Gene Conversion of Two Functional Goat \( \alpha \)-Globin Genes Preserves Only Minimal Flanking Sequences*

Eric A. Schon‡, Suzanne M. Wernke, and Jerry B. Lingrel

From the Departments of Biological Chemistry and of Microbiology and Molecular Genetics, University of Cincinnati College of Medicine, Cincinnati, Ohio 45267

(Received for publication, December 29, 1981)

We have determined the complete nucleotide sequence of the nonallelic adult goat \( \alpha \)- and \( ^1\alpha \)-globin genes and, as is the case with the duplicated human \( \alpha \)-globin genes, these two genes are highly homologous to each other. Such high homology (98%) has most likely been preserved via a gene conversion mechanism. The conversion unit in goats is only about 900 base pairs in length, and contained within this short region are all the known signals required for accurate and efficient transcription, with the CCAAT box adjacent to the 5' boundary of the conversion unit and the poly(A) addition site adjacent to the 3' end. This conversion unit is also flanked by a 23-base-pair direct repeat "boundary sequence," vestiges of which are also observable in the human and mouse \( \alpha \)-globin genes and pseudogenes. These direct repeats imply that a transposition-like event may have been responsible for the insertion of an ancestral \( \alpha \)-like sequence into a new chromosomal locus, and that this insertion event and subsequent gene duplication may have predated the mammalian radiation.

Adult goat hemoglobin contains two nonallelic \( \alpha \)-globin chains, \( \alpha_1 \) and \( \alpha_2 \). These two chains differ at four amino acid positions: \( \alpha_1 \) has the structure designated (Gly 19, Ala 26, Leu 113, Asn 115), while \( \alpha_2 \) has the structure designated (Ser 19, Thr 26, His 113, Ser 115). In addition, there is an allele of \( \alpha_2 \), designated \( \alpha_2^A \), in which Asp 75 has been replaced by Tyr 75 (Huisman et al., 1967a; Huisman et al., 1967b; Huisman et al., 1968; Adams et al., 1968; Garrick and Huisman, 1968).

Sheep, which are closely related to goats, may also have two nonallelic \( \alpha \)-globin loci. The sheep \( \alpha_1 \) chain differs from the goat \( \alpha_1 \) chain at only one position (Thr 104 in the sheep versus Ser 104 in the goat) (Wilson et al., 1968), although a variant has also been described, with Asp replacing Gly at position 15 (Vestri et al., 1980). Vestri et al. (1980) have recently observed what they believe to be the second nonallelic sheep \( \alpha \)-globin, designated \( \alpha_2 \), presumably corresponding to the goat \( \alpha_1 \) locus, which has a substitution of Leu to His at position 113 or 114.

Finally, Wilson et al. (1970) have observed that Barbary sheep also contain two nonallelic \( \alpha \)-globin loci, in which one locus is identical with the goat \( \alpha \) chain and the other is identical with the goat \( \alpha_1 \) chain except for position 26, which contains an alanine residue (as does the \( ^1\alpha \) chain) instead of a threonine residue.

The two nonallelic \( \alpha \) loci in both goats and sheep do not produce equal amounts of protein. The \( \alpha_1:\alpha_2 \) ratio in goats is 3:1 (Adams et al., 1960), while the \( \alpha_1:\alpha_2 \) ratio in sheep is either 2:1 or 4:1 (Vestri et al., 1980). There are other mammalian species with nonallelic \( \alpha \)-globin loci which also do not appear to produce equal amounts of globin from each locus.

Several interesting features of \( \alpha \)-globin genes have been observed. First is the extremely high homology that exists between the duplicated human \( \alpha \)-globin gene pairs, \( \alpha_1 \) and \( \alpha_2 \) (Lieberhaber et al., 1980; Liebhaber et al., 1981). Gene conversion has been proposed as a mechanism to explain this unusual observation. Second, it has recently been found that, in the mouse, \( \alpha \)-globin-like sequences (i.e. \( \alpha \)-pseudogenes) are present on more than one chromosome (Leder et al., 1981; Popp et al., 1981), and that such dispersion may be explained by a transposition-like mechanism. Finally, all \( \alpha \)-globin genes sequenced to date have an unusually high G + C content, much greater than the average C + G content of mammalian DNA, or of the \( \beta \)-like globin genes.

We have previously reported the isolation of a clone that hybridizes to mouse \( \alpha \)-globin cDNA, which was derived from a partial Eco RI library of goat genomic DNA contained in the Escherichia coli bacteriophage \( \lambda \) Charon 4A (Haynes et al., 1980). In this communication, we report the isolation and complete nucleotide sequencing of the duplicated nonallelic goat \( \alpha \)- and \( ^1\alpha \)-globin genes and define the precise position of the areas of homology that define the conversion unit of these two genes. We also report the presence of cloning artifacts, including a "fusion gene" containing sequences from the 5' end of the \( \alpha \) gene and the 3' end of the \( ^1\alpha \) gene.

**EXPERIMENTAL PROCEDURES**

Isolation of Phage DNA and Subcloning of Restriction Fragments—Selection of recombinant phage containing goat genomic globin sequences was as described by Robbins et al. (1979). Propagation of recombinant phage and isolation of phage DNA was as reported by Blattner et al. (1977) and Maniatis et al. (1978). Agarose gel electrophoresis was as performed by Robbins et al. (1979), while polyacrylamide gel electrophoresis in Tris/borate/EDTA buffer was performed as described by Maniatis et al. (1975). Bam HI-cut DNA...
Chiron 4A and \textit{Hae} III-cut pBR322 DNA were used as markers for both 5' agarose and polyacrylamide electrophoresis. Fragments sequenced for sequencing were obtained by digestion of the DNA with appropriate restriction enzyme(s) (New England Biolabs, Bethesda Research Laboratories, or Boehringer Mannheim) followed by electrophoresis on agarose or polyacrylamide; subsequent elution from gel slices was described by Haynes et al. (1980), or in some cases, by direct phenol and phenol/chloroform extraction of fragments separated in low melting temperature agarose (Bio-Rad or Sigma).

The 4.0-kb\textit{ Bam} HI-\textit{Eco RI} fragment from clones 218L and 222L containing the 3'-globin gene, the 1.9-kb \textit{Bam} HI-\textit{Eco RI} fragment from clone 222U containing the 5'-globin “fusion” gene, and the 5.3-kb \textit{Bam} HI-\textit{Eco RI} fragment from clone 222U were ligated to the large \textit{Bam} HI-\textit{Eco RI} fragment of the bacterial plasmid pBR322 as described by Mulligan et al. (1979), and then propagated by the method of Wensink et al. (1974). Bacteria harboring recombinant plasmids were selected on the basis of ampicillin resistance and tetracycline sensitivity. In a similar manner, the 1.3-kb \textit{Eco RI} fragment from clones 218L and 222L, containing the 3'-globin gene was subcloned into the \textit{Eco RI} RI site of the bacterial plasmid pBR322, and bacteria containing this recombinant plasmid were selected on the basis of ampicillin and tetracycline resistance and chloramphenicol sensitivity. Growth of strains carrying recombinant plasmids and amplification and isolation of plasmid DNA were described by Chiron et al. (1980). All work involving recombinant DNA molecules was performed in accordance with the National Institutes of Health guidelines for recombinant DNA research.

End-labeling of DNA and DNA Sequencing—Approximately 10 pmol of each desired fragment was labeled at its 5' termini with DNA polymerase I (Klenow fragment, New England Nuclear or Boehringer Mannheim) and an appropriate [\textsuperscript{32}P]dNTP (Amersham, 410 Ci/mmol) as described by Haynes et al. (1980). Complementary strands were end-labeled at their 5' termini with T4 polynucleotide kinase (Bethesda Research Laboratories) and [\textsuperscript{32}P]ATP (Amersham, 2000 Ci/mmol) according to the procedures described by Maxam and Gilbert (1980). Fragments thus labeled at two ends were converted to fragments labeled at one end only by cutting with a suitable restriction enzyme(s) (New England Biolabs, Bethesda Research Laboratories) and [\textsuperscript{32}P]ATP (Amersham, 2000 Ci/mmol) according to the procedures described by Maxam and Gilbert (1980).

Computer Applications—The DNA sequences were analyzed by the program of Queen and Korn (1980). Divergence analysis was performed by a program written by F. Fuller, Harvard Medical School, based on procedures described by Perler et al. (1980), and kindly supplied to us by A. Efstatriadis, Harvard Medical School.

RESULTS

Restriction Maps and Sequencing Strategies—As shown in Fig. 1, two groups of clones hybridizing to mouse 3'-globin cDNA probe were obtained. Group I, containing the clones designated 11, 261, 213L, 213U, 218L, and 218U, all have similar 5' ends (i.e. the \textit{Eco RI} fragments, 6.0 and 2.4 kb in length), but have slightly varying lengths in the sizes of their 3' \textit{Eco RI} fragments. Group II is made up of only two clones, 222U and 222L, which have a different set of similar 5' \textit{Eco RI} fragments (1.3 and 4.3 kb in length), but have the same type of variation in their 3' \textit{Eco RI} fragments as do the clones in Group I.

Since the genomic library from which these clones were isolated was prepared from a partial digestion of DNA using the restriction endonuclease \textit{Eco RI}, the possibility existed that an artifact of cloning may have occurred in which two \textit{Eco RI} fragments which are not contiguous in the goat genome were ligated together prior to insertion in the \textit{λ} Chiron 4A vector. If this were the case, then an \textit{Eco RI} site should mark the boundary where two such genomic fragments had been ligated. Both genomic mapping of the region spanning the postulated “cloning artifact boundary” and examination of DNA fragments on either side of this boundary were needed to resolve this question. A \textit{Bam HI} digest of goat genomic DNA was electrophoresed on 0.5% agarose and transferred to nitrocellulose paper according to the method of Southern (1975) and was probed with a nick-translated 1.6-kb fragment obtained from within the 5.3-kb \textit{Eco RI-}\textit{Bam HI} fragment of clone 222U. As shown in Fig. 2b, only one hybridizing band was obtained, having a size of 10.8 kb. This result is consistent with the interpretation that the Group I clones (which have a 10.8-kb \textit{Bam HI} fragment in the region spanning the “cloning artifact boundary,” including the 1.6-kb probe segment) are the true chromosomal representations of this area of the goat genome and that the Group II clones (which have a shorter 8.8-kb \textit{Bam HI} fragment in this region) contain a cloning artifact in their 5' regions. In addition, the \textit{Eco RI} fragments on either side of the postulated “cloning artifact boundary” were sequenced. As shown in Fig. 1, were isolated and cut with \textit{Alu I} and \textit{Hae III}. The 3' \textit{Eco RI} fragment of both the Group I and Group II clones (i.e. the 9.3-kb \textit{Eco RI} fragment) displayed very similar restriction patterns with both \textit{Alu I} and \textit{Hae III}, while the 5' end \textit{Eco RI} fragment (i.e. the 2.4-kb fragment in the Group I clones and the 4.3-kb fragment in the Group II clones) had totally different restriction patterns (data not shown), again indicating that the \textit{Eco RI} site at the dotted line in Fig. 1 is indeed most likely such a boundary where two dissimilar genomic fragments were ligated during the cloning process.

While the 5' ends of the Group II clones had been determined to be the result of an artifact of cloning, it became clear that the heterogeneity in the sizes of the \textit{Eco RI} fragments in the 3' ends of both Group I and Group II clones was the result of a different artifact: that occurring during clone propagation after plaque purification. This type of artifact is the result of deletions or “popping out” of segments of DNA between two highly homologous or repeated DNA sequences in the regions flanking the deletion. Such deletions can be seen most clearly in the variable length of the rightmost \textit{Eco RI} fragment(s) in the Group I clones. In addition, a second type of deletion gave rise to two populations of clones during propagation from plaque-purified phage: two bands would appear after centrifugation in cesium chloride, rather than the expected single band. The lower (L) band contained the full-length “parental” clone, while the upper (U) band contained a shorter “daughter” clone harboring a deletion between two highly homologous segments of DNA. Clones 213L and 218L, 218U, and 222U and 222L were obtained in this manner. The nature of the deletion in 222U is analyzed further below.

The 3'-globin-hybridizable regions were determined by hybridizing nick-translated mouse 3'-globin cDNA probe to clone fragments that had been immobilized on nitrocellulose filters after single and double digestions with \textit{Eco RI}, \textit{Bam HI}, and \textit{HindIII}. These hybridizing regions were localized to the 4.0-kb \textit{Bam HI-}\textit{Eco RI} fragments in clones 11, 213L, 218L, \textit{Bam HI}, and \textit{HindIII} fragments in clones 213L, 218L, and 222L, as shown in Fig. 1.
Fig. 1. Goat α-globin-containing clones. The clones recovered from a partial Eco RI library of goat genomic DNA (Robbins et al., 1979) are shown. L and R indicate the orientation of the left and right arms, respectively, of the Charon 4A vector. Fragment sizes (in kilobase pairs) between adjacent restriction sites are marked on the clones as shown. The α-globin genes (shaded boxes) shown in clones 218L, 222L, and 222U have been mapped to the positions indicated; the open boxes in the other clones denote the presumed location of the α genes in these clones (these regions hybridize to mouse α-globin cDNA) by analogy to their position in the three above-mentioned clones. The gene marked 'α/α' is a fusion gene formed by deletion during phage propagation in vitro. The dotted line at the third Eco RI site from the left arm of the Group II clones marks the boundary of a presumed cloning artifact arising during ligation of partially Eco RI-digested genomic DNA to the Charon 4A arms during preparation of the library.

HindIII-Eco RI α-globin-hybridizable region were subcloned into the Eco RI site of plasmid pBR325. Preliminary sequencing studies localized one α-globin gene within a 900-bp Kpn I-Pst I fragment in the 4.0-kb Bam HI-Eco RI subclone of clone 218L, as shown in Fig. 3. Using this 900-bp Kpn I-Pst I fragment as a nick-translated probe in experiments involving Bam HI-cut and Eco RI-cut genomic DNA, it was determined that the goat α-globin genes lie within one 6.0-kb Bam HI fragment, but on two Eco RI fragments 9.3 and 1.3 kb in length (see Fig. 2a). Because it contains the 10.8-kb Bam HI fragment noted above and the 9.3- and 1.3-kb Eco RI fragments present in genomic DNA, as well as a truncated Bam HI fragment at least 5.3 kb long (see Fig. 1), clone 218L is therefore most likely to be the authentic and complete chromosomal representation of the region containing the goat α-globin genes.

The two α-globin genes localized within the 4.0-kb Bam HI- Eco RI and 1.3-kb Eco RI-Bam HI fragments were sequenced according to the strategies shown in Fig. 3. Note that more than two-thirds of the sequences were obtained on both strands, and that sequences were obtained through most of the restriction sites shown in Fig. 3. However, sequence was not obtained through a few sites outside the coding regions. The entire sequence of the two nonallelic goat α-globin genes, 'α' and 'α', was obtained, including the protein coding regions, the intervening sequences, and the immediate 5' and 3' flanking regions.

The Protein Coding Regions—The complete nucleotide sequences for 'α' and 'α' are presented in Fig. 4. The protein sequence encoded in the three coding blocks of the 'α' gene agree completely with the published 'α' amino acid sequence (Huisman et al., 1968; Dayhoff, 1972). In particular, the four amino acids Gly 19, Ala 26, Leu 113, and Asn 115 distinguish this globin from its 'α' nonallelic counterpart, while the pres-
Fig. 2. Maps and sequence strategies in the goat \( \alpha \)-globin gene loci. An expansion of the globin-hybridizable regions of clones 218L (top) and 222U (bottom) is shown (Fig. 1 for the overall clone maps). Numbers above and below the clones show fragment sizes, in base pairs, between adjacent restriction sites. The \( \alpha \)-globin genes are denoted by the alternating shaded boxes (exons) and unshaded boxes (introns). The hatched area flanking the genes designates the region of high homology between \( \alpha \) and \( \beta \) that extends beyond the coding region for 122 bp in the 5' direction and 139 bp in the 3' direction. The gene shown at the bottom is a "fusion gene," arising as a propagation artifact in clone 222U via deletion of 3.6 kb of intervening DNA (the region between the dotted lines) normally found in clone 222L. The sequencing strategy is shown below each gene. Solid arrows denote fragments labeled at their 3' ends; the dotted arrow denotes a fragment labeled at its 5' terminus.

Fig. 4. Complete nucleotide sequence of the goat \( \alpha \)- and \( \beta \)-globin genes and their flanking regions. The nucleotide sequences are displayed 5' to 3'. The amino acids encoded by the \( \alpha \) gene are shown immediately above the \( \alpha \) sequence, and are numbered in parentheses below the sequences; only the three amino acid differences found in \( \beta \) relative to \( \alpha \) are shown below the \( \alpha \) sequence. Areas of major nucleotide homology are enclosed within boxes; small areas of homology in the essentially nonhomologous 5' and 3' flanking regions are not boxed. The 5' flanking region is numbered above the sequences, using the presumed cap site as nucleotide 1. Various restriction sites are marked above (for \( \alpha \)) and below (for \( \beta \)) the sequences, to assist in comparison to the maps in Figure 3. A number of interesting sequence features are **overscored** or marked in **bold lines**. These include the CACCT sequence found in adult \( \beta \)-globin genes, the CCAAT box, theATA box, the presumed mRNA cap site, the sequence TTTCGTG presumed to be important in ribosome binding, the AATAAA poly(A) addition signal, and the presumed polyadenylation site (for review see Efstratiadis et al., 1980). Other sequences unique to the goat \( \alpha \)-globin genes are also marked, including the simple sequence (AE)\( _5 \) around position -580 in \( \alpha \), the 5' and 3' boundary sequences, the 7-bp deletion in IVS-I of \( \beta \) (see Fig. 4). This arrangement is consistent with the "slipped mispairing" model for the creation of deletions (Efstratiadis et al., 1980). The second IVS of both \( \alpha \) and \( \beta \) is 103 bp in length, with only one nucleotide difference be-

---

1. D. D. Luse, unpublished observations.
The % G + C and % homology were calculated for each region indicated, using the alignment shown in Fig. 4. Gaps in one sequence relative to another have been ignored in the homology calculations. The 5' and 3' "flanking, homologous" regions extend 5' from the cap site and 3' from the poly(A) site, respectively, to the points where the high homology between 'a' and 'b' ceases (i.e. the end of the sequences enclosed in boxes). The 5' and 3' "flanking, nonhomologous" regions extend from the regions enclosed by boxes further in the 5' and 3' directions, respectively.

<table>
<thead>
<tr>
<th>Region</th>
<th>GC content</th>
<th>homology</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' Flanking, nonhomologous</td>
<td>54</td>
<td>50</td>
</tr>
<tr>
<td>5' Flanking, homologous</td>
<td>79</td>
<td>75</td>
</tr>
<tr>
<td>5' Untranslated</td>
<td>54</td>
<td>49</td>
</tr>
<tr>
<td>Intron I</td>
<td>64</td>
<td>63</td>
</tr>
<tr>
<td>Intron II</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>Intron III</td>
<td>60</td>
<td>61</td>
</tr>
<tr>
<td>3' Untranslated</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>3' Flanking, homologous</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>3' Flanking, nonhomologous</td>
<td>54</td>
<td>59</td>
</tr>
<tr>
<td>Flanking regions</td>
<td>56</td>
<td>61</td>
</tr>
<tr>
<td>Transcribed regions</td>
<td>67</td>
<td>66</td>
</tr>
<tr>
<td>Nonhomologous regions</td>
<td>54</td>
<td>50</td>
</tr>
<tr>
<td>Homologous regions</td>
<td>68</td>
<td>67</td>
</tr>
<tr>
<td>Overall</td>
<td>61</td>
<td>64</td>
</tr>
</tbody>
</table>

Between them. As can be seen in Table I, both intervening sequences are GC-rich (nearly 80% G + C), and each IVS in 'a' is highly homologous to its counterpart in 'b' (79% for IVS-I and 99% for IVS-II). There are both introns contain a number of short GC-rich repeated elements, although we cannot ascribe any functional meaning to this observation; two of these, marked "a" (CCGGGCC) and "b" (GCCCC), are shown in Fig. 4 for illustrative purposes. In addition, the lengths of these two introns are shorter than those reported for other alpha-globin introns, including mouse a1 (122 and 134 bp for IVS-I and IVS-II, respectively; Nishioka and Leder, 1979), human a1 (117 and 141 bp; Leibhaber et al., 1980) and a2 (117 and 132 bp; Leibhaber et al., 1980).

The 5' and 3' Noncoding Regions—As can be seen in Table I and Fig. 4, the 5' and 3' noncoding regions can be divided into two distinct groups. 1) There is a homologous region that is essentially identical in both 'a' and 'b', extending 122 bp 5' from the ATG initiation codon and 129 bp 3' from the TAA stop codon. The 5' region of both genes, containing 122 bp of flanking sequence (96% homology), contains those sequences that have been deemed important in transcription and translation initiation—the CCAAT box, the ATA or Hogness-Goldberg box, the presumed mRNA cap site, the CTTCTCTCTCTC sequence presumed to be involved in ribosome binding, and the ATG initiation codon (for review, see Efstratiadis et al., 1980)—while the 3' region of both genes, containing 139 bp of flanking sequence (98% homology), contains the hexanucleotide AATAAA believed to be a signal for polyadenylation (Proudfoot and Brownlee, 1976), as well as the presumed poly(A) addition site, denoted as shown in Fig. 4 by analogy to other known alpha-globin genes. 2) Although these two genes are highly homologous in the above-mentioned 5' and 3' regions, this homology drops off precipitously in the noncoding regions extending further in both the 5' and 3' directions. In fact, this nonhomologous region begins so abruptly that a clear line of demarcation, or "homology boundary," can be drawn. The appearance of two "islands" of high homology embedded in a "sea" of surrounding nonhomologous sequence is made even more striking in light of the sequence organization at the "homology boundaries." Fig. 5a compares the sequences near the 5' and 3' "homology boundaries" which are delineated by the brackets in Figure 4. Close inspection of these four regions reveals that there is a sequence of 23 bp that is homologous at the boundaries of both genes, i.e. not only are the 5' and 3' boundaries of 'a' homologous to the 5' and 3' boundaries, respectively, of 'b', but more strikingly, the 5' boundaries of both 'a' and 'b' are homologous to the 3' boundary regions of both genes. Note that although the boundary sequence at the 3' end of both genes lies completely within the 3' homologous region just inside the 3' homology boundary, the 5' boundary sequence straddles the 5' homology boundary, with 16 of the 23 bp inside the homologous region and the other 7 bp outside of this region. In spite of this discontinuity at the 5' boundary sequence, the 5' boundary of 'a' is 70% homologous to the 3' boundary of 'b' (see Fig. 5a), which is significantly higher than the 31% homology expected for the comparison of two random sequences corrected for the nucleotide distribution in these boundary regions (Shen et al., 1981). On the other hand, the 5' boundary of 'a', which is obviously related to 'b', is only 43% homologous to the 3' boundary of 'b' due mainly to the discontinuity noted above.

The occurrence of two 23-bp homologous sequences in the vicinity of both ends of two alpha-globin genes demands explanation, not only in terms of postulating a role for these sequences, but also in terms of their preservation during gene duplication.

The nonhomologous region also contains a number of other interesting sequence features that have been marked in Fig. 4. These include: (a) the sequence CACCTCT which appears twice in tandem in 'a' at position -119 to -131, and which has been suggested to be specific to adult beta-globins (Lacy and Maniatis, 1980; Efstratiadis et al., 1980). This sequence also appears in 'b' about 265 bp 3' to its TAA stop codon. (b) Approximately 580 bp 5' to the cap site of 'a' there appears a 30-bp pentadecamer of the simple sequence (AG)(AG)12 of whose bases are perfectly complementary to 11 of the 12 bases of the simple sequence (GT), found immediately after the poly(A) addition site of both genes. This latter sequence is also present as the sequence (TG) at an analogous position at the 3' end of the human alpha-globin gene and as (TG)11 in the large intron of the human y-globin gene; it has been suggested that simple sequences may be associated with the mechanism by which genes are duplicated or converted (Proudfoot and Maniatis, 1980; Slightom et al., 1980; Shen et al., 1981). (c) The 11-bp simple sequence -5' AGAGAGAGAAAG 3'-found 81 bp 3' to the TAA stop codon of 'a' is complementary to the sequence-3'TCTCTCTCTC 5'-found 405 bp 3' to the stop codon of the goat y globin gene. Although the meaning of this observation is unclear, the fact that y is a b-like globin gene (Schon et al., 1981) may implicate these two complementary simple sequences in the coordinate regulation of the a-like and b-like gene clusters.

G + C Content—As is the case with the known sequence of other alpha-globin genes, the entire alpha-globin locus is GC-rich, compared to the average G + C content of 39.5% reported for human DNA and 42.8% for sheep DNA (Sober, 1968). As can be seen in Table I, the GC content of the flanking 5' homologous regions (79% and 75%, respectively, for 'a' and 'b') containing two regions believed to be important for transcription initiation—the CCAAT and ATA boxes—is considerably higher than that of the 5' nonhomologous region (54% and 60% for 'a' and 'b', respectively), the 3' homologous region (64% and 64%), or the 3' nonhomologous region (54% and 64%).

1 R. A. Schon and J. B. Lingrel, unpublished results.
Fig. 5. Boundary sequence homologies. a, comparison of the 5' and 3' sequences of the goat a  and a2 globin genes in the area surrounding the 23-bp boundary sequences. b, comparison of the 5' and 3' sequences of the human a1 and a2 globin genes in the areas analogous to the goat boundary sequences shown in a, c, comparison of the goat 5' and 3' boundary sequence regions to the analogous regions of the human a1 (Liebhaber et al., 1981), a2 (Liebhaber et al., 1980), and mouse a1 (Nishioka and Leder, 1979) genes, and the human a4e1 (Proudfoot and Maniatis, 1980) and mouse a4e3/4e30.5 (Nishioka et al., 1980; Vanin et al., 1980) pseudogenes. Asterisks are placed between homologous pairs of nucleotides when sequences at similar locations (i.e. the 5' flanking regions or the 3' flanking regions) are being compared. Vertical lines are drawn between nucleotides only where all of the nucleotides at that position are identical. Dashed lines above the sequences in (c) denote the extent of the high homology between the two goat a genes; dotted lines denote the approximately 750 bp of a gene sequence, including the ATG start codon and TAA stop codon.

Formation of a Fusion Gene—As noted above, a number of clones were propagated harboring deletions. Close inspection of these clones (213U, 218U, and 222U) indicated that they most likely contained a “fusion” gene, composed of the 5' end of a and the 3' end of a2, with the deletion of intervening DNA arising via the possible mechanisms, for example, of slipped mispairing or unequal crossing over. For this region, the 1.9-kb Bam HI-Eco RI fragment of clone 222U, which hybridized to a mouse a-globin cDNA probe, was subcloned into pBR322 cut with Bam HI and Eco RI, for further mapping and sequencing studies (see Fig. 3).

Both mapping and sequence analysis confirmed that clone 222U indeed contains a fusion gene: the region of about 1.0 kb from the Bam HI site through Exon III is the same as that of a, while the 0.3-kb region extending from around Exon III to the Eco RI site is the same as that of a2. Approximately 3.6 kb of intervening DNA, extending from the 3' end of a to the 3' end of a2, was apparently deleted from this clone to give rise to this fusion gene. Because the sequences of a and a2 are so homologous, it was impossible to pinpoint the exact location of the cross-over between the two genes. However, nucleotide sequencing of the fusion gene (sequence not shown) confirmed the mapping data, and narrowed the cross-over point to a 196-bp region between Asn 115 in a and the G found 5 bp inside the 3' boundary sequence of a2.

This fusion gene resulting from clone propagation in vitro appears similar to cross-over-generated deletions obtained in the propagation of the human a-globin clones (Lauer et al., 1980) and of the closely linked chicken b-like globin clones (Villepontreau and Martinson, 1981). In addition, it is similar to a number of deletions associated with certain b- and a-thalassemias. For example, Baird et al. (1981) have localized the site of cross-over between the human b- and b-globin gene sequences resulting in Lepore Boston globin, a b-beta fusion gene, to a region between the 3' end of Exon II and the first 15 bp of Intron II, while Lepore Hollandia contains a cross-over between codons 22 and 50 and Lepore Baltimore between codons 50 and 86 (Bunn et al., 1977). The identification of fusion genes in a-thalassemias is made more difficult due to the high homologies of the a1 and a2 genes, but the identification of triplicated a-globin loci in humans (Goossens et al., 1980; Higgs et al., 1980; Lie-Injo et al., 1981) indicates that the in vivo formation of such fusion genes most likely does occur, probably via nonhomologous crossing over; these results are also consistent with the deletions associated with a-thalassemia 2 (Embrey et al., 1979) resulting in the formation of a single a-globin gene, rather than the normal two.

Divergence Analysis—We have used the procedures described by Perler et al. (1980) and Efstatiadis et al. (1980) to compare the coding regions of pairs of genes for divergence of
Goat α-Globin Genes

**Table II**

<table>
<thead>
<tr>
<th>Divergence analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected % divergence</td>
</tr>
<tr>
<td>Silent sites</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Goat α/goat βα</td>
</tr>
<tr>
<td>Goat α/human α2</td>
</tr>
<tr>
<td>Goat α/mouse α1</td>
</tr>
<tr>
<td>Goat α/rabbit α</td>
</tr>
<tr>
<td>Goat α/α-chicken α</td>
</tr>
<tr>
<td>Goat α/goat βα</td>
</tr>
<tr>
<td>Human α2/human β</td>
</tr>
<tr>
<td>Mouse α1/mouse βα</td>
</tr>
<tr>
<td>Rabbit α/rabbit β1</td>
</tr>
<tr>
<td>Chicken αα/chicken β</td>
</tr>
</tbody>
</table>

The corrected per cent divergence of the indicated gene pairs is based on codon comparisons as described by Perler et al. (1980). The codons were aligned to maximize homology while at the same time minimizing the number of gaps required to achieve such alignment. Raw divergence values were corrected for multiple base change events as described by Holmqvist (1972) and Kimura and Ohta (1972). Published sequences were used in comparisons involving the mouse α1 (Nishioka and Leder, 1979), human α2 (Liehbaber et al., 1980), rabbit α (Heindell et al., 1978), chicken αα (Richards and Wells, 1980; Dodgson et al., 1981), human β (Lawn et al., 1980), mouse ββ (Konkel et al., 1979), rabbit β (Hardison et al., 1979), and chicken β (Richards et al., 1979). Divergence values from Efstratiadis et al. (1980) differ slightly with the values shown here, most likely due to differences in the choice of alignments. Values above 100% are not a mathematical artifact, but rather reflect the nature of Perler’s (1980) methodology when applied to highly divergent pairs of sequences.

DISCUSSION

We have determined the complete nucleotide sequence of the two nonallelic goat α-globin genes, αа and αα. These two genes are highly homologous to each other (99%) not only in their coding regions, but also in their intervening sequences and their immediate 5’ and 3’ flanking sequences as well, extending 122 bp 5’ to the ATG initiation codon and 139 bp 3’ to the TAA stop codon. However, this homology, which covers approximately 900 bp, and which essentially defines the extent of a gene conversion event, ends abruptly in both the 5’ and 3’ directions beyond these two points, and declines to background homology (~26%) in the remaining sequenced area. Moreover, there appears to be a 23-bp sequence at or near the “homology boundary” of both genes which is homologous at both the 5’ and 3’ borders of each gene (e.g. the 5’ and 3’ boundary sequences of αа are 70% homologous to each other, and those of αα are 43% homologous).

The high overall homology of the two goat α genes, including flanking and intervening sequences, provides strong evidence for the role of gene conversion in maintaining sequence identity between closely linked pairs of genes (for one model of gene conversion, see Zimmer et al., 1980). This mechanism, which has been subsumed under the rubric of coincidental or concerted evolution, appears to be operating on other pairs of globin genes, such as human α1 and α2 (Liehbaber et al., 1981) and αа and αγ (Slighmot et al., 1980), as well as on the two closely linked adult β-globin genes, ββ and ββ, found in the “single” haplotype Hbb mouse (Weaver et al., 1981). Close proximity of two genes, however, does not automatically result in concerted evolution. Both the mouse βpα—βpα pair in the Hbb “diffuse” haplotype (Konkel et al., 1979) and the adult chicken αα—αα gene pair (Dodgson et al., 1981) demonstrate that genes need not correct each other simply because they are near each other, while conversely, gene conversion appears to be operating at two loci for the two pairs of hsp70 heat shock genes in Drosophila located about 500 kb apart (Brown and Ish-Horowicz, 1981).

It is important to differentiate between this 900-bp “conversion unit” and the larger entity denoted by the term “duplication unit.” The conversion unit includes the transcribed region of the α gene, plus flanking sequences extending in the 5’ and 3’ directions to the “homology borders,” which are at, or near, a pair of duplicated 23-bp “boundary sequences.” The mapping data, on the other hand, clearly show that the duplication unit that gave rise to the two nonallelic goat α-globin genes appears today to be 3.6 kb in length (e.g. the distance between the two identically positioned Pst I sites at the 3’ end of each gene—see Fig. 3).

The appearance of a pair of homologous sequences at the borders of both of the goat α-globin genes may have implications for the dispersion of single copy gene sequences throughout the genome, especially in light of the results of Leder et al. (1981) and of Popp et al. (1981) involving the movement of mouse α-like globin genes to different chromosomes. These workers have shown that α-like gene sequences are present on more than one chromosome and have proposed transposition as a likely mechanism of dispersion. In addition, Van Arsdell et al. (1981) have recently shown that three small nuclear RNA pseudogenes dispersed in the human genome are also flanked by three different pairs of direct repeats and have postulated transposition-like insertion events to account for this result. In the case of both these small nuclear RNAs and a mouse αβ-globin pseudogene that has lost both of its intervening sequences (Nishioka et al., 1980; Vanin et al., 1980), a mechanism involving reverse transcription from mRNA to cDNA prior to reinsertion of the cDNA into a new chromosomal locus has been proposed. The appearance, therefore, of the vestiges of a direct repeat at the boundaries of the goat α-globin genes takes on a new significance, since direct repeats are created at the border where an exogenous (transposed) element is inserted into a host segment of DNA (for a review, see Calos and Miller, 1980).

Fig. 6 shows a proposed scenario of events that could have resulted in the duplication of the goat α-globin genes. The most unusual aspect of this model is the role of a transposition-like event for insertion of a sequence containing an ancestral β-globin transcription unit (i.e. not only a β-globin gene, but also, perhaps, the associated 5’ and 3’ flanking sequences necessary for transcription of that gene) into a new chromosomal location, resulting in the formation of a short duplicated sequence—the 23-bp boundary sequence—at the “host-insertion” boundary (Fig. 6, a and b). Subsequent unequal crossing...
important in regulation, at least since, besides containing all the currently identified regions.

Fig. 6. Proposed origin of the duplicated goat α-globin genes. a, initial insertion of the ancestral α-globin sequence into a segment of host DNA, perhaps via a transposition-like mechanism. b, this insertion results in duplication of host DNA at the host-insertion boundary (small open boxes flanking the α sequence). c, unequal crossing-over event at a repeated sequence (open boxes enclosing the letter R) flanking the α-globin sequence. d, the resulting duplicated α-globin gene loci. Note that in this scenario original host DNA sequences marked A and B are also duplicated.

over, perhaps at a repeated sequence (R in Fig. 6) flanking the inserted α sequence, would then result in a chromosome containing two α genes, each having a duplicated unit length equal to the distance between these repeats (Fig. 6, c and d). Note that the 23-bp duplicated repeat could not have been such a cross-over point, since that would have given rise to an uneven number of "daughter" repeats (i.e. one or three), but not the even number of repeats (i.e. four) that we see flanking the two goat α genes. The length of the original "conversion unit" was presumably identical with that of the original "duplication unit." However, as Weaver et al. (1981) have pointed out, subsequent sudden large scale mutations, such as insertions and deletions, which destroy homology at analogously positioned sites in a duplication unit, could cause gene conversion to cease at that position. In other words, later mutations during evolution would "whittle down" the extent of the conversion unit from its original "duplication unit" length to the shorter length we see today. (Note that if such mutation occurred at or near a structural gene—in an intron, for example—conversion would then cease to operate in the gene locus as well; Weaver et al. (1981) see this as a mechanism to explain the high homology of the mouse β'β' pair versus the much less homologous βαβα′′′pair.) In this scenario, the original host DNA sequences, marked A and B in Fig. 6, would also be duplicated. We have not yet searched these flanking regions for the presence of either short repeated elements (i.e. the R sequences) or longer regions of homology (i.e. the A and B flanking regions), although a cursory examination of the flanking sequences shown in Fig. 4 does not reveal the presence of such repeats.

An important consequence of this view of conversion is its implications for gene regulation. Note that both α and βα are "functional" genes. If the conversion unit has shrunk due to ostensibly random mutations outside the two α loci, but the two α genes are still functional, then one could speculate that the conversion unit necessarily still contains those sequences important in regulation, at least so far as transcription is concerned; in other words, the conversion unit has fortuitously become congruent with the "transcription unit" in this particular instance. This idea attains greater significance upon examination of the conversion unit of the goat α-globin genes, since, besides containing all the currently identified regions believed to be important in transcription (i.e. ATA box, cap site, AATAAA hexanucleotide sequence, and poly(A) addition site), it also contains the CCAAT box in the 5′ boundary sequence of these two functional genes (see Figs. 4 and 5a). Thus, the goat α gene loci may provide the first naturally occurring in vivo evidence that the CCAAT box defines the 5′ boundary of sequences required for accurate and efficient transcription, thereby corroborating the results obtained with in vitro deletion experiments (Grosschedl and Birnstiel, 1980; Benoist and Chambon, 1981; Dierks et al., 1981). Note, by the way, that the conversion unit does not contain the sequence CACCCT found around position -120 in the α gene. Of course, other as yet unidentified sequences further upstream to the CCAAT box or downstream from the poly(A) site may still be important in gene transcription, but the fact nevertheless remains that the conversion unit is remarkably congruent with only those regions that contain signals currently known to be important for transcription.

Since the preponderance of evidence points to the origin of the α-globin gene duplication prior to the mammalian radiation (Dayhoff, 1972; Zimmer et al., 1980), one should expect to see evidence of the goat duplication/conversion units in other mammalian species as well. This is an important consideration, especially in light of the fact that, as opposed to the short 900-bp conversion unit in goats, the human α-globin genes share about 1.8 kb of homology in their 5′ regions.

In this regard, the data of Lauer et al. (1980), Proudfoot and Maniatis (1980), and Nishioka and Leder (1979) are instructive. Based on the mapping of HindIII and Pvu II sites, the duplication unit for the DNA segment containing the human α1 and α2 genes is 3.8 kb, which is remarkably close to the 3.6 kb duplication unit we have established for the goat α genes. (The duplication unit encompassing the human α1 and α2 genes is slightly larger—4.3 kb.) Even more to the point are the sequencing data at the 5′ and 3′ ends of the human α-globin genes. Michelson and Orkin (1980) have noted that the entire 3′ untranslated regions of α1 and α2 are unexpectedly divergent, while Proudfoot and Maniatis (1980) have shown that the homology between these two genes extends 15 bp to the poly(A) addition site, ending at a simple sequence, (TG)n, flanked by the pentanucleotide GCCTG. Interestingly, the point where Proudfoot and Maniatis (1980) claim that the human α1 and α2 genes diverge is precisely within the proposed goat 5′ boundary sequence (see Fig. 5c), although we do not believe that this sequence was the cross-over point for duplication of α-globin genes in the goat. Specifically, the pentanucleotide GCCTG, which these authors believe may be associated with the duplication unit that gave rise to both the human and mouse duplicated α-globin genes, is also present in the duplicated goat α-globin genes (sequence GCCTC rather than GCCTG), again at the goat 3′ boundary sequence. Moreover, comparison of the 5′ and 3′ flanking regions of the human α2 and α1 genes shows that, as is the case with the goat α and βα genes, the 5′ flanking sequence shares an unusually large amount of homology with the 3′ flanking sequence, beginning almost exactly at the same place where the analogously positioned goat boundary sequence begins (Fig. 5b). Unfortunately, there is only a limited amount of published 5′ and 3′ sequence for comparing both human α genes, so it is difficult to be sure that the human α loci truly exhibit vestiges of the boundary sequence that is more clearly observable in goats, especially since comparison of the 5′ and 3′ flanking regions of the mouse α1-globin gene (Nishioka and Leder, 1979) reveals only a barely observable vestige of a boundary sequence. However, when the 5′ flanking regions of the correspondingly positioned goat, human, and mouse α genes and pseudogenes are aligned
as shown in Fig. 5c, and the 3′ flanking regions of these species are also similarly aligned, the evidence is highly persuasive that this boundary actually does exist and is common to all three species. Note that in both humans and mice there is a dropoff in the interspecies homology at the same places where the goat α(α/α) homology disappears (the interspecies homology, of course, does not drop off at this boundary, due to the larger conversion units present in the human and mouse α-globin loci relative to the conversion unit in goats).

For example, at the boundary sequences shown in Fig. 5c, goat α is 94% homologous to human α2 at the 5′ boundary and 61% homologous at the 3′ boundary (i.e., the portions of the boundary sequences containing the dotted lines), while outside the boundaries these values drop off to 36% and 33%, respectively. Similarly, goat α is 73% homologous to mouse α1 at the 5′ boundary and 82% at the 3′ boundary, while outside these regions the homology drops off to 14% and 42%, respectively.

We do not think that the evidence of vestiges of this homology boundary shown in Fig. 5c is artifactual. One might argue that any globin gene which required sequences for transcription that extended only from the CCAAT box to the poly(A) addition site would contain homologies to these two end points, but would lose homology rapidly in the regions beyond these points, since no selective pressure could be expected to preserve them in evolution. In other words, could the homology dropoff shown for three species in Fig. 5c actually be evidence of convergent evolution? We believe this is an unlikely explanation, mainly because there are sequences in both the 5′ and 3′ immediate flanking regions of all three species which are also highly homologous, but which do not appear to be required for transcription (e.g., sequences between that CCAAT box and the ATG initiation codon but not including the ATGA box, cap site, or ribosome binding site). In fact, the 5′ flanking region from the ATG initiation codon to the CCAAT box of goat α is 89% homologous to the analogous region of human α2 and 62% homologous to that of mouse α.

This fact thus points to a common origin for the duplicated α-globin genes in humans, mice, and goats, and, by extension, to a common origin for the boundary sequence that is most clearly observable in the goat α gene loci.

It should be noted that this interspecies homology difference at the 5′ end of the human and mouse α-like genes had also been observed by Proudfoot and Maniatis (1980), who noted that the homology dropoff was immediately adjacent to the CCAAT box. The availability of the goat α-globin sequences now gives a new perspective as to why these homology variations are positioned as we see them today, i.e., these variations may not be involved in gene duplication per se, but may be evidence of an original transposition-like event.

Our analysis of the duplicated goat α-globin genes gives new insight into the possible origin and transcriptional control of these genes. First, a strong possibility exists that a transposition-like event was responsible for the insertion of an ancestral α-like sequence into the goat locus as we see it today, as evidenced by the preservation, through gene conversion, of a 23-bp direct repeat flanking the two α genes. Second, vestiges of this insertion event are present in other α-globin genes sequenced to date, implying that this insertion and subsequent gene duplication predated the mammalian radiation. And, lastly, the fact that the conversion unit in goats is so small, extending less than 150 bp beyond the coding regions of the α genes, has been fortuitous, since it suggests the possibility that the converted region contains all the important signals for accurate and efficient transcription, with the CCAAT box defining the 5′ end of such required sequences and the poly(A) addition site defining the 3′ end.

Acknowledgments—We wish to thank F. Fuller and A. Efstratiadis for providing a copy of the computer program to calculate evolutionary divergence, and to H. Chan for use of his Apple II computer. We also wish to thank B. Mach for providing us with the mouse α-globin-containing cDNA plasmid. We are very grateful for the technical assistance of Mr. Donald VanLeeuwen, Ms. Gail Sears, Ms. Jane Selegue, Mr. Stephan Glasser, and Mr. Michael Amur. We are also grateful to Dr. Steven G. Shapiro for extremely helpful discussions and to Ms. Rose Alden for expert assistance in the preparation of this manuscript.

REFERENCES


Dayhoff, M. O., ed (1972) Atlas of Protein Sequence and Structure, Vol. 5, The National Biomedical Research Foundation, Silver Spring, MD


Konkel, D. A., Maizel, V. J., Jr., and Leder, P. (1979) Cell 18, 865-873


Maniatis, T., Jeffrey, A., and Van de Sande, J. (1975) Biochemistry...
Gene conversion of two functional goat alpha-globin genes preserves only minimal flanking sequences.
E A Schon, S M Wernke and J B Lingrel


Access the most updated version of this article at http://www.jbc.org/content/257/12/6825

Alerts:
• When this article is cited
• When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 0 references, 0 of which can be accessed free at http://www.jbc.org/content/257/12/6825.full.html#ref-list-1