The EMBO Journal Vol.2 No.8 pp.1339–1343, 1983

The primary structure of the duck α^D-globin gene: an unusual 5′ splice junction sequence

Cebrail Erbil* and Jürgen Niessing

Physiologisch-Chemisches Institut I, Lahnberge, D-3550 Marburg, FRG

Communicated by D. Gallwitz
Received on 25 April 1983

The complete nucleotide sequence of the duck minor α^D-globin gene including the flanking regions has been determined. A unique structural feature of the α^D-globin gene is a GC instead of the invariant GT dinucleotide at the 5′ end of the second intervening sequence. The 1013 base pair long gene has otherwise all the characteristics normally attributed to a functional globin gene. Indirect evidence suggests that the α^D-globin gene is expressed *in vivo*.

Key words: IVS 2 5′ splice site/GT to GC transition

Introduction

Three α-like globin chains, designated π, α^D and α^A are expressed during avian development. At the early embryonic stage the π-globin polypeptide is made while the minor α^D-globin chain and the major α^A-globin chain are synthesized throughout late embryonic and adult life (Brown and Ingram, 1974).

Recently, we isolated and partially characterized an α-globin gene-containing recombinant phage, DαG-1, from a duck (*Cairina moschata*) DNA recombinant library (Niessing *et al.*, 1982). These results and more recent studies (Erbil and Niessing, in preparation) revealed that phage DαG-1 carries the three linked α-like globin genes. Their transcriptional orientation is 5′-π-α^D-α^A-3′, and the individual genes are separated by 2.0–2.2 kb of DNA. The complete nucleotide sequences of the duck α^A-globin gene (Erbil and Niessing, 1982) and the π-globin gene (Erbil and Niessing, in preparation) have been determined.

Here we present the entire primary structure of the minor α^D-globin gene. A unique feature of this gene is a T to C transition at intervening sequence 2 (IVS 2) which destroys the GT dinucleotide invariant at 5′ splice sites (Breathnach and Chambon, 1981; Mount, 1982). Accordingly, the Breathnach-Chambon rule, which states that all introns start with GT and end with AG is violated by the α^D-globin gene. There is no reason to think that this globin gene represents a pseudogene (for review see Little, 1982).

Results

DNA sequencing

The recombinant bacteriophage DαG-1, which carries the three α-like duck globin genes was isolated from a lambda library of duck (*C. moschata*) DNA (Niessing *et al.*, 1982). A 2.1-kb *Bgl*II restriction fragment and a 4.4-kb *Bam*HI restriction fragment, both containing the entire α^D-globin gene, were derived from DαG-1 and subcloned into the plasmid pBR322 for sequencing. The restriction map of the α^D-globin gene, together with the approach used to determine the gene sequence including 100 nucleotides of the 5′-flanking region and 72 nucleotides of the 3′-flanking region, is shown in Figure 1. The entire sequence of both DNA strands was determined by the Maxam and Gilbert (1980) procedure and the sequences at all restriction sites were overlapped.

The 5′- and 3′-non-coding and flanking regions

The complete nucleotide sequence of the duck α^D-globin gene is shown in Figure 2. The sequenced region begins 100 bp 5′ to the capping site and extends over the 1013 bp long α^D-globin gene, ending 72 bp 3′ to the poly(A) addition site. The 5′ and 3′ borders of the α^D-globin gene were determined by S1 mapping (Berk and Sharp, 1977; Weaver and Weissmann, 1979) as previously described for the α^A-globin gene (Erbil and Niessing, 1982). The capping site was localized by hybridization of total duck globin mRNA to a 5′ ^{32}P-labeled minus strand of a fragment extending from the *Eco*RI site at amino acid position 23 to the *Pst*I site in the 5′-flanking region. S1 nuclease-protected fragments were electrophoresed along with a sequencing ladder of the end-labelled *Eco*RI-*Pst*I fragment (results not shown). The poly(A) addition site was determined by the same S1 mapping procedure using a 3′ ^{32}P-labeled minus strand of a 219-bp *Ava*I-*Hae*III fragment (see Figure 1).

The 5′-non-coding region starts 42 bp 5′ of the translated initiation codon ATG. As in all globin mRNAs studied so far the capping site is characterized by the dinucleotide AC (Efstratiadis *et al.*, 1980). The 42 bp 5′-non-coding region of the α^D-globin gene exceeds the size of the duck α^A-globin gene 5′-non-coding region by six nucleotides and is 58% homologous to it in the region of overlap.

Highly conserved sequences, which are found in most eucaryotic genes, are also retained in the 5′-flanking region of the duck α^D-globin gene: the TATAA box (Goldberg, 1979; Breathnach and Chambon, 1981) 30 bp upstream of the mRNA cap site and the pentanucleotide CCACC which might represent an equivalent of the CCAAT box (Efstratiadis *et al.*, 1980) is located 80 bp 5′ of the capping site. A similar sequence (CCAGC) is also found in the α^A-globin gene 5′-flanking region at an identical position relative to the capping site (Erbil and Niessing, 1982).

In the 46 bp long 3′-non-coding region (not including the termination codon TGA) the putative polyadenylation signal AATAAA (Proudfoot and Brownlee, 1976) is present 14 bp 5′ to the poly(A) addition site. The complete 3′-non-coding region is strictly conserved in the duck *C. moschata* α^D-globin gene sequence (Figure 2) and in the Peking duck α^D-globin cDNA sequence (Frankis and Paddock, 1982).

The protein coding region

Figure 2 shows the nucleotide sequence of the duck α^D-globin gene and the deduced α^D-globin amino acid sequence. When compared with the chicken α^D-globin (Dodgson *et al.*, 1981) there are 26 amino acid changes (underlined in Figure 2) which represent a homology of 81.5%. A similar degree of homology (87%) was obtained in a comparison of the major α^A-globin of chicken and duck (Erbil and Niessing, 1982).

The α^D-globin amino acid sequence has also been determined from cloned cDNAs of another duck strain, the Pek-

*To whom reprint requests should be sent.

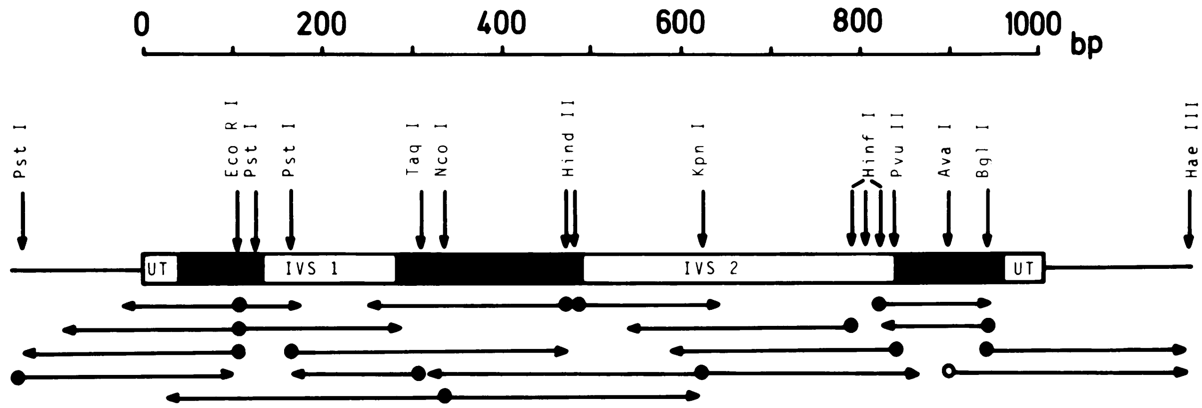

Fig. 1. Restriction enzyme map and sequencing strategy used to determine the DNA sequence of the duck α^D-globin gene. Only those restriction enzyme sites used in deriving the sequence are indicated. The direction of transcription is 5′ to 3′ from left to right. The mRNA coding region of the gene (filled boxes) and intervening sequences (open boxes) as well as the 5′- and 3′-non-coding sequences (open boxes, UT) are indicated. The horizontal arrows below the map denote the regions of the DNA that were sequenced. Restriction fragments were labeled at their 5′ ends (filled circles) or 3′ ends (open circle).

```
                                                        -100                -80
                                                   5'   CTGCGTGGCCTCAGCCCCTCCACCCCTCCACGCTGATAAGATAA
                               -30                       -1  capping site                          +42
GGCCAGGGCGGGAGCGCAGGGTGCTATAAGAGCTCGGCCCCGCGGGTGTCTCCACC ACAGAAACCCGTCAGTTGCCAGCCTGCCACGCCGCTGCCGCC
 1                                   10                                       20
met leu thr ala glu asp lys lys leu ile val gln val trp glu lys val ala gly his gln glu glu phe gly
ATG CTG ACC GCC GAG GAC AAG AAG CTC ATC GTG CAG GTG TGG GAG AAG GTG GCT GGC CAC CAG GAG GAA TTC GGA
                 30
ser glu ala leu gln arg
AGT GAA GCT CTG CAG AG  GTGTGGGCTGGGCCCAGGGGGCACTCACAGGGTGGGCAGCAGGGAGCAGGAGCCCTGCAGCGGGTGTGGGCTGGG

                                                                                  met phe leu ala tyr
ACCCAGAGCGCCACGGGGTGCGGGCTGAGATGGGCAAAGCAGCAGGGCACCAAAACTGACTGGCCTCGCTCCGGCAG G ATG TTC CTC GCC TAC
             40                              50                                60
pro gln thr lys thr tyr phe pro his phe asp leu his pro gly ser glu gln val arg gly his gly lys lys
CCC CAG ACC AAG ACC TAC TTC CCC CAC TTC GAC CTG CAT CCC GGC TCT GAA CAG GTC CGT GGC CAT GGC AAG AAA
                                 70                             80
val ala ala ala leu gly asn ala val lys ser leu asp asn leu ser gln ala leu ser glu leu ser asn leu
GTG GCG GCT GCC CTG GGC AAT GCC GTG AAG AGC CTG GAC AAC CTC AGC CAG GCC CTG TCT GAG CTC AGC AAC CTG
             90
his ala tyr asn leu arg val asp pro val asn phe lys
CAT GCC TAC AAC CTG CGT GTT GAC CCT GTC AAC TTC AAG GCAAGCGGGGACTAGGGTCCTTGGGTCTGGGGGTCTGAGGGTGTGGG

GTGCAGGGTCTGGGGGTCCAGGGGTCTGAGTTTCCTGGGGTCTGGCAGTCCTGGGGGCTGAGGGCCAGGGTCCTGTGGTCTTGGGTACCAGGGTCCTGG

GGGCCAGCAGCCAGACAGCAGGGGCTGGGATTGCATCTGGGATGTGGGCCAGAGGCTGGGATTGTGTTTGGAATGGGAGCTGGGCAGGGGCTAGGGCCA

GGGTGGGGGACTCAGGGCCTCAGGGGGACTCGGGGGGGGACTGAGGGAGACTCAGGGCCATCTGTCCGGAGCAGGGGTACTAAGCCCTGGTTTGCCTTG
    100                                 110                                 120
    leu leu ala gln cys phe gln val val leu ala ala his leu gly lys asp tyr ser pro glu met his ala
CAG CTG CTG GCA CAG TGC TTC CAG GTG GTG CTG GCC GCA CAC CTG GGC AAA GAC TAC AGC CCC GAG ATG CAT GCT
                    130                                 140
ala phe asp lys phe leu ser ala val ala ala val leu ala glu lys tyr arg   TERM.
GCC TTT GAC AAG TTC TTG TCC GCC GTG GCT GCC GTG CTG GCT GAA AAG TAC AGA TGA GCCACTGCCTGCACCCTTGCACC

TTCAATAAAGACACCATTACCAC AGCTCTGTGTCTGTGTGTGCTGGGACTGGGCATCGGGGGTCCCAGGGAGGGCTGGGTTGCTTCCACACATCC  3'
                        ▲poly(A) addition site
```

Fig. 2. Nucleotide sequence of the duck α^D-globin gene. The nucleotide sequence of the plus strand of the gene is displayed in the 5′ to 3′ orientation. The deduced amino acid sequence is shown on a line above the coding sequence. Differences in the amino acid sequence of the duck and chicken α^D-globin are underlined. The unusual C residue at IVS 2 position 2 is marked by a vertical arrow. The long horizontal arrows at the IVS 1 5′ splice junction and within IVS 1 show the location of a repeat sequence (see also Figure 4). Putative regulatory sequences ('CCACC', TATAA, AATAAA) are underlined, the location of the mRNA capping site and the polyadenylation site are indicated.

ing duck (Ben Tahar and Scherrer, in preparation; Frankis and Paddock, in preparation). There are only three amino acid differences between the α^D-globin of *C. moschata* and the Peking duck: at position 11 (Val→Thr), position 13 (Val→Leu) and at position 129 (Leu→Met). Thus, in the α^D-globin gene of *C. moschata* essential amino acids are conserved and there are no replacements at important functional sites.

Intervening sequences

In all the vertebrate globin genes studied thus far the coding region is interrupted by two intervening sequences. In

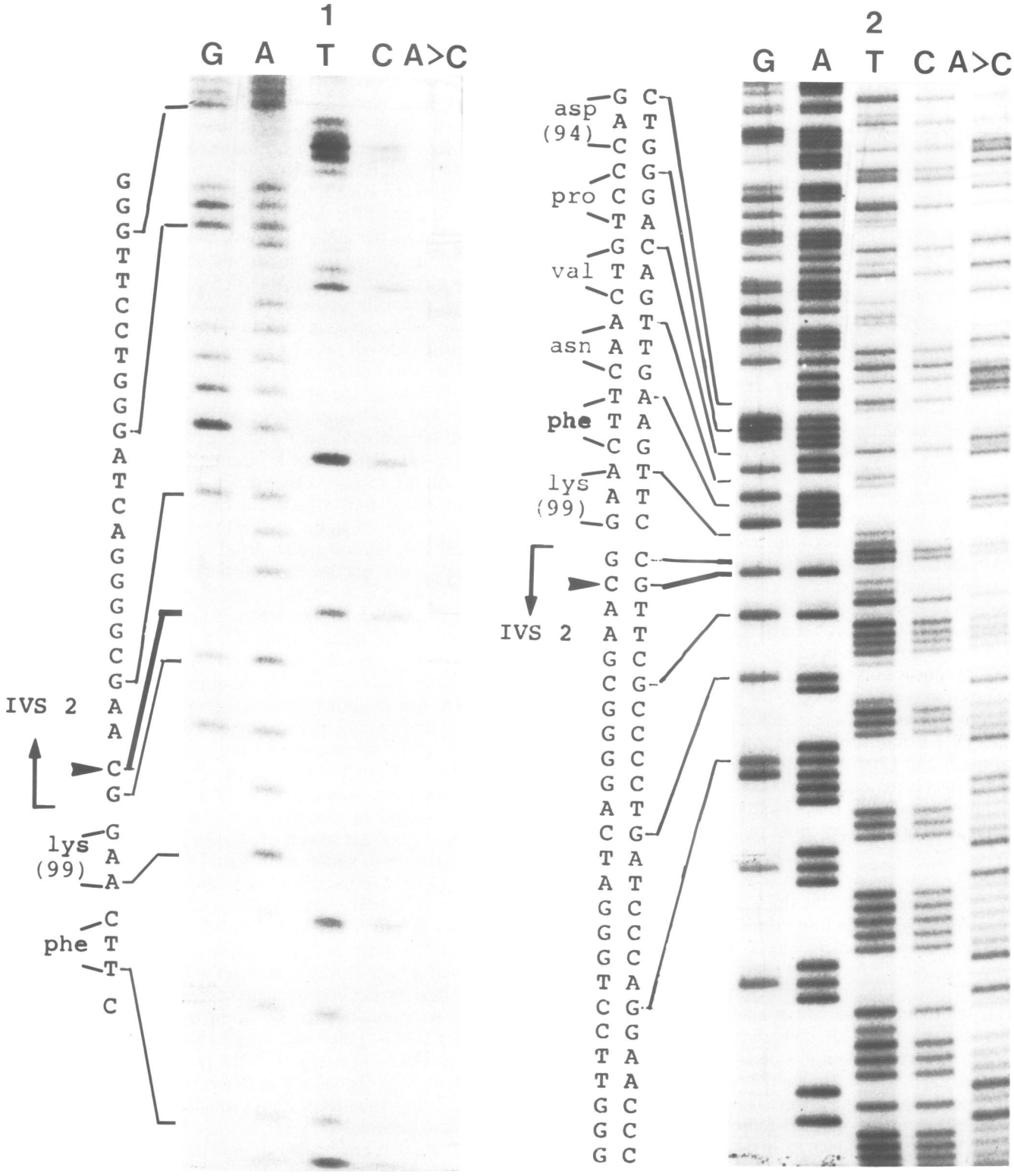

Fig. 3. Transition of the conserved GT dinucleotide to GC at the IVS 2 5′ splice site. Sequencing ladders of the plus strand **(left)** and minus strand **(right)** DNA covering the region around the IVS 2 5′ splice site are shown. The C residue is indicated by a horizontal arrow.

the human α2-globin gene (Liebhaber *et al.*, 1980), human ζ-globin gene (Proudfoot *et al.*, 1982), mouse α-globin (Nishioka and Leder, 1979), chicken α^A- and α^D-globin gene (Dodgson *et al.*, 1981) as well as in the duck α^A-globin gene (Erbil and Niessing, 1982) and duck π-globin gene (Erbil and Niessing, in preparation) IVS 1 consistently is localized within codon 31 while IVS 2 is always found between codon 99 and 100. In all cases, codon 31 is arginine, codon 99 is lysine and in five out of six α-like globin genes, codon 100 is leucine. As can be seen from Figure 2, the locations of both introns in the α^D-globin gene are identical to those found in all other α-like globin genes: the 152 bp long IVS 1 is found within codon 31 (arginine) and the 347 bp long IVS 2 between codon 99 (lysine) and 100 (leucine).

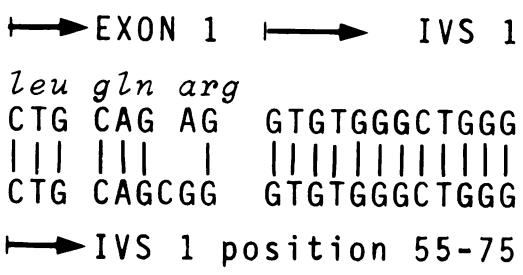

Fig. 4. Comparison of a repeat sequence identified at the exon 1-IVS 1 boundary and within IVS 1 (see figure 2).

Inspection of the second intervening sequence of the α^D-globin gene reveals a surprising finding: at the 5′ end of IVS 2, the invariant GT dinucleotide is converted into a GC, thereby violating the GT-AG rule (Breathnach and Chambon, 1981) (see Figure 2). Figure 3 shows the sequencing ladder of both DNA strands covering the region of interest. The DNA sequences of the minus and plus strands both clearly show that the 5′ end of IVS 2 starts with the dinucleotide GC. The same 5′ splice junction sequence was determined in a different α^D-globin gene subclone derived from DαG-1. This is the only sequence abnormality within the entire gene region.

Another interesting structural feature of the IVS 1 exon-intron boundary is an almost perfect duplication of the 20 bp long splice junction sequence encompassing the last eight nucleotides of exon 1 and the first 12 nucleotides of IVS 1 (see Figures 2 and 4). In the duplicated sequence, which is located 55 nucleotides downstream of the 5′ splice junction, an additional C residue is inserted and an A to G transition is located at the position corresponding to the first nucleotide of codon 31 (Figure 3). Despite these changes, the duplicated sequence within IVS 1 would seem to be a perfect 5′ splice junction (Breathnach and Chambon, 1981; Mount, 1982). This finding raises interesting questions related to the as yet unknown mechanisms responsible for the selection of a particular splice site while leaving other potential splice sites inactive.

Discussion

The DNA sequence of the duck α^D-globin gene including the regulatory signals for transcriptional and translational initiation and termination as well as the deduced α^D-globin amino acid sequence are clearly compatible with normal function. The T to C transition at IVS 2 position 2, which destroys the GT dinucleotide invariant at all 5′ splice sites sequenced so far, appears to be of considerable interest and is unprecedented. A recent compilation of 139 5′ splice site boundaries and 130 3′ splice site boundaries clearly shows that – among functional genes – there is no exception to the GT-AG rule (Breathnach and Chambon, 1981) according to which all intervening sequences invariably start with GT and end with AG.

Two types of specific β^o-thalassaemic DNA lesions are caused by a G to A transition of the conserved GT dinucleotide at the 5′ splice site of IVS 1 (Orkin *et al.*, 1982) and IVS 2 (Treisman *et al.*, 1982). In the latter case, the β^o-gene transcript has been shown to be abnormally spliced. One form of α-thalassaemia is associated with a pentanucleotide deletion at IVS 1 position 2–6 (Orkin *et al.*, 1981) which leads to the inactivation of the functional splicing site (Felber *et al.*, 1982). Moreover, splicing is completely abolished upon site-specific mutagenesis leading to a GT to GG change at the 5′ splice site of an adenovirus IVS (Montell *et al.*, 1982), or after a GT to AT transition at IVS 2 position 1 of the rabbit β-globin gene (Wieringa *et al.*, 1983). Yeast actin gene intron deletions including the T residue of the conserved GT dinucleotide also lead to the production of unspliced actin mRNA (Gallwitz, 1982). All these studies show that the integrity of the invariant GT dinucleotide is absolutely required for efficient RNA splicing. It has not yet been demonstrated, however, whether a transition of the invariant GT to GC, as in the case of the duck α^D-globin gene, will lead to an inactivation of the authentic splice site. In this context, it is interesting to note that mutagenic obliteration of the 5′ splice site of rabbit β-globin gene IVS 2 leads to the activation of three normally unused splice sites. In one of these cryptic splice sites, splicing did not occur 5′ to the usual GT dinucleotide but rather 5′ to a GC sequence (Wieringa *et al.*, 1983). This finding suggests that, in principle, a GC doublet might replace the 'invariant' GT dinucleotide without concomitant inactivation of the authentic splice site.

Additional evidence for the use of a GC instead of the conserved GT dinucleotide is inferred from the alternative splicing of the mouse αA-crystallin gene transcript (King and Piatigorsky, 1983). It is of importance, therefore, to establish whether the α^D-globin gene is expressed *in vivo*. Attempts to reach a conclusion are hindered by the fact that globin RNA samples are not available from the individual duck from which the DNA recombinant library had been prepared. Moreover, we do not know whether the α^D-globin gene represents an allelic variant of a normal counterpart. Finally, we cannot exclude the unlikely possibility that the point mutation had been introduced during cloning of the genomic DNA. On the other hand, we know (Niessing, unpublished data) that normal levels of mature α^D-globin mRNA exist in duck erythroblasts and that the molar ratio of α^A-, α^D- and β-globin chains is the same in erythrocytes from duck and chicken (Brown and Ingram, 1974).

In the human β-globin gene cluster, specific thalassaemic mutant genes are strongly linked to patterns of restriction site polymorphism within and around the mutant gene (Orkin *et al.*, 1982). We have used four different restriction enzymes (*Eco*RI, *Bam*HI, *Hind*III, *Bgl*II) for genomic blotting experiments in which DNAs isolated from six unrelated duck individuals were compared with that of the recombinant bacteriophage D αG-1. With all the restriction enzymes used identical restriction fragments carrying the α^D- and α^A-globin gene were obtained for the six different DNA samples as well as for D αG-1 (Niessing, unpublished data). Thus, at least for these restriction enzymes, DNA polymorphism in the region of the α^D-globin gene is not detectable. In addition, the α^D- and α^A-globin gene very probably occur only once per haploid genome.

Thus, it is reasonable to assume that the α^D-globin gene represents a functional globin gene. Analysis of the RNA splicing pattern after introduction of the α^D-globin gene into HeLa cells will help to answer this question.

Materials and methods

Materials

Restriction enzymes, *Escherichia coli* DNA polymerase I (Klenow fragment), T4 polynucleotide kinase, T4 DNA ligase, S1 nuclease and calf intestine phosphatase were purchased from Boehringer, BRL or New England Biolabs. [γ-^{32}P]ATP (3000 Ci/mmol) and [α-^{32}P]dTTP (3000 Ci/mmol) were obtained from Amersham.

Gene isolation and sequencing

The α^D-globin gene was isolated from the recombinant bacteriophage D αG-1 (Niessing *et al.* 1982) either as a 2.1-kb *Bgl*II fragment or as a 4.4-kb *Bam*HI fragment and subcloned in pBR322. DNA fragments were 5′ end-labeled using T4 polynucleotide kinase or 3′ end-labeled by filling in 5′ overhangs using the Klenow fragment of *E. coli* DNA polymerase I. After strand separation or secondary restriction enzyme cleavage, the end-labeled fragments were sequenced according to Maxam and Gilbert (1980) and analysed on thin (0.3 mm) urea-acrylamide gels.

S1 nuclease mapping of RNA

S1 mapping of the cap site and poly(A) addition site was carried out according to the procedure of Berk and Sharp (1977) as modified by Weaver and Weissmann (1979). Hybrids were digested with S1 nuclease (BRL) at a concentration of 260 U/ml, 780 U/ml and 1300 U/ml for 60 min at 30°C and protected fragments were analysed on sequencing gels along with a sequencing ladder (Maxam and Gilbert, 1980) of the corresponding end-labelled DNA strand.

Acknowledgements

We gratefully acknowledge the expert technical assistance of S. Schnell. We thank Dr. K. Scherrer and Dr. G. Paddock for sending their manuscripts prior to publication. This work was supported by the Deutsche Forschungsgemeinschaft (SFB 103).

References

Berk,A.J. and Sharp,B.A. (1977) *Cell,* **12**, 721-732.
Breathnach,R. and Chambon,P. (1981) *Annu. Rev. Biochem.,* **50**, 349-383.
Brown,J.L. and Ingram,V.M. (1974) *J. Biol. Chem.,* **249**, 3960-3972.
Dodgson,J.B., McCune,K.C., Rusling,D.J., Krust,A. and Engel,J.D. (1981) *Proc. Natl. Acad. Sci. USA,* **78**, 5998-6002.
Efstratiadis,A., Posakony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spritz,R.A., DeRiel,J.K., Forget,B.G., Weissman,S.M., Slightom,J.L., Blechl,A.E., Smithies,O., Baralle,F.E., Shoulders,C.C. and Proudfoot, N.J. (1980) *Cell,* **21**, 653-668.
Erbil,C. and Niessing,J. (1982) *Gene,* **20**, 211-217.
Felber,B.K., Orkin,S.H. and Hamer,D.H. (1982) *Cell,* **29**, 895-902.
Frankis,R. and Paddock,G.V. (1982) *J. Mol. Biol.,* **157**, 681-686.
Gallwitz,D. (1982) *Proc. Natl. Acad. Sci. USA,* **79**, 3493-3497.
Goldberg,M. (1979) Ph.D. Thesis, Stanford University, Stanford, CA.
King,R.C. and Piatigorsky,J. (1983) *Cell,* **32**, 707-712.
Liebhaber,S.A., Goossens,M.J. and Kan,Y.W. (1980) *Proc. Natl. Acad. Sci. USA,* **77**, 7054-7058.
Little,P.F.R. (1982) *Cell,* **28**, 683-684.
Maxam,A.M. and Gilbert,W. (1980) *Methods Enzymol.,* **65**, 499-560.
Montell,C., Fisher,E.F., Caruthers,M.H. and Berk,A.J. (1982) *Nature,* **295**, 380-384.
Mount,S.M. (1982) *Nucleic Acids Res.,* **10**, 459-472.
Niessing,J., Erbil,C. and Neubauer,V. (1982) *Gene,* **18**, 187-191.
Nishioka,Y. and Leder,P. (1979) *Cell,* **18**, 875-882.
Orkin,S.H., Goff,S.C. and Hechtman,R.L. (1981) *Proc. Natl. Acad. Sci. USA,* **78**, 5041-5045.
Orkin,S.H, Kazazian,H.H., Antonarakis,S.E., Goff,S.C., Boehm,C.D., Sexton,J.P., Waber,P.G. and Giardina,P.J.V. (1982) *Nature,* **296**, 627-631.
Proudfoot,N.J., Gil,A. and Maniatis,T. (1982) *Cell,* **31**, 553-563.
Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature,* **263**, 211-214.
Treisman,R., Proudfoot,N.J., Shandler,M. and Maniatis,T. (1982) *Cell,* **29**, 903-911.
Weaver,R.F. and Weissmann,C. (1979) *Nucleic Acids Res.,* **7**, 1175-1193.
Wieringa,B., Meyer,F., Reiser,J. and Weissmann,C. (1983) *Nature,* **301**, 38-43.

Note added in proof

We have obtained further evidence for correct splicing of the duck α^D-globin gene transcript in a transient expression system. Identical DNA fragments, protected against S1 nuclease, were obtained for authentic duck globin mRNA and for RNA from cells (HeLa cells and L cells) which had been transfected with an expression vector carrying the α^D-globin gene. Splicing occurs at the exon 2-intron 2 boundary and at the intron 2-exon 3 boundary. After submission of this manuscript a GT to GC transition has also been reported for the IVS 2 of the chicken α^D-globin gene (Dodgson,J. and Engel, J.D. (1983) *J. Biol. Chem.,* **258**, 4623-4629).