Volume 10 Number 6 1982 Nucleic Acids Research

The DNA sequence of the 5' flanking region of the human β-globin gene: evolutionary conservation and polymorphic differences

Nikos Moschonas, Ernie de Boer and Richard A.Flavell

Laboratory of Gene Structure and Expression, National Institute for Medical Research, Mill Hill, London NW7 1AA, UK

Received 2 February 1982; Accepted 5 March 1982

ABSTRACT

We have determined the DNA sequence of a 1464 bp segment immediately flanking the 5' side of the human β-globin gene. The sequence shows little similarity to the corresponding regions of the ε- or γ-globin genes. There is about 75% homology, however, between the 5' extragenic regions of the β-globin genes of man, goat and rabbit respectively. The mouse β minor globin gene, but not the mouse β major globin gene, also shares this extensive homology. A short segment of simple sequence DNA is found from about 1418 to 1388 bp upstream from the human β-globin gene which consists of repeats of the sequence (TTTTA). Similar DNA sequences are also found at several sites in the large intron of the β-globin gene. We have compared the DNA sequence of the 5' extragenic region of the normal β-globin gene with the same segment of the β-globin gene of a patient with β^{o}thalassaemia. Of the two nucleotide differences observed, one generates a polymorphic HinfI site present 990 bp upstream from the β-globin gene in the thalassaemic β-globin gene and absent in the normal gene. A second β^{o}thalassaemic β-globin gene which has the same molecular defect as the above mentioned case, however, lacks this HinfI site. It is therefore not yet clear whether this HinfI site will have any value in prenatal diagnosis of β^{o}thalassaemia.

INTRODUCTION

The comparison of the DNA sequence of a number of eukaryotic genes has revealed several conserved DNA sequences; the analysis of the expression of genes lacking such sequences has in turn identified the role of certain of these with transcription and RNA processing events, for a review see (1). Until now the sequence comparison of the 5' extragenic regions of eukaryotic genes has been restricted to the one to two hundred nucleotides contiguous with the respective genes. Since evidence for a role in transcription of sequences further upstream is accumulating (1, 2, 3), it is necessary that these DNA sequences are available for comparison. In this article we present the DNA sequence of a 1464 bp segment of the 5' extragenic regions of the human β-globin gene. This sequence has been compared with that of the β-globin gene from an individual with β^{o}thalassaemia, and with corresponding

0305-1048/82/1006-2109$2.00/0

sequences of several other β-globin genes and the human ∈- and γ-globin genes.

MATERIALS AND METHODS

The general procedures for the isolation and characterization of the globin genes and the Maxam-Gilbert (2) sequence determinations have been described previously (3). The genes studied are described in the Results and Discussion section.

Sequence comparisons were performed in two ways. The first used the Sequence Analysis System (SEQ), Stanford University, 1981. programme. In most comparisons a minimal homology of 75% was chosen with a minimal sequence length of 5 and loop size of 3. The second approach used a computer programme (see ref 4) which generates a matrix of dots, each of which indicate a homologous segment. A number of criteria for the presence of homology (and hence the presence of a dot on the matrix) were used. The comparisons that we show presented here use a perfect match of 4 nucleotides per dot.

GENERAL APPROACH

In this article, we discuss the DNA sequence of four human β-globin genes.

1. The first derives from a normal Dutch individual and was originally cloned from a sample of placental DNA as a cosmid (5). It has subsequently been subcloned as a 1.8 kb BamHI fragment in pAT153 (see 6). We refer to this gene as Nβ-globin gene -1.

2. The second derives from a sample of foetal liver DNA and was originally cloned by Lawn et al. (7) in λ phage charon 4A. It was subsequently subcloned by Maniatis and his colleagues as the 4.4 kb PstI fragment containing the β-globin gene in pBR322. We refer to this as Nβ-globin gene -2. The sequence of the β-globin gene from this clone is presented in Lawn et al. (8).

3. The third β-globin gene was cloned from DNA from whole blood of an Italian patient with $\beta^{o}/\delta\beta^{o}$thalassaemia (3) as a 7.5 kb HindIII fragment which contains the β^{o}thalassaemic β-globin gene. The original fragment was cloned in λ charon 21A and subcloned in the HindIII site of pAT153 (3). We refer to this gene as β^{o}-globin gene -1.

4. The fourth β-globin gene derives from spleen DNA of an Italian patient with homozygous β^{o}thalassaemia and was cloned as described for the

other thalassaemic patient (3). We refer to this gene as β^{o}globin gene -2.

RESULTS AND DISCUSSION

The DNA sequence of the 5' extragenic region of the human β-globin gene

The segment of DNA sequenced and the sequencing strategy employed is shown in Fig. 1(a). Fig. 2 shows the DNA sequence. Immediate inspection of the sequence shows a few interesting features. First, a segment of simple sequence DNA is located at the 5' end of the DNA sequence determined, which consists of several repeats derived from the basic unit $(TTTTA)_n$. The central portion of this region is a perfect repeat, but at the extremities the sequence differs by one or more nucleotides from this basic unit. We have determined the DNA sequence of this region for 3 β-globin genes. In two cases (Nβ-globin gene -1 and β^{o}-globin gene -1) the sequence is identical. In the third case, namely Nβ-globin gene -2 cloned by Lawn et al. (7) there is one fewer TTTTA repeat present. Since this difference is detectable by gel electrophoresis we have examined this point for β^{o}-globin gene -2. Judged by gel electrophoresis this gene has the same number of repeats as in Nβ-globin gene -1 and β^{o}-globin gene -1 (not shown). It is not clear whether the deficiency of one TTTTA repeat in Nβ-globin gene 2 occurred during cloning in bacteria or existed in the DNA from the original sample used for cloning by Lawn et al. (7) Clearly such a TTTTA repeat can be deleted by unequal crossing over, either in man or in E.coli. We have searched the remainder of the β-globin gene sequence of Lawn et al. (8) for the TTTTA repeated structure. Several short sequences similar to this repeating unit can be detected within the large intron. (Table I). This type of structure has much in common with the basic type of repeating unit seen in satellite DNAs but, in the case described here the total repeated segment only contains a few copies of the repeat. In addition to this type of structure, other areas of simple sequence DNA are evident e.g. alternating TA interspersed with TG and CA.

Comparison of the 5' flanking DNA sequence of human β-globin genes from a normal and a thalassaemic individual

Prior to the comparison of the structure of a given gene region from different individuals, it was speculated that considerable polymorphic differences may exist (see e.g. ref 9). Polymorphic differences can be utilized to identify a given allele of a gene if these are in linkage disequilibrium with that allele. This is particularly useful if the allele is

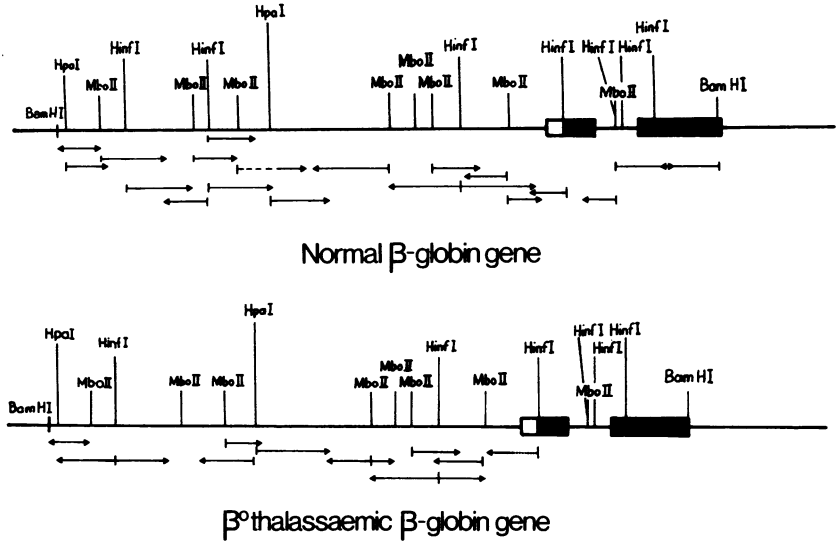

Fig. 1
The sequencing strategy for the 5' extragenic regions of the

a) normal human and
b) β^{o}thalassaemic β-globin gene.

The first two exons of the gene are shown as boxes; the filled areas encode the β-globin protein and the open area the 5' untranslated region.

linked with a mutated form of a human gene such as the human β-globin gene in β-thalassaemia. Indeed, Kan and colleagues (10) have utilized the linkage disequilibrium between a polymorphic HpaI site at the 3' side of the human β-globin gene and the human β-globin gene from patients with sickle cell anaemia to perform prenatal diagnosis of sickle cell anaemia. Unfortunately, this approach has not yet been fruitful for the diagnosis of the important hereditary anaemia β-thalassaemia, although in one case such a diagnosis has been carried out successfully (11).

We have recently cloned two β-globin genes from patients who both have a form of β^{o}thalassaemia which results from a stop codon in the β-globin gene at the position encoding amino acid residue 39 of the β-globin protein (3);Kan and his colleagues have identified a similar case, also in an Italian patient (12). These preliminary indications suggest that this may be a common form of β^{o}-thalassaemia and we have therefore determined the DNA sequence of the 5' regions of the β-globin gene of one of these patients to search for polymorphic differences. Fig. 1b shows the strategy followed.

```
        10         20         30         40         50         60         70         80         90        100
GGATCCAGTT TCTTTTGGTT AACCTAAATT TTATTTCATT TTATTGTTTT ATTTTATTTT ATTTTATTTT ATTTTGTGTA ATCGTAGTTT CAGAGTGTTA

       110        120        130        140        150        160        170        180        190        200
GAGCTGAAAG GAAGAAGTAG GAGAAACATG CAAAGTAAAA GTATAACACT TTCCTTGCTA AACCGACATG GGTTTCCAGG TAGGGGCAGG ATTCAGGATG

       210        220        230        240        250        260        270        280        290        300
ACTGACAGGG CCCTTAGGGA ACACTGAGAC CCTACGCTGA CCTCATAAAT GCTTGCTACC TTTGCTGTTT TAATTACATC TTTTAATAGC AGGAAGCAGA

       310        320        330        340        350        360        370        380        390        400
ACTCTGCACT TCAAAAGTTT TTCCTCACCT GAGGAGTTAA TTTAGTACAA GGGGAAAAAG TACAGGGGGA TGGGAGAAAG GCGATCACGT TGGGAAGCTA

       410        420        430        440        450        460        470  G     480        490        500
TAGAGAAAGA AGAGTAAATT TTAGTAAAGG AGGTTTAAAC AAACAAAATA TAAAGAGAAA TAGGAACTTG AATCAAGGAA ATGATTTTAA AACGCAGTAT

       510        520        530        540        550        560        570        580        590        600
TCTTAGTGGA CTAGAGGAAA AAAATAATCT GAGCCAAGTA GAAGACCTTT TCCCCTCCTA CCCCTACTTT CTAAGTCACA GAGGCTTTTT GTTCCCCCAG

       610        620        630        640        650        660        670        680        690        700
ACACTCTTGC AGATTAGTCT AGGCAGAAAC AGTTTAGATG TCCCCAGTTA ACCTCCTATT TGACACCACT GATTACCCCA TTGATAGTCA CACTTTGGGT

       710        720        730        740        750        760        770        780        790        800
TGTAAGTGAC TTTTTATTTA TTTGTATTTT TGACTGCATT AAGAGGTCTC TAGTTTTTTA TCTCTTGTTT CCCAAAACCT AATAAGTAAC TAATGCACAG

       810        820        830        840        850        860        870        880        890        900
AGCACATTGA TTTGTATTTA TTCTATTTTT AGACATAATT TATTAGCATG CATGAGCAAA TTAAGAAAAA CAACAACAAA TGAATGCATA TATATGTATA

       910        920        930        940        950        960        970        980        990       1000
TGTATGTGTG TATATATACA CATATATATA TATATTTTTT TTCTTTTCTT ACCAGAAGGT TTTAATCCAA ATAAGGAGAA GATATGCTTG GAACTGAGGT

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
AGAGTTTTCA TCCATTCTGT CCTGTAAGTA TTTTGCATAT TCTGGAGACG CAGGAAGAGA TCCATCTACA TATCCCAAAG TGAATTATGG TAGACAAAAC

      1110       1120  T    1130       1140       1150       1160       1170       1180       1190       1200
TCTTCCACTT TTAGTGCATC AACTTCTTAT TTGTGTAATA AGAAAATTGC GAAAACGATC TTCAATATGC TTACCAAGCT GTGATTCCAA ATATTACGTA

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
AATACACTTG CAAAGGAGGA TGTTTTTAGT AGCAATTTGT ACTGATGGTA TGGGGCCAAG AGATATATCT TAGAGGGAGG GCTGAGGGTT TGAAGTCCAA

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
CTCCTAAGCC AGTGCCAGAA GAGCCAAGGA CAGGTACGGC TGTCATCACT TAGACCTCAC CCTGTGGAGC CACACCCTAG GGTTGGCCAA TCTACTCCCA

      1410       1420       1430       1440       1450       1460
GGAGCAGGGA CGGCAGGAGC CAGGGCTGGG CATAAAAGTC AGGGCAGAGC CATCTATTGC TTA
```

Fig. 2
The DNA sequence of the 5' extragenic region of the human β-globin gene. Two polymorphic differences were found between the sequence of the normal and thalassaemic β-globin gene region. These are at nucleotide 474 (= -990) and 1123 (-341) and are indicated above the relevant part of the DNA sequence.

Only two nucleotide differences have been found in this comparison (although a segment of about 107 nucleotides of the β^{o}thalassaemic sequence has not been determined), aC->G transversion at position -990 and a C->T transition at -340 (Fig. 2). Interestingly, the first of these deletes a HinfI site in the thalassaemic individual. This cannot be a cloning artefact since we have shown the absence of this HinfI site in the genomic DNA of this patient by Southern blotting HinfI-cut DNA and hybridizing the blots with a probe for the human β-globin genes; (T. de Lange, unpublished). The latter observation suggests that HinfI could possibly be used for the diagnosis of this form of β-thalassaemia. We have therefore analyzed two normal β-globin genes and another β^{o}thalassaemic β-globin gene (with the same genetic lesion as in this case). In all these cases, however, the HinfI site is present (not shown). We are currently analyzing a series of genomic DNAs from the thalassaemic individuals to see how widespread this HinfI polymorphism is.

TABLE

Presence of DNA sequences related to TTTTA in the human β-globin gene region

	Sequence	Position	Coordinates
1.	TTTGG TTAAC(CTAAA)TTTTA TTTCA TTTTA(TTG) TTTTA TTTTA TTTTA TTTTA TTTTA TTTTG	5' extragenic	-1449 to -1388
2.	TTTTA GTTTC TTTTA TTTGC TGTTC	intron	+638 to +662
3.	TTTTC TTTTG TTTAA TTCTT GTTCT TTTTT TTTTC TTCTC	intron	+669 to +713
4.	(CCCTA)CTTTA TTTTC TTTTA TTTTT	intron	+874 to +895
5.	CTTTC TTCTT TTAAT(ATACT)TTTTT GTTTA TCTTA(T)TTCTA	intron	+1007 to +1046

The conservation of the DNA sequence in the 5' flanking region of the β-globin gene is remarkable and reminiscent of the results recently obtained (e.g. ref 3) which show very few polymorphisms in the DNA sequence of the β-globin genes of different individuals.

Comparison of the DNA sequence of the 5' flanking regions of related β-globin genes

The availability of the DNA sequence of the 5' flanking regions of a number of β-globin genes permits the comparison of these sequences and thus a search for DNA sequence conservation. About 400 nucleotides of 5' flanking DNA sequence has been determined for the rabbit β-globin gene (13) and around 200 nucleotides for the mouse β major and β minor globin genes (4) and the goat (14) β-related globin genes. We have compared these DNA sequences with the corresponding region of the human β-globin gene; the results are shown in Figs. 3 to 6. About 75% homology can be seen between the human and rabbit sequences from about -370 to the cap site; 75% homology is also found between the goat and human sequences throughout the published sequence of the goat 5' region (14). In the case of the human/rabbit sequence comparison, we have performed the homology search in two ways. In the first, a computerized dot-matrix comparison was carried out (Fig. 3A; see ref. 4) and in the second, we have used the SEQ programme (see Materials and Methods). In the latter case, short sequences are matched and printed when the homology

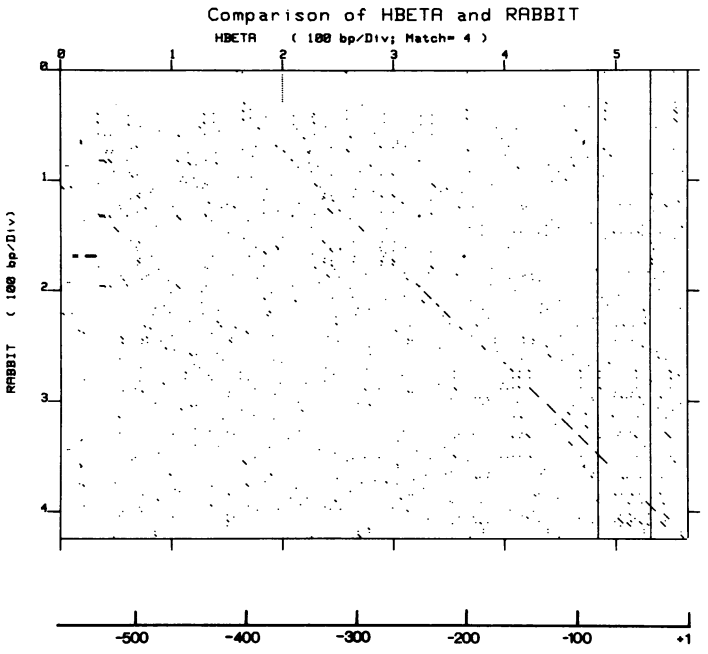

Fig. 3
Dot matrix comparison of the 5' extragenic regions of the human (H Beta) and rabbit β-globin genes.
A computer programme (see 4) was used to generate a matrix of dots indicating segments of homology. The human DNA sequence is from this paper, the rabbit sequence from ref. 13. Each dot represents a homologous segment of 4 nucleotides in length. The two vertical lines at the right of the figure indicate the CAAT and ATA boxes.

falls within chosen limits of minimal homologous sequence length. % mismatch and so on. The results of this comparison of the human 5' sequences with those of the rabbit, goat and mouse β-globin genes is shown in Figures 4, 5 and 6 respectively. Both the mouse β major and β minor genes are homologous to the human β-globin gene, but the homology is significantly less than in the rabbit and goat comparisons (about 65% for both mouse (genes) versus 75% for the goat and rabbit homology). Van Ooyen et al., (15) noted previously that the rabbit and human β-globin gene mRNA coding sequences were more similar to each other than are the mouse and rabbit sequences. The divergence time for all the ma-mals discussed here is considered to be the same, that is about 85 million (MY). The rate of sequence divergence calculated from the rabbit and goat comparison is therefore about 0.15% per MY

Fig. 4
Alignment of the human (H) and rabbit (R) 5' extragenic β-globin gene sequence to display homology. The sequence was aligned with the help of the SEQ search computer programme as described in the text and Materials and Methods. *denotes a region of non-homology.

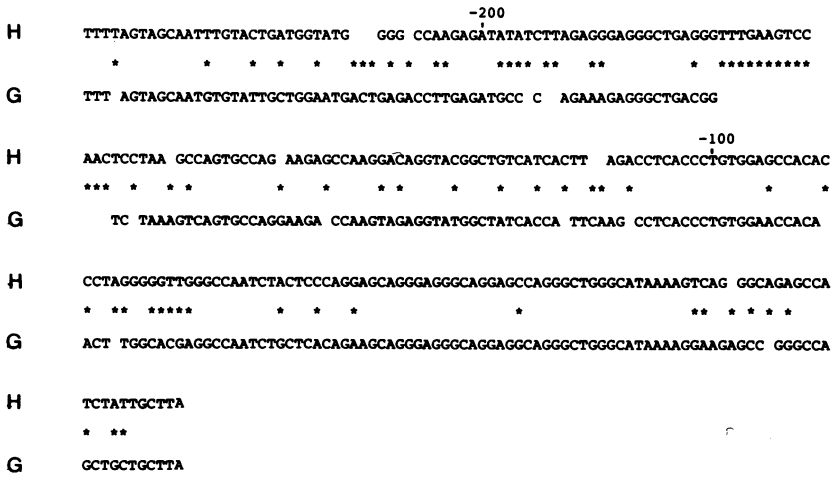

Fig. 5
Sequence conservation between the human (H) and goat (G) β-globin genes (see Fig. 4 for details).

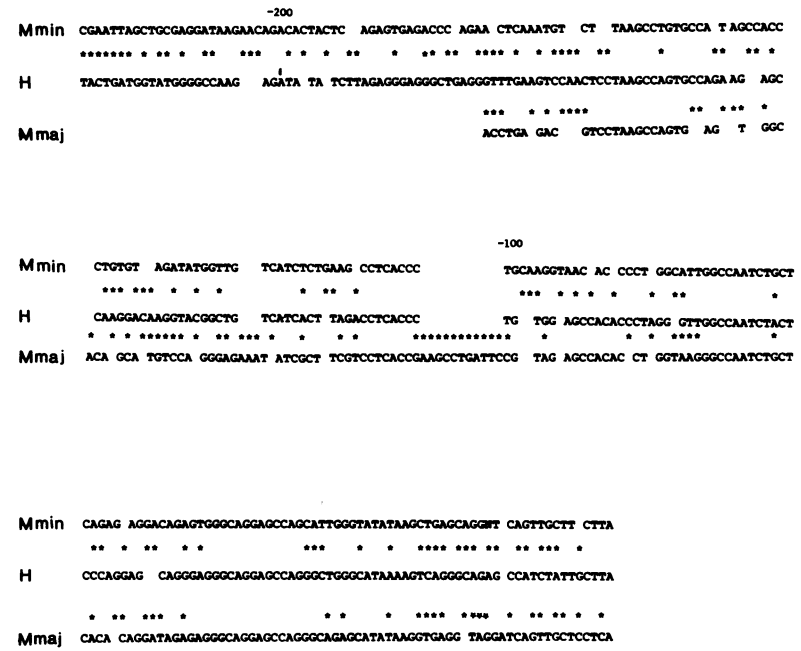

Fig. 6
Sequence conservation between the human (H) and mouse β major (M maj) and β minor (M min) globin genes.

per single lineage, whereas the value calculated for the mouse genes is about 0.25% per MY. Both numbers are significantly lower than the values (0.5% - 1% per MY) normally used for the rate of change of DNA sequences where little selective pressure is thought to be exerted (see e.g. 17). This discrepancy can be explained in two general ways. First, the rate of change of nucleotide sequence per unit time might differ between different gene combinations, between different regions of the genome in a given animal or, more likely, between different animals. The fact that the rate of change differs significantly between the mouse and, say, rabbit supports this notion. It is possible, as has been suggested repeatedly, that the generation time of the organism might influence the rate of accumulation of mutations at a given site in the genome. Specifically, the generation time of the mouse is significantly shorter than that of the other mammals studied here. Alternatively, this discrepancy could reflect DNA sequence conservation as a result of selection pressure. The rate of sequence divergence in the 5' extragenic regions seems to be significantly less than the rate of change of the DNA

sequence of the large intron of the β-globin gene which has been reported, both in the case of the rabbit-mouse comparison (15) or the rabbit-human comparison (our unpublished comparison). This might reflect differential rates of mutation at the two sites (albeit unlikely!) as discussed above and/or the effect of selection. This selection might act to conserve the 5' flanking region. In addition, a high proportion of the DNA sequences differences seen in the large intron is the result of deletions and insertions. It is also possible, therefore, that this type of sequence change is tolerated in an intron, but prohibited in the 5' flanking regions. This type of mutation would obviously have a serious effect on the functioning of DNA sequence elements whose position with respect to one another must be kept constant (e.g. promotor elements). The DNA sequences from -100 up to and including the ATA box at -31 have been shown to play a role in the transcription of globin (19, 13, 20, 27) and other genes (e.g. 21, 22, 23). Sequence conservation in these regions has been given ample coverage in the literature and need not be reiterated here. There are two other regions which show apparant DNA sequence conservation in all the five mammalian β-globin genes that we have compared here. One of these is an approximately 13 base pair region localized at about -170 to -157 in the case of the mouse β-major globin gene, but at about -160 to -147 in the other genes. The consensus sequence for this region is $TC^{C}_{T}TAAG^{C}_{T}CA^{G}_{T}TGCCA$. The region downstream from -120 is also conserved in all these β-globin genes. The sequence at about -160 is not conserved (data not shown) between the human β-globin gene or the other human β-related globin genes (γ and ϵ) for which DNA sequence information is available in these regions (see 18, 24 and 25). This sequence conservation seen in the β-globin genes might then reflect a role for these DNA sequences in the functioning of the β-globin gene specicically rather than a general sequence element for all genes. Functional analysis of the DNA sequence requirement for the expression of the β-globin gene in erythroid cells should help to elucidate this point.

Note added

After this manuscript was prepared a paper by Spritz (26) appeared which also describes the simple sequence DNA (TTTTA) to the 5' side of the human β-globin gene.

ACKNOWLEDGEMENTS

N.M. was the grateful recipient of an EMBO long term fellowship. We are grateful to P. Gillett and H. Bud of this Institute and M. Waterfield

and G. Scrace of Imperial Cancer Research Fund for help with computer analyses of DNA sequences. This work was supported by the British Medical Research Council.

REFERENCES

1. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
2. Maxam, A. M. and Gilbert, W. (1980) Methods in Enzymology, vol. 65, Grossman, L. and Moldave, K. (Eds.), Academic Press, New York, 499-560.
3. Moschonas, N., deBoer, E., Grosveld, F. G., Dahl, H. H. M., Wright, S. Shewmaker, C. K. and Flavell, R. A. (1981) Nucl. Acids Res. 17, 4391-4401.
4. Konkel, D. A., Maizel, J. V. Jr. and Leder, P. (1979) Cell 18, 865-873.
5. Grosveld, F. G., Dahl, H. H. M., deBoer, E. and Flavell, R. A. (1981) Gene 13, 227-237.
6. Twigg, A. J. and Sherratt, D. (1980) Nature 283, 216-218.
7. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. and Maniatis, T. (1978) Cell 15, 1157-1174.
8. Lawn, R. M., Efstratiadis, A., O'Connell, C. and Maniatis, T. (1980) Cell 21, 647-651.
9. Flavell, R. A., Little, P. F. R., Kooter, J. M. and deBoer, E. (1979) in: Models for the Study of Inborn Errors of Metabolism. Hommes, F.A. (Ed.) Elsevier/North Holland Biomedical Press, Amsterdam, 355-366.
10. Kan, Y. W. and Dozy, A. M. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 5631-5635.
11. Little, P. F. R., Annison, G., Darling, S., Williamson, R., Camba, L. and Modell, B. (1979) Nature 285, 144-147.
12. Trecartin, R. F., Liebhaber, S. A., Chang, J. C., Lee, K. Y., Kan, Y. W., Furbetta, M., Angius, A. and Cao, A. (1981) J. Clin. Invest (in press).
13. Dierks, P., van Ooyen, A., Mantei, N. and Weissmann, C. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 1411-1415.
14. Haynes, J. R., Rosteck, P. Jr. and Lingrell, J. B. (1980) Proc. Natl. Acad. Sci. 77, 7127-7131.
15. van Ooyen, A., van den Berg, J., Mantei, N. and Weissmann, C. (1979) Science 206, 337-344.
16. Slightom, J. L., Blechl, A. E. and Smithies, O. (1980) Cell 21, 627-638.
17. Jeffreys, A. J. (1981) In: Genetic Engineering II, Williamson, R. (Ed.) Academic Press, New York, pp.1-48 Vol 2.
18. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B., Weissman, S.M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, G. C. and Proudfoot, N. J. (1980) Cell 21, 653-668.
19. Grosveld, G. C., Shewmaker, C., Jat, P. and Flavell, R. A. (1981) Cell 25, 215-226.
20. Grosveld, G. C., deBoer, E., Shewmaker, C. K. and Flavell, R. A. (1981) Nature 295, 120-126.
21. Grosschedl, R. and Birnstiel, M. L. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 1432-1436.
22. Benoist, C. and Chambon, P. (1981) Nature 290, 304-310.
23. McKnight, S. L., Gavis, E. R., Kingsbury, R. and Axel, R. (1981) Cell 25, 385-398.

24. Baralle, F. E., Shoulders, C. C., Goodbourn, S., Jeffreys, A. and Proudfoot, N. J. (1980) Nucl. Acids Res. 8, 4393-4404.
25. Shen, S., Slightom, J. L. and Smithies, O. (1981) Cell 26, 191-203.
26. Spritz, R. (1981) Nucl. Acids Res. 9, 5037-5047.
27. Mellon, P., Parker, V., Gluzman, Y. and Maniatis, T. (1981) Cell, 27, 279-288.