Volume 9 Number 19 1981 Nucleic Acids Research

Duplication/deletion polymorphism 5'- to the human β globin gene

Richard A.Spritz

Departments of Medical Genetics and Pediatrics, Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706, USA

Received 15 June 1981

ABSTRACT

DNA sequence analysis of the human β globin locus has identified an array of simple tandem repeated sequences upstream from the β globin structural gene. Comparison of several cloned human β globin alleles demonstrated a high frequency of sequence heteromorphism at this site apparently due to duplication or deletion of single units of the repeat array. At least two such duplication/deletion events are necessary to account for the observed variation. No other sequence variation was observed, suggesting that duplication/deletion events within the tandem repeat array may be at least 13 to 14 times more frequent than nucleotide substitutions in the surrounding DNA.

INTRODUCTION

Tandem arrays of relatively simple repeated sequences are widespread among eukaryotes, and in many species constitute a substantial portion of the genome (rev. in 1). Sequence homogeneity among members of a family of repeats is probably maintained by strand slippage and mispairing during DNA replication (2,3) or by non-homologous recombination (4,5), which may be particularly frequent in such regions. A "hotspot" for spontaneous mutations in the *lac I* gene of *E. coli* has been found to consist of an array of short simple repeated sequences, with frequent duplication or deletion of single copies of the repeating unit at this site (6).

DNA sequence analysis has identified a similar array of short simple tandemly repeated sequences in the region approximately 1300 nucleotides 5'- to the human β globin gene. Comparison of the sequences of a number of cloned β globin alleles demonstrates considerable sequence heteromorphism in this region, consisting of duplication or deletion of single copies of the basic repeat unit. At least two such duplication/deletion events have probably occurred at

this site, one involving a normal β globin allele and another involving a β^{o}-thalassemic allele with a nonsense mutation of codon Gln_{39} of the structural gene. No other sequence variation was observed in this region, suggesting that duplication/deletion events may be considerably more frequent at the β globin polymorphic site than are nucleotide substitutions in the surrounding DNA.

MATERIALS AND METHODS

Source of Human β Globin Clones.

HβG1 and HβG2 (7) derive from a normal abortus of non-Mediterranean extraction, and represent the two β globin alleles from that individual. C6 was derived from a Greek Cypriot patient with typical β^{+}-thalassemia due to a mutation in the small intervening sequence of the β globin gene (8). βT2-1 is from an Italian patient also with typical β^{+}-thalassemia, although this allele apparently does not contain the same mutation as that identified in C6 (A. Bank, personal communication). β^{+}5.2 and β12 were isolated from individuals of Moroccan and mixed Scottish-Irish/Italian ethnic origins, respectively, both apparently heterozygous for β^{o} and β^{+}-thalassemia alleles. The alleles represented by β^{+}5.2 and β12 are β^{o}-thalassemic genes, with nonsense mutations at codon Gln_{39} (R. Spritz, unpublished observation; A. Bank, personal communication). Portions of these clones were propagated as plasmids in the *E. coli* K-12 strains HB101 or C600 under P1 containment conditions.

Restriction Endonuclease Cleavage Analysis and Nucleotide Sequencing.

DNA was cleaved and fragments were isolated by standard methods. Fragments were labelled at their 3'- termini with α^{32}[P]-dGTP and *E. coli* DNA polymerase (Klenow fragment), and sequenced by a modification of the method of Maxam and Gilbert (9).

RESULTS

The nucleotide sequence of several cloned DNA fragments covering the region from approximately 1050 to 1500 nucleotides 5'- to the cap site of the human β globin gene was determined, and is presented in Fig. 1. The nucleotide sequences of the normal β globin clone, HβG1, and β^{+}-thalassemic clone, C6, were determined completely in this

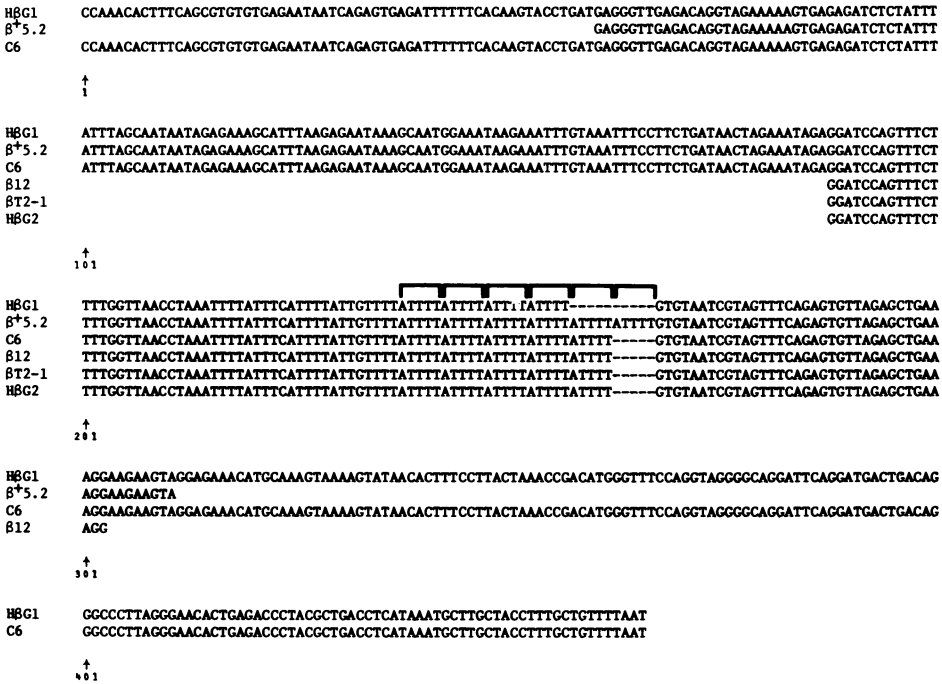

Figure 1. Partial nucleotide sequences of six cloned human β globin alleles. The sequence presented is centered approximately 1300 nucleotides 5'- to the cap site of the human β globin gene, and represents the coding DNA strand with respect to the β globin structural gene. Individual clones are described in the text. Brackets identify ATTTT repeats.

region, and portions of the DNA sequences of the other β globin clones were also determined. The DNA in the region from nucleotides 76 to 354 of the sequence determined is extremely rich in A-T base pairs (76%), with numerous runs of oligoadenylate or oligothymidylate residues. However, the sequence of the region from nucleotides 238 to 262 (±5) is even more unusual, consisting entirely of a series of perfect tandem repetitions of the sequence ATTTT. Moreover, the sequence upstream from the tandem repeat array also contains a number of similar imperfect repeats.

Comparison of the six β globin alleles analyzed in this region revealed remarkable heterogeneity in the number of perfect ATTTT repeats at this site. C6, β12, βT2-1, and HβG2 each contain five

tandem repetitions of this sequence, while HβGl and β^{+}5.2 contain four and six ATTTT repeats, respectively (Fig. 2). With the exception of these differences, the nucleotide sequences of the six alleles are completely identical in the regions sequenced although extensive nucleotide sequences were compared only for HβGl, C6, and β^{+}5.2.

DISCUSSION

This report describes a site with a high degree of sequence polymorphism approximately 1300 nucleotides upstream from the human β globin gene. The polymorphism occurs within an array of perfect

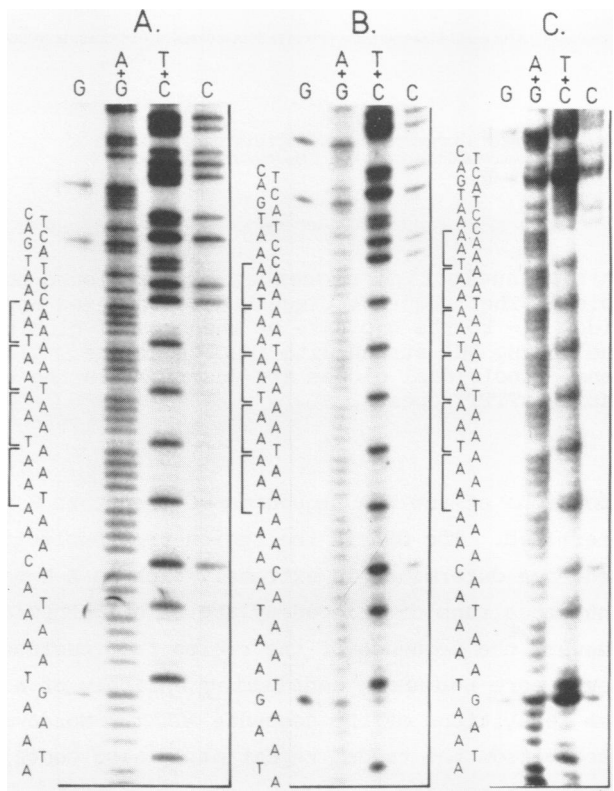

Figure 2. DNA sequencing gels showing variation in the number of repeat units in the tandem repeat array. The nucleotide sequences of HβGl (A), C6 (B), and β^{+}5.2 (C) were all determined from the 3'-end-labelled BamHI restriction sites at position 189.

tandem repeats of the simple oligonucleotide ATTTT. The modal number of repeat units in this array appears to be five. However, one allele each with six and four repeats have also been identified, apparently the result of duplication and deletion of single repeat units, respectively. It is unlikely that these two variant alleles represent the reciprocal products of a single duplication/deletion event, since the allele with six repeats, $\beta^{+}5.2$, is a β^{o}-thalassemic allele and the allele with four repeats, HβG1, is a normal allele. Only by interchromosomal misalignment of the tandem repeats in a heterozygous carrier of the β^{o}-thalassemic allele with simultaneous crossing-over within the tandem repeat array could a single event produce the observed sequence differences. Thus, the observed variants probably are the products of at least two duplication/deletion events; one involving a normal allele and another involving a β^{o}-thalassemic allele.

There is no apparent relationship between the sequence structure at the β globin tandem repeat array and the β-thalassemic phenotype. Transcription of HβG1, C6, and $\beta^{+}5.2$, containing four, five, and six copies of the ATTTT repeat unit, respectively, appears normal *in vitro* (8,10), and transcription of C6 appears normal *in vivo* (11). Although both of the β^{+}-thalassemic alleles, C6 and βT2-1, contain five copies of the repeat unit in the tandem array, a normal β globin allele, HβG2, also contains five repeat units at this site. However, another normal allele, HβG1, contains only four copies of the repeat unit. The other two alleles analyzed, $\beta^{+}5.2$ and β12, are both β^{o}-thalassemic alleles with nonsense mutations at codon Gln_{39} of the β globin structural gene and contain five and six repeat units, respectively (R. Spritz, unpublished observation; A. Bank, personal communication). This indicates that this particular β^{o}-thalassemic mutation may be associated with either five or six ATTTT repeats in the tandem array, and this suggests that this mutation most likely predates the most recent duplication/deletion event involving these alleles.

No divergent nucleotides were observed in the DNA which surrounds the tandem repeat array, although only for HβG1 and C6 (466 nucleotides) and $\beta^{+}5.2$ (246 nucleotides) was enough sequence determined to permit extensive comparisons. These data are in good agreement with an estimated polymorphism frequency of 0.005 per

nucleotide in non-selected DNA (12). The observed sequence homogeneity is unlikely to be the product of selection via some protein product, since numerous translation termination codons occur in all reading frames on both DNA strands. Furthermore, no major transcripts initiate in this region *in vitro* (8), although numerous promoter-like ATAAAA sequences occur on the anti-sense strand within the tandem repeat array. Although selection by an as-yet undefined mechanism remains formally possible, such a phenomenon seems unlikely to account for the lack of variation observed.

The frequency of sequence variation within the tandem repeat array thus appears to be considerably greater than that in the surrounding sequences. Kimura and Ohta (13) have shown that for selectively neutral alleles which change states by single steps in a positive or negative direction (such as duplication or deletion of single repeat units within a tandem repeat array)

$$H_o = 1/\sqrt{1 + 8N_e\nu} \quad ,$$

where H_o represents the proportion of homozygotes, N_e is the effective size of the population, and ν is the mutation rate. Because H_o also equals the sum of the squares of the observed allelic frequencies, for the β globin polymorphism

$$H_o = (4/6)^2 + (1/6)^2 + (1/6)^2 = 1/2 \quad .$$

Therefore,

$$\nu = 3/8\ (N_e^{-1}) \quad .$$

However, Kimura and Crow (14) have shown that for ordinary selectively neutral mutations

$$H_o \simeq 1/(4N_e\nu + 1) \quad .$$

Although no such mutations were observed in the region sequenced, if we assume that at the level of DNA sequence $H_o \geq 0.9$, for example, then

$$\nu \leq 1/36\ (N_e^{-1}) \quad .$$

Thus, the rate of duplication/deletion events within the tandem repeat sequence array is at least 13 to 14 times the rate of nucleotide substitutions in the surrounding DNA. A similar tandem array of simple oligonucleotide repeats in the lac I gene of E. coli has also been shown to be a hotspot for frameshift mutations caused by duplication or deletion of single repeat units (6).

Simple repeating sequences are a ubiquitous feature of eukaryotic DNA, and may comprise up to 65% of the genome in some species (rev. in 1). Satellite DNAs consist primarily of tandem arrays of simple repeated sequences of oligonucleotide length (1, 15-20), although tandem arrays of longer repeats have been identified in many species (1,21-33). The reiterated histone (rev. in 34) and ribosomal RNA genes (rev. in 35,36) are themselves organized in repeating tandem arrays. Moreover, regions of DNA adjoining the ribosomal (35-44) and 5S (45-48) genes and the class-switch region of immunoglobulin heavy chain genes (49-52) contain arrays of simple repeated sequences, and regions of simple alternating dinucleotide sequence have been identified adjacent to the histone genes (34,53-55) and within the human γ globin genes (56). All of these arrays of repeating sequences exhibit considerable heterogeneity of total length due to variablility in the number of repeat units per array. Alternating dinucleotide sequences have also been identified near the mouse α globin (57) and immunoglobulin V-region (58) genes, although sequence heterogeneity in these regions has yet to be established. Variability in the number of repeat units thus appears to be a general feature of tandem simple repeating sequence arrays in eukaryotic DNA, and may account for a significant portion of the total variation in the genome.

Two general mechanisms have been proposed to account for duplication/deletion events among repeated sequences. Strand slippage and mispairing between tandem repeat units on the leading and trailing strands could occur during the DNA replication process (2,3). It is this mechanism which perhaps best accounts for duplication or deletion of integral numbers of short repeat units in a tandem array (6,59). Alternatively, duplication or deletion of integral numbers of repeat units might also result from nonhomologous recombination events

between repeat units on sister chromatids or homologous chromosomes (4,5), since misalignment of DNA segments might be particularly frequent in regions composed of tandem repeat sequences. Kedes has suggested that such regions of simple sequence composition may be the initial nucleation sites in the recombination process (34). This mechanism seems most likely to account for duplication and deletion of relatively large pieces of DNA, such as those observed to produce duplication or deletion of sequences *in vitro* (60-62) and *in vivo* (61, 63-65), and Chambers *et al.* (66) have suggested that binding of at least seven homologous nucleotide pairs may be required for misalignment of subrepeats and nonhomologous crossing-over to occur. Both of these processes would tend to "homogenize" repeated DNA sequences, due to dilution of rare sequence variants by the predominant sequence in the gene pool.

Farabaugh and co-workers (6) have observed variation similar to that described here in an array of simple oligonucleotide repeats in the *lac I* gene of *E. coli*. Moreover, Arnheim and Kuehn (43) have observed *rec*A-independent deletion of repeat elements in a cloned mouse ribosomal gene spacer. Therefore, it is possible that the observed duplication and deletion events at the β globin tandem repeat array occurred during cloning or subsequent propagation in bacteria. However, multiple isolates of several of the human β globin plasmids analyzed have yielded identical sequences at this site, and in no instance was a mixed DNA sequence observed. As the tandem repeat array therefore appears to be stable in the plasmid form, the observed sequence variation seems unlikely to have arisen during propagation in *E. coli*, and probably accurately reflects heteromorphism in the genome of the individuals sampled. Studies are currently in progress to investigate the potential biological significance of this sequence variation.

ACKNOWLEDGMENTS

This investigation was supported by grants from the N.I.H., and was carried out in accordance with the N.I.H. guidelines for recombinant DNA research. I thank Y.W. Kan, Art Bank, Tom Maniatis, and Jeff Ross and colleagues for providing subclones of β^{+}5.2, βT2-1 and HβG2, HβG1, and β12, respectively. I particularly thank Fred Blattner for the generous use of his laboratory facilities in the

early phase of this work. I also thank Oliver Smithies and James Crow for constructive criticism and Sandy Keller for expert secretarial assistance. This is paper number 2509 from the Laboratory of Genetics, University of Wisconsin.

REFERENCES

1. Skinner, D.M. (1977) BioScience 27, 790-796.
2. Kornberg, A., Bertsch, L.L., Jackson, J.F. and Khorana, H.G. (1964) Proc. Nat. Acad. Sci. USA 51, 315-323.
3. Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. and Inouye, M. (1966) Cold Spring Harbor Symp. Quant. Biol. 31, 77-84.
4. Smith, G.P. (1976) Science 191, 528-536.
5. Perelson, A.S. and Bell, G.I. (1977) Nature 265, 304-310.
6. Farabaugh, P.J., Schmeissner, U., Hofer, M. and Miller, J.H. (1978) J. Mol. Biol. 126, 847-863.
7. Lawn, R.M., Fritsch, E.F., Parker, R.C., Blake, G., and Maniatis, T. (1978) Cell 15, 1157-1174.
8. Spritz, R.A., Jagadeeswaran, P., Choudary, P.V., Biro, P.A., Elder, J.T., de Riel, J.K., Manley, J.L., Gefter, M.L., Forget, B.G. and Weissman, S.M. (1981) Proc. Nat. Acad. Sci. USA, 78, 2455-2459.
9. Maxam, A.M. and Gilbert, W. (1980) Meth. Enz. 65, 499-560.
10. Parker, C.S., Goosens, M. and Kan, Y.W. (1980) Clin. Res. 38, 488A.
11. Benz, E.J., Jr., Spritz, R.A., Scarpa, A.L., Tonkonow, B.L., Pearson, H.A., Ritchey, A.K, Weissman, S.M. and Forget, B.G. (1981) Pediatr. Res., in press.
12. Ewens, W.J., Harris, H. and Spielman, R. (1981). Proc. Natl. Acad. Sci. USA. In press.
13. Kimura, M. and Ohta, T. (1975) Proc. Nat. Acad. Sci. USA 72, 2761-2764.
14. Kimura, M. and Crow, J. F. (1964) Genetics 49, 725-738.
15. Bond, H.E., Flamm, W.G., Burr, H.E. and Bond, S.B. (1967) J. Mol. Biol. 27, 289-302.
16. Gall, J.D. and Atherton, D.D. (1974) J. Mol. Biol. 85, 633-664.
17. Skinner, D.M. Beattie, W.G., Blattner, F.R., Stark, B.R. & Dahlberg, J.E. (1974) Biochem. 13, 3930-3937.
18. Biro, P.A., Carr-Brown, A., Southern, E.M. and Walker, P.M.B. (1975) J. Mol. Biol. 94, 71-86.
19. Salser, W., Bowen, S., Browne, D., El Aldi, F., Federoff, N. Fry, K., Heidell, H., Paddock, G., Poon, R., Wallace, B. & Whitcome, P. (1976) Fed. Proc. Fed. Am.Soc. Exp. Biol. 35, 23-35.
20. Fry, K. and Salser, W. (1977) Cell 12, 1069-1084.
21. Manuelidis, L. (1978) Chromosoma 66, 1-21.
22. Kurnit, D.M. (1979) Hum. Genet. 47, 169-186.
23. Botchan, M.R. (1974) Nature 251, 288-292.
24. Maio, J.J., Brown, F.L. and Musich, P.R. (1977) J. Mol. Biol. 117, 637-655.
25. Musich, P.R., Maio, J.J. and Brown, F.L. (1977) J. Mol. Biol. 117, 657-677.
26. Manuelidis, L. and Wu, J.C. (1978) Nature 276, 92-94.

27. Rosenberg, H., Singer, M. and Rosenberg, M. (1978) Science 200, 394-402
28. Donehower, L. and Gillespie, D. (1979) J. Mol. Biol. 134, 805-834.
29. Singer, D. and Donehower, L. (1979) J. Mol. Biol. 134, 835-842.
30. Bedbrook, J.R., Jones, J., O'Dell, M., Thompson, R. & Flavell, R.B. (1980) Cell 19, 545-560.
31. Bedbrook, J.R., O'Dell, M. and Flavell, R.B. (1980) Nature 288, 133-137.
32. Donehower, L., Furlong, C., Gillespie, D. and Kurnit, D. (1980) Proc. Nat. Acad. Sci. USA 77, 2129-2133.
33. Moss, B., Winters, E. and Cooper, N. (1981) Proc. Nat. Acad. Sci. USA 78, 1614-1618.
34. Kedes, L. (1979) Ann. Rev. Biochem. 48, 837-870.
35. Wellauer, P.K., Reeder, R.H., Dawid, I.B., and Brown,D.D. (1976) J. Mol. Biol. 105, 487-505.
36. Reeder, R.H., Brown, D.D. Wellauer, P.K. and Dawid, I.B. (1976) J. Mol. Biol. 105, 507-516.
37. Wellauer, P.K., Dawid, I.B., Brown, D.D. and Reeder, R.H.(1976) J. Mol. Biol 105, 461-486.
38. Arnheim, N. and Southern, E.M. (1977) Cell 11, 363-370.
39. Botchan, P., Reeder, R.H. and Dawid, I.B. (1977) Cell 11, 599-607.
40. Wellauer, P.K. and Dawid, I.B. (1977) Cell 10, 193-212.
41. Wellauer, P.K. and Dawid, I.B. (1978) J. Mol. Biol. 126, 769-782.
42. Arnheim, N. (1979) Gene 7, 83-96.
43. Arnheim, N. and Kuehn, M. (1979) J. Mol. Biol. 134, 743-765.
44. Arnheim, N., Krystal, M., Schmickel, R., Wilson, G., Ryder, O. and Zimmer, E. (1980) Proc. Natl. Acad. Sci. USA 77, 7323-7327.
45. Brownlee, G.G., Cartwright, E.M. and Brown, D.D. (1974) J. Mol. Biol. 89, 703-718.
46. Carroll, D. and Brown, D.D. (1976) Cell 7, 467-475.
47. Federoff, N.V. and Brown, D.D. (1978) Cell 13, 701-716.
48. Korn, L.J. and Brown, D.D. (1978) Cell 15, 1145-1156.
49. Davis, M.M., Kim, S.K. and Hood, L.E. (1980) Science 209, 1360-1365.
50. Dunnick, W., Rabbits, T.H. and Milstein, C. (1980) Nature 286, 669-675.
51. Sakano, H., Maki, R., Kurosawa, Y., Roeder,W. and Tonegawa, S. (1980) Nature 286, 676-683.
52. Kataoka, T., Miyata, T. and Honjo, T. (1981) Cell 23, 357-368.
53. Overton, G.C. and Weinberg, E.S. (1978) Cell 14, 247-257.
54. Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H.O. and Birnsteil, M.L. (1978) Cell 14, 655-671.
55. Sures, I., Lowry, J. and Kedes, L.H. (1978) Cell 15, 1033-1044.
56. Slightom, J.L., Blechl, A.E. and Smithies, O. (1980) Cell 21, 627-638.
57. Nishioka, Y. and Leder, P. (1979) Cell 18, 875-882.
58. Nishioka, Y. and Leder, P. (1980) J. Biol. Chem. 255, 3691-3694.
59. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R., M. O'Connell, C., Spritz, R.A., deRiel, J.K., Forget, B.

Weissman, S., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21, 653-668.
60. Russell, R.L., Abelson, J.N., Landy, A., Gefter, M.L., Brenner, S. and Smith, J.D. (1970) J. Mol. Biol. 47, 1-13.
61. Fritsch, E.F. Lawn, R.M. and Maniatis, T. (1980) Cell 19, 959-972.
62. Lauer, J., Shen, C.-K. J. and Maniatis, T. (1980) Cell 20, 119-130.
63. Kan, Y.W., Dozy, A.M., Varmus, H.E., Taylor, J.M., Holland, J.P., Lie-Injo, L.E., Ganesan, J. and Todd, D. (1975) Nature 255, 255-256.
64. Goosens, M., Dozy, A.M., Embury, S.H., Zachariades, Z., Hadjiminas, M.G., Stamatoyannopoulos, G. and Kan, Y.W. (1980) Proc. Natl. Acad. Sci. USA 77, 518-521.
65. Higgs, D.R., Old, J.M., Pressley, L., Clegg, J.B. & Weatherall, D.J. (1980) Nature 284, 632-635.
66. Chambers, C.A., Schell, M.P. and Skinner, D.M. (1978) Cell 13, 97-110.