Structural analysis of interspersed repetitive polymerase III transcription units in human DNA

J.Pan, J.T.Elder, C.H.Duncan and S.M.Weissman

Departments of Human Genetics and Molecular Biophysics and Biochemistry, Yale University School of Medicine, New Haven, CT 06510, USA

Received 5 November 1980

ABSTRACT

The nucleotide sequences of two cloned fragments of human DNA which function as templates for RNA polymerase III *in vitro* confirm their identities as members of the *Alu* family of human interspersed repetitive DNA sequences (1,2). The interspersed and repetitive nature of these sequences in the genome was demonstrated by hybridization of nick-translated DNA from one of these clones to total genomic DNA and to DNA of individual random clones from a λ Ch4A-based human genomic library.

Short, direct terminal repeats of non-conserved sequence flank the 300-nucleotide *Alu* family conserved sequence. Within the *Alu* family sequence is found a 40-nucleotide region which is directly repeated 135 nucleotides downstream. This 40 nucleotide sequence is found once in the murine B1 interspersed repetitive sequence family (8). This and other evidence indicates that the human *Alu* family resembles a partial duplication of the murine B1 sequence.

INTRODUCTION

In the course of screening libraries of cloned human genomic DNA for sequences complementary to low molecular weight cytoplasmic RNA, we noted that a large fraction of genomic clones hybridized weakly to low molecular weight cytoplasmic non-polyadenylated RNA from KB cells, a human tumor-derived cell line. Most of these clones were not complementary to known low molecular weight RNA such as transfer RNA and 5S RNA. To further analyze the sequences homologous to cytoplasmic RNA from HeLa cells, we selected ten representative clones from a plasmid-based, partial human genomic library that showed reproducible hybridization with radioactive RNA.

DNA from eight of these clones was found to direct transcription of discrete low molecular weight RNA *in vitro* in a polymerase III transcription system. One of these clones was chosen for further investigation, including DNA sequence analysis and comparison of nucleic acid homology with other clones of genomic DNA.

While this work was underway, it was found that the human non-α globin gene cluster contained several DNA sequences that either served as polymerase III transcription units *in vitro* or showed sequence homology with other DNA segments that did promote transcription (3). DNA sequence comparisons showed that the globin genomic segments were homologous to the interspersed repetitive family of DNA sequences recently characterized by Schmid and co-workers (1, 2), now known as the "*Alu* family" of interspersed repetitive sequences. Our results confirm that the polymerase III templates in the DNA that we have studied are homologous to those of the globin clones and are representative of the highly reiterated "*Alu* family" DNA sequences. The frequency of detection of these DNA sequences in genomic DNA libraries and the spacing between them in the human non-α globin gene cluster indicate that they are interspersed very widely in human genomic DNA. The results from the non-α globin gene cluster (3, 4) confirm the deductions by hybridization experiments with total human DNA that such interspersed DNA sequences may be directly linked to unique DNA sequence in the genome (1,5, 6, 7).

We present the DNA sequence of two of these interspersed repetitive polymerase III templates and compare them to the previously sequenced (2) BLUR 8 *Alu* family clone obtained as a cloned fragment from rapidly reannealed S1-treated human DNA. One of these templates, on the plasmid pJP53, was originally selected by hybridization to KB cell low molecular weight RNA (LMW-RNA). The other template on plasmid pA36γ (3), maps about 2 kb upstream from the $^{G}\gamma$ globin gene. We note several intriguing structural features of these sequences and compare them as a group to the sequences of the members of the murine B1 family of interspersed repetitive DNA sequences recently published by Georgiev and co-workers (8). Certain features of the sequence of the $^{G}\gamma$ globin RNA Polymerase III transcription template were previously noted (9), and the complete DNA sequence of this template and surrounding regions has been submitted for publication elsewhere (25).

MATERIALS AND METHODS

Tissue Culture

DNA Preparation.

Plasmid DNA was purified by the cleared lysate procedure (10), followed by equilibrium density gradient centrifugation in ethidium bromide/cesium chloride (11). Bacteriophage λ DNA was prepared essentially according to Maniatis *et al.* (12). High molecular weight human placental DNA, prepared by the method of Blin and Stafford (13), was provided by Dr. B. G. Forget.

RNA Preparation.

KB cell RNA was labeled with ^{32}P by growth for 24 hours in medium containing phosphate at 2% of its normal concentration using 10 mCi ^{32}P-orthophosphate per 10^8 cells. Isolation of labeled RNA was performed as described in the following paper.

Nucleic Acid Transfer and Hybridization Procedures.

Complete restriction enzyme digests of human DNA were subjected to agarose gel electrophoresis and transferred either to nitrocellulose by the method of Southern (14) or to diazobenzyloxymethyl-paper according to Alwine *et al.* (13). Hybridizations and washes of DNA transfers onto DBM-paper were performed according to Alwine *et al.* (15). Hybridization of DNA transfers to nitrocellulose were performed according to Tuan *et al.* (16). Hybridization probes were prepared by nick translation (17) of isolated DNA fragments in the presence of α-^{32}P-labeled deoxyribonucleoside triphosphates (New England Nuclear, Amersham).

DNA Sequencing Procedures.

DNA sequencing was performed by the Maxam-Gilbert chemical degradation protocol (18), by the Maat-Smith dideoxynucleotide extension method (19) and by partial snake venom phosphodiesterase digestion followed by two-dimensional chromatography (15).

Enzymes.

*Eco*RI (20), *Bam* HI (21), and *Bgl* II (22) were purified by established procedures. All other restriction endonucleases were obtained from New England Biolabs, Miles Laboratories or Bethesda Research Laboratories and were used in the buffers recommended by the supplier. DNA polymerase I was from New England Biolabs or Boehringer Mannheim Biochemicals.

Containment.

All recombinant DNA procedures were performed in accordance with the current NIH Guidelines for Recombinant DNA Research.

RESULTS

Isolation and transcription of genomic clones homologous to low-molecular-weight cytoplasmic RNA.

A partial genomic library of human DNA was constructed in the bacterial plasmid pBR322 (23) by Drs. P. A. Biro and P. V. Choudary and kindly provided to us for these studies. The human DNA inserts in this library are 10 to 12 kb long and are bounded by *Bam* H1 and *Eco*RI restriction enzyme sites.

To screen these clones, nonpolyadenylated cytoplasmic RNA was prepared from KB cells grown in the presence of ^{32}P-orthophosphate for 24 hours. After electrophoresis through a 6% polyacrylamide gel, the 5S RNA was excised and discarded and the remainder of the ^{32}P-labeled RNAs of chain length less than 700 were combined, eluted from the polyacrylamide matrix, and used as radioactive probe against the genomic DNA clones. A large proportion of clones reacted weakly with this radioactive RNA probe. Eleven colonies that had reacted in the initial screen were recovered and tested again with the probe. Ten of the colonies proved positive. Of these clones, one was probably a tRNA gene, since it bound a cellular RNA, approximately 75 bases long, that contained pseudouridine (data not shown). This clone was not further analyzed.

Eight of the remaining nine DNA preparations were active as templates in Wu's *in vitro* RNA polymerase III transcription system (24), as described in the following paper. Radioactive RNA was prepared from one clone, pJP53, by large-scale synthesis in the presence of 250 uCi of $[\alpha\text{-}^{32}P]$ GTP. The resultant RNA was purified by gel electrophoresis and a portion was then used as a radioactive probe against Southern blots of restriction digests of DNA from each of the eight plasmids. These experiments demonstrated homology between this RNA transcript and a segment of DNA in each of the isolated plasmids (data not shown). To check this, a second plasmid (JP72) was used as template for RNA transcription. This transcript was isolated in similar fashion and hybridized to Southern blots of the same restriction enzyme digests of each of the other DNA clones.

The JP 53 and JP 72 transcripts hybridized to the same restriction fragments in each of the 8 clones studied, suggesting that the templates for polymerase III transcription in each clone are homologous to each other (data not shown).

Certain properties of the *in vitro* transcript produced by RNA polymerase III from the pA36 gamma plasmid template have been discussed (3, 25).

DNA Sequence of pJP53 and pA36γ Polymerase III Templates.

A restriction map of the pJP53 insert was generated and is shown in Fig. 1. The RNA polymerase III template of pJP53 was localized to within specific *Hpa*II and *Hinf*I fragments by Southern blot analysis, using the pJP53 *in vitro* transcript as probe (Fig. 1). The DNA sequence of the *Hpa*II fragment was subsequently determined by the Maxam-Gilbert protocol and confirmed by the Maat-Smith protocol and is shown in Fig. 2. The localization of the pA36γ transcription template was previously described (3). The DNA sequence of the *Alu* family segment of the A36γ template was determined by similar methods.

Representation in Genomic DNA

When *Eco* RI digests of genomic DNA were electrophoresed in 1% agarose gels, transferred to nitrocellulose paper and hybridized with the RNA polymerase III transcript template fragment generated by *Eco*RI/*Bam*H1/*Bgl*II triple digestion

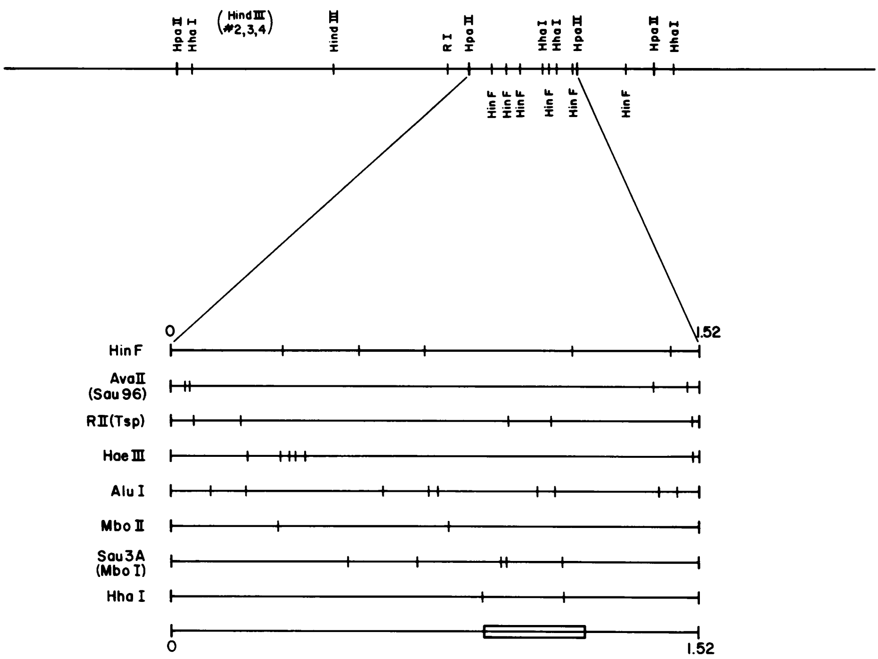

Fig. 1. Restriction endonuclease map of the DNA insert in clone JP53.

The upper horizontal line represents a segment of human genomic DNA inserted in a plasmid. The lower portion of the figure is a detailed restriction map of the segment sequenced and shown in Fig. 3. Vertical dash marks represent the location of restriction endonuclease cleavage sites. The open rectangle in the bottom line represents the location of the template for the polymerase III transcript described in the text.

of the plasmid pA36γ (2), a highly heterodispersed and strong pattern of hybridization is obtained (Fig. 3A, lane 3).

In contrast, probes derived from regions of the human non-α gene complex known not to contain highly repetitive sequences (i.e., pβPst or pδ Pst [4]) hybridize only to the expected discrete bands on parallel strips of nitrocellulose from the Southern blot of the same gel (Fig. 3A, lanes 1 and 2).

```
         0        20        30        40        50        60        70        80
ACCGGCCGGTGTTGTCTGCCATCTGCAGACCAGCCCTGCATAGGCTCAGGACCAATGACTGTGGACCTGGGTGTGCATAT

        90       100       110       120       130       140       150       160
GTAGTCCCTGCCACTGTGGTGAATTGCAAATCAGAGTTTGCAGCTACAGTTGTGTGTTTAGGCTTTGATGCAGGCTGATA

       170       180       190       200       210       220       230       240
CCTCATAATCACTGAGTTGTTGTTTTCCCAGTTGTACTATCTTGTGCCTGGACAGTAGCTGTTCTTGGCCTTTTTTCTTT

       250       260       270       280       290       300       310       320
GTGCCTCCTGCTCAGTTACCCCATTAGAGACTTCGGAGACTGACCCTGAATGACTAACTATTGTCTCCAAGAAGAACTGG

       330       340       350       360       370       380       390       400
AGGCCAATCCATGACTCTCCGTGGCCATTTTTCTTAAGACAGAGGCCTGCTTCAATTCTTGACGTATTTAGGGCCCCTGA

       410       420       430       440       450       460       470       480
ATTAAAAACTTGTGCTCTTACCTGATGTCAAGAAGCACAAAACTCAGATTGCCTCATCCTTTGGACAAGACCTCTTGGAC

       490       500       510       520       530       540       550       560
TTTGATGGTGTCTCAGGTACCCTCAACTTTGCTGATCTGGTCAGTTTTCCGTGGTCCCCACACTAAGAGTCATTCTAACT

       570       580       590       600       610       620       630       640
TGATTGCATCATGCAATTATTAGGCTCTTTATGATATCTGACTTCGTTTTTAAAGTAGCTTAAAATTTTTTACCAAAGTA

       650       660       670       680       690       700       710       720
AATTTTTATCAAAATCAAACAATTAAAGTTAAAAGAAAAATAAAACAAAAACCAAACACAAAATAGCAGTTTCCTGATCC

       730       740       750       760       770       780       790       800
ACTTCTCCCTAACTCTATTGACTCAAATGCTAGCTCTTGGTTTATAAAATTTTATAGCTTTTTTGTTTTTTGTTGGTTTT

       810       820       830       840       850       860       870       880
ATTATGAAGATGAGGATTTAGCATACTTACATAATTCAACATCCTTGCTCCCCCTCCTGTTTTCCAAGTAAAATTATATA

       890       900       910       920       930       940       950       960
AAAATGTTTAGATAAGGGCTAGGCGCGGGTTCACGCCTGTAATCCCAGCATTTTGGGAGGCTGAGACGGGTGGATCATGA

       970       980       990      1000      1010      1020      1030      1040
GGTCAGGAGATCGAGACCATCCTGGCTAACATGGTGAAACCCCGTCTCTACTAAAAATACAAACAACCAGCCAGCCGAGT

      1050      1060      1070      1080      1090      1100      1110      1120
GTGGTGGTGGGCACCTGTAGTCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATGACTTGAACCTGGAGGTGGAGCTTGCA

      1130      1140      1150      1160      1170      1180      1190      1200
GTGAGCCAACGATCGCGCCACTGTCATCATCATGGGTGACAGAGAGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAA

      1210      1220      1230      1240      1250      1260      1270      1280
AAAGTTTAGATAAAACAATGTAAAATGTTTTCCTAATTTTCCACCTAGCATTTTGTGCTTACATTTCCTTTCTTGTTCAG

      1290      1300      1310      1320      1330      1340      1350      1360
AATGTTTTGCTTTCTAGAGTTAATACTTAACTCATTTTTTTCCTTACTTGATTGGTTGTCTGTGTTCCTATCACTATGTT

      1370      1380      1390      1400      1410      1420      1430      1440
CAAACTCCACAACAACAATGACTGTTATTTCTCGGACCAAAGCAAGAAGCATCAGCTTTTCATTATTCTTGGAAACACTC

      1450      1460      1470      1480      1490      1500      1510      1520
CTTCGAGAGTCCTCTCTCCTATTGGAAGCTGTGCTCTGGGTGTTCTCTACAGACTGGTCCCCTGGCCCTTCTCTTTACCTGTCTA
```

Fig. 2. Nucleotide sequence of the DNA segment of clone pJP53 containing the template for polymerase III transcription.

The sequence reads from the 5' to the 3' direction from left to right on the restriction map shown in Fig. 1. The section of the sequence corresponding to the highly reiterated interspersed Alu family DNA segment is shown in larger letters and extends from residue 897 through residue 1203, including the run of 25 adenylic acids at the 3' end of the Alu template.

To quantitate the frequency of occurrence of these sequences in the genome, DNA was prepared from eight randomly selected clones from a human gene library constructed in the bacteriophage λ Charon 4A (26) and kindly provided for these studies by Drs. P. A. Biro and P. V. Choudary. Each cloned DNA was digested with EcoRI, electrophoresed on 0.9% agarose gels, blotted onto DBM-paper, and hybridized with the 225 bp HaeIII "G" fragment of pA36γ labeled

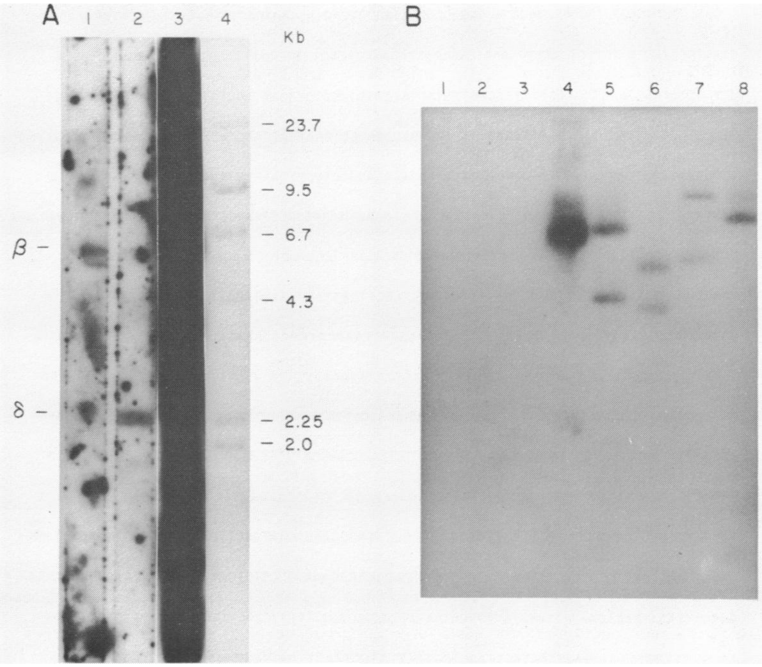

Fig. 3. Genomic representation of Alu family sequences.

A. Genomic Southern blotting

Human placental DNA, prepared by the method of Blin and Stafford (13) was digested with an excess of EcoRI enzyme, subjected to electrophoresis in an 0.8% agarose gel, and transferred to nitrocellulose as described (16). Adjacent strips were hybridized for 36 hr in 3xSSC at 65°C against (lane 1) β Pst, containing the human β globin gene (4); (lane 2) δ Pst, which wholly contains the human δ globin gene (4); (lane 3) the BamH1/BglII/EcoRI triple digest fragment of pA36 γ which contains the in vitro RNA polymerase III transcript (2). Lane 4 contains ^{32}P end-labeled HindIII fragments of λ DNA as size markers. Probe specific activities were between 1 and $1.5x10^7$ cpm, and hybridizations contained $2x10^5$ cpm/ml of hybridization solution. The final hybridization wash was in 0.2xSSC at 65°C for 30 min.

B. Analysis of individual clones.

Eight randomly chosen clones from a human gene library constructed from partial EcoRI digests of human DNA were digested to completion with EcoRI, subjected to electrophoresis in 0.8% agarose, and transferred to DBM-paper as described (18). The paper was hybridized in 10 ml 50% formamide hybridization solution containing 10% dextran sulfate (18), for 24 hr at 42°C. The probe specific activity was $4x10^7$ cpm/ug and the probe concentration was $2x10^4$ cpm/ml hybridization solution.

with ^{32}P by nick-translation (17). This probe maps entirely within the A36γ RNA polymerase III transcription template as shown in Ref. 3. The results are shown in Fig. 3B. Of the 8 clones studied, 4 clones gave more than one positive band, 1 gave a single positive band, and three gave no visible positive bands. Assuming that no EcoRI sites fall inside the repetitive sequence family, and that no two sequences lie in any one fragment, there are 9 positive bands in about (8x11 kb) = 88 kb of genomic insert DNA. Therefore, this crude method yields an average occurrence of interspersed repetitive sequence family approximately once every 10 kb in genomic DNA.

DISCUSSION

Genomic representation of in vitro RNA polymerase III template sequences.

Previous studies of human DNA (1, 6, 7) have demonstrated a DNA sequence, about 300 nucleotides in length, which is interspersed throughout the genome with longer tracts of single-copy DNA sequences. The exact distance between any two of these sequences is quite variable, however, and neighboring sequences may be in an inverted orientation with respect to each other.

The inverted repeated sequences reanneal with zero-order kinetics, while the remainder reanneal at intermediate Cot values (1) and are referred to as interspersed repetitive DNA sequences. Many of these "snapback" and interspersed repetitive DNA sequences, when isolated from reannealed total human DNA and trimmed with S1 nuclease, contain a discrete site for the Alu I endonuclease. Therefore, they have been named the "Alu family" of interspersed repetitive DNA sequences. Thermal stability analysis (27) and DNA sequencing (28) indicate that individual members of this sequence family in the human differ from each other by an average of 12 to 15 percent.

Recently, two genomic DNA clones have been described (3) which are derived from the human non-α-globin gene complex and which serve as templates in Wu's soluble RNA polymerase

III system (24). The DNA sequence presented here (A36γ, Fig. 4) confirms that one of these transcription templates found approximately 2 kb upstream from the human $^{G}\gamma$ globin gene is a member of the *Alu* family of interspersed repetitive DNA sequences. Fritsch, Lawn and Maniatis (4) have identified a nonglobin repeat sequence appearing 7 times in a span of 65 kilobases within the human β-like globin gene cluster, all of which can be demonstrated to be representatives of the *Alu* family (4, and our unpublished results).

Furthermore, of nine clones isolated from a partial human genomic library by hybridization to *in vivo* ^{32}P-labeled, non-5S LMW-RNA, eight proved to be active *in vitro* templates for RNA polymerase III. α-^{32}P-GTP-labeled transcripts from two different clones hybidized to the same restriction fragments of all 8 active templates, confirming that all the template sequences shared nucleic acid homology. Comparison of the DNA sequence derived from the pJP53 transcription template with those of the BLUR clones (2) confirmed its membership in the *Alu* family.

A nick-translated probe prepared from a restriction fragment mapping within the pA36 γ transcription template detects a highly heterodisperse distribution of *Eco*RI restriction fragments by genomic Southern blot analysis (Fig. 5A, lane 3). DBM-paper transfer and hybridization analysis (15) of *Eco*RI-restricted individual genomic clones yields an average distance between human *Alu* family sequences of about 10,000 base pairs. Assuming a haploid genome size of 3x10^{9} kb, there would be 3x10^{5} copies of the sequence per haploid genome, in agreement with the value estimated by Houck, Rinehart, and Schmid (1).

We conclude that the *in vitro* RNA polymerase III templates studied here all contain members of the *Alu* family of interspersed repetitive DNA sequences, as judged by hybridization behavior of the templates and by DNA sequence analysis. Furthermore, a high proportion of all *Alu* family sequences are capable of carrying out this template function. The transcriptional properties of these sequences are considered in greater detail in the following paper.

Structural features of cloned interspersed repetitive DNA sequences: Length and boundaries of the sequence

The DNA sequences to be compared in this analysis are shown in Fig. 4. We note that the transcription templates of pA36γ and pJP53 each contain a region, roughly 300 nucleotides in length, over which they share 79% homology (44 substitutions, 9 insertions and 10 deletions in 281 positions, % homology = [1-([insertions + deletions + substitutions]÷ total number of positions)]x100.) The JP53 sequence, in turn, demonstrates 78% homology (40 substitutions, 10 insertions, and 10 deletions in 274 positions) with the BLUR clone 8 of Rubin *et al.* (2).

The 5' end of the conserved DNA sequence maps to the tetranucleotide GGCT starting at positions 897 in pJP53 (Fig. 3A) and at position 1 of the A36γ and JP53 sequences as shown in Fig. 4. The published sequence of BLUR clone 8 (2) does not extend far enough back from the canonical *Alu* I site to include this tetranucleotide.

The 3' end of the conserved region is located at a tract of deoxyadenylic acid residues (the oligo(dA) tract) starting at positions 283 in JP53, 282 in A36γ and 282 in BLUR 8, as numbered in Fig. 4. The length of this tract is 25 nucleotides in JP 53, 12 nucleotides in A36γ and 8 nucleotides in BLUR 8. This oligo(dA) tract is a conserved element of all human *Alu* families studied to date (2, 28), as is the tetranucleotide TCTC immediately preceding the tract. Upstream and downstream (in the transcriptional sense) from the conserved sequence, homologies are no greater than expected by random chance alone. Thus, the boundaries of the *Alu* family conserved region are quite distinct. Furthermore, the total length of the *Alu* family sequence excluding the oligo(dA) tract is well-conserved, with insertions balancing out deletions (Fig. 4).

The *Alu* family sequence seems unlikely to code for peptides of any length in the absence of extensive RNA splicing, as terminators occur in every reading frame in A36γ, JP53 and BLUR clone 8.

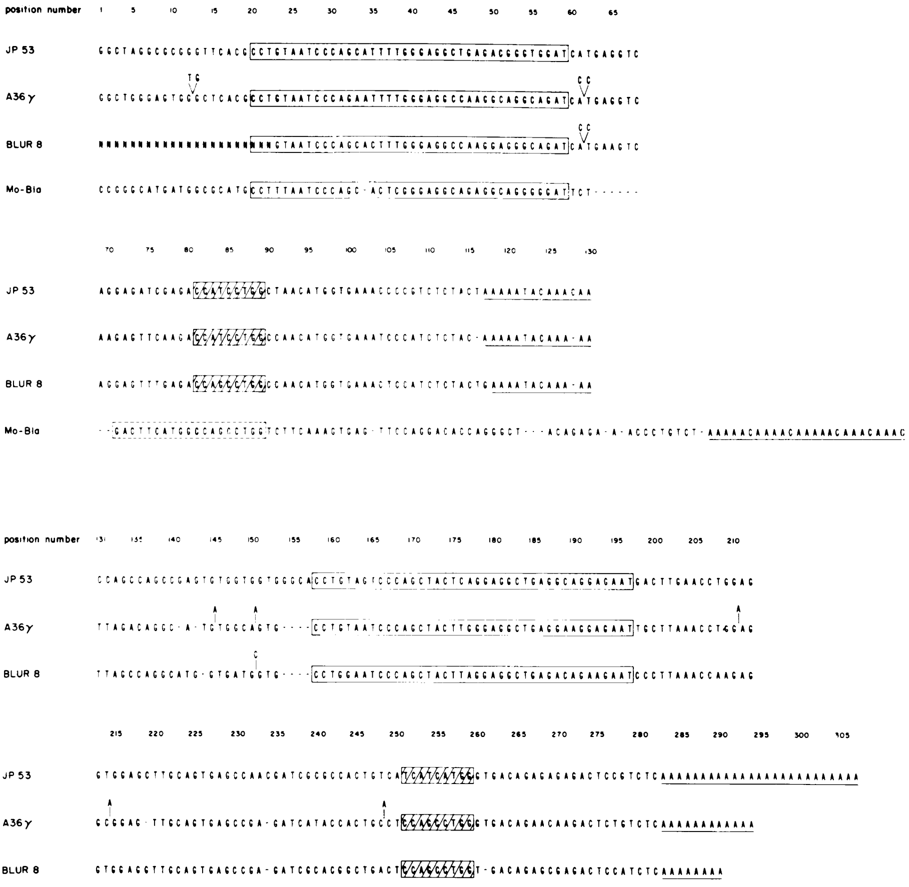

Fig. 4. Comparison of RNA Polymerase III templates, Alu family DNA Sequence and Murine Bla Clone.

The strands synonymous to the pJP53 and pA36γ in vitro transcripts are shown. With reference to the JP53 sequence, insertions are indicated above the line, and deletions are shown as dashes (-). Locations of insertions and deletions were determined by inspection to achieve maximum homology. N denotes undetermined nucleotide at this position. The sequence information has been partitioned in unequal parts to emphasize the dimeric nature of the human sequences.

Open boxes enclose the 40 nt conserved direct repeat found twice in human and once in murine Alu family sequences.

Cross-hatched boxes surround a sequence of nine bases displaying a less perfect identity with a nonanucleotide found in the second half of the sequence.

Dashed box indicates region of greatest homology between mouse and human other than the 40 nt conserved repeat.

Underlined sequences denote highly A-rich regions.

Short, direct repeats flank the Alu family sequence.

Recently, Bell, Pictet and Rutter have observed a direct repeat of 19 base pairs flanking an Alu family member located about 6 kb from the 3' end of the human insulin gene (28). We observe a perfect direct repeat of 10 nucleotides flanking the pJP53 sequence, as shown in Fig. 5. The 5' end of the conserved Alu family sequence, defined by the tetranucleotide GGCT, is separated from the repeat sequence by a single G residue, while the 3' copy of the repeat is directly contiguous with the oligo (dA) sequence. In the pA36 Alu family sequence, a 17-nucleotide direct repeat with one mismatch is separated from the 5' end of the conserved Alu family sequence by a single A residue but follows the 3' end separated from the oligo(dA) tail by 11 nucleotides (Fig. 5). The direct repeat sequences are not themselves homologous, suggesting that they arose from a duplication of sequences flanking the Alu family segment (see below).

The internal structure of the transcription template: a 40 nucleotide sequence forms the basis of a partially dimeric structure.

Analysis of the pJP53, pA36γ, and BLUR clone 8 Alu family sequences revealed a sequence of 40 nucleotides which is directly repeated about 135 nucleotides downstream, in the transcriptional sense. In Fig. 4, these sequences have been boxed.

Transcriptionally distal to each copy of the 40 nucleotide repeated sequence are sequences which are not repeated within the Alu family conserved sequence, except for: 1) a sequence of 9 nucleotides beginning at position +85 in pA36γ which is repeated imperfectly 165 nucleotides downstream,

```
JP53                 GTTTAGATAAG[GGCT...300 nt...AAA]GTTTAGATAAA
A36γ             AAGATTCACTTGTTAGA[GGCT...291 nt...AAA]GAGAGATTCAAAAGATTCACTTGTT(T)AG
Insulin (28)      AAAACAAGCAGGAGA[GGCT...314 nt...AAA]AAAACAAGCAGGAG
```

Fig. 5. Direct repeats flanking human Alu family conserved sequence.

Alu family conserved sequence is bracketed. Direct repeats are underlined.

An insertion in the 3' copy of the A36γ repeat is in parentheses.

at position +250 (in crosshatched boxes in Fig. 4), and 2) an A-rich region 15 nucleotides long starting at position +118 in pA36γ and repeated 164 nucleotides downstream in the oligo(dA) tract (underlined in Fig. 4).

The 40 nucleotide conserved repeat is the region which contains the previously noted homologies between RNAs of Chinese hamster cells, human Alu family sequences and papovaviral origins of replication (3). The sequences of the left-and right-hand copies of the 40 nucleotide conserved repeat from pA36γ pJP53 and BLUR clone 8 have been aligned in Fig. 6, "a" through "f".

Within the conserved repeat there is a six nucleotide variable segment located from positions 15 through 20 which is flanked on either side by regions of greater sequence conservation (Fig. 6k). Within the six nucleotide variable segment, the homology between all left-hand or all right-hand copies of the repeat is greater than the homology between the left-hand and right-hand repeats within any one clone. (Compare sequences a-c to d-f in positions 15 through 20 in Fig. 6.)

```
position within region        1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
a. JP53 left        +20       C  C  T  G  T  A  A  T  C  C  C  A  G  C  A  T  T  T  T  G  G  G  A  G  G  C  T  G  A  G  A  C  G  G  G  T  G  G  A  T     +59
b. A36γ left        +22       C  C  T  G  T  A  A  T  C  C  C  A  G  A  A  T  T  T  T  G  G  G  A  G  G  C  C  A  A  G  G  C  A  G  G  C  A  G  A  T     +61
c. BLUR8 left       +22       N  N  N  G  T  A  A  T  C  C  G  A  G  C  A  C  T  T  T  G  G  G  A  G  G  C  C  A  A  G  G  A  G  G  G  C  A  G  A  T     +61
d. JP53 right       +158      C  C  T  G  T  A  G  T  C  C  C  A  G  C  T  A  C  T  C  A  G  G  A  G  G  C  T  G  A  G  G  C  A  G  G  A  G  A  A  T     +197
e. A36γ right       +156      C  C  T  G  T  A  A  T  C  C  C  A  G  C  T  A  C  T  T  G  G  G  A  G  G  C  T  G  A  G  G  C  A  G  G  C  G  G  A  T     +195
f. BLUR8 right      +157      C  C  T  G  G  A  A  T  C  C  C  A  G  C  T  A  C  T  T  A  G  G  A  G  G  C  T  G  A  G  A  C  A  G  A  A  G  A  A  T     +196
g. Mo-Bla           +19       C  C  T  T  T  A  A  T  C  C  C  A  G  C  -  A  C  T  C  G  G  G  A  G  G  C  A  G  A  G  G  C  A  G  G  G  G  G  A  T     +58
h. Mo-Blb           +19       C  C  T  T  T  A  A  T  C  C  C  A  G  C  -  C  C  T  C  G  G  G  A  G  G  C  A  G  A  G  G  C  A  G  G  C  G  G  A  T     +58
i. Mo-Blc           +20       C  C  T  T  T  A  A  T  C  C  C  A  G  C  -  C  C  T  C  G  G  G  A  G  G  C  A  G  A  G  G  C  A  G  G  C  G  G  A  T     +59
j. ≥8/9 identical             *  *  *     *  *  *  *  *  *  *  *  *  *           *        *  *  *  *  *  *        *  *     *     *  *           *  *
k. variable domain                                                     *  *  *  *  *  *
```

Fig. 6. Sequence comparison of the 40 nt conserved repeat in human and mouse.

Boxed numbers refer to sequence numbering in Fig. 7. Deletions, placed by inspection to maximize sequence homology, are indicated by dashes (-). Undetermined nucleotides are denoted by an N. Restriction sites for the AluI enzyme are underlined.

In this, the six nucleotide variable region resembles the sequences which follow the conserved repeat. The 9 base pair direct repeat starting at positions 85 and 250 differs from the 40-nucleotide repeat in that the inter-repeat spacing of the former is about 28 nucleotides greater than for the latter. The first copy of the (dA)-rich region differs from the second in that residues other than dA occur in the first copy (i.e., a T at position 122 and a dC at position 124).

The murine interspersed repetitive sequence B1 resembles half of the human Alu family transcriptional template

Kramerov *et al.* (29) have selected pBR322 plasmid-based genomic clones from mouse DNA complementary to snapback double-stranded hnRNA, designated ds-RNA B (30). One-quarter of all clones screened were positive, suggesting that these sequences are highly reiterated in the murine genome. The repetitive DNA sequences contained in these clones were grouped into two mutually-exclusive subclasses, designated B1 and B2, based on the inability of ds-RNA eluted from nitrocellulose filters bearing only B1 repetitive sequences to bind to other filters bearing only B2 sequences (29). The B1 sequence-bearing regions from three independent clones were sequenced (8), and denoted B1a, b and c. Each of the three clones contains a highly conserved homology region 135 nucleotides in length (Fig. 3 of Ref. 8).

Beginning 20 nucleotides downstream from the 5' end of the B1 conserved sequence (Fig. 4) is a 39 nucleotide region displaying strong homology (Fig. 6g, h, i) to the human *Alu* family 40 nucleotide conserved repeat. The distance between the 5' end of the interspersed repetition sequence and the 5' end of the strongly conserved 40 nucleotide region is the same in man and mouse (Fig. 4). The best match with the human conserved repeat is derived by placing a deletion at position 15 (Fig. 6) of the mouse 39 nucleotide sequence. This deletion falls within the 6 nucleotide variable domain discussed above. Thus aligned, the homology between B1a and the righthand JP53 repeat is (34/41)x100%=83%.

Seventy-one nucleotides separate the mouse 39 nucleotide conserved region from a highly A-rich tract, preceded by the trinucleotide TCT, which resembles the oligo(dA) tract found at the 3' end of the human *Alu* family sequences. Within this 71 nucleotide region, there is a 19 nucleotide domain beginning at position 63 in the B1 sequence which matches in 17 of 19 positions with nucleotides beginning at position 75 in the A36γ sequence and which overlaps the 9 base pair direct repeat at position 85 in the human *Alu* family sequence. This sequence displays homology with a sequence identified by Fowlkes and Shenk (31) as a possible intragenic RNA polymerase III promoter, as discussed in the following paper.

Nucleotide sequences extending 90 to 100 nucleotides upstream from the beginning of the B1a, b and c sequences (Fig. 4B, Ref. 7) do not match each other, nor do they match any human *Alu* family sequences. Sequences downstream from the B1 sequences seem to represent much longer versions of deoxyadenylate-rich tracts than have been found to date in human *Alu* family sequences, containing more residues other than deoxyadenylate within them (Fig. 4A, Ref. 7).

Finally, neither the deoxyadenylate tract, the 19 nucleotide domain, nor the 39 nucleotide conserved region is repeated within the B1 sequence. Thus, the sequence data lead to the conclusion that the human *Alu* family sequence strongly resembles in many features a duplication of the first 135 nucleotides of mouse B1 sequence. This conclusion is verified by the electronmicroscopic studies of duplex DNA lengths in the so-called "foldback" DNA from human (7) and mouse (32), as well as non-"foldback" interspersed repetitive DNA in man (7). While Cech and Hearst conclude that the number average length of duplex stems is 0.5-0.8 kb, examination of their length histograms (Fig. 4, Ref. 43) reveals that the most frequently observed (modal) length is 175± 50 nt in foldback DNA. Similarly, while Bell and Hardman (33) report the number average length of foldback DNA sequences in the hamster to be 0.9 kb, inspection of their length histograms indicates that the modal length of double-stranded structures is between 0.1 and 0.2 kb.

In man, however, the most frequently observed length of duplex DNA in both inverted and interspersed repeated DNA is 280±25 nucleotides (Figs. 2 and 8, Ref. 6). Assuming that renaturation of A-T "tails" is poor under the hybridization and spreading conditions used for electron microscopy, the observed lengths of duplex DNA are in quantitative agreement with those expected if the interspersed and inverted repeat DNAs observed by electron microscopy in human and mouse are due to sequences of the *Alu* and B1 families, respectively. The lengths of inverted and interspersed repetitive DNA observed for hamster resemble those found for mouse rather than those found for man.

How are the similarities between the multiple copies of this highly repetitive interspersed DNA sequence maintained throughout the chromosomes in one species, and how could the duplication (or deletion of one-half molecule) be transmitted throughout the entire genome of the organism? The mechanism of unequal crossing over during mitosis or meiosis for the evolution of satellite DNAs (34) seems inapplicable because it leads to tandem as opposed to interspersed sequence repetition. Selective pressures operating independently in the sequences in each copy of the *Alu* family might maintain the similarity of sequences throughout the genome. However, it is difficult to imagine how a uniform change in all copies of the *Alu* sequence could occur on the basis of independent selection of changes in each copy separately. We demonstrate the presence of direct repeats of non-conserved DNA flanking the *Alu* sequence as reported by Bell *et al.* (28) in two additional *Alu* family sequences. A flanking direct repeat of host DNA has been a consistent feature of all sites of transposable element insertion studied to date (36). However, while the length of the direct repeat is fixed for all known transposable elements (36), the direct repeats flanking the three *Alu* family members compared in Fig. 5 are 10, 14, and 16 nucleotides long. Moreover, 11 nucleotides separate the 3' flanking copy of the pA36γ direct repeat and the conserved *Alu* family sequence. The *Alu* family sequence itself is much shorter than any known transposable element, being closer in size to the δ sequences "left behind" by the transposable TY1 element of yeast (35). The conserved Alu family sequence also lacks the terminal inverted repeat beginning with the nucleotide TG common to proretroviruses, TY1 and the copia elements of Drosophila (36). All these facts suggest that the Alu family sequence is not a prototypical transposable element as currently defined. Nevertheless, a transpositional mechanism seems attractive as an explanation for the interspersion of repetitive sequences throughout the genome, and for the distribution of major changes in these sequences throughout the interspersed repetitive sequences over evolutionary time.

ACKNOWLEDGMENTS

J. T. Elder is supported by a predoctoral fellowship from the Insurance Medical Scientist Scholarship Fund. This research was also supported by Grant #2 PO1 CA 16038-06, awarded by the National Cancer Institute, Department of Health and Human Services, and by the Comprehensive Cancer Center for Connecticut at Yale, Grant #CA 16359.

ABBREVIATIONS:

kb, kilobases; nt, nucleotides; DBM, diazobenyloxymethyl; LMW-RNA, low-molecular-weight RNA; dA, deoxyadenosine; μCi, microcurie; GTP, guanosine triphosphate.

REFERENCES

1. Houck, C. M., Rinehart, F. P. and Schmid, C. W. (1979) J. Mol. Biol. 132, 289-306.
2. Rubin, C. M., Houck, C. M. Deininger, P. L., Friedmann, T. and Schmid, C. W. (1980) Nature 284, 372-374.
3. Duncan, C., Biro, P. A., Choudary, P. V., Elder, J. T., Wang, R. R. C., Forget, B. G., deRiel, J. K. and Weissman, S. M. (1979) Proc. Natl. Acad. Sci. USA 76, (10), 5095-5099.
4. Fritsch, E. F., Lawn, R. M. and Maniatis, T. (1980) Cell 19, 959-972.
5. Bishop, J. O. and Freeman, K. B. (1973) Cold Spring Harbor Symp. Quant. Biol. 38, 707.
6. Schmid, C. W. and Deininger, P. L. (1975) Cell 6, 345-358.

7. Deininger, P. L. and Schmid, C. W. (1976) J. Mol. Biol. 106, 773-790.
8. Krayev, A. S., Kramerov, D. A., Skryabin, K. G., Ryskov, A. P., Bayev, A. A. and Georgiev, G. P. (1980) Nucl. Acids Res. 8,(6), 1201-1215.
9. Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L., and Schmid, C. W. (1980) Proc. Natl. Acad. Sci. USA 77,(3) 1398-1402.
10. Clewell, D. B. and Helinski, D. R. (1970) Biochem. 9, 4428-4440.
11. Clewell, D. B. and Helinski, D. R. (1969) Proc. Natl. Acad. Sci. USA 62, 1159-1166.
12. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. and Efstratiadis, A. (1978) Cell 15, 687-701.
13. Blin, N. and Stafford, D. W. (1976) Nucl. Acids Res. 3, 2303-2308.
14. Southern, E. M. (1975) J. Mol. Biol. 98, 503-517.
15. Alwine, J. C., Kemp, D. J., Parker, B. A., Reiser, J., Renart, J., Stark, G. R., and Wahl, G. M. (1979) Meth. in Enzymol. 68, 220-242.
16. Tuan, D., Biro, P. A., deRiel, J. K., Lazarus, H. and Forget, B. G. (1979) Nucl. Acids Res. 6(7) 2519-2544.
17. Maniatis, T., Jeffrey, A., and Kleid, D. G. (1975) Proc. Natl. Acad. Sci. USA 72 (3) 1184-1188.
18. Maxam, A. M. and Gilbert, W. (1980) Meth. in Enzymol. 65, 499-559.
19. Maat, J. and Smith, A. J. H. (1978) Nucl. Acids Res. 5, 4537-4545.
20. Yoshimuri, R. N., Dissertation, 1970 (University of California, San Francisco).
21. Greene, P. J., Heynecker, H. L., Bolivar, F., Rodriguez, R. L., Betlach, M. C., Covarrubias, A. A., Backman, K., Russell, D. J., Tait, R. and Boyer, H. W. (1978) Nucl. Acids Res. 5, 2373-2378.
22. Duncan, C. H., Wilson, G. A., Young, F. (1978) J. Bacteriol. 134, 338-344.
23. Bolivar, F., Rodriguez, R. L., Greene, P. J., Belach, M. C., Heyneker, N. L., Boyer, H. W., Crosa, J. H., and Falkow, S. (1977) Gene 2, 95-113.
24. Wu, G.-J. (1978) Proc. Natl. Acad. Sci. USA 75, 2175-2179.
25. Duncan, C. H., Jagadeeswaran, P., Wang, R. R. C. and Weissman, S. M. (1980) Gene, submitted for publication.
26. Blattner, F. R., Williams, B. G., Blechl, A. E., Denniston-Thompson, K., Faber, H. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L., and Smithies, O. (1977) Science 196, 161-169.
27. Deininger, P. L. and Schmid, C. W. (1979) J. Mol. Biol. 127, 437-460.
28. Bell, G. I., Pictet, R., and Rutter, W. J. (1980) Nucl. Acids Res. 8 (18), 4091-4109.
29. Kramerov, D. A., Grigoryan, A. A., Ryskov, A. P. and Georgiev, G. P. (1979) Nucl. Acids Res. 6(2), 697-712.

30. Kramerov, D. A., Ryskov, A. P. and Georgiev, G. P. (1977) Biochem. Biophys. Acta 475, 461-476.
31. Fowlkes, D. M., and Shenk, T. (1980) Cell, in press.
32. Cech, T. R. and Hearst, J. E. (1975) Cell 5, 429-446.
33. Bell, A. J. and Hardman, N. (1977) Nucl. Acids Res. 4(1), 247-268.
34. Smith, G. P. (1976) Science 191, 528-535.
35. Faraborough, P. J., and Fink, G. R. (1980) Nature 286, 352-356.
36. Temin, H. M. (1980) Cell 21, 599-600.