Volume 17 Number 14 1989 Nucleic Acids Research

Unique sequence organization and erythroid cell-specific nuclear factor-binding of mammalian $\theta 1$ globin promoters

Jae-Ho Kim, Chun-Yuan Yu, Arnold Bailey, Ross Hardison[1] and C.-K. James Shen*

Department of Genetics, University of California, Davis, CA 95616 and [1]Department of Molecular and Cell Biology, Pennsylvania State University, University Park, PA 16802, USA

Received May 12, 1989; Accepted June 2, 1989

ABSTRACT

The θ1 globin gene is an α globin-like gene, and started to diverge from the other members of the α globin family 260 million years ago. DNA sequencing and transcriptional analysis indicated that it is functional in erythroid cells of the higher primates, but not in prosimians and rabbit. The θ1 promoter region of higher primates including man consists of GC-rich sequences characteristic of housekeeping gene promoters, and CCAAT and TATA boxes located further upstream. It is shown here that the housekeeping gene promoter-like region of human θ1 contains two tandemly arranged, GC-rich motifs (GC-I and GC-II). Of these, GC-II interacts with nuclear factor(s) present in the globin-expressing, erythroleukemia cell line K562, before and after hemin induction. GC-I, however, interacts with nuclear factor(s) only present in hemin-induced K562 cells. These factors are different from previously reported erythroid cell-specific factors, and are not detectable in non-erythroid Hela cells. Furthermore, the sequence of the motif GC-I and its location relative to ATG codon have been conserved among all known mammalian θ1 globin genes. Finally, and most interestingly, the CCAAT box of θ1 is contained within a 38 bp internal segment of Alu repeat sequence. Immediately upstream from this CCAAT box-containing Alu repeat segment is a 241 bp Alu repeat pointing in the opposite direction. The conservation of this novel arrangement among the higher primates suggests that an inserted Alu family repeat and its flanking genomic sequence have co-evolved, for at least 30 million years, to provide the canonical CCAAT and TATA promoter elements of the θ1 globin genes in higher primates.

INTRODUCTION

The θ1 globin gene identified in the past few years is located at the very 3' end of the primate α globin gene cluster, approximately 3 kb downstream of the adult α1 globin gene (1-7). Contrary to galago (4) and rabbit (5), the structures of the θ1 globin genes of human, orangutan, and olive baboon, (2,3,7) suggest that they are functional in higher primates. They are all split into three exons with the potential to code for a globin polypeptide 141 amino acids long. Unlike the ξ and α globin genes, they use TGA, not TAA, as the termination codon, and AGTAAA, not AATAAA, as the polyadenylation signal. At approximately 220 bp and 260 bp upstream of the ATG codon are the canonical CCAAT and TATA promoter elements (Fig.1).

There are multiple initiation sites mapped at upstream of the human θ1 globin gene, and the transcription is regulated in a cell type- and tissue-specific way (7,8). These earlier studies have demonstrated the transcriptional activities of the θ1 gene in human fetal erythroid tissues and in a globin-expressing erythroleukemia cell line (K562), but not in non-erythroid cells like the Hela cell line. The 5' ends of the polyadenylated θ1 transcripts have been mapped to three locations upstream of the ATG codon (7,8). Of these, the location of the farthest initiate site (7; Fig. 1A) suggests that it is initiated by the canonical CCAAT and TATA boxes. The cluster of heterogeneous initiation sites located at 6-39 bp upstream of the ATG codon is contained within GC-rich sequences (Fig. 1A), and transcriptional initiations from this cluster have been suggested (8) to be controlled by their immediately upstream region, by mechanism(s) similar to house-keeping genes like dihydrofolate reductase gene (DHFR) or hypoxanthine phosphoribosyl transferase gene (HPRT) (9 and references therein). More recently, correctly spliced θ1 transcripts have also been detected in adult bone marrow cells from patients with sickle cell anemia (T. Ley, personal communication).

Thus, it appears that θ1 globin gene provides an unique system to study erythroid cell-specific gene regulation. In the following, we present a case in which specific Alu repeat sequences were recruited during evolution to provide RNA polymerase II - dependent promoter sequence of the θ1 globin gene, and have since been fixed in the genomes of the higher primates including human. We also show that the sequences immediately upstream of the cluster of heterogeneous initiation sites are composed of two adjacent GC-rich motifs, both of which interacting with specific nuclear factor(s) present in erythroid K562 cells but not non-erythroid Hela cells. The sequence of one of the two motifs is also conserved among the promoter regions of all known mammalian θ1 globin genes.

MATERIALS AND METHODS

DNA Sequencing.

DNA sequencing was carried out by the chemical degradation method of Maxam and Gilbert (10). The DNA fragments covering the θ1 promoter regions were isolated from cosmid clone cα3'Bg (11) kindly provided to us by Doug Higgs.

Cell Cultures.

Both Hela and K562 cells were grown in RPMI 1640 medium supplemented with 10% fetal bovine serum, 50 units/ml of penicillin, and 50 μg/ml of streptomycin (all from Gibco). For hemin induction, K562 cells were maintained in the above medium containing 30-50 μM hemin (Sigma) for 3 days.

Preparation of Nuclear Extracts.

For preparation of nuclear extracts, 1-5 liters of cells were harvested at the density of 5-7 x 10^5 cells/ml. Procedures for preparing nuclear extracts from Hela cells, K562 cells, or hemin-induced K562 cells are essentially the same as those described by Dignam et al. (12). The concentrations of proteins in our nuclear extract preparations vary from 2 μg/μl to 7 μg/μl.

DNase I footprinting.

Binding of nuclear factors to specific DNA sequences were assayed by DNase I footprinting technique (13). To prepare DNA probes used in the DNase I footprinting assay, a 264 bp Bam HI - Acc I restriction fragment, which contains the DNA region from 193 bp upstream of the human θ1 ATG codon to 71 bp downstream of it, was labeled with ^{32}P at either the Bam HI end or the Acc I end by Klenow polymerase, and purified by polyacrylamide gel electrophoresis (14).

The nuclear extracts were first preincubated with 1.5 µg of poly dI-dC (Pharmacia) in 50 µl of 50 mM KCl, 25 mM HEPES (pH 7.9 at 4°C), 0.05 mM EDTA, 0.5 mM DTT, 10% glycerol, 1 mM PMSF, and 0.2 µg/ml Leupeptin. After 10 minutes at room temperature, 1 ng of the DNA probe (approximately 10^4 cpm) was added to the reaction mixture. After another 2 hrs, the reaction mixture was adjusted to 5 mM $MgCl_2$ followed by the addition of DNase I (Worthington, grade DPFF). Digestion was allowed to proceed at room temperature for 1 minute, and then stopped by the addition of 100 µl of 0.1 M Tris-HCl (pH 8.0), 0.1 M NaCl, 1% SDS, 10 mM EDTA, 100 µg/ml tRNA and 100 µg/ml proteinase K. The mixture was incubated at 37°C for 15 minutes, extracted with phenol-chloroform, precipitated with ethanol, dried, and resuspended in deionized formamide. The digestion products were then analyzed by denaturing 6% or 8% polyacrylamide - 8M urea gels and autoradiography (14).

RESULTS

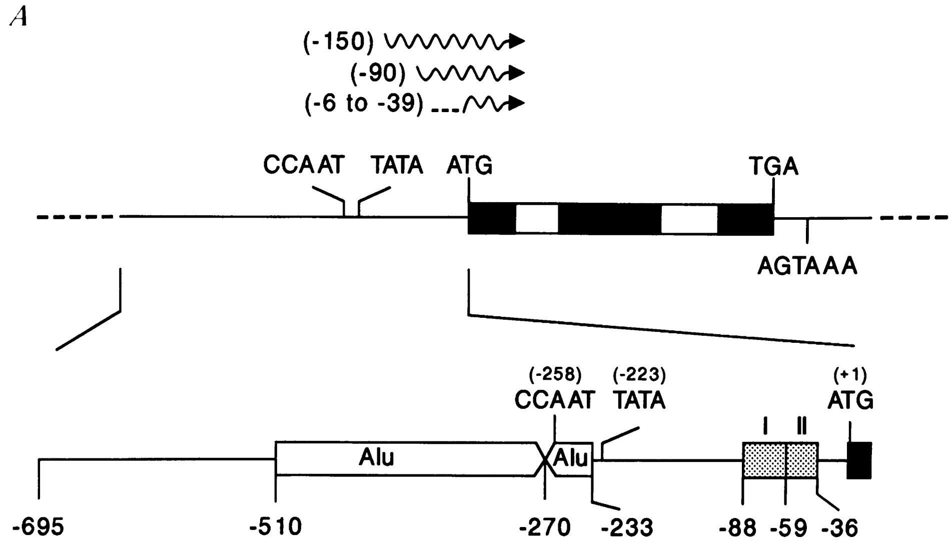
A Mosaic Arrangement of the CCAAT-TATA Boxes of Human θ1 Globin Gene and Alu Family Sequences.

In order to further understand the promoter organization and functioning of the human θ1 globin gene, the entire α1-θ1 intergenic DNA (3 kb) have been sequenced and analyzed (A. Bailey, unpublished results). 693 bp of these sequences upstream of the human θ1 ATG codon are shown in Fig. 1B.

To our surprise, it was found that the CCAAT promoter box is buried within a novel arrangement of Alu repeat sequences. As indicated in Fig. 1A, and shown in more details in Fig. 1B, the CCAAT box is contained within a 38 bp segment which shares 90% homology to positions 112-149 of the genomic consensus Alu sequence (15), but in an inverted orientation relative to the direction of θ1 transcription. Furthermore, immediately upstream from this CCAAT box-containing Alu segment is another truncated Alu repeat homologous to the consensus Alu sequence from position +1 to +251 (Fig.1B). This 241 bp Alu segment points oppositely to the 38 bp, CCAAT box-containing Alu segment, and overlaps with the later by one base pair (Fig.1B). It contains both of the RNA polymerase III - dependent promoter elements (Fig.1B), i.e. the anterior box A and the posterior box B, that are typical for Alu family repeats (16-18).

Conservation of the Mosaic CCAAT-Alu Repeat Arrangement Among θ1 Globin Genes of Higher Primates.

Interestingly, this mosaic arrangement of Alu repeats and CCAAT plus TATA promoter boxes is well conserved among the higher primates including orangutan and olive baboon (Fig.2), whose θ1 genes are also apparently functional (2,3). These two primates started to diverge from



B

```
-695  5'- NNNTCAGGCA   CCTCACAGTT   GTTATCCGTT   TAATTCTCAC   AATCTGAGAA

          GAAACTGTCA   CCCTCATTTT   ATATAATAAA   TGAGAAAACA   GACTCGGCCA

-595      AGTGTCACAA   TAGAATCAAG   AGGCAGAATA   AACTGACTTC   CAATGCCAAA
                                                                 Box A
          TCCATGCCGA   AATTCAGTGC   TATAATAATG   TACATGGCCG   GGCGCGGTGG
                                                    *****   **********
                                                    1
-495      TTCACGCCTG   TAATCCCAGA   ACTTTGGGAG   GCTGAGGCGG   GAGGATCACC
          C*********   *********C   **********   **C*******   *C********
                         Box B
          TGAGGTCGGG   AGTTTGAGAT   CAGCCT----AACA   CGGTGAAACC   CTGTCTCTAC
          -******A**   ****C****C   ******GGCC****   T*********   *C********

-395      TAAAAATACA   AAA-TT------GGCAT   GGTGGCATGC   ACCTGTGATC   CCAGTTACTC
          **********   ***A**AGCCTG***G*   ******GCAT   G*****A***   ****C*****

          GGGAGGCTGA   GGCAGGAGAA   TCGTTTGAAC   CCGGGAGGCG   GAGGTTGCAG
          **********   **********   ***C******   **********   **********
                                    149
                                        ****A   *G***A**T*   ********A*
-295      TGAGCCGGAA   TGGCGCCACT   GCACTCACCG   CACCCGGCCA   ATTTTTGTGT
          *******AG*   *T********   ******
                         112         251
          **********   ***
•         TTTTAGTAGA   GACTAAATAC   CATATAGTGA   ACACCTAAGA   CGGGGGGCCT

-195      TGGATCCAGG   GCGATTCAGA   GGGCCCCGGT   CGGAGCTGTC   GGAGATTGAG

          CGCGCGCGGT   CCCGGGATCT   CCGACGAGGC   CCTGGACCCC   CGGGCGGCGA

-95       AGCTGCGGCG   CGGCGCCCCC   TGGAGGCCGC   GGGACCCCTG   GCCGGTCCGC

          GCAGGCGCAG   CGGGGTCGCA   CGGCGCGGCG   GGTTCCAGCG   CGGGGATGGC

+6        GCTGTCCGCG   GAGGACCGGG   CGCTGGTGCG   CGCCCTG  -3'
```

Fig. 1 Promoter organization of the human θ1 globin gene. (A) General diagram of the human θ1 gene. The gene is shown to the right with the three exons indicated by black boxes. Also shown are the initiation codon (ATG), the termination codon (TAG), the polyadenylation signal

human approximately 15 and 30 million years ago (19-21). In addition to DNA sequence comparison (Fig.2), partial restriction enzyme cleavage mapping has also confirmed the conservation of the novel arrangement of the two Alu repeat sequences in the promoter regions of orangutan and olive baboon θ1 globin genes (A. Bailey, unpublished results).

However, the mosaic arrangement could not be found for the non-primate mammals including galago (4) and rabbit (5) whose θ1 genes have been inactivated recently (3). It is well documented that there has been a complete turn-over and homogenization of Alu family members after the divergence of human and non-primate mammals, and again after the divergence of galago and the other higher primate species (15 and references therein). Thus, it is difficult to ascertain whether the inactivation of the rabbit and galago θ1 globin genes, but not those of the higher primates, was due to the loss of a similar promoter arrangement composed of Alu type repeats specific for the two species, or due to evolutionary demands for maintenance of functional θ1 genes in higher primates. In any case, the novel promoter arrangement described above is most likely the result of insertion and scrambling of two Alu repeats at upstream of the θ1 globin genes of higher primates, and subsequent co-evolution, possibly by positive selection, of the repeats and its immediately flanking sequences to generate the canonical CCAAT-TATA promoter elements. The generation of the CCAAT box within the short, truncated Alu segment was due to a substitution of A to G at position 136 of the Alu consensus sequence. Interestingly, the corresponding region of the 241 bp Alu repeat also has a CCAAT sequence (Fig. 1B). Thus, some kind of gene conversion may have occurred between these two Alu repeat segments during evolution. A computer search of more than 200 Alu repeat sequences in the GenBank has revealed only one other Alu family member having the CCAAT sequence (unpublished results).

(AGTAAA), and the canonical CCAAT and TATA promoter boxes. The numbers in the parentheses indicate the positions relative to the A(+1) of ATG codon. The wavy lines represent RNAs initiated from three locations, as mapped previously by S1 nuclease digestion and primer extension assay (7, 8). The two open arrows represent the two truncated Alu segments pointing opposite to each other, one of which contains the CCAAT box (see Fig. 1B and text). The two stippled boxes represent the two GC-rich motifs, GC-I and GC-II, that interact with nuclear factors present in K562 and/or hemin induced K562 cells (see Figs. 4, 5 and text for details).
(B) Nucleotide sequence of the human θ1 promoter region, from 42 bp downstream to 692 bp upstream of the ATG initiation codon. The sequences were determined by Maxam and Gilbert method (10). Only the sense strand is shown with the numbers representing the locations of the nucleotides, upstream (–) or downstream (+), relative to the ATG (+1) codon. The two canonical promoter elements, CCAAT and TATA, and the potential SP1 factor binding site, GGGCGG, are indicated by bold face letters. The two oppositely oriented Alu repeat segments, one from nucleotide -510 to -270 and the other from -270 to -233, are each aligned and boxed together with the homologous blocks of Alu genomic consensus sequence. The Alu genomic consensus from 1 to 251 is shown below the sequence of the 241 bp Alu repeat segment, while the antisense sequence of Alu genomic consensus from 149 to 122 is shown above the 38 bp Alu repeat segment. "*" represent homologous nucleotides, "-" represent deletions, and different nucleotides are individually indicated. The anterior and posterior promoter boxes A and B of the Alu family repeats are also indicated.

```
            -358
            |
Human       ATCCCAGTTACTCGGGAGGCTGAGGCAGGAGAATCGTTTGAACCCGGG
Orangutan   -----------------A-T---------------------------A
Baboon      -------C----------------------------C-------T---

                                          ....-----><---...
            AGGCGGAGGTTGCAGTGAGCCGGAATGGCGCCACTGCACTCACCGCACCCGG
            -----A--------------------C-----------------------T-
            -------------C--------A---C-------------------------

                                                        -211
                                                           |
            CCAATTTTTGTGTTTTTAGTAGAGACTAAATACCATATAGTGAACACC
            ------------------------------A-------G---------
            ------------------------------A--T----G---------
```

Fig. 2 Nucleotide sequence comparison of DNA regions surrounding the canonical promoter elements of human, orangutan, and olive baboon θ1 globin genes. The nucleotide sequence from -358 to -211 upstream of the ATG codon of human θ1 globin gene(Fig. 1B) is compared to the orthologous genomic regions of orangutan and olive baboon (2, 3). The nucleotides are indicated wherever they are different from the human. The two oppositely oriented arrows represent the Alu segments described in text and Fig. 1.

A 30 bp GC-Rich Motif GC-I Is Conserved at Upstream of all Mammalian θ1 Globin Genes.

In addition to the above mosaic arrangement of CCAAT-TATA-Alu family repeats, sequence comparison has identified another novel DNA element that is located in between the TATA box and the ATG codon of all known mammalian θ1 globin genes. As shown in Fig. 3, alignment of DNA sequences upstream of different mammalian θ1 globin genes reveals a highly conserved, 30 bp GC-rich sequence motif, which we termed GC-I. In the three higher primates including human, GC-I is located at 59 bp to 88 bp upstream of the ATG codon of θ1 (Fig.1A; Fig.3). It is located at 56 bp to 85 bp, and 74 to 104 bp upstream of the ATG codons of the galago and rabbit θ1 genes, respectively. The pattern of conservation is most obvious when the sequences of the higher primates and those of the other two mammals are compared. Approximately 90% of homology exists within the GC-I motif. On the contrary, outside of it, little similarity could be found (Fig.3).

Erythroid Cell-Specific Nuclear Factors Interact with the GC-I, and Another GC-rich Motif (GC-II).

In order to examine whether GC-I interacts with specific DNA-binding factors, nuclear extracts were prepared from Hela cells, K562 cells, and hemin-induced K562 cells. Protein binding to the motif was then assayed by DNase I footprinting technique as described in

MATERIALS AND METHODS.

A 264 bp Bam HI-Acc I fragment covering DNA region from 193 bp upstream to 71 bp downstream of the θ1 ATG codon was used as the probe for DNase footprinting. As exemplified in Fig. 4A, there is no detectable protein binding to this probe in Hela nuclear extract (lane 2, fig. 4A). However, nuclear extracts prepared from both K562 and hemin-induced K562 cells protected a GC-rich region from 35 bp to 57 bp upstream of the ATG codon (lane 3, Fig. 4A, and lane 1, Fig. 4B). This motif, which we termed GC-II, is located immediately downstream of GC-I (Figs. 1A and 5). Interestingly, in addition to GC-II, the nuclear extract of hemin-induced K562 cells also protected a region coinciding exactly with the conserved 30 bp motif GC-I (Lane 2, Fig. 4B). The protection of GC-I in induced K562 extract is accompanied by the presence of DNase I hypersensitive sites at the boundary of GC-I and GC-II (arrows, Figs. 4B and 5). These data are summarized in Fig. 5.

DISCUSSION

By sequence analysis and protein-DNA interaction assays, we have identified several interesting DNA elements upstream of the human θ1 globin gene. This is not surprising since multiple transcriptional initiating sites from upstream of the human θ1 gene have been identified in erythroid cells of different developmental lineages (7, 8; T. Ley, personal communication), suggesting the existence of multiple upstream promoter elements.

Two of such elements, GC-I and GC-II, are adjacent to each other. Although both motifs are highly GC-rich (Fig. 5), they do not contain any apparent SP1-binding site (22). Furthermore, the Hela nuclear extract we have prepared gave as clear footprints as K562 extract at several SP1-binding sites upstream of the human α globin gene (data not shown). Motif GC-II, which is located at 35 bp to 57 bp upstream of ATG, is protected from DNase I digestion by nuclear extracts prepared from K562 cells with or without hemin induction (Fig.4). Since this nuclear factor(s) is present only in K562 cells in which the θ1 globin gene is expressed, but not in Hela cells in which θ1 is inactive, it very likely plays a functional role in the transcriptional initiation of the θ1 gene in erythroid cells. This possibility is particularly intriguing because the DNA region it protected, 35-57 bp upstream of the ATG codon, is located immediately upstream of the set of heterogeneous sites of initiation of θ1 (Figs. 1A & 5). The lack of CCAAT and TATA boxes, and the high GC content in the vicinity of this set of heterogeneous θ1 initiation sites have led to the proposal (8) that they are generated by mechanisms similar to house-keeping genes. Mutagenesis and transcriptional studies have shown that sequences located at 27 to 39 bp upstream of the heterogeneous initiation sites of mouse HPRT are required for expression of the gene in vivo (9). In the case of θ1, the set of heterogeneous initiation sites spans a region of 33 bp, with the prominent sites mapped at 6 bp, 9 bp, 19 bp and 28 bp, respectively, upstream of the ATG codon

```
Human     (-118) GGCCCTGGAC CCCCGGGCGG CGAAGCTGCG GCGCGGCGCC CCCTGGAGGC CGCGGGACCC CTGGCCGGTC CGCGCAGGCG CAGCGGGGTC (-29)
Orangutan (-118) *******CG* ***A****** ***G*****A ********** ********** ********** **A******* ********** *G******A* (-29)
Baboon    (-118) *******C** ***A***T** ***G*****A ********** ******T*** ********** ********** ********** *****C**G* (-29)
Galago    (-115) *CTAAGACCG GAG*CTCG*C AAG*TAGCGC **A******* *****C**** *TG******* ACC*GGC*GT *CTA*GC*A* AGCTCCC*CT (-26)
Rabbit    (-135) *CG*GGC*G* A*A**T*TC* G****GGC*T ****CA**** ***-A***** ********** ——(73 bp of GC rich repeats)—
                                           C                           C     G
```

(8; Fig. 5). Thus, the location of GC-II relative to the heterogeneous initiation sites of θ1 is very similar to that of the mouse HPRT promoter relative to the heterogeneous initiation sites of HPRT gene. This suggests that DNA-protein interaction(s) at GC-II, in combination with RNA polymerase II and other factors, may direct the heterogeneous transcriptional initiation of θ1 in erythroid cells.

Immediately upstream of GC-II is the conserved GC-rich motif GC-I, initially identified by sequence analysis (Fig.3). The conservation of this motif at upstream of all cloned mammalian θ1 globin genes suggests that it may also play some role in their expression. Although the rabbit θ1 and galago θ1 genes are apparently pseudogenes due to DNA mutations in their coding regions (4, 5), these inactivation events probably have occurred recently (3). The evolution time elapsed since these inactivation events may thus not be long enough to allow much divergence among different species of an once functionally important DNA region. The relative abundance of nuclear factor(s) interacting with GC-I in hemin-induced K562 extract (Fig. 4) suggests that DNA-protein interaction at GC-I may play important role(s) in the hemin-induced transcriptional enhancement of human α-like globin genes in K562 cells. Alternatively, this nuclear factor(s) may be related to the differentiation process of K562 cells after hemin induction. The unique sequences of GC-I and GC-II suggest that the K562 cell-specific nuclear factors interacting with the two motifs are different from the other erythroid cell-specific transcription factors reported previously (23-26). The functional roles of motif GC-I and GC-II, and the factor(s) interacting with them await further biological studies.

In this communication, we have also presented a novel mosaic organization of the θ1 CCAAT promoter box of higher primates and two Alu family repeat sequences. The human genome is interspersed with approximately 5 x 10^5 to 10^6 copies of Alu family repeats (27, 28). The typical length of these repeats is 300 bp although cloned Alu family members 260 bp and 600 bp long (15,29,30) have also been identified. Detailed sequence analysis of cloned Alu repeats have divided them into several subclasses, each one with diagnostic base substitutions relative to the others (31-33). It appears that the Alu family repeats have been generated from several conserved and ancestral genes, which replaced each other during overlapping periods of evolution. Alu family repeats are also present in abundance in other primates including chimpanzee, gorilla, gibbon, orangutan, New World monkeys, Old World monkeys, and prosimians (15 and references therein). Retroposition has been proposed to be the mechanism of

Fig. 3 Alignment of sequences upstream of five mammalian θ1 globin genes. The nucleotide sequences immediately upstream of the θ1 genes of human (7), orangutan (2), baboon (3), galago (4), and rabbit (5) are aligned for comparison. The numbers in the parentheses indicate locations relative to ATG codon. The human sequence is shown on top. The "*"s denote bases of the other mammalian species that are identical to man. The sequences in the boxed region indicate that the 30 bp GC-rich motif, GC-I, is conserved even between the primates and the rabbit, which have diverged from each other 85-120 million years ago.

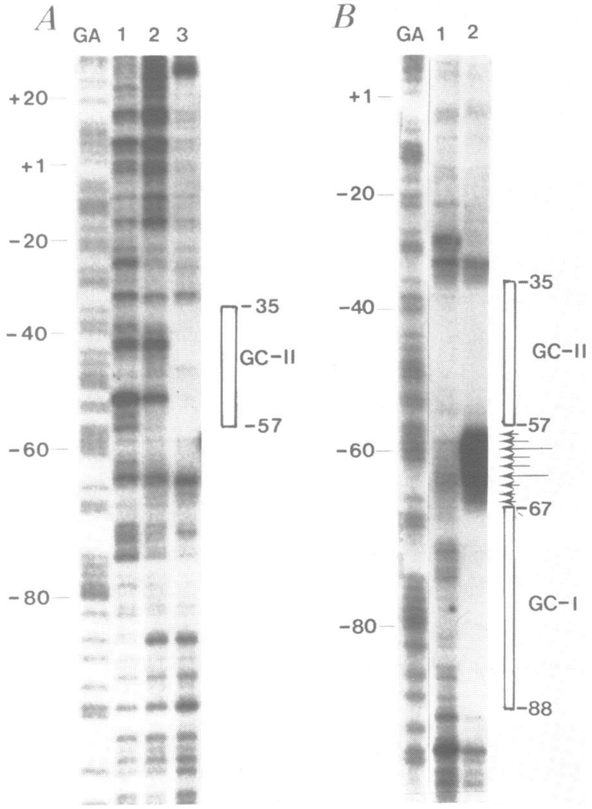


Fig. 4 DNase I footpring analysis of promoter region immediately upstream of human θ1 ATG codon. (A) Comparison of nuclear extracts of Hela cells and K562 cells. The BamHI-AccI restriction fragment, which contains the human θ1 promoter region, was labeled with ^{32}P at its BamHI end and analyzed by DNase I footprinting in nuclear extracts prepared from Hela cells or K562 cells (see MATERIALS AND METHODS for more details). Lane GA, G + A sequencing markers prepared from the labeled fragment; lane 1, control DNase I digestion in the absence of nuclear extract; lane 2, DNase I digestion in the presence of 80 μg of Hela nuclear extract; lane 3, DNase I digestion in the presence of 80 μg of K562 nuclear extract. The numbers to the left of the panel indicate nucleotide positions relative to the ATG codon (+1). The open box to the right of the panel represents the motif, GC II, that is protected from DNase I digestion by the K562 extract. There is also a partial protection of the region from -30 to +30 in the K562 extract. Protection of the GC-II motif could also be observed at the amount of 40 μg of K562 extract, but 20 μg of K562 extract only gave a partial protection (data not shown). Under the conditions used in lanes 2 and 3, both Hela and K562 nuclear extract gave clear footprints of SP1-binding sites and CCAAT box region upstream of the human α globin gene (data not shown). (B) Comparison of nuclear extracts of K562 cells and hemin-induced K562 cells. Lane GA, G + A sequencing markers; lane 1, DNase I digestion in the presence of 80 μg of K562 extract; lane 2, DNase I digestion in the presence of 70 μg of hemin-induced K562 extract. Both nuclear extract protect the motif GC-II, while motif GC-I is protected from DNase I digestion only by the hemin-induced K562 extract. The arrows on the right side of the panel indicate bands resulting from enhanced DNase I cleavages in hemin induced-K562 extract.

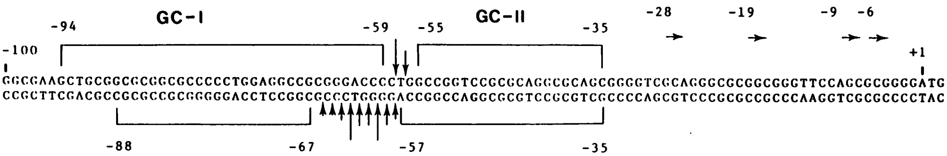

GC-I
GC-II
-100
-94
-59
-55
-35
-28
-19
-9
-6
+1
GGCGAAGCTGCGGCGCGGCGCCCCCTGGAGGCCGCGGGACCCCTGGCCGGTCCGCGCAGCCGCAGCGGGGTCGCAGGGCGCGGCGGTTCCAGCGCGGGATG
CCGCTTCGACGCCGCGCCGCGGGGACCTCCGGCGCCCTGGGGACCGGCCAGGCGCGTCCGCGTCGCCCCAGCGTCCCGCGCCGCCCAAGGTCGCGCCCCTAC
-88
-67
-57
-35

Fig. 5 A summary of the DNase I footprints of the promoter region immediately upstream of human θ1 ATG codon. Brackets indicate the ranges of DNase I footprints of motif GC-I in hemin-induced K562 extracts, and of motif GC-II in both K562 and induced K562 extracts. The vertical arrows indicate enhanced DNase I cleavages in induced K562 extract. The major sites of the heterogeneous transcriptional initiation (8) are also indicated by horizontal arrows. The footprint results were obtained by using the BamHI-AccI fragment labeled at either the BamHI end (Fig. 4), or at the AccI end (data not shown) as the probe.

amplification and genomic insertion of the Alu family repeats and other interspersed repeat families (34,35 and references therein).

The function(s), if there is any, of the Alu family repeats is unknown. Many cloned Alu repeats contain RNA polymerase III - dependent promoter, and can be efficiently transcribed in vitro (16-18, 36, 37). However, neither the small Alu RNAs nor the small ribonucleoproteins assembled in vitro on these RNAs (38) have been convincingly demonstrated to exist in vivo. Several deletions, duplications, and other types of genetic rearrangements causing inherited human diseases have been shown to be the results of recombination between Alu repeats (11, 39-41). The dynamics of the possible genomic turnover of Alu family repeats is further supported by the great difference in their copy numbers among different primates (28). A recent study has shown that most, if not all, of the length polymorphisms among the adult α globin loci of human, orangutan, and Old World monkeys are either insertion/deletion of Alu repeats, or gross genomic rearrangements immediately flanking the Alu repeats (42).

The unique promoter organization of the primate θ1 globin genes, as shown in Fig. 1, provides an interesting example of mosaic arrangement of Alu family sequences, which are RNA polymerase III-dependent templates, and the CCAAT and TATA boxes which are essential promoter elements for many RNA polymerase II - dependent genes. The possible interactions between RNA polymerase II - dependent and RNA polymerase III - dependent transcription processes have been noted in other studies (43-45). For example, both polymerases are able to initiate transcriptions of c-myc gene from the same site but terminate at different locations (46,47). Upstream transcriptional initiations, in vitro and in vivo, from nucleotides other than the canonical sites have been observed for different human globin genes (48-50). Certain upstream initiation events of the human ε, γ, and β globin genes appear to be carried out by RNA polymerase III (51-53). It is not clear at the present time how the transcriptional regulation of θ1 globin gene is related to the arrangement shown in Fig. 1. However, some intriguing possibilities can be considered. The CCAAT box-containing, inverted Alu repeat segment corresponds to the central AT-rich region of the Alu consensus, and thus provides two poly-T stretches in between the

CCAAT and TATA boxes (Fig.1B). Poly-T stretch is known to be efficient transcriptional terminator for RNA polymerase III (54). Thus, transcription processes initiated by RNA polymerase III from the 241 bp Alu repeat would terminate right in between the CCAAT and TATA boxes. This may impose a negative effect on θ1 initiation by interfering with the formation of a polymerase II - dependent preinitiation complex in the vicinity of the two promoter elements. Alternatively, the polymerase III - dependent transcription initiated from the Alu repeat may render the chromatin surrounding the CCAAT and TATA boxes in an open configuration that is accessible to RNA polymerase II transcription. These speculations could be tested in the future by DNA transfection studies.

On one end of the spectrum, it has been proposed that the highly abundant, interspersed repetitive sequences of eukaryotes have no function, but to survive on their ability to amplify and disperse in the eukaryotic genomes (55,56). On the other end, it has been proposed that repetitive sequences may be involved in transcriptional regulation in eukaryotic cells (57,58). Our study here brings up the possibility that certain repeats may provide promoter elements, as well as play regulatory role(s), for RNA polymerase II - dependent genes in the eukaryotic genomes. In some aspects, this evolutionary recruitment of Alu repeats is similar to the insertions of retroviral LTRs, and their activating the genes downstream of the insertion sites, e.g. the insertional oncogenesis by avian leukosis viruses (59).

ACKNOWLEDGEMENTS

One of the authors (C.-K.J.S.) thanks Ian Mattaj, Patrick Charnay, and their colleagues for teaching him techniques of protein-DNA interactions in an EMBO workshop. We thank Doug Higgs for providing the cosmid clone cα3'Bg, and Jeng-Pyng Shaw for helpful suggestions on DNase I footprinting. We also thank Carl Schmid for his help with computer analysis. This research was supported by NIH grants DK 29800 (C.-K.J.S.) and DK 27635 (R.H.), and NIH Research Career Development Awards to C.-K.J.S. and R.H.

*To whom correspondence should be addressed

REFERENCES

1. Marks, J., Shaw, J.-P. & Shen, C.-K.J. (1986) Proc. Nat. Acad. Sci. USA 83, 1413-1417.
2. Marks, J., Shaw, J.-P. & Shen, C.-K.J. (1986) Nature (London) 321, 785-788.
3. Shaw, J.-P., Marks, J. & Shen, C.-K.J. (1987) Nature (London) 326, 717-720.
4. Sawada, I. & Schmid, C.W. (1986) J. Mol. Biol. 192, 693-709.
5. Cheng, J.-F., Raid, L. & Hardison, R.C. (1986) J. Biol. Chem. 261, 839-848.
6. Clegg, S.D. (1987) Nature (London) 329, 465-466.
7. Hsu, S.-L., Marks, J., Shaw, J.-P., Tam, M., Higgs, D.R., Shen, C.-C. & Shen, C.-K. J. (1988) Nature (London) 331, 94-96.
8. Leung, S.-O., Proudfoot, N.J. & Whitelaw, E. (1987) Nature (London) 329, 551-556.
9. Melton, D.W., McEwan, C., McKie, A.B. & Reid, A.M. (1986) Cell 44, 319-328.

10. Maxam, A.M. & Gilbert, W. (1977) Proc. Nat. Acad. Sci. USA 74, 560-564.
11. Nicholls, R.D., Fischel-Ghodsian, H. & Higgs, D.R. (1987) Cell 49, 369-378.
12. Dignam, J.D., Lebovitz, R.M. & Roeder, R.G. (1983) Nucleic Acids Res. 11, 1475-1489.
13. Galas, D.J. & Schmitz, A. (1978) Nucleic Acids Res. 5, 3157-3170.
14. Maniatis, T., Fritsch, E.F. & Sambrook, J. (1982) in Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, New York.
15. Schmid, C.W. & Shen, C.-K.J. (1985) in Molecular Evolutionary Genetics, ed. R.J. McIntyre, Plenum Publishing Co., New York, pp. 323-358.
16. Paolella, G., Lucero, M.A., Murphy, M.H. & Barelle, F.E. (1983) EMBO J. 2, 691-696.
17. Perez-Stable, C., Ayres, T.M. & Shen, C.-K.J. (1984) Proc. Nat. Acad. Sci. USA 81, 5291-5295.
18. Perez-Stable, C. & Shen, C.-K.J. (1986) Mol. Cell. Biol. 6, 2041-2052.
19. Gingerich, P. & Schoeninger, M. (1977) J. Hum. Evol. 6, 483-505.
20. Andrews, P. (1985) Nature (London) 314, 498-499.
21. Fleagle, J. (1986) in Major Topics in Primate and Human Evolution, eds. Wood, B.A., Martin, L. & Andrews, P., Cambridge University Press, pp. 130-149.
22. Jones, K.A., Kadonaga, J.T., Rosenfeld, P.J., Kelly, T.J. & Tjian, R. (1987) Cell 48, 79-89.
23. Mantovani, R., Malgaretti, N., Giglion, B., Comi, P., Cappllini, N., Nicolis, S. & Ottolenghi, S. (1987) Nucleic Acids Res. 15, 9349-9364.
24. Barnhardt, K.M., Kim, C.G., Banerji, S.S. & Sheffery, M. (1988) Mol. Cell. Biol. 8, 3215-3226.
25. Evans, T., Reitman, M. & Felsenfeld, G. (1988) Proc. Nat. Acad. Sci. USA 85, 5976-5980.
26. deBoer, E., Antoniou, M., Mignotte, V., Wall, L. & Grosveld, F. (1988) EMBO J. 7, 4203-4212.
27. Schmid, C.W. & Jelinek, W.R. (1982) Science 216, 1065-1070.
28. Hwu, H.R., Roberts, J.W., Davidson, E.H. & Britten, R.J. (1986) Proc. Nat. Acad. Sci. USA 83, 3875-3879.
29. Hess, J.F., Fox, M., Schmid, C.W. & Shen, C.-K.J. (1983) Proc. Nat. Acad. Sci. USA 80, 5970-5974.
30. Hess, J.F., Perez-Stable, C., Wu, G., Weir, B., Tinoco, I. Jr. & Shen, C.-K.J. (1985) J. Mol.Biol. 184, 7-21.
31. Willard, C., Nguyen, H.T. & Schmid, C.W. (1987) J. Mol. Evol. 26, 180-186.
32. Britten, R.J., Baron, W.F., Scott, D.B. & Davidson, E.H. (1988) Proc. Nat. Acad. Sci. USA 85, 4770-4774.
33. Jurka, J. & Smith, T. (1988) Proc. Nat. Acad. Sci. USA 85, 4775-4779.
34. Singer, M.F. (1982) Cell 28, 433-434.
35. Weiner, A.M., Deininger, P.L. & Efstratiadis, A. (1986) Annu. Rev. Biochem. 55, 631-661.
36. Elder, J.T., Pan, J., Duncan, C.H. & Weissman, S.M. (1981) Nucleic Acids Res. 9, 1171-1189.
37. Fuhrman, S., Deininger, P.L., LaPorte, P., Friedman, T. & Geiderschek, E.P. (1981) Nucleic Acids Res. 9, 6439-6456.
38. Shen, C.-K.J. & Maniatis, T. (1982) J. Mol. Appl. Genet. 1, 346-360.
39. Lehrman, M.A., Schnieder, W.J., Sudhof, J.C., Brown, M.S., Goldstein, J.L. & Russell, D.W. (1985) Science 227, 140-146.
40. Lehrman, M.A., Goldstein, J.L., Russell, D.W. & Brown, M.S. (1987) Cell 48, 827-835.
41. Rougher, F., Simmler, M.C., Page, C. & Weissenbach, J. (1987) Cell 51, 417-425.
42. Shen, C.C., Bailey, A., Kim. J.-H., Yuan, C.-Y., Marks, J., Shaw, J.-P., Klisak, I., Sparkes, R. & Shen, C.-K.J. (1989) Prog. Clin. Biol. Res. in press.
43. Bark, C., Weller, P., Zabielski, J., Janson, L. & Pettersson, U. (1987) Nature 328, 356-359.
44. Carbon, P., Murgo, S., Ebel, J.-P., Krol, A., Tebb, G. & Mattaj, I.W. (1987) Cell 51, 71-79.
45. Krol, A., Carbon, P., Ebel, J.-P. & Appel, B. (1987) Nucleic Acids Res. 15, 2463-2478.
46. Bentley, D. & Groudine, M. (1986) Mol. Cell. Biol. 6, 3481-3489.

47. Chung, J., Sussman, D.J., Zeller, R. & Leder, P. (1987) Cell 51, 1001-1008.
48. Allan, M., Grindlay, L., Stefani, L. & Paul, J. (1982) Nucleic Acids Res. 10, 5133-5147.
49. Grindlay, G.J., Lanyon, W.G., Allan, M. & Paul, J. (1984) Nucleic Acids Res. 12, 1811-1820.
50. Hess, J.F., Perez-Stable, C., Deisseroth, A. & Shen, C.-K.J. (1985) Nucleic Acids Res. 13, 6059-6074.
51. Carlson, D.P. & Ross, J. (1983) Cell 34, 857-864.
52. Kollias, G., Sekeris, C.E. & Grosveld, F.G. (1985) Nucleic Acids Res. 13, 7993-8005.
53. Carlson, D.P. & Ross, J. (1986) Mol. Cell. Biol. 6, 3278-3282.
54. Korn, L.J. & Brown, D.D. (1978) Cell 15, 1145-1156.
55. Doolittle, W.F. & Sapienza, C. (1980) Nature (London) 284, 601-603.
56. Orgel, L.E. & Crick, F.H.C. (1980) Nature (London) 284, 604-607.
57. Britten, R.J. & Davidson, E.H. (1969) Science 165, 349-357.
58. Davidson, E.H. & Britten, R.J. (1979) Science 204, 1052-1059.
59. Neel, B.G., Hayward, W.S., Robinson, H.L., Fang, J. & Astrin, S.M. (1981) Cell 23, 323-334.