Ruminant globin gene structures suggest an evolutionary role for Alu-type repeats

John C.Schimenti and Craig H.Duncan*

Room IDR 704, Children's Hospital Research Foundation, Elland and Bethesda Avenues, and University of Cincinnati Medical School, Graduate Program in Developmental Biology, Cincinnati, OH 45229, USA

Received 24 October 1983; Revised and Accepted 12 December 1983

ABSTRACT

Bovine fetal and adult globin genes were cloned and subjected to DNA sequence analysis. Both of these genes contained insertions of Alu-type repetitive DNA within their introns. Comparison of cow and goat β -type globin genes indicates that intragenic DNA insertions played a role in their evolution. These data support the theory that Alu-type repeats maintain genetic diversity by inhibiting gene conversion.

INTRODUCTION

Mammalian genomes contain abundant, short, interspersed, repetitive DNA sequences (1, for reviews see 2,3). Such a collection was first described in the human genome and named the "Alu family" (4). Subsequent studies found sequences which hybridize to Alu family DNA in the DNA of all other primates tested (5-7). Analogous but distinct DNA elements, referred to as Alu-type repeats, have been found in animals representing non-primate mammalian orders. Rodent species including Chinese hamster (8), mouse (9-11), and rat (12) share two repetitive DNA families, named B1 and B2. Alu family members resemble a tandem repeat of the B1 sequence, but appear unrelated to B2.

Two artiodactyl genomes, from the goat (13) and the cow (14), contain indistinguishable families of short repeats which share a 40 bp region of homology to Alu family, the remainder of the sequence being unrelated to either rodent or primate repeats. Rabbits, from the order Lagomorpha, also possess a family of short repeats in chromosomal DNA (15). Published sequence from one member of this family shows no detectable homology to any of the above DNA elements (16), except for a 10 bp RNA Polymerase III promoter sequence.

Bell et al. (17) suggested that Alu family sequences are transposable elements, accounting for their widespread dispersal in the human genome. This has been confirmed by the discovery of Alu-type DNA insertions in tandemly repeated arrays of DNA sequences (18,19). Weiner et al. (20) and

Jagadeeswaran et al. (21) proposed that RNA Polymerase III transcripts of Alu-type sequences (22-25) function as intermediates in transposition events, a cDNA copy being inserted via a staggered break into genomic DNA. Most recently, Miyake et al. (26) suggested that Alu family sequences can form a mobile composite with another, longer repetitive sequence element known as Kpn (27).

Many functional roles have been ascribed to Alu-type elements, including origins of DNA replication (1), end points for duplication events (28), modulators of chromatin structure (25), hot spots for recombination (29) or catalysts of chromosome rearrangement (30). There is little evidence to substantiate any of these hypotheses. We have focused our studies on a possible evolutionary role for Alu-type repeats. To this end we have cloned β -type globin genes from the bovine genome and subjected them to sequence analysis. Comparison of these data with similar data from the goat globin genes (13) indicate that Alu-type repeats were involved in the evolution of these globin gene complexes. The data provide strong support for the hypothesis of Hess et al. (31) that Alu-type repeats stabilize gene structure and maintain diversity by inhibiting gene conversion.

MATERIALS AND METHODS

DNA

The genes described in this report were isolated from a bovine cosmid library, constructed in this laboratory. The methods of construction, clone isolation and mapping will be described in a future publication. The plasmid pGy5, a fetal goat globin probe obtained from P. Liberator and J. Lingrel, was used as a radiolabelled probe in this procedure.

DNA Sequence Analyses

Fragments of DNA containing globin genes were cloned in M13 mp10 and M13 mp11 vectors (32). ss DNAs were used as sequencing templates according to the dideoxy termination method of Sanger et al. (33). Both strands of the DNA fragments were sequenced, except for several short regions which gave completely unambiguous readings on a single strand.

The DNA sequences were analyzed on a Basis microcomputer, using the programs described by Fristensky et al. (34).

RESULTS

In situ hybridization of a bovine library to a goat fetal globin probe allowed us to isolate a set of cosmid clones containing globin genes. When

γ gene
 -226 TTTAGTAGCAGTTTGTATTGCTGGAATGACTGAGACCTAGAGATGCCCAAGAAAGAGGGCTGAAGGTTTAA
 -225 TTTAATAGCAATTTGTATTGCTGGAATGACTGTGATCTAGAGATGCCCAAGAAAGAGGGCTGATGGTCTAA
 β gene
 -156 AGTCAGTCCAGGAAGACTAAGGAGAGGTATGACTGTCACTTCAAGCCTCACCTGTGGAACCAACC
 -155 AGTCAGTCCAGGAAGACCAAGGAGAGCTATGACTATCATCGTTCAAGCCTCACCTGTGGAACCAACC
 -86 TTGGCCTGAGCCAATCTGCTCACAGAAGCAGGGAGGTGAGGAGGCAGGGCTAAGCATAAAAGGAAGAGCT
 -85 TTGGCATGAGC-AATCTGCTCACAGAAGCAGGGAGGGCAGGAGGCAGGGCTGGGCATAAAAGGAAGAGCT
 cap Me
 -16 GGGCCAGCTGCTGCTTACACTTGCTTCTGACACAACCGTGTTCACTAGCAACTACACAAACAGACCCAT
 -16 GGGCCAGCTGCTGCTTACACTTGCTTCTGACACAACCGTGTTCACTAGCAACTACACAAACAGACCCAT Me
 tLeuSerAlaGluGluLysAlaAlaValThrSerLeuPheAlaLysValLysValAspGluValGlyGly
 55 GCTGAGCGCTGAGGAGAAGGCTGCCGTCACTCCCTATTGCCAAGGTGAAAGTGGATGAAGTTGGTGGT
 55 GCTGACTGCTGAGGAGAAGGCTGCCGTCAACCGCTTTTGGGCAAGGTGAAAGTGGATGAAGTTGGTGGT
 tLeuThrAlaGluGluLysAlaAlaValThrAlaPheTrpGlyLysValLysValAspGluValGlyGly
 GluAlaLeuGlyAr
 125 GAGGCCCTGGCAGGTAGGTATCCTGCTTACAAGACAGGTTTCTGCTGCTGCTGCTAAGTCGCTTCAGT
 125 GAGGCCCTGGCAGGTAGGTATCCCACTTACAAGGCAGGTTT-----
 GluAlaLeuGlyAr
 195 CGTGTCCGACTCTGTCCGACCCATAGACGGCAGCCATCAGGCTTCAACCGTCCCTGGGATTCTCCAGGC

 265 AAGAACATTGGAGTGGGTGCCATTTCTTCTCCAATGCATGAAAGTGAAAGTGAAAGTGAAGTCGCTC

 335 AGTCATGTCCCATGGACTGCAGCGATCCCATGGACTGCGCCTTCCAGGCTCCTCCGTCCATGGGATTTT

 405 CCAGGCAAGAGTACCGGAGTGGGTGCCATTGCCTTCTCCGACAAGACAGGTTTAAGGAGAGTGAAATGC
 -----AAGGAGAGTGAAATGC
 475 ACCTGGGCGTGTGAGGACAAAGCCGTCCTGAGATTCTGAAAGCTGCTGACTTCCTCTGACCTTGTGCTG
 183 ACCTGGGCGTGTGAGGACAGAGCCGTCCTGAGATTGAGAGAGCTGCTGGCTTCTCTGACCTTGTGCTG
 gLeuLeuValValTyrProTrpThrGlnArgPhePheGluSerPheGlyAspLeuS
 545 TTTTCTCCCCCTAGGCTGCTGTTGTCTACCCCTGGA CTGAGGTTCTTTGAGTCCTTTGGGGACTTGT
 253 TTTTCTCCCCCTAGGCTGCTGTTGTCTACCCCTGGA CTGAGGTTCTTTGAGTCCTTTGGGGACTTGT
 gLeuLeuValValTyrProTrpThrGlnArgPhePheGluSerPheGlyAspLeuS
 erSerAlaAspAlaIleLeuGlyAsnProLysValLysAlaHisGlyLysLysValLeuAspSerPheCy
 615 CCTCTGCCGATGCCATTTTGGGAAACCTAAGGTGAAGGCCCATGGCAAGAGGTGCTGGACTCCTTCTG
 323 CCACTGCTGATGCTTTATGAACAACCTAAGGTGAAGGCCCATGGCAAGAGGTGCTAGATTCTTTAG
 erThrAlaAspAlaValMetAsnAsnProLysValLysAlaHisGlyLysLysValLeuAspSerPheSe
 sGluGlyLeuLysGlnLeuAspAspLeuLysGlyAlaPheAlaSerLeuSerGluLeuHisCysAspLys
 685 TGAGGGCCTGAAGCAA CTGATGACCTCAAGGGTGCCTTTCCTTCGCTGAGTGAGCTGCACTGTGATAAG
 393 TAATGCGATGAAGCATCTCGATGACCTCAAGGGCCTTTCCTTCGCTGAGTGAGCTGCACTGTGATAAG
 rAsnGlyMetLysHisLeuAspAspLeuLysGlyThrPheAlaAlaLeuSerGluLeuHisCysAspLys

γ LeuHisValAspProGluAsnPheArg
755 CTGCACGTGGATCCTGAGAACTTCAGGGTGAGTTTGTGGAATCCTCAATATTCTC-----TTTTTATG
463 CTGCATGTGGATCCTGAGAACTTCAAGGTGAGTTTGTGGAATCCTCAGTGTTCTCCTTCTTTTATG
 β LeuHisValAspProGluAsnPheLys
818 GTCAAGCTTGTGT CATGGGAGAGGGCTGA CTGCCAGGACGTAGTTTAGAATGGAGAAGATGTATTCTGG
533 GTCAAGCTCATGT CATGAGGAGAAAGCTGAATGGCAGGACACAGTTTAGAATGGAGAAGAGGTATTCTGG
888 TTAAAGTGCTAAGGACTCCTCAGAACCATTTAGATTCTTTTAATCTCTTGCTCAATAATCATTTCTT
603 TTAGATTACTAAGGACTCCTCAGAACCCTTAGACTCTTTTAACCTCTTGTTCAACCAAGTATTCTT
958 CTGATTCATTCTTGTTCTCTGTTGTCTGCAATGTCTTCTCTTTTAATTATACITCTT-----GAGTGTT
673 CTGATTCATTCTTGTTCTCTGTTGTCTGCAATGTC--CTCTTTTAATTATATTTTATTCTTGGGGT
1023 CAATTTGAAAAAAGATTTCTTCATCTACTTTAAACATAATATCTAATATTTTCCCCTTATCTGTTCTT
741 TAATTTGAAAAAAT--TATATATCAACTTTAAAAATTGTATCTAATATTTCCCCTTATCTGTTCTT
1093 TCAAGGGATAAAATGTGTATTGCTTTTGAAGTGATTCAAAATAATAAAAAATGATAACAAGTTCTGGAT
811 TCAAGGAATAAGATGTTCTATTGCTTTTGAATGATTCAAAATAATAAAAAATATAACAAGTTCTGGAT
1163 TAAGATAGAAAGAAAGAAACATTTCTAAATATAAATTCAAGCTGCTATGGGTACCTTCATATCAGCAGTA
881 TAAGTTAGAAAGAGAGAAACATTTCTAAATATATATTCCGGGAAGATATAGGTAGATTCAATCAGTAGTA
1233 GCATCTATACTTCAGTCATCTTTGTGCTAATACCTAGGGGCAAGCTTGGGATGAATCTGAAGTACTCT
951 ACAACTTCATTCAGTCATCTTTGTGCTTATA-TCTACGGTCACAGCTTGGGATAAGACTGAAATACCT
1303 GAATCTAACCTGGG-----
1020 GAATCTAACCTTGATTTCCCTCATAGCTCAGTTGGTTAAGCATCTGCCTGCAATGCAAGAGATCCAGT
GOAT β GAATCTAACCTTGGACTTCTCTCATAGCTCGGTTGGTAAAGAGTCTGCCTGCAATGCAAGGATCCAGT

1090 TCGATTCTGGGTGGGAA--GGATGGCTGGAGAAGGGATAGGC-ACCCACTCCAATATTCTTGGGTTTC
GOAT β TCGATTCTGGGTGAGGAAAGATGGCTGGAGAAGGGATAGGCTACCCACTCCAATATTCTTGTGCTTC

1157 CCTTGTGGCTCAGCTGGTAAAGAATCTGCCTGCAATGTGGGAGACCCAGCTTCT---ATCCCTGAGTTTG
GOAT β CCTTGTGGCTCAGCTGGTAAAGAATCTGCCTGCAATGCGGGAGACCTGGCTTCTCTATCCATGGGTTGG

1224 GAAGATCCCTGGAGAAGGGAAAGGCTACCCA CTCCAATATTCTGGCTGGAGAAATCTATGGA CTGTAG
GOAT β GAAGATCCCTGGAGAAGGGAAAGGCTACCTCTCCAATATTCTGGCTGGAGAAATCCATGGA CTGTAT

1294 AGTC-ATGGGGTTGCAAGAAATCAGACAGATTGAGAGACTCTCACTTCACTCACCTGCACTAACCTGC
GOAT β AGTCCATAGGGTTGCAAGAGT CAGACATGACTGAGCAACTTCACTTTACTAACCTGCACTAACCTGC

LeuLeuGlyAsnValLeuValValValLeuAlaArgArgPheG1
1335 CCTTGCTTAATGTCTTTTCCACACAGCTCCTGGGCAACGTGCTGGTGGTGTGCTGGCTCGCCGCTTTGG
1363 CCTTGCTTAATGTCTTTTCCACACAGCTCCTGGGCAACGTGCTAGTGGTGTGCTGGCTCGCAATTTTGG

LeuLeuGlyAsnValLeuValValValLeuAlaArgAsnPheG1
ySerGluPheSerProGluLeuGlnAlaSerPheGlnLysValValThrGlyValAlaAsnAlaLeuAla
1405 CAGTGAATTCTCCCCGAGCTGCAGGCTAGCTTTCAAGAGTGGTGACTGGTGTGGCCAATGCCCTGGCC
1433 CAAGGAATTCAACCCGGTGCTGCAGGCTGACTTTCAAGAGTGGTGGCTGGTGTGGCCAATGCCCTGGCC
yLysGluPheThrProValLeuGlnAlaAspPheGlnLysValValAlaGlyValAlaAsnAlaLeuAla

```

γ      HisArgTyrHisTer
1475   CACAGATATCACTAAGCTCCCCTTCCTGATTTTCAGGAAAGGCTTTTCATCCTCAGAGCCCAAAACTG
1503   CACAGATATCATTAAGCTCCCTTCCTGCTTCCAGGAAAGGTTTTTTCATCCTCAGAGCCCAAGATTG
β      HisArgTyrHisTer

1545   AATATGGAAAAATTATGAAGCGTTTTGTGCATCTTGCCTCTGCCTAATAAAGACATTTATTTTCATTGCA
1573   AATATGGAAAAATTATGAAGTGTTTTGAGCATCTGGCCTCTGCCTAATAAAGACATTTATTTTCATTGCA

1615   CTGGTGTATTTAAATTATTTCACTGTCTCTTACTCAGATGGGCACACGGGAGGGCAAAGCACTGAAGACA
1643   CTGGTGTATTTAAATTATTTCACTGTCTCTTACTCAGATGGGCACATGGGAGGGCAAACACTGAAGACA

1685   TAAAGAAATGAAGGCCTAAGTTGACACTTTGAGAAAATACATCAGTATCTTGGACCCCATGACAGGAGTG
1713   TAAAGAAATGAAGGGCTA-GTCGAGACCTTGAGAAAATATATCAGTATCTTGGACCCCATGACAGGAGTG

1755   GTTGTACACAGCTGATGTTATTGAAAAACAGGCTCTGCTCCTTACTCTTACTTTCCTTTAAAGAATTC
1782   GTTGAAATAGCTGATGTTAT-GGAAAAACAGGCTTTGCTCCTTAGCCTTACTCTCCCTTAAAGAATTC

```

Figure 1. Sequences of bovine fetal (γ) and adult (β) globin genes. The γ gene is shown on the top lines, while the β gene is on the bottom lines. Coding sequences are marked with the three letter amino acid code. Blanks have been inserted to maximise alignment between the two sequences. Locations of Alu-type insertions are indicated by long stretches of blanks. Duplicated oligonucleotides flanking the γ gene insertion are marked by underlines. The sequence of the Goat β^A gene in the region corresponding to nucleotides #1020-1363 of the bovine β gene is included for comparison.

the probe was hybridized to these purified cosmid DNAs, two distinct regions showed intense binding. Other DNA regions exhibited lesser hybridization signals, indicating the existence of a family composed of related genes or pseudogenes. The global arrangement and structure of these sequences are under study. The present report is limited to the two intensely hybridizing regions. This DNA was cleaved with appropriate restriction enzymes and the resulting fragments subcloned in M13 vectors for sequence analysis.

The sequence data showed that (Fig. 1) one of these regions contained the bovine fetal (γ) globin gene while the other contained the bovine adult (β) globin gene. Both of these genes contained two introns in locations analogous to those of all other studied mammalian globin genes. The exons predicted amino-acid sequences which were in complete agreement with those deduced for bovine γ globin protein (35). The predicted amino acid sequence for the β globin was substantially in agreement with that deduced for bovine β globin type A chain (36). However, amino acid residues 55 and 72, assigned to aspartic acid (in Fig. 5 of Ref. 36), are predicted as asparagine in the DNA sequence. The order of amino acid residues of tryptic peptide T14b (see Fig. 5, Ref. 36) was later revised (see Fig. 5, Ref. 37). The DNA sequence agrees with the revised ordering for this peptide.

Both of these bovine globin genes contained insertions of repetitive DNA

Figure 2. Structures of ruminant globin genes. The five genes discussed in this paper are shown. Introns are marked by thinner horizontal lines while exons are marked by thicker lines. Locations and orientations of intragenic insertions are indicated. Structures of goat globin genes were taken from Schon et al. (13).

within their introns. In this respect, they were similar to goat globin genes and distinct from other mammalian globin genes. The repetitive DNAs were inserted at varying positions in ruminant globin genes (Fig. 2). The bovine β and caprine β^A or β^C gene insertions shared the same location near the 3' end of the second intron. This indicated that these three genes were all derived from a common precursor in the cow-goat ancestor and that this ancestral gene already contained the inserted Alu-type sequence within its second intron. This point was illustrated by a comparison between the goat β^A and cow β genes through this region (Fig. 1, bases 1020-1369 of the β sequence). The insertions started and stopped at the same nucleotide in the two genes, and their sequences had the same degree of homology (88%) as did the rest of the second introns. In contrast, the fetal genes contained independent insertions, the goat γ insertion at the 5' region of the second intron, while the cow γ insertion was in the first intron. These insertions occurred after the divergence of the cow and goat, because the two γ genes derived from the same precursor in the cow-goat ancestor. The best evidence for this was the presence of identical 5 bp deletions (base no. 1014 of the γ sequence in Fig. 1) in both cow and goat γ genes, relative to β genes. Consideration of the overall homology between these genes (Table I) also supports this conclusion.

DNA insertions within the bovine globin genes showed a high degree of sequence homology to repetitive DNA elements elsewhere in the bovine genome. Watanabe et al. (14) presented the sequence of a 650 base pair region, situated approximately 2 kb 5' to the bovine corticotropin- β -lipotropin precursor gene. This sequence contained three copies of a 120 bp consensus sequence. The first and second copies of the 120 bp repeat were joined by a tandem triplet of the sequence CACTTT, while the second and third copies were

Table I. Ruminant globin gene sequences were compared for divergence. Alu-type insertions were removed from the sequences before comparison. The sequences were also compared for nucleotide changes which cause alterations in the amino acid structures of the encoded proteins as compared to those which are silent at the amino acid level. When two nucleotide changes were involved in the alteration of a single amino acid, both were counted.

<u>% DIVERGENCE</u>			
<u>Overall</u>	(1768 bp)	<u>Introns</u>	(718 bp)
Cow β /Cow γ	9.3% (165)	Cow β /Cow γ	11.3% (81)
Goat β^A /Goat γ	10.3% (182)	Goat β^A /Goat γ	12.0% (86)
Cow γ /Goat γ	5.7% (100)	Cow γ /Goat γ	5.4% (39)
Cow β /Goat β^A	6.8% (121)	Cow β /Goat β^A	7.2% (52)
<u>Coding Blocks</u>	(438 bp)	<u>Flanking Regions</u>	(612 bp)
Cow β /Cow γ	9.6% (42)	Cow β /Cow γ	6.9% (42)
Goat β^A /Goat γ	11.0% (48)	Goat β^A /Goat γ	7.8% (48)
Cow γ /Goat γ	5.7% (25)	Cow γ /Goat γ	5.9% (36)
Cow β /Goat β^A	5.9% (26)	Cow β /Goat β^A	7.0% (43)
<u>CODING BLOCK ALTERATIONS</u>			
	<u>Changes resulting in amino acid substitution</u>		<u>Changes silent for amino acids</u>
Cow β /Cow γ	32		10
Goat β^A /Goat γ	39		9
Cow γ /Goat γ	15		10
Cow β /Goat β^A	16		10

joined by a 100 bp sequence which began with (AGC)₇ (see line b, Fig. 2, Ref. 14). These authors demonstrated that there were 10⁵ copies of this sequence in the bovine genome, including three other locations in the vicinity of the bovine corticotropin- β -lipotropin precursor gene.

The bovine globin gene insertions were also members of this repetitive sequence family. This point was illustrated by a comparison of these sequences with the region 2 kb 5' to the bovine corticotropin- β -lipotropin precursor gene (Figs. 3&4). The γ gene insertion matched closely the region containing the first and second copies of the 120 bp repeat, including (CACTTT)₃ and (AGC)₅. The β gene insertion was more complex, consisting of an

Figure 3. Schematic structure of bovine globin gene insertions. The bottom line represents the 600 base repeat sequence 5' to the bovine corticotropin- β -lipotropin precursor gene, with the raised portions designating the 120 bp consensus repeat units. The bold numbers below the triangles correspond to those used in line b, Fig. 2, of Ref. 14. The upper lines and arrows represent the globin gene insertions, which have been aligned for maximum homology with the lower sequence. The smaller numbers used here correspond to those of Fig. 1. Note that both upper and lower portions of the S-shaped β insertion were homologous to the region from bases #350-460 in the sequence of Ref. 14. The γ gene insertion was 291 bp long while the β gene insertion was 310 bp long.

imperfect duplication of 80 bp linking the second and third 120 bp repeats, as well as the third repeat itself (Fig. 4). The first copy of this duplication was, however, truncated at the 3' end while the second copy contained a 18 bp deletion within the 120 bp repeat (Fig. 3). In addition, the orientation of the repeat units with respect to gene coding sequence was opposite in the two globin genes. Alu-type insertions usually are flanked by tandem oligonucleotide repeats of pre-existing genomic DNA at the insertion sites (17). The sequence data (Fig. 1) demonstrated that the γ gene insertion was flanked by a 13 bp repeat. This oligonucleotide appeared only once within the matching region of the β globin gene. The β gene insertion, in contrast, was not flanked by recognizable oligonucleotide repeats.

Because of the close evolutionary relationship between cows and goats, we compared the nucleotide sequences of their pristine globin genes, that is, genes from which insertion elements had been removed. This analysis (Table 1) illustrated the exceptional degree of sequence conservation which has been maintained in these species. The close homology was not limited to regions of DNA which code for amino acids. The introns also matched very well throughout their lengths. As discussed later, the selective pressures operating on these genes were illuminated by an additional analysis on the 438 bp of coding sequence to distinguish changes which result in amino acid substitutions from those which are cryptic for protein structure (Table 1).

DISCUSSION

Ruminant globin genes form an excellent system for comparative studies of molecular evolution. Caprine globin gene loci have been extensively

γ GENE INSERTION SEQUENCE

167 GCTGCTGCTGCTGCTAAGTCGTTCACTCGTGTCCGACTCTGTGCGACCCCATAGACGGCAGCCC
 335 GCTGCTGCTGCTGCTAAGTCGTTCACTCGTGTCCGATTCTGTGCGACTCCATAGATTGCAGCCC
 REPEAT 5' TO CORTICOTROPIN-β-LIPOTROPIN GENE

232 ATCAGGCTTACCGTCCCTGGGATTCTCCAGGCAAGAACATTGGAGTGGGTTGCCATTTCTTCT
 270 ACCAGGCTCCGCCATCCCTGGGATTCTCCAGGCAAGAACACTGGAGTGGGTTGCCATTTCTTCT

297 CCAATGCATGAAAGTGAAAAGTGAAAAGTGAAGTCGCTCAGTCATGTCCCATGGACTGCAGCGAT
 205 TCAATGCATGAAAGTGAAAAGTGAAAGGGAAGTCGCTCAGTCATGTTCCGACTCTTAGAGAC----

362 CCCATGGACTGC-GCCTTCCAGGCTCCTCCGTCCATGGGATTTTCCAGGCAAGACTACCGGAGTG
 144 CCCATGGACTCGGGCTACCAAGGCTCCTCCGTCCGTGGGATTTTCCAGGCAAGACTACTAGAGTG

426 GGGTGCATTGCTTCTCCG (ACAAGACAGGTTT)
 79 GGGTGCATTGCTATTCCTCA

β GENE INSERTION SEQUENCE

1034 ATTTCCCTCATAGCTCAGTTGGTTAAGCATCTGCCTGCAATGCAAGAGATCCAGTTCGATTCTT
 344 CCGGGTGGCTCAGACAGTAAAGCGTCTGTCTACAATGTGAGAGACCTGGGTTGATCCCT
 REPEAT 5' TO CORTICOTROPIN-β-LIPOTROPIN GENE

1099 GGGTCGGAAGGATGGCTGGAGAAGGGATAGGCA-CCCACTCCAGTATTCTTGGGTTTCCCTTGT
 404 GGGTTGGGAAGATTCCCTGGAGAAGGAAATGGCAACCCACTCCAGTACTCTTGCCTT 460

1163 GGCTCAGCTGGTAAAGAATCTGCCTGCAATGTGGGAGCCAGCTTCTATCCCTGAGTTTGAAG
 350 GGCTCAGACAGTAAAGCGTCTGTCTACAATGTGAGAGACTGGGTTTATCCCTGGTTTGGGAAG

1228 ATCCCTCGAGAAGGAAAAGGCTACCACTCCAGTATTCTGGCTGGAGAAATCTATGGACTGTA
 415 ATTCCCTCGAGAAGGAAATGGCAACCCACTCCAGTACTCTTGCCTTGAAAATCCCATGGATGGAA

1293 GAGT-----CATGGGTTGCAAAGAATCAGACACGATTGAGAGACTCTCACT
 480 GAGCTTGGTGCAGGCTACTATCCATGGGGTCACAAAGAGTCGGGCACGACTGAGCGACT-TACA

1340 TCACT
 544 TTACA

Figure 4. Sequence comparison of repetitive DNA insertions within bovine globin gene introns with DNA sequences 5' to the bovine corticotropin-β-lipotropin precursor gene. The bovine γ and β globin gene insertion DNA sequences were fitted to the sequence described (14) by using the computer programs of Fristensky et al. (34). The upper lines represent the globin insertion sequences according to the numbering system of Fig. 1 while the lower lines use the numbering system of Watanabe et al. (line b, Fig. 2, Ref. 14). Note that the Watanabe et al. sequence is reversed and complemented in the comparison with the γ gene insertion. The 120 bp consensus unit repeat regions are marked by underlines and overlines. The 13 bp repeated oligonucleotide which flanks the γ gene insertion is set off by parentheses.

studied by Lingrel and coworkers (13,38-43). The ovine globin genes have been characterized to a much lesser degree (44). The present work is the initial report on bovine globin gene structures. The fossil record suggests a common ancestor for these species as recently as 18-20 million years ago (45), a

figure which is corroborated by comparison of myoglobin amino acid sequences (46) and satellite DNA sequences (47). Despite this close relationship, patterns of bovine and caprine globin gene expression differ substantially. Goats produce three non-embryonic β -type globin chains. The earliest, fetal (γ) globin, is supplanted by juvenile (β^G) globin synthesis just before birth. During the first year of life juvenile globin declines and is replaced by adult (β^A) globin production (48,49). The cow, in contrast, makes only two non-embryonic β -type globins. In this case fetal (γ) globin synthesis is replaced directly by adult (β) globin synthesis, just before birth (50-53).

It is unlikely, however, that the varying locations of repetitive intragenic DNA insertions are responsible for those distinct patterns of gene expression. The γ globin genes differ most in insertion structure and location, but their temporal expression periods are identical, lasting from late embryonic life to just before birth in both species. The β genes, which do differ in temporal expression, have identical insertion locations.

It is equally unlikely that the multiple insertions of Alu-type repeats described here arose as chance events in a manner free of selective pressure. The probability of that occurrence can be estimated by assuming that ruminant genomes, like the human genome (54), contain one Alu-type repeat per 4200 bp. The target size of the fetal globin gene introns is 718 bp. Therefore, the odds of independent insertions within the goat and cow γ globin genes are about $(718 \div 4200)^2 = 3\%$. Actually the odds are reduced further because these insertions have occurred since the divergence of goat and cow 20 MYr ago while Alu-type sequences have certainly been spreading through ruminant genomes for a longer period. Furthermore, these low odds can be coupled to the 17% $(718 \div 4200)$ chance for an Alu-type insertion in the β globin gene of the cow-goat ancestor. Thus, it is necessary to conclude that selective pressure favored those ruminants which by chance received insertions in their globin genes.

In accordance with the ideas of Hess et al. (31), we propose that this selective force was the inhibition of gene conversion, resulting in the maintenance of diversity between γ and β globin genes. Gene conversion, first described in yeast and fungi, involves the unilateral transfer of sequence information between two similar but non-identical genes. The process is usually visualized as a gene correction or master slave process whereby a divergent duplicated gene is changed back to an exact replica of its original. Recent studies have demonstrated that gene conversion is an active process in several mammalian gene systems, including immunoglobulin genes (55-58), histocompatibility genes (59-60), globin genes (61-66) and possibly sn RNA

genes (67). The best characterized examples are in human globin gene loci. The fetal globin genes, γ^G and γ^A , arose via duplication of a 5kb DNA sequence about 34 MYr ago. Since then regions flanking the genes have diverged by 14% while a 1.5 kb region, including the first and second exons plus the first and most of the second intron, is virtually identical. Statistical analysis of these sequences led to the conclusion that a gene conversion event occurred at this locus about 1 MYr ago (61,62). The human α_1 and α_2 globin genes are part of a 4 kb duplication which has also undergone gene conversion (63-65). Sequence analysis of this region led to the original proposal (31) that Alu sequences, by inserting regions of non-homology in duplication blocks, inhibit gene conversion.

This principle clearly applies to ruminant globin genes. These gene clusters were formed by duplication events which resulted in 4-6 kb stretches of homologous DNA flanking and encompassing the globin genes (12, 37, 41-44). Subsequent to this duplication, fetal and adult globins assumed their distinct oxygen binding capacities by accumulating amino acid differences over evolutionary time. These correspond to single nucleotide alterations scattered through the exons. The selective pressure to accumulate these alterations is evident from an analysis of protein coding sequences within the ruminant globin genes (Table I). Each pairwise comparison showed 9 or 10 changes which were silent at the amino acid level. Nucleotide alterations which resulted in amino acid substitution showed a different pattern. Comparison of cow γ to goat γ or cow β to goat β^A gave 15 or 16 such alterations. In contrast, the comparison of cow β to cow γ or goat β to goat γ showed more than 30 alterations in each case. It is apparent that the maintenance of this sequence diversity over evolutionary time is desirable. A gene conversion event could at one stroke potentially eliminate many of these useful nucleotide differences; conversely, inhibition of gene conversion at certain loci would be an adaptive process. The different locations and structures of inserted DNA between fetal and adult genes break up the long stretches of homology which are conducive to gene conversion. The goat juvenile (β^C) and adult (β^A) insertion structures can also be explained by this theory. The only significant difference between them is a 60 bp deletion within the β^C insertion (13). Again, it is logical that this deletion was favored by natural selection because it interfered with gene conversion between the β^A and β^C genes.

If Alu-type insertions inhibit gene conversion within a localized region of DNA, then sequences proximal to insertions should diverge more rapidly than

distal sequences. This prediction is verified by sequence homology comparisons in regions proximal or distal to insertion elements in ruminant globin genes. Coding blocks and introns are more divergent than flanking sequences for both cow and goat β/γ gene pairs (Table I). Studies are underway to extend these observations further into flanking DNA regions.

A deletion in one allele of a diploid cell should have the same inhibitory effect on gene conversion which we have attributed to the insertions in the bovine globin genes. Lawrence et al. found that a deletion, cycl-1, of the yeast cytochrome c gene did inhibit gene conversion frequencies by a factor of five (68). However, deletions in another yeast gene, HIS4, had no effect on the frequency of gene conversion (69). At present it is difficult to ascertain which of these results can be generalized to other yeast genes, or if gene conversion in yeast is truly a good model for gene conversion in mammalian cells.

Alu-type elements can be viewed as genetic fixatives. By random insertion into duplicated regions of DNA they protect diverging sequences from the homogenizing effects of gene conversions. This concept will be important to understanding the forces which led to present mammalian genome structure. It is probable that intergenic DNA regions involved in gene regulation are subject to selective pressures as stringent as those operating on the genes themselves. This implies that regulatory regions will often be found in conjunction with Alu-type insertions, even though the insertions per se have no role in gene control. Such an insertion could protect a neighboring regulatory region from gene conversion. Conversely, a random insertion into a favorable genomic location could initiate the formation of a regulatory sequence during subsequent evolution. A possible example of this occurs in the human β -type globin locus, where two Alu family elements are found in a region thought to be essential for the fetal to adult globin switch (70).

Ruminant globin gene loci, which have undergone extensive changes in recent evolutionary time, form an ideal system in which to test these ideas. Structural analysis of intergenic Alu-type repeats and neighboring DNA regions may help to pinpoint genomic structures which are involved in the temporal regulation of gene expression.

ACKNOWLEDGMENTS

The authors thank M. Morabito, S. Spence, R. Ganschow, and R. Palmer for advice and discussion. Technical assistance from J. Eichhold, and M.

Loescher's secretarial skills were greatly appreciated.

This research was supported by the following grants to C.H.D.: from the American Cancer Society, #NP-409; from March of Dimes Birth Defects Foundation, Basil O'Connor Starter Research Grant #5-354; and from National Institutes of Health, # HL15996-11 to the Comprehensive Sickle Cell Center of Cincinnati.

*To whom reprint requests should be sent

REFERENCES

1. Jelinek, W.R., Toomey, T.P., Leinwand, L., Duncan, C.H., Choudary, P.V., Biro, P.A., Weissman, S.M., Rubin, C.M., Houck, C.M., Deininger, P.L. and Schmid, C.W. (1980) *Proc. Natl. Acad. Sci. USA* 77, 1398-1402.
2. Jelinek, W.R. and Schmid, C.W. (1982) *Ann. Rev. Biochem.* 51, 813-844.
3. Schmid, C.W. and Jelinek, W.R. (1982) *Science* 216, 1065-1070.
4. Houck, C.M., Rinehart, F.P. and Schmid, C.W. (1979) *J. Mol. Biol.* 132, 289-306.
5. Grimaldi, G., Queen, C. and Singer, M.F. (1981) *Nucleic Acids Res.* 9, 5553-5569.
6. Dhruva, B.R., Schenk, T. and Subramanian, K.N. (1980) *Proc. Natl. Acad. Sci. USA* 77, 4514-4518.
7. Houck, C.M. and Schmid, C.W. (1981) *J. Mol. Evol.* 17, 148-155.
8. Haynes, S.R., Toomey, T.P., Leinwand, L. and Jelinek, W.R. (1981) *Mol. Cell. Biol.* 1, 573-583.
9. Krayev, A.S., Kramerov, D.A., Skryabin, K.G., Ryskov, A.P., Bayev, A.A. and Georgiev, G.P. (1980) *Nucleic Acids Res.* 8, 1201-1215.
10. Krayev, A.S., Markusheva, T.V., Kramerov, D.A., Ryskov, A.P., Skryabin, K.G., Bayev, A.A. and Georgiev, G.P. (1982) *Nucleic Acids Res.* 10, 7461-7475.
11. Kalb, V.F., Glasser, S., King, D. and Lingrel, J. (1983) *Nucleic Acids Res.* 11, 2177-2187.
12. Page, G.S., Smith, S. and Goodman, H.M. (1981) *Nucleic Acids Res.* 9, 2087-2104.
13. Schon, E.A., Cleary, M.L., Haynes, J.R. and Lingrel, J.B. (1981) *Cell* 27, 359-369.
14. Watanabe, Y., Tsukada, T., Notake, M., Nakanishi, S. and Numa, S. (1982) *Nucleic Acids Res.* 10, 1459-1469.
15. Shen, C.-K.J. and Maniatis, T. (1980) *Cell* 19, 379-391.
16. Suske, G., Wenz, M., Cato, A.C.B. and Beato, M. (1983) *Nucleic Acids Res.* 11, 2257-2271.
17. Bell, G.I., Pictet, R. and Rutter, W.J. (1980) *Nucleic Acids Res.* 8, 4091-4109.
18. Grimaldi, G. and Singer, M.F. (1982) *Proc. Natl. Acad. Sci. USA* 79, 1497-1500.
19. Kominami, R., Muramatsu, M. and Moriwaki, K. (1983) *Nature* 301, 87-91.
20. Van Arsdell, S.W., Denison, R.A., Bernstein, L.B., Weiner, A.M., Manser, M. and Gesteland, R.F. (1981) *Cell* 26, 11-17.
21. Jagadeeswaran, P., Forget, B.G., Weissman, S.M. (1981) *Cell* 26, 141.
22. Duncan, C.H., Biro, P.A., Choudary, P.V., Elder, J.T., Wang, R.R.C., Forget, B.G., deRiel, J.K. and Weissman, S.M. (1979) *Proc. Natl. Acad. Sci. USA* 76, 5095-5099.

23. Di Segni, G., Carrara, G., Tocchini-Valentini, G.R., Shoulders, C.C. and Baralle, F.E. (1981) *Nucleic Acids Res.* 9, 6709-6722.
24. Fuhrman, S.A., Deininger, P.L., LaPorte, P., Friedmann, T. and Geiduschek, E.P. (1981) *Nucleic Acids Res.* 9, 4333-4350.
25. Duncan, C.H., Jagadeeswaran, P., Wang, R.R.C. and Weissman, S.M. (1981) *Gene* 13, 185-196.
26. Miyake, T., Migita, K. and Sakaki, K. (1983) *Nucleic Acids Res.* 11, 6837-6846.
27. Shafit-Zagardo, B., Maio, J.J. and Brown, F.L. (1982) *Nucleic Acids Res.* 10, 3175-3193.
28. Barsh, G.S., Seeburg, P.H. and Gelinas, R.E. (1983) *Nucleic Acids Res.* 11, 3939-3958.
29. Jagadeeswaran, P., Tuan, D., Forget, B.G. and Weissman, S.M. (1982) *Nature* 296, 469-470.
30. Calabretta, B., Robberson, D.L., Barrera-Saldana, H.A., Lambrou, T.P. and Saunders, G.F. (1982) *Nature* 296, 219-225.
31. Hess, J.F., Fox, M., Schmid, C. and Shen, C.-K.J. (1983) *Proc. Natl. Acad. Sci. USA* 80, 5970-5974.
32. Messing, J., Crea, R. and Seeburg, P.H. (1981) *Nucleic Acids Res.* 9, 309-321.
33. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
34. Fristensky, B., Lis, J. and Wu, R. (1982) *Nucleic Acids Res.* 10, 6451-6463.
35. Babin, D.R., Schroeder, W.A., Shelton, J.R., Shelton, J.B. and Robberson, B. (1966) *Biochemistry* 9, 1297-1310.
36. Schroeder, W.A., Shelton, J.R., Shelton, J.B., Robberson, B. and Babin, D.R. (1967) *Arch. Biochem. Biophys.* 120, 124-135.
37. Schroeder, W.A., Shelton, J.R., Shelton, J.B., Apell, G., Huisman, T.H.J., Smith, L.L., and Carr, W.R. (1972) *Arch. Biochem. Biophys.* 152, 222-232.
38. Haynes, J.R., Rosteck, P., Jr. and Lingrel, J.B. (1980) *Proc. Natl. Acad. Sci. USA* 77, 7127-7131.
39. Cleary, M.L., Haynes, J.R., Schon, E.A. and Lingrel, J.B. (1980) *Nucleic Acids Res.* 8, 4791-4802.
40. Robbins, J., Rosteck, P., Jr., Haynes, J.R., Freyer, G., Cleary, M.L., Kalter, H.D., Smith, K. and Lingrel, J.B. (1979) *J. Biol. Chem.* 254, 6187-6195.
41. Cleary, M.L., Schon, E.A. and Lingrel, J.B. (1981) *Cell* 26, 181-190.
42. Shapiro, S.G., Schon, E.A., Townes, T.M. and Lingrel, J.B. (1983) *J. Mol. Biol.* 169, 31-52.
43. Lingrel, J.B., Haynes, J.R., Schon, E.A., Cleary, M.L. and Gallagher, P.M. (1981) in Organization and Expression of Globin Genes, Stamatoyannopoulos, G. and Nienhuis, A.W. Eds., pp. 59-68, Alan R. Liss, Inc., New York.
44. Kretschmer, P.J., Coon, H.C., Davis, A., Harrison, M. and Nienhuis, A.W. (1981) in Organization and Expression of Globin Genes, Stamatoyannopoulos, G. and Nienhuis, A.W., Eds., pp. 59-68, Alan R. Liss, Inc., New York.
45. Gentry, A.W. (1978) in Evolution of African Mammals, Maglio, V.J., and Cooke, H.B.S., Eds., pp. 540-572, Harvard University Press, Cambridge, MA.
46. Romero-Herrera, A.E., Lehmann, H., Joysey, K.A. and Friday, A.E. (1973) *Nature* 246, 389-395.
47. Reisner, A.H. and Bucholtz, C.A. (1983) *EMBO J.* 2, 1145-1149.
48. Kitchen, H. and Brett, I. (1974) *Ann. N.Y. Acad. Sci.* 241, 653-671.
49. Huisman, T.H.J. (1974) *Ann. N.Y. Acad. Sci.* 241, 549-555.

-
50. Nijhof, W., Ronda, G.J. and Jonxis, J.H.P. (1972) *Biol. Neonate* 21, 120-129.
 51. Gabuzda, T.G., Silver, R.K., Chui, L.C. and Lewis, H.B. (1970) *Br. J. Haematol.* 19, 621-633.
 52. Congote, L.F. and Solomon, S. (1978) *Biochemistry* 17, 1160-1165.
 53. Lavrijsen, K. and Verwilghen, R.L. (1983) *Hemoglobin* 7, 159-179.
 54. Rinehart, F.P., Ritch, T.G., Deininger, P.L. and Schmid, C.W. (1981) *Biochemistry* 20, 3003-3010.
 55. Ollo, R. and Rougeon, F. (1983) *Cell* 32, 515-523.
 56. Schreier, P.H., Bothwell, A.L.M., Mueller-Hill, B. and Baltimore, D. (1981) *Proc. Natl. Acad. Sci. USA* 78, 4495-4499.
 57. Krawinkel, U., Zobebelein, G., Bruggemann, M., Radbruch, A. and Rajewsky, K. (1983) *Proc. Natl. Acad. Sci. USA* 80, 4997-5001.
 58. Bentley, D.L. and Rabbitts, T.H. (1983) *Cell* 32, 181-189.
 59. Weiss, E.H., Mellor, A., Golden, L., Fahrner, K., Simpson, E., Hurst, J. and Flavell, R.A. (1983) *Nature* 301, 671-674.
 60. Pease, L.R., Schulze, D.H., Pfaffenbach, G.M. and Nathenson, S.G. (1983) *Proc. Natl. Acad. Sci. USA* 80, 242-246.
 61. Slightom, J.L., Blechl, A.E. and Smithies, O. (1980) *Cell* 21, 627-638.
 62. Shen, S., Slightom, J.L. and Smithies, O. (1981) *Cell* 26, 191-203.
 63. Liebhaber, S.A., Goossens, M. and Kan, Y.W. (1981) *Nature* 290, 26-29.
 64. Lauer, J., Shen, C.K.J. and Maniatis, T. (1980) *Cell* 20, 119-130.
 65. Zimmer, E.A., Martin, S.L., Beverly, S.M., Kan, Y.W. and Wilson, A.C. (1980) *Proc. Natl. Acad. Sci. USA* 77, 2158-2162.
 66. Schon, E.A., Wernke, S.M. and Lingrel, J.B. (1982) *J. Biol. Chem.* 257, 6825-6835.
 67. Weiner, A.M. and Denison, R.A. (1982) *C.S.H. Symp.* 47 (P2), 1141-1150.
 68. Lawrence, C.W., Sherman, F., Jackson, M. and Gilmore, R.A. (1975) *Genetics* 81, 615-629.
 69. Fink, G.R. and Styles, C.A. (1974) *Genetics* 77, 231-244.
 70. Maeda, N., Bliska, J.B. and Smithies, O. (1983) *Proc. Natl. Acad. Sci. USA* 80, 5012-5016.