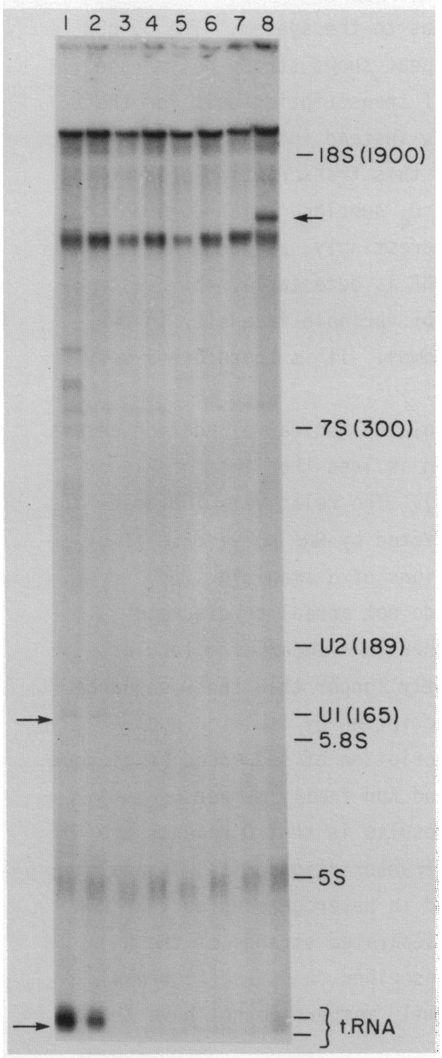
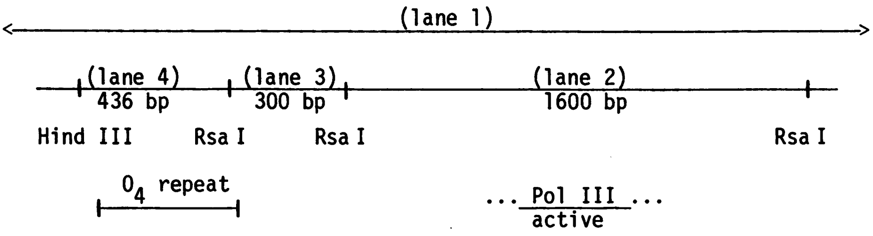


Figure 3. *In vitro* RNA Pol III transcription of non-Alu family members. P^{32}-labeled RNA was transcribed from plasmid or restriction fragment templates using a HeLa RNA Pol III extract (40) and run on a 4% polyacrylamide/7M urea gels. Lane 1, O_4 plasmid; lane 2, RsaI-1600 bp; lane 3, RsaI-300 bp; lane 4, Hind III/ Rsa-436 bp; lane 5, O_5 plasmid; lane 6, EcoRI/HpaII-2.6 kbp (an internal fragment of the O_5 clone, 2.6 kb, containing the O_5 repeat); lane 7, no DNA added; lane 8, BLUR 2-Alu family plasmid. Arrows indicate specific transcripts (several of the longer bands in lane 1 probably correspond to the pBR vector as similar bands are observed). Restriction fragments of the O_4 insert map as follows: Hind III/RsaI-436 bp, RsaI-300 bp, RsaI-1600 bp. The results of this experiment are depicted schematically on this map of the O_4 insert.



by the no DNA control (lane 7) are endogenous to the system. The 0_4 pBR subclone (lanes 1 and 2) of the 0 family repeat shows strong polymerase III transcripts, Figure 3. However, the Pol III transcription unit for the 0_4 subclone is not the 0 family repeat but maps instead into an adjacent fragment (lane 2, Figure 3). Additionally, this transcription product hybridizes to a restriction fragment of the 0_4 subclone which does not contain the 0 repeat (data not shown). Interestingly, although no hybridization to this band by total human DNA is detectable, the restriction fragment anneals to a number of bands of variable intensity in restriction digests of human DNA (data not shown). It is therefore a very low copy number interspersed repeat.

0 family repeats are transcribed in isolated nuclear elongation assays (Figure 4). The level of this transcription is less than that of Alu or Kpn I (*vide infra*) family members (Figure 4). The relatively low transcriptional activity of 0 appears to be directed by RNA polymerase II by virtue of its sensitivity to low concentrations of α amanitin.

By Northern blot analysis 0 subclones do not anneal to discrete length cytoplasmic or nuclear RNAs, but rather to a smear of molecular weights including RNAs which are significantly longer than the 0 sequence (unpublished). In agreement with the transcription by isolated nuclei (Figure 4), the steady state level of transcription of sequences homologous to the 0 family is lower than that of Alu and Kpn family members.

The simplest interpretation of these results is that 0 repeats are coincidently transcribed as part of longer transcription units, in analogy to Alu family members which are interspersed in heterogeneous nuclear RNA (1,17). In support of this suggestion the separated strands of the 0 repeats are approximately symmetrically transcribed *in vitro* (Figure 4). According to this interpretation, most 0 family members do not have their own specific promoter.

4A) The Kpn I family: structure

Base sequence comparisons show that one family of repeats isolated here is closely related to the primate Kpn I family of sequences (Figure 5) which in turn is known to be related to an equivalent repeat family in mouse, the Bam H1 family (20). Of particular relevance to our present findings, the Bam H1 family is known to consist of length variants which have truncated 5' ends (21,22). The full length mouse Bam H1 family, as well as the primate Kpn family, is about 7 kb in length (reviewed in 20). However, discrete fragment lengths of about 500 bp and 1000 bp from the 3'

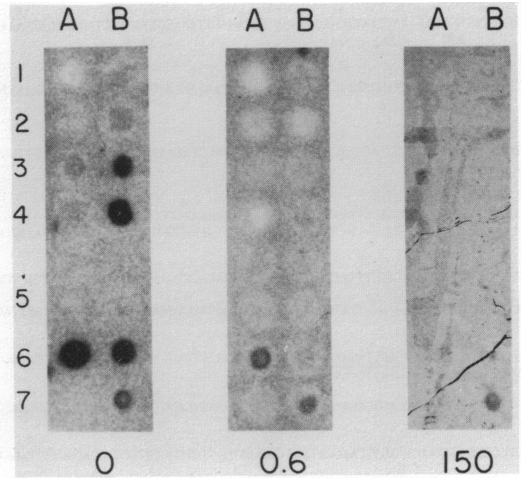


Figure 4. Hybridization of ^{32}P-RNA made in isolated nuclei with and without α amanitin to DNA from various repetitive DNA sequence families. RNA labeled with α-^{32}:-UTP was hybridized to DNA spotted onto nitrocellose. The filters were washed and exposed to X-ray film. 0 = no α amanitin, 0.6 = 0.6 μg/ml α amanitin and 150 = 150 μg/ml α amanitin. 2.5 μg of DNA from plasmid clones (A6 & 7, B6), 50 ng of DNA from 18S rDNA, (B7), and 1.0 μg of DNA from M13 clones (A1-5; B1-5) were spotted onto nitrocellulose. The identities of the DNAs and interpretation of the dots are as follows: 1A and 2A; M13 clones of complementary strands of the "K" repeat - demonstrating K is not significantly transcribed. 3A and 4A; M13 clones of complementary strands of the "O" repeat - demonstrating that both strands are lightly and approximately symmetrically transcribed by RNA polymerase II. 5A; M13 mp 8 - a negative control (see also 5B). 6A; Blur 8 - a pBR 322 clone of an Alu repeat demonstrating that this sequence is transcribed by both RNA polymerase II and III. 7A; pBR 322 - a negative control. 1B and 2B; M13 clones of $(CA)_{17}$ and $(GT)_{17}$ respectively - demonstrating that both strands are lightly and approximately symmetrically transcribed by RNA polymerase II. 3B and 4B; M13 clones of complementary strands of a Kpn family repeat (H Kpn-E13) - demonstrating that both strands are transcribed by RNA polymerase II. Probing by genomic DNA gave the same relative intensity for 3B and 4B so that 4B merely contains more DNA than 3B. Based on this additional control the strands are approximately symmetrically transcribed. 5B; M13 mp 9 - a negative control, see also 5A. 6B; Blur 16 a pBR 327 clone of the 3' end of the Kpn family (see Figure 6) - demonstrating that Kpn is transcribed by RNA polymerase II, see also 3B and 4B. 7B; 18 S rDNA an RNA polymerase I control which is insensitive to α amanitin.

end of the Bam H1 sequence have been observed as solitary dispersed units (21-23).

The complete sequence of one short human Kpn I family member and a partial sequence of a longer 2 kb member are compared to a partial sequence of a monkey Kpn I family member and a mouse Bam H1 family consensus sequence

```
                                                                                                           100
H KPN-E13 GGATCCCTTCCTTACACCTTGCACAAAAATTAATTCAAGATGGATTAAAGACTTAAACGTTAGACCTAAAACCATAAAAACCCTAGAAAAAAAACCTAGG
AGM       A  C     T  C       ATG      G                           T  A A   C                         --

                                                                                                           200
H KPN-E13 CATTACCATTCAGGACATAGGCATGGGCAAGGACTTCATGTCTAAAACACCAAAAGCAATGGCAACAAAAGACAAAATTGACAAAATGGGGGATCTAATT
AGM         A            GG             A         A   C          A    T          C             T--- T

                                                                                                           300
H KPN-E13 ACTAAAGAGCTTCTGCACAGCAAAAGAAACTACCATCAGAGTGAACAGGCAACCTACAAAATGGAAAGAAACCTTTTCGCAACCTACTCATCTGACAAAG
AGM             -G     A           T      T      C      G      T          G GA   --    T    G    C

                                                                                                           400
H KPN-E13 GGCTAATATCCAGAATCTACAATGA-GTCAAACTTGTTTATAAGAAAAAACAAGAACCCC----------------------ATCAAAAGGTGGCAAAGG
AGM        T                    G     T   ----    C   -        ---CCCCATCAAAAAGGGGGCATATCGC       A CA

                                                                                                           500
H KPN-E13 ACATGAACAGACGCTTCTCAAAAGTAAGTATCTTTATGCAGTTTCCAAAAAACACCTGAAAA--TGCTCGTCATCACTGGCCATCAGAGAAATGCAAATC
AGM        T          A           A  AC  --       ACCAA -   --   A        A     ---  T    T  CT
MOUSE                GAATTCTCACCCGAGGA-----TTATCGAATAGCTGAGAAGCACCTGAAAAAATTTTCAACATCCTTAGTCATCATAGAAATACAAATC

                                                                                                           600
H KPN-E13 AAAACCACAATGAGATACCATCTCACACCAGTTAGAATGGCAATCATTAAAAAGTCAGGAAACAACAGGTACTGGAGAGGATGTGGAGAAATAGGAACAC
H KPN-10  catgggccaatcattctagagtgaactcaagatcccctaatgg                         A G                            -
AGM                                           G   TG  T         C    -        A G    C    C           G    TG
MOUSE     AAAACAACCCTGAGATTCCATCTCACACCAGTCAGAATGGGTAAGATCAAAAATTCAGGTGACAGCAGATGCTGGCAAGGATGTGGAGAAAGGGAACAC

                                                                                                           700
H KPN-E13 TTTTACACTGTTGGTGGGACTGTAAACT-AGTTCAACCATTGT-GGAAGTCAGTGTGGCCATTCCTCCAGGGATCTAGAACTAGAAATACCATTTGACCC
H KPN-10                     G      T C              -   - A                 A      G
AGM         C                A A    T -            A A            A  --------------------------------
MOUSE     TCCTCCATTGTTGGTGGGATTGCAAGCTT-GTACAACCACTCT-GGAAATCATTCTGGC-AGTCCTC-AGAAAATTGGACATGGTACTACCGGAGGATCC

                                                                                                           800
H KPN-E13 AGCCA--TCCCATTACTGGGTATATACCCAAAGGATAAATCATGCTGCTATAAA-GACACATGCACATGTATGTTTATTGTGGCACTATTCACAATAGCA
H KPN-10    TA -- T      T                   A  TAT          G  G             A        C     A ACG
AGM          A --C      C                 A --       TG A    G  -                         CA        T     T
MOUSE     AGCAATACCCC--TCCTGGGCATATATCCAGAAGAT-GCCCCAACTGG-ATGAAGGACACATGCTCCACTATGTTCATAGCAGCCTTATTTATAATAGCC

                                                                                                           900
H KPN-E13 --GACTTGGAACCAACCCAAATGTCCAACAATGAT-GACTGGATAAAGAAA-TGTGGCACATATACACCATGGCATACTGT-CAGCCATAAAAATGAATG
H KPN-10  AA C                   C   TA      A           A   A A             T     A     A C            A G
AGM       AA  T          A       C   TT      A               A             T       A   A A G            A
MOUSE     AGAAGCTGGAAAGAACCTAGATGCCCCTCAACAGAGGAATGGATACAGAAAATGTGGTACATTTACACAATGGAGTACTATTCAGCAATTAAAAAGAATG

                                                                                                          1000
H KPN-E13 AGTTCATGTCCTTTGTAGGGACATGGATGAAG-TGGAAACCATCATTCTTAGCAAACTGGCGCAAGGACAGAAAAC-CAAACACCGCATGTTCTCACTCA
H KPN-10                TC                C                C        -----             -        A   A
H BLUR 16 -------------             -    AC       G                 ATT         A     -                   T
AGM                      C                C                C           A  G A         -        T
MOUSE     AATTTATGAAATTCTTAGGCAAATGGATGGACCTGGAGGGCATCATCCTGAGTGAGGTAACAGACTCACAAAAGAAACTCACACAATATGTACTCACTGA

                                                                                                          1100
H KPN-E13 TAAGTGGGAATTGAACAATGAGAACACATGGACACAGGGAAGGGAACATCACACACTGGGGCCTGTTGGGGGTGGGGGCAAGGGGAGGGATAGCATTAG
H KPN-10           G                     T     TA       G    T        A     CA    CA      T
H BLUR 16   G                                -                  T      A      T  AGT     AGG
AGM                     G                               T    A        T      A     T             A   G
MOUSE     TAAGTGGATATTAACCCAAGA          Mouse Diverges

                                                                                                          1200
H KPN-E13 GAGATATACCTAATGC----------TAAA-------TGACAGGTTGATGGGTGCAGCACACCAACATGGCACATGTATACATATGTAACAAACCTGCAC
H KPN-10     TA          -----------    -------   TG                          A    G T     C       T
H BLUR 16                 ----------    -------     A   A
AGM         C C   A       ATGTGGGATT    GTCTAGA       G              A    C        G      TG

                                                                                                          1300
H KPN-E13 GTTCTGCACATGTACCCTAAAACTTAAAGTATAATAATAATAATAATAATAAATAAATAAATACAATAAAATAAAATTTCCTT
H KPN-10                   - G          TAGAACAAAAAAAAAaagatcccctaatgggatttgccaagatgtctcctcc
H BLUR 16    G            A GGGG        ATAATAATAATAAAA
AGM                     T  C G G        AAAAAAAAAAAAAAAATGCTGAAAAAAATTGAATAAAGCTT
```

Figure 5. Sequence of human Kpn I family members. Human genomic Kpn sequences E13 and 10 and the previously described human Kpn BLUR 16 (16) are presented in the figure. Genomic sequence flanking clone Kpn-10 is in lower case letters and direct repeats (positions 630 and 1250) are indicated by arrows. Blanks indicate homology between the primate sequences, -, indicates a deletion. For comparison we have presented a portion of the African green monkey Kpn I sequence of Lerman et al. (25) and a consensus mouse Bam HI sequence mouse Bam HI sequence derived from several sources (20,21,22,23).

(Figure 5). In addition, a fragment of a renatured member of the Kpn I family which was adventitiously cloned in our base sequence study of renatured human repeats is also included in this comparison (16). The evolutionary comparison of Figure 5 shows 67% homology between primates and mouse. Approximately 2/3 of the differences between mouse and primates can be regarded as species-specific, i.e. positions at which all human and monkey sequences are different than the corresponding mouse sequence. The 30 bp gap in the African green monkey sequence (positions 660-690) may be either a pecularity in this particular member of the monkey Kpn I family member or a specific difference between the human and monkey Kpn I families.

Individual cloned members of the Kpn I family differ by a number of point mutations (Figure 5). The number of point mutations determined by base sequence comparison agrees with the value estimated by thermal stability studies of renatured DNA (Table I). Again, the sequence diversity exhibited by the few cloned sequences appears typical of that for the family as a whole. Transitions (57%) are more frequent then transversions (43%), although not to the same extent as observed for the O and Alu families.

The various Kpn I family members which have been partially sequenced have an A-rich 3' end. The one short member of the family which is sequenced in entirety (670 bp) is flanked by 15 bp short direct repeats (Figure 5). Restriction mapping shows that a partially sequenced member of the family which shares the A rich 3' end is 2 kb in length (Figure 5). A third cloned member which also includes the A rich 3' end is at least 4 kb long (data not shown). Restriction mapping shows that all elements contained within the 2 kb repeat are also represented in this longer variant. As described for the O family (Figure 2) S1 hybridization has been used to determine if any of these Kpn I family length variants were predominant. Unlike the results reported for clone O, all attempts to detect renaturation products of Kpn I family members which have discrete lengths were unsuccessful. Instead of bands (see Figure 2, for example) those hybridizations routinely resulted in a smear of non-specific DNA fragment lengths (data not shown). Presumably, no particular 3' end length variants, such as that observed in Figure 5, is an abundant subgroup of the primate Kpn I family. This contrasts with Wilson and Storb's finding of discrete length variants in the mouse Bam H1 family (21).

Maio and colleagues (26,27) describe the Kpn I family as consisting of at least six distinct families of sequences largely in recognition of the number of distinct Kpn I restriction fragments which have been identified.

Several of these are now known to be part of a single Kpn I family. Our data suggests that the human Kpn I family of sequences is composed of variants that are truncated to various extents on their 5' end. This interpretation agrees with the previous finding that the equivalent family of mouse sequences (the Bam H1 family) consists of precisely such truncated variants (20-23). Although we believe this to be the most general structure for Kpn I family members, Thayer and Singer (48) report a scrambled Kpn I family member which is flanked by direct repeats.

The extreme length heterogeneity of Kpn I family members precludes an accurate estimate of the mass fraction of the genome which is occupied by the Kpn I family (Table I). Clearly, full length (6.4 kb?, (46)) members of the family are the exception, as is the 670 bp member of the family reported here. Consequently, estimates of the copy number of this family also vary from 6,400 to 50,000 copies, depending on precisely which variants are being studied (12,46). The 50,000 fold repetition (Table I) has been determined for a Kpn subclone (Blur 16) which should be especially abundant as it is located on the extreme 3' end of the consensus sequence (Figure 5). Results on the monkey Kpn I family (20) and the present results suggest that the average length of Kpn I family members could be roughly 2 kb. If, in the absence of more precise data, 2 kb is taken as the average length of members of this family, then the Kpn I family corresponds to about 4% of the human genome (Table I). The data of Kole et al. (24) suggest that the Kpn I family comprises 3 to 6% of the human genome.

4B) Kpn I family: transcriptional activity

In parallel with the study of the transcriptional activity of the O family the transcription of the Kpn I family was also investigated by Northern blot hybridization and *in vitro* transcription by nuclei. Transcription with isolated nuclei shows that most of the Kpn I family transcription is directed by RNA polymerase II (Figure 4). However, both strands of the sequence are transcribed, suggesting that these sequences are non-specifically transcribed as part of longer transcription units. By Northern blot analysis three groups have detected *in vivo* Kpn transcripts in nuclear RNA and at a reduced level in cytoplasmic RNA (24-26). We have repeated this observation by the use of complementary single strand M13 probes (as described in the parallel experiment of Figure 6) and find that both strands are represented in both nuclear and cytoplasmic RNA (data not shown). These *in vivo* results qualitatively confirm the finding cited above that transcripts from isolated nuclei hybridize to both strands. This implies that most but not necessarily all Kpn I transcripts are non-specific. Kole et al. (24) have clearly demonstrated the existence of several discrete length (i.e. specific) transcripts of Kpn I family members in cytoplasmic RNA. Assuming that in addition to these specific transcripts there is a large mass fraction of non-specific transcription products there is no contradiction between our results and those of Kole et al. In contrast, Shafit Zagardo et al. (27) report a strong asymmetry in the transcription of Kpn I sequences. To quantitatively determine the transcriptional asymmetry of the Kpn I family the following experiment was performed. Cytoplasmic and nuclear RNAs, immobilized on nitrocellulose, were annealed with M13 clones of the complementary strands of H Kpn-E13 (sequence depicted in Figure 5). The specific activity and amounts of the complementary probes were identical. Nuclear RNA (2 µg/filter) was symmetrically hybridized by the complementary subcloned strands (4730 and 4300 cpm above background, 30 cpm). Cytoplasmic RNA (10 µg/filter) also hybridized symmetrically (210 and 220 cpm above background, 30 cpm). The discrepancy between these findings and those of Shafit-Zagardo et al. (27) may be attributed to our use of M13 subclones of complementary strands and their technically much more difficult separation of strands by gel electrophoresis. It is also possible that the Kpn I clones which they employed represent different regions of the family than the 1.3 kb clone (H Kpn-E13) used here.

The filter hybridization experiment described above also indicates that the abundance of Kpn I transcripts is reduced 100 fold in cytoplasmic RNA relative to nuclear RNA. This is similar to the relative abundance of Alu transcripts in nuclear and cytoplasmic RNAs (1,17).

5) The Poly (CA) family: structure and transcriptional activity

During the course of this investigation two members of the poly (CA) family of sequences have been isolated: one from a randomly selected genomic clone; the second by use of the first clone as a hybridization probe. The base sequences of these two repeats are:

5' ATTCTAATAACACCT$(AC)_{17}$CACTTCTTTCCAGA 3' and

5' CAATATAAATACATG$(CA)_{26}$TTTAATTAACGGTAA 3'

These are members of the poly (CA) family of repeats which is present in all eukaryotic DNAs (28,29). In contradication to Roger's interpretation (45), scrutiny of known members of this family does not convince us that short direct repeats or any other noteworthy sequence features flank the poly

(CA) run. Examples cited by Rogers (45) as possible direct repeats (e.g. 24 bp in reference 28) are purine (or pyrimidine) rich regions which may be regarded as similar rather than repeated sequences.

Poly (CA) is not transcribed *in vitro* by RNA polymerase III (conditions of Figure 3, data not shown). It is marginally transcribed by isolated nuclei and the sensitivity of this transcription to α amanitin shows that it is directed by RNA polymerase II, Figure 4. As both strands, i.e. poly (TG) and poly (CA) are transcribed, these transcription products are non-specific "read throughs" of the poly (CA) regions (Figure 4). By Northern blot analysis discrete length cytoplasmic RNAs against a background smear, hybridize to either poly (CA) or poly (TG) (Figure 6). The nuclear RNAs contain a smear of RNA lengths which hybridize to poly (CA) and poly (TG) (Figure 6). Although the significance of these discrete length cytoplasmic RNAs has not been investigated, three points might be noticed: First, the lengths of these RNAs exceed known lengths of poly (CA) runs. Second, both poly (CA) and poly (TG) are found in these transcription products suggesting their transcription is non-specific. Third, poly (CA) and poly (TG) are present in heterogeneous length molecules in hn RNA. We suggest that poly (CA) and poly (TG) are adventitiously included in longer transcription units and that in the case of a few particular mRNAs they survive RNA processing.

6) K repeats and their relative transcriptional inactivity

More detailed restriction mapping of several of the parental clones which contained the non-Alu family repeats described above revealed the presence of additional repetitive sequences. One such repeat, which occurs 250 bp 5' to a poly (CA) stretch, has been studied in more detail. The base sequence of a region containing this new repeat (called the K element) is:

```
              ------->       <--------
5'      AGCTCATCTGTCCACTGAAGATGCTTGGACAGAGTTAGGAATGCTTCCTG

        GGAGAGGTAACATTTGAGACTTTCCTGGAAGAATGGTCAGAGTAAACCAA

        GTAAGTAGGAATGGAAAGAGGATGGGAGGCCCCAGCTTCCCAGAGGCATA
                                       ---------->
        AGGTGAGGANGNCCCTATGCATTCAGATGTGGCCCACCCTGGGGTCTGGT

        GGACTAAAGNCTTGGACACCCCAGATCAGCCTTAGTGGGATGAGGCAGGA
                                             <----------
        AAGACAGCTGAGGGTCAGAACCCAGGCAGGTCCAATGCCAGGGTGGGCAT

        TTCGAGTTGGTGAGACATTTCACCCTGGTGCCAAGCT  3'
```

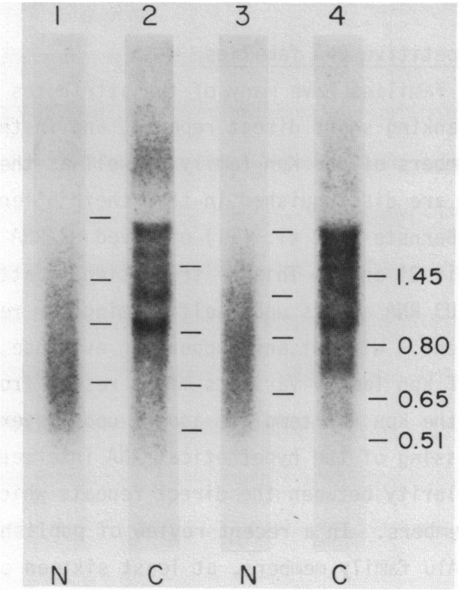


Figure 6. *In vivo* RNA transcripts of the poly (CA). 2-3 μg nuclear RNA and 10-15 μg cytoplasmic RNA were formamide denatured and separated by electrophoresis through agarose gels containing formaldehyde and transferred to nitrocellulose. Duplicate blots of nuclear and cytoplasmic RNA were hybridized to ^{32}P-labeled M13 clones containing opposite strands of each family of repeat. Size markers are as indicated; N indicates lanes containing nuclear RNA, C indicates lanes containing cytoplasmic RNA. Lane 1 and 2, poly (CA); lane 3+4, poly (GT).

Except for perfect eleven nucleotide and eight nucleotide inverted repeats this region does not exhibit any notable features. Restriction mapping suggests that the repeat is wholly contained in the above sequence which also includes flanking regions.

The relatively rare K repeats are transcriptionally inactive compared to the repeat families described above. Under the conditions of Figure 3 there is no discernable RNA polymerase III activity *in vitro*. Transcripts from isolated nuclei do not contain abundant K repeats (Figure 4). K transcripts are detectable as a smear of hn RNA by Northern blot analysis but only after relatively long (ca. 10 fold) exposures (data not shown) compared to the data of Figure 6. K is transcribed at a much lower level than the other repeat families studied here.

DISCUSSION

1) Structure of repetitive DNA families

The O and Kpn I families have many of the attributes of retrogenes; an A rich 3' end, flanking short direct repeats, and in the case of O, precise 5' ends. Members of the Kpn family as well as the equivalent Bam Hl family in rodents are distinguished in that their 5' end is variably truncated (20-23). Bernstein et al. (11) observed U3 RNA pseudogenes which are truncated on their 3' ends. This 3' truncation is attributed to the secondary structure U3 RNA adopts upon self priming for reverse transcription by its 3' end. Without any supporting evidence, we speculate that the 5' truncations of Kpn family variants might result from either the secondary structure the Kpn RNA template adopts upon reverse transcription or degradative processing of the hypothetical RNA intermediate.

There is a similarity between the direct repeats which flank Alu, Kpn I and O family members. In a recent review of published direct repeats which flank Alu family members, at least sixteen out of twenty pairs of direct repeats could be represented as beginning with (or perhaps being preceded) by one or more A residues (48). The remaining four pairs of direct repeats may or may not share this feature, depending upon sequence interpretations. The 5' situated direct repeats observed in this work [aagatccctaatgg for H Kpn 10, agtaac for O_5 and atgacgaatga for O_4] also obey this rule (Figures 1 and 5). Because all of these families end in an A rich 3' end there is some ambiguity as to whether these 5'-ward A residues should be regarded as part of the direct repeat or merely A residues which precede the direct repeat. Regardless of which interpretation is correct, the implication is that insertion of these repetitive elements is not completely random but has a sequence preference.

The mechanism by which poly (CA) is dispersed is presumably different than that described for retrogenes. Comparison of regions flanking duplicate alpha-like globin genes suggest that such sequences are generated *in situ*, perhaps by template slippage during DNA replication (30). Poly (CA) is not the only example of dispersed tandem repeats. The intervening sequences of the two embryonic alpha-like globin genes are occupied by tandemly repeated oligonucleotides (31). Blot hybridization suggests that this tandem repeat occurs elsewhere in the genome (unpublished). A number of other examples of tandemly repeated sequences occur upstream from known genes (32,33) and repetitive polypyrimidine tracts are also found in all eukaryotic DNAs (34). Apparently there are at least two broad classes of interspersed repeats in human DNA: retrogenes and dispersed tandem repeats.

2) Transcriptional activity of repetitive DNA families

The high repetition frequency of Alu family members has been explained by the presence of an internal RNA polymerase III promoter within the dispersed members of the family (4). Any one of the family is presumed competent to direct the dispersion of succeeding generations of the family. The promoter type cannot be decisive in determining the numerical success of a repeat sequence family. The abundant Kpn I family members appear in this and other studies to be transcribed by RNA polymerase II (24,26,44). Kole et al. (24) observe discrete length Kpn transcripts which implies that one or more Kpn master sequences has retained its transcriptional competence. Presumably the truncated Kpn retrogenes abandon their extragenic RNA polymerase II promoter when they disperse to new chromosomal sites. The equally abundant poly (CA) family is probably generated by a transcription independent mechanism. In this work the one tentative example of a new family of repeats which is transcribed by RNA polymerase III is a very low copy number family.

The relative abundances with which the K, O, Kpn I and Alu families are expressed in hn RNA reflect their repetition frequencies. Is the level of expression in hn RNA attributable to their repetition frequencies or is their repetition frequency attributable to the level of expression in hn RNA? As we do not believe that an hn RNA precursor can account for the well-known structures of translocated repeats we conclude that the relative abundance of these repeats in hn RNA non-specifically mirrors their genomic abundance. The repeated sequences Kpn I, O and K are less abundantly represented in cytoplasmic RNA than in nuclear RNA. Presumably, non-specific transcripts of these repeats are removed during the processing of hn RNA into mature mRNA as also seems to be the case for Alu family transcripts in hn RNA (1, 17).

3) Genomic abundance of interspersed repeats

Together, the Kpn I and Alu families account for a major fraction of the human genome ($\gtrsim$ 10%). The best estimate is that at least 20% of the human genome consists of repetitive DNA (1,35). Included in this fraction is an unknown amount of non-interspersed clustered satellite-like repeats which might constitute $\approx$ 5% of the human genome. It therefore seems likely that the Kpn I and Alu families of repeats account for the majority of the interspersed repeats in the human genome by both mass and number.

The method by which the original ten randomly selected human clones were surveyed should be especially sensitive to families of interspersed repeats which are as abundant as the Alu, Kpn I and poly (CA) families of repeats. The substantial literature on interspersed repeats in human DNA does not identify any additional repeat families which are as hyperabundant as the Alu family. (To date there are 30 sequenced Alu family members.) Other hyperabundant families certainly do not exist. It is significant that the next most numerically abundant families detected here (Kpn I and poly (CA) are both already well documented in the literature. It is unlikely that there are many other undiscovered 50,000 fold repeat families in the human genome. Less is known about the lower abundance classes of repeats. The O and K families might be relatively abundant compared to the majority of the remaining interspersed repeat families. Cot analysis, while rather inexact for this purpose, suggests that much of the remaining repetitive human DNA would be at least 500 fold repetitive (35) so that the O and K families are probably not atypical of these remaining repeats. From the considerations given above a very rough estimate is that 10% of the human genome or 2.5×10^8 bp (2.5×10^9 bp x 0.1) would fall into this class of unassigned interspersed repeats. Taking O and K as typical, these unidentified families might include on the order of $\sim 10^3$ members. Each sequence for the purpose of round numbers might have a length of ∿ 250 bp (Table I). In agreement with these approximations known families of genes and pseudogenes for small RNAs in human and in rodent typically seem to contain ∿ 200 to 2,000 members (3 and references therein). According to these values, there could be about 1000 $(2.5 \times 10^8)/(250 \times 1000)$ families of relatively low abundance interspersed repeats in the human genome. Although these estimates are approximate, they are sufficiently accurate to show that human DNA, like DNA from lower eukaryotes, (Introduction) might then contain a rather large number of different families of short interspersed repeats (2). It is significant that many of the subcloned repeats studied here revealed the presence of additional repeat elements (e.g. the Pol III transcription unit in O_4 and K in a poly (CA) subclone and others not reported here) upon analysis by more sensitive procedures. The human genome may be a tangle of repetitive elements. However unlike lower eukaryotes, the interspersed repeats in human DNA are dominated in number by a single sequence family, Alu, and in mass by two sequence families, Alu and Kpn I.

ACKNOWLEDGEMENTS

The authors thank Drs. P. Yau and M. Bradbury for their gifts of Hela cells and Mr. C. Perez-Stable and Dr. J. Shen for their gift of the polymerase III extract. We also thank Dr. D. Lowensteiner for constructing an O clone and Mr. Marc Beal for assisting with the base sequence determination of that clone.

This work was supported by USPHS grants GM 21346 (CS) and GM 29090 (LL) and by a University of California Chancellor's Patent Fund Award to Mr. Paulson.

*To whom correspondence should be communicated

REFERENCES

1. Schmid, C.W. and Jelinek, W.R. (1982) Science 216, 1065-1070.
2. Davidson, E.H., Galau, G.A., Angerer, R.C. and Britten, R.J. (1975) Quart. Rev. Biol. 106, 773-790.
3. Van Ardsell, S.W., Denison, R.A., Bernstein, L.B. and Weiner, A.M. (1981) Cell 26, 11-17.
4. Jagadeeswaran, P., Forget, B.G. and Weissman, S.M. (1981) Cell 26, 141-142.
5. Grimaldi, G. and Singer, M.F. (1982) Proc. Natl. Acad. Sci. USA 79, 1497-1500.
6. Kominami, R., Muramatsu, M. and Moriwaki, K. (1983) Nature 301, 87-89.
7. Lemishka, I.R. and Sharp, P.A. (1982) Nature 300, 330-335.
8. Hess, J.F., Fox, G.M., Schmid, C. and Shen, C.-K.J. (1983) Proc. Natl. Acad. Sci. in press.
9. Ullu, E., Murphy, S., Melli, M. (1982) Cell 29, 195-202.
10. Sharp, P.A. (1983) Nature 301, 471-472.
11. Bernstein, L.B., Mount, S.M. and Weiner, A.M. (1983) Cell 32, 461-472.
12. Schmid, C.W., Fox, G.M., Dowds, B., Lowensteiner, D., Paulson, K.E., Shen, C.-K.J. and Leinwand, L. (1983) in Perspectives on Genes and the Molecular Biology of Cancer, D.L. Robberson, and G.F. Saunders, Eds.
13. Rinehart, F.P., Ritch, T.G., Deininger, P.L. and Schmid, C.W. (1981) Biochemistry 20, 3003-3010.
14. Houck, C.M., Rinehart, F.P. and Schmid, C.W. (1979) J. Mol. Biol. 132, 289-306.
15. Bonner, T.I., Brenner, D.J. Neufeld, B.R. and Britten, R.J. (1973) J. Mol. Biol. 81, 123-135.
16. Deininger, P.L., Jolly, D.J., Rubin, C.M., Friedmann, T. and Schmid, C.W. (1981) J. Mol. Biol. 151, 17-33.
17. Jelinek, W.R. and Schmid, C.W. (1982) Ann. Rev. Biochem. 51, 813-844.
18. Weiner, A.M. (1980) Cell 22, 209-218.
19. Shen, C.-K.J. and Maniatis, T. (1980) Cell 19, 379-391.
20. Singer, M.F., Thayer, R.E., Grimaldi, G., Lerman, M.I. and Fanning, T.G. (1983) Nucleic Acids Res. 11, 5739-5745.
21. Wilson, R. and Storb, U. (1983) Nucleic Acids Res. 6, 1803-1819.
22. Fanning, T.G. (1983) Nucleic Acids Res. 11, 5073-5091.
23. Gebhard, W., Meitinger, T., Höchtl, J. and Zachau, H.G. (1982) J. Mol. Biol. 157, 453-471.

24. Kole, L.B., Haynes, S.R. and Jelinek, W.R. (1983) J. Mol. Biol. 165, 257-286.
25. Lerman, M.I., Thayer, R.E. and Singer, M.F. (1983) Proc. Natl. Acad. Sci. 80, 3966-3970.
26. Shafit-Zagardo,B.,Brown, F.L., Zavodny, P.J. and Maio, J.J. (1983) Nature 304, 277-280.
27. Shafit-Zagardo, B., Brown, F.L., Maio, J.J. and Adams, J.N. (1980) Gene 20, 397-407.
28. Miesfeld, R., Krystal, M. and Arnheim, N. (1981) Nucleic Acids Res. 9, 5931-5447.
29. Hamada, H., Petrino, M.G. and Kakungat, . (1982) Proc. Natl. Acad. Sci. 79, 6465-6469.
30. Sawada, I., Beal, M.P., Shen, C.-K.J., Chapman, B., Wilson, A.C. and Schmid, C.W. (1983) Nucleic Acids Res. 11, 8087-8101.
31. Proudfoot, N.J., Gill, A. and Maniatis, T. (1982) Cell 31, 553-563.
32. Maroteaux, L., Helig, R., Dupret, D. and Mandel, J.L. (1983) Nucleic Acids Res. 11, 1227-1245.
33. Kominami, R., Urano, Y., Mishima, Y., Muramatsu, M., Moriwaki, K. and Yoshikura, H. (1983) J. Mol. Biol. 165, 209-228.
34. Birnboin, H.C., Sederoff, R.R. and Paterson, M.C. (1979) Eur. J. Biochem. 98, 301-307.
35. Schmid, C.W. and Jelinek, W.R. (1982) Science 216, 1065-1070.
36. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1983) Molecular Cloning A Laboratory Manual, Cold Spring Harbor Lab. N.Y.
37. Messing, J., Crea, R. and Beeburg, P.H. (1981) Nucleic Acids Res. 9, 309-323.
38. Maxam, A. and Gilbert, W. (1980) Methods in Enzymology 65, 499-560.
39. Benz, E., Wydro, R.M., Nadel-Ginard, B. and Dina, D. (1980) Nature 288, 665-669.
40. Wu, G.-J. (1980) J. Biol. Chem. 255, 251-258.
41. Smith, G.E. and Summers, M.D. (1980) Ann. Biochem. 109, 123-129.
42. Lehrach, H., Diamond, D., Wozney, J.M. and Boedtker, H. (1977) Biochemistry 6, 4743-4759.
43. Kafatos, F.C., Jones, C.W. and Efstratiadis, A. (1979) Nucleic Acids Res. 7, 1541-1552.
44. DiGiovanni, L., Haynes, S.R., Misra, R. and Jelinek, W.R. (1983) Proc. Natl. Acad. Sci. in press.
45. Rogers, J. (1983) Nature 305, 101-102.
46. Adams, J.W., Kaufman, R.E., Kretschner, P.J., Harrison, M. and Nienhuis, A.W. (1980) Nucleic Acids Res. 9, 6113-6128.
47. Fuhrman, S.A., Deininger, P.L., LaPorte, P., Friedmann, T. and Geiduschek, E.P. (1981) Nucleic Acids Res. 9, 6439-6456.
48. Thayer, R.E. and Singer M.F. (1983) Mol. Cell Biol. 3, 967.
49. Schmid, C.W. and Shen, C.-K. J. (1983) in Molecular Evolutionary Genetics, R.J. MacIntyre, Eds., Plenum Publishing Co., NYC, in preparation.