Unusual sequence homology at the 5' ends of the developmentally regulated β^A -, β^C -, and γ -globin genes of the goat

(globin gene regulation/DNA sequence determinations)

JOEL R. HAYNES, PAUL ROSTECK, JR., AND JERRY B. LINGREL

Department of Biological Chemistry, University of Cincinnati College of Medicine, 231 Bethesda Avenue, Cincinnati, Ohio 45267

Communicated by Philip Leder, August 29, 1980

ABSTRACT We have constructed a recombinant library of goat DNA and have isolated clones containing the β^A -, β^C -, and γ -globin genes and a presumptive embryonic globin gene. These genes are switched on and off during development and, thus, provide a model system for the study of gene regulation. To identify regions that may be involved in this switch, we determined the sequence of the areas surrounding the 5' ends of the β^A -, β^C -, and γ -globin genes. Surprisingly, the sequences of the β^A - and β^C -globin genes are identical, beginning with the translation initiation codon and extending 131 base pairs (bp) in the 5' direction. At this point, two nucleotide differences are seen and only six additional differences are encountered when the sequences are extended 144 bp further in the 5' direction. Furthermore, the nucleotide sequence of the 5' end of the γ -globin gene is very similar to those of the β^A - and β^C -globin genes. Only three nucleotide differences are found in the γ -globin gene within the 131-bp region in which the β^A - and β^C -globin genes are identical. We conclude that this identical region cannot contain regulatory signals that mediate the switch from β^C to β^A expression.

In prokaryotes, sequences functioning as regulatory recognition sites for gene expression are located near the origin of transcription (1-3). Although analogous regions may exist at similar positions in eukaryotic genes, such sequences have not yet been identified. One approach to locating regulatory regions in eukaryotic DNA is to examine the sequences surrounding the transcription initiation sites of a set of developmentally regulated genes. The globin genes of the goat other than the α -globin genes are expressed at different times during development and, thus, provide such a system.

During development in goats, embryonic chains are the first β -globin-like chains to appear and are synthesized in the yolk sac. At approximately 25 days after conception, these chains are replaced by γ -globin chains. Just before birth the γ -globin chains are replaced by β^C -globin chains, which in turn are replaced by β^A -globin chains during the first year of postnatal life (4, 5). We have reported the isolation of recombinants containing the β^A -, β^C -, and γ -globin genes (hereafter designated β^A -, β^C -, and γ genes) and a presumptive embryonic globin gene from a λ Charon 4A library of goat DNA (6, 7). In this study, we describe the nucleotide sequences adjacent to the 5' ends of the goat β^A -, β^C -, and γ genes. Surprisingly, the β^A and β^C genes have identical sequences surrounding their 5' ends, beginning with the translation initiation codon and extending 131 base pairs (bp) in the 5' direction until the first nucleotide difference is observed. Furthermore, the sequence surrounding the 5' end of the γ gene differs at only three nucleotide positions within the region where the β^A and β^C genes are identical.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Restriction Enzyme Mapping. Development of the restriction endonuclease cleavage maps for clones containing the β^A -, β^C -, and γ genes and of the *Hae* III map for the 1.2-kilobase (kb) *Hind*III/*Bam*HI fragment from the 5' end of the β^A gene are described elsewhere (6). *Hae* III digestion of the 1.2-kb *Hind*III/*Bam*HI fragments from the β^C and γ genes yielded fragment patterns similar to that observed for the same fragment of the β^A gene. The β^C fragment was found to contain one additional *Hae* III site whereas the γ fragment contained two extra sites. The positions of these new sites were determined from DNA sequence data presented here and elsewhere (6). Agarose and polyacrylamide gel electrophoresis was performed as described (7).

Subcloning Restriction Fragments. The 1.2-kb *Hind*III/*Bam*HI fragments from the 5' ends of the β^A -, β^C -, and γ genes were ligated to the large *Hind*III/*Bam*HI fragment of the bacterial plasmid pBR322 as described by Mulligan *et al.* (8). Ligated DNAs were cloned by the method of Wensink *et al.* (9). Clones harboring recombinant plasmids were selected on the basis of ampicillin resistance and tetracycline sensitivity. Growth of strains carrying recombinant plasmids and amplification and isolation of plasmid DNAs was as described (10). All work involving recombinant DNA molecules was performed in accordance with the National Institutes of Health guidelines for recombinant DNA research.

DNA Labeling and Nucleotide Sequence-Determination Strategy. The 1.2-kb fragments containing the β^A -, β^C -, and γ 5' ends were excised from 0.5 mg of each recombinant plasmid by *Hind*III/*Bam*HI double digestion followed by electrophoresis on a 1.0% agarose gel. Fragments were recovered by electroelution from gel slices as described (6). Ten micrograms of an isolated 1.2-kb *Hind*III/*Bam*HI fragment from each gene was digested with *Pvu* II and labeled at all 3' termini with DNA polymerase I (Klenow fragment; Boehringer Mannheim) and [α^{32} P]dATP as described (6). Because *Pvu* II digestion leaves blunt ends, the labeling utilized the 3'→5'-exonuclease function of the enzyme in which the 3'-terminal G residue and penultimate A residue are removed. At this point the enzyme adds a 32 P-labeled A residue by the 5'→3'-polymerase function. The use of reaction buffer containing no potassium and room-temperature incubation (25°C) promotes the 3'→5'-exonuclease function (11). In our hands *Pvu* II sites were more efficiently labeled by [α^{32} P]dATP than by the more direct 3'-terminal exchange with [α^{32} P]dGTP.

The doubly end-labeled fragments were digested with *Hinf*I to produce fragments labeled on one end only, and these were separated by electrophoresis on a long (39 cm) 4% (wt/vol) polyacrylamide gel and isolated as described (6). Sequences

Abbreviations: bp, base pair(s); β^A -, β^C -, and γ genes; β^A -, β^C -, and γ -globin genes; kb, kilobase(s).

were determined according to the method of Maxam and Gilbert (11). Complementary sequence data for the γ gene were obtained by the above rationale, except that the *Pvu* II ends were labeled on their 5' termini with T4 polynucleotide kinase and [$\gamma^{32}\text{P}$]ATP (12). Complementary sequence data for the β^A and β^C fragments were obtained by cleaving the 1.2-kb fragments with *Hae* III and labeling their 3' ends with DNA polymerase I and [$\alpha^{32}\text{P}$]dGTP or labeling their 5' ends with T4 polynucleotide kinase and [$\gamma^{32}\text{P}$]ATP. *Hae* III fragments with radioactive label on both ends were isolated from a 2% (wt/vol) agarose gel and either cut with a second enzyme or strand-separated, followed by electrophoresis on 4–6% polyacrylamide gels. Strand separation was accomplished by dissolving the fragments in a 0.15 M NaOH/1 mM EDTA and incubating at 37°C for 2 min. The solution was then chilled in an ice bath, diluted with an equal volume of ice-cold water, and electrophoresed on a polyacrylamide gel at 4°. Before sequencing, all isolated fragments were precipitated twice from ethanol, washed twice with 95% ethanol, vacuum dried, and dissolved in 30 μl of H_2O . DNA sequence-determination gels were as described by Sanger and Coulson (13).

RESULTS AND DISCUSSION

In an attempt to identify regions that may be involved in the switching of the goat globin genes, we have determined the nucleotide sequences surrounding the 5' ends of the β^A , β^C , and γ genes. Restriction maps (6) of these genes, as determined from overlapping clones, are shown in Fig. 1. Unfortunately clones that link the β^A , β^C , or γ genes to one another have not been identified. Moreover, as the clones we have isolated cover 77 kb of DNA, it appears that intergenic distances for the globin genes of the goat are significantly greater than those observed for the rabbit (14) and for human globin genes (15). However, linkage of the β^A and β^C genes to at least two defective globin genes (β^X and β^Z) has been found (Fig. 1). Details of these defective genes will be presented elsewhere. An additional gene termed " β^Y " is found upstream from the defective β^X gene, and

details concerning its structure also will be presented elsewhere.

In order to facilitate sequence determinations of the 5' regions of the globin genes, the 1.2-kb *Hind*III/*Bam*HI fragments containing the 5' third of each of the β^A , β^C , and γ genes were subcloned into the bacterial plasmid pBR322. Detailed restriction maps for each of these fragments are shown in Fig. 2. The position of the small intervening sequences for each of the globin genes was determined by sequence analysis through the first coding block into the intervening sequence. Their lengths were all calculated to be approximately 135 bp. Each intervening sequence begins after amino acid position 29, which is 340 bp upstream from the *Bam*HI site at amino acid position 98. Since only 207 bp of the 340-bp sequence is needed to code for amino acids 30–98, each intervening sequence must be approximately 135 bp. These are similar in size and position to the small intervening sequences of the mouse (16–18) and rabbit (14, 18) globin genes.

Nucleotide sequences surrounding the 5' ends of the β^A , β^C , and γ genes were determined from the *Hae* III and *Pvu* II sites shown in Fig. 2. The arrows indicate the strands and the directions in which sequences were determined. The use of both 3'- and 5'-end labeling techniques enabled us to obtain sequence information for both the coding and noncoding strands for the majority of the sequence data presented.

The sequence for the entire first coding block and for 278 nucleotides proximal to the translation initiation codon for the β^A , β^C , and γ genes is shown in Fig. 3. The site labeled "Cap", located 52 bp upstream from the ATG codon, is considered the transcription-initiation site. The assignment is based on the finding that the 5' end, or cap sequence, of the 15S RNA precursor to β globin mRNA maps at this position (19). Because no larger β -globin precursor has been identified (10), it is assumed that the 15S RNA represents the initial transcript of this gene. Further evidence that the cap sequence is the initiation site of transcription is provided by *in vitro* studies. RNA transcription is initiated at this position when the mouse β^{maj} globin

FIG. 1. Restriction endonuclease cleavage maps of genomic regions containing the goat β^A , β^C , and γ genes. Restriction maps for the genomic regions containing the β^A , β^C , and γ genes were determined from nine independently isolated clones (6), and the regions contained in each of these clones is indicated by the bars below each map. Linkage of the β^A and β^C genes to other β -globin-like genes is demonstrated (see text). Restriction endonuclease cleavage sites: $\downarrow$, *Eco*RI; $\uparrow$, *Bam*HI; $\blacktriangledown$, *Hind*III.

FIG. 2. Restriction endonuclease cleavage maps of subcloned globin gene fragments. The 1.2-kb *Hind*III/*Bam*HI fragments containing the 5' third of the β^A , β^C , and γ genes were subcloned into *Hind*III/*Bam*HI-cleaved pBR322 DNA. Restriction maps for these fragments and positions of the first and second coding blocks (blackened areas) are indicated. The DNA sequence-determination strategy is shown by the arrows below each map. Solid arrows show the directions and distances of sequence determinations in the coding strand whereas dotted arrows show the direction and distances of sequence determinations in the noncoding strand. Restriction endonuclease cleavage sites: ∇ , *Hind*III; $\dagger$, *Bam*HI; $\bullet$, *Hae* III; $\blacktriangle$, *Pvu* II.

gene is added to a RNA polymerase II *in vitro* transcription system (20).

The Hogness-Goldberg box (21), a sequence preceding many eukaryotic mRNA genes (14-18, 22-26), is located at nucleotide positions -24 to -33 in all three goat globin genes. This se-

quence is thought to be involved in transcription, and Groschedl and Birnstiel have found that it may function in directing the RNA polymerase to the transcription-initiation site (28).

A comparison of the 5'-nucleotide sequences of the β^A , β^C , and γ genes shows a very high degree of conservation between these genes in areas that do not code for globin amino acid sequences. In fact, the β^A and β^C genes are identical beginning with the translation-initiation codon and extending 131 bp in the 5' direction. At this point, two nucleotide differences are found between these genes, and only six additional nucleotide differences are observed in the next 144 bp in the 5' direction. The sequence of the γ gene is very similar to those of the β^A and β^C genes in that it contains only three nucleotide differences within the 131-bp region wherein the β^A and β^C genes are identical.

The reason for this high degree of conservation in the 5'-end sequences is unknown. It is striking in lieu of the fact that the mouse β^{maj} and β^{min} globin genes, two genes that are expressed at the same developmental stage although at different levels, have significantly different sequences surrounding their 5' ends (16). These differences include insertions, deletions, and single nucleotide changes. Thus, it seems unlikely that all of these nucleotide positions in the goat genes would be conserved because they are required for a specific function. It should be noted that the first coding-block regions are not nearly as highly conserved between these genes. Table 1 shows the percent homology calculations between the β^A , β^C , and γ genes within the 5'-untranslated, the first coding-block, and the second coding-block regions. The 5'-untranslated region is defined here as the sequence beginning with the translation initiation codon and extending to a point 100 bp upstream from the cap site.

FIG. 3. Comparison of nucleotide sequences surrounding the 5' ends of the β^A , β^C , and γ genes. The nucleotide sequences surrounding the 5' ends of the β^A , β^C , and γ genes include the entire first coding-block regions and 278 bp proximal to the translation-initiation codon, which is indicated by "Met". Conserved nucleotide positions are enclosed in boxes.

Table 1. Percent homology between regions of the β^A , β^C , and γ genes

Regions	β^A/β^C	β^A/γ	β^C/γ
5' untranslated	98.7	96.1	96.1
First coding block	84.0	81.0	68.6
Second coding block	95.6	90.7	92.3

Sequence data for the three genes in the second coding block has been reported (6). In the three possible two-way comparisons, these genes are more homologous to one another in the 5'-untranslated region than in either of the two coding blocks. The second coding block is conserved to a greater extent between these genes than is the first coding block, which is not unexpected since the second exon codes for the heme-binding region of the globin chains.

It is interesting to note that the translation-initiation codon not only serves as the border between the 5'-untranslated and first coding-block regions but also is the point dividing the conserved and diverged sequences. It also should be noted that a 9-bp deletion is found at this point in the β^C gene. This deletion also is shown by the protein sequence data (29) and therefore, is not an artifact of cloning. A similar deletion of 6 bp is observed at this point in the β^A and γ genes when they are compared to the β^Y gene (an unidentified goat globin gene; unpublished data), the mouse β^{maj} gene, and the rabbit β gene. Such deletions occurring at the translation-initiation codon may be the result of unequal crossing-over at this point between sister chromatids. The fact that two different-sized deletions are observed suggest that they were caused by independent events. Unequal crossing-over near the translation-initiation codon not only is a possible explanation for deletions arising in this area but also can explain the formation of chimeric genes in which the 5' end of one gene is recombined with the coding region of another, giving rise to two genes with identical 5' ends and at the same time to an extra gene copy in one of the two sister chromatids involved in the recombination. The extra copy of a particular gene could then evolve into a nonfunctional pseudogene because only one copy of a particular gene is required for survival.

If the assumption that transcription is initiated at the cap site is correct, then it must follow that the nucleotide sequences involved in the switch from β^C - to β^A -globin synthesis are not located at the immediate 5' ends of these genes because these sequences are completely identical in this region. However, it is possible that the differences in the 5' flanking-region, which do occur between the β^A and β^C genes 80 bp upstream from the cap site and beyond, are somehow involved in the developmental regulation. Likewise, the extensive nucleotide differences encountered in the first coding block also may serve as regulating signals mediating the developmental switch. Finally, the two sequence differences around the Hogness-Goldberg box in the γ gene also may be involved in the switch from γ to β^C and β^A gene expression. It is probable that the immediate 5'-end regions are required for transcription initiation even if they do not serve a regulatory function, as it has been shown that the Hogness-Goldberg box and the region immediately upstream exert distinct effects on the *in vitro* transcription of the histone H2A gene (28). Sequences responsible for transcription initiation also may be located within the gene, as is the case for a *Xenopus* 5S ribosomal gene (30, 31) and a tRNA gene (32), both of which are transcribed by RNA polymerase III.

Our finding that the sequences that mediate the switch from β^C - to β^A -globin expression are not located within the 131-bp identical regions of these genes is in line with studies of hemo-

globin synthesis in certain hemoglobinopathies (33) and with analysis of globin gene deletion mutations (15), which suggest that regions downstream from the human γ and α genes may have an effect on the regulation of γ gene expression.

It is generally assumed that the phenomenon of hemoglobin switching is mediated at the level of transcription. However, it is possible that the γ , β^A , and β^C genes are transcribed throughout development and that post-transcriptional events are responsible for determining the levels of each of the globin mRNAs. It is known in sheep that switching between the synthesis of the β^A - and β^C -globins is accompanied by a change in the type of mRNA present in the cytoplasm (34). This suggests that globin gene expression is not regulated at the level of translation but does not rule out the possibility that unequal processing of mRNA precursors is responsible for the differential synthesis of the β^A - and β^C -globin chains. If this were true, then sequence differences responsible for unequal processing must occur between the two precursors. This possibility can be tested by determining the rates of synthesis of each of the precursors at different developmental stages.

Note Added in Proof. It recently has been shown by heteroduplex analysis (35) that extensive sequence homology occurs in the 5' flanking regions of the sheep β^A and γ -globin genes.

We are grateful for the excellent technical assistance of Mr. Donald VanLeeuwen and Ms. Gail Sears. This work was supported by Grants AM 20119 and PCM 76-80222 from the National Science Foundation and by Grant GM 10999 from the U.S. Public Health Service.

- Franklin, N. C. (1978) *Annu. Rev. Genet.* **12**, 193-221.
- Dickson, R. C., Abelson, J., Barnes, W. M. & Reznikoff, W. S. (1975) *Science* **187**, 27-35.
- Dickson, R. C., Abelson, J., Johnson, P., Reznikoff, W. S. & Barnes, W. M. (1977) *J. Mol. Biol.* **111**, 55-75.
- Garrick, M. D., Manning, R. F., Reichlin, M. & Mattioli, M. (1973) *The Biochemistry of Gene Expression in Higher Organisms*, eds. Pollack, J. K. & Lee, J. W. (Australia and New Zealand Books, Atarmon, Australia), pp. 379-392.
- Kitchen, H. & Brett, I. (1974) *Ann. N.Y. Acad. Sci.* **241**, 653-671.
- Haynes, J. R., Rosteck, P., Jr., Schon, E. A., Gallagher, P. M., Burks, D. J., Smith, K. & Lingrel, J. B. (1980) *J. Biol. Chem.* **255**, 6355-6367.
- Robbins, J., Rosteck, P., Jr., Haynes, J. R., Freyer, G., Cleary, M. L., Kalter, H. D., Smith, K. & Lingrel, J. B. (1979) *J. Biol. Chem.* **254**, 6187-6195.
- Mulligan, R. C., Howard, B. H. & Berg, P. (1979) *Nature (London)* **277**, 108-114.
- Wensink, P. C., Finnegan, D. J., Donelson, J. E. & Hogness, D. S. (1974) *Cell* **3**, 315-325.
- Haynes, J. R., Kalb, V. F., Rosteck, P., Jr. & Lingrel, J. B. (1978) *FEBS Lett.* **91**, 173-177.
- Donelson, J. E. & Wu, R. (1972) *J. Biol. Chem.* **247**, 4654-4660.
- Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499-560.
- Sanger, F. & Coulson, A. R. (1978) *FEBS Lett.* **87**, 107-110.
- Hardison, R. C., Butler, E. T., III, Lacy, E., Maniatis, T., Rosenthal, N. & Efstratiadis, A. (1979) *Cell* **18**, 1285-1297.
- Fritsch, E. F., Lawn, R. M. & Maniatis, T. (1979) *Nature (London)* **279**, 598-603.
- Konkel, D. A., Maizel, J. V. & Leder, P. (1979) *Cell* **18**, 865-873.
- Konkel, D. A., Tilghman, S. M. & Leder, P. (1978) *Cell* **15**, 1125-1132.
- van Ooyen, A., van den Berg, J., Mantei, N. & Weissmann, C. (1979) *Science* **206**, 337-344.
- Weaver, R. F. & Weissmann, C. (1979) *Nucleic Acids Res.* **7**, 1175-1193.

20. Luse, D. S. & Roeder, R. G. (1980) *Cell* **20**, 691–699.
21. Goldberg, M. (1979) Dissertation, (Stanford Univ., Stanford, CA).
22. Lai, E. C., Stein, J. P., Catterall, J. F., Woo, S. L. C., Mace, M., Means, A. R. & O'Malley, B. W. (1979) *Cell* **18**, 829–842.
23. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979) *Cell* **18**, 545–558.
24. Nishioka, Y. & Leder, P. (1979) *Cell* **18**, 875–882.
25. Tsujimoto, Y. & Suzuki, Y. (1979) *Cell* **18**, 591–600.
26. Ziff, E. B. & Evans, R. M. (1978) *Cell* **15**, 1463–1475.
27. Gannon, F., O'Hare, K., Perrin, F., LePennec, J. P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B. & Chambon, P. (1979) *Nature (London)* **278**, 428–434.
28. Grosschedl, R. & Birnstiel, M. L. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 1432–1436.
29. Dayhoff, M. O. & Eck, R. V. (1968) in *Atlas of Protein Sequence and Structure* (The National Biomedical Research Foundation, Silver Spring, MD), pp. 154–158.
30. Bogenhagen, D. F., Sakonju, S. & Brown, D. D. (1980) *Cell* **19**, 27–35.
31. Sakonju, S., Bogenhagen, D. F. & Brown, D. D. (1980) *Cell* **19**, 13–25.
32. Kressmann, A., Hofstetter, H., Capua, E. D., Grosschedl, R. & Birnstiel, M. L. (1979) *Nucleic Acids Res.* **7**, 1749–1763.
33. Schroeder, W. A. & Huisman, T. H. J. (1974) *Ann. N.Y. Acad. Sci.* **241**, 70–79.
34. Benz, E. J., Steggles, A. W., Geist, C. E. & Nienhuis, A. W. (1978) *J. Biol. Chem.* **253**, 5025–5032.
35. Kretschmer, P. J., Coon, H. C., Davis, A., Harrison, M. & Nienhuis, A. W. (1980) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **39**, 1784.