Volume 15 Number 22 1987 Nucleic Acids Research

Unusual features of CpG-rich (HTF) islands in the human α globin complex: association with non-functional pseudogenes and presence within the 3' portion of the ζ gene

N.Fischel-Ghodsian, R.D.Nicholls and D.R.Higgs

MRC Molecular Haematology Unit, Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK

Received September 8, 1987; Revised and Accepted October 26, 1987

SUMMARY

We have characterised a cluster of CpG rich (HTF) islands in the α-globin complex and report here two unusual features: The human embryonic ζ2-globin gene is associated with an HTF island within its 3' portion rather than at the 5' end. Furthermore at least two non-functional pseudogenes within the cluster (ψζ1 and ψα2) are associated with CpG rich islands.

INTRODUCTION

Many vertebrate genes include segments (0.5-3.0 kb) of unmethylated CpG-rich DNA at their 5' ends, extending upstream and downstream of the site of initiation of transcription and often encompassing one or two exons (1,2). It has been estimated that the mammalian genome contains approximately 30,000 such segments, referred to as Hpa II-tiny fragment (HTF) or CpG rich islands (3), which corresponds well to the estimated number of functional genes (20,000-50,000). HTF islands can be distinguished from other regions of the genome by their susceptibility to cleavage by enzymes that are sensitive to the presence of CpG methylation in their recognition sequences. The identification of HTF islands has been successfully used to isolate expressed sequences in general and the 5' ends of candidate genes for some human genetic diseases (4,5).

Most known HTF islands have been identified fortuitously following the isolation of specific genes and hence current assertions about such sequences may be biased. However, recently, the correspondence between HTF islands and expressed sequences has been tested prospectively with success suggesting that the identification of HTF islands will be useful in predicting the location of genes (4,5). Despite this, at present little is known about the function of HTF islands, or the mechanisms by which they evolve and are maintained.

In this study we have analysed the distribution and evolution of HTF islands within the human α-globin gene cluster. This multigene family which

has arisen by duplication and divergence of an ancestral globin gene, includes the two α globin genes (α2 and α1), an embryonic globin gene (ζ2), three pseudogenes (ψζ1, ψα1 and ψα2) (6,7) and a gene (θ1) (8) whose function is yet to be determined. A previous report showed that the human α genes have HTF islands at their 5' ends but the ψα1 gene appears to have lost its HTF island during evolution (2). In this study we have identified three additional HTF islands; at the 5' end of the θ1 gene, within the 3' portion of the ζ2 gene and an island spanning the 3' end of the ψζ1 and 5' end of ψα2. These findings are relevant to the evolution and function of HTF islands in general and have important implications for the strategies used to search for genes through their association with HTF islands.

MATERIALS AND METHODS

Analysis of DNA sequence

A large proportion of the human α globin gene complex has been previously sequenced. The sequenced region extends from 770 bp 5' of the ζ2 mRNA cap site to the 3' end of the 3' HVR excluding one segment of 4.6 kb between the two ζ-like genes and another of 6.6 kb between the α1 gene and 3' HVR that have not yet been sequenced (see figure 1). The sequence data base was searched for restriction sites corresponding to the methylation sensitive restriction enzymes Hpa II (CCGG), Hha I (CGGG) and eighteen rare-cutting, CpG-containing enzyme sites listed by Van Ommen and Verkerk (9). Segments of the α globin complex where such sites occurred frequently were further analysed in steps of 200 bp for G+C content, and the ratio of observed to expected (dependant on G+C content) CpG content to assess the degree of CpG suppression within such segments of DNA.

Methylation studies

DNA was obtained from semen, fetal brain, teratocarcinoma cells (10), peripheral blood leukocytes, and the human erythroid line K562 (11). DNA extraction and Southern blot analyses were as previously described (12) except that probes were often labelled by the random primer method (13). The ζ-α globin genotype was determined for each sample to distinguish chromosomes with ζ2-ψζ1 and ζ2-ζ1 arrangements (14). To assess the methylation status of a particular site, limit digests obtained with Hind III, Pst I or Pvu II were further digested with the enzyme under examination. In some instances a direct comparison between methylation sensitive (Hpa II, CCGG) and methylation insensitive (Msp I, CCGG or $C^{m}CGG$) enzymes were made. The probes used were a (0.45 kb) Bam HI/EcoRI fragment

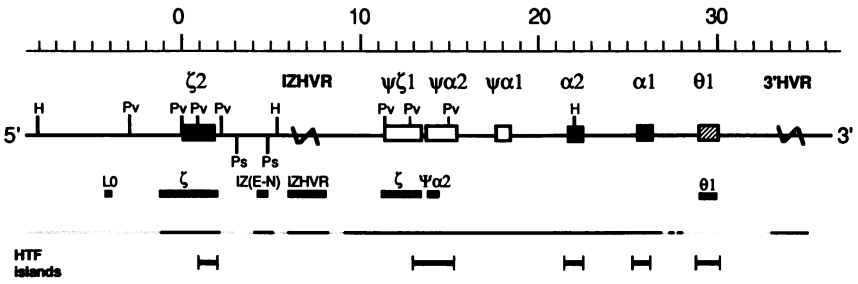

Figure 1: Map of the α-globin complex. Numbers are in kilobases, 0 being the cap site of the ζ2 m-RNA. ■, genes; □, pseudogenes; ▨, gene with undetermined function; ∧∧, hypervariable regions; restriction enzymes used in double digests with methylation sensitive enzymes are H, HindIII; Pv, PvuII; Ps, PstI; The probes used are shown by the black rectangles. The continous black lines below the probes show the sequenced DNA while the stippled lines indicate the non sequenced DNA. The positions of HTF islands are shown below.

(LO) from a subclone of the 6.8 kb Bam HI fragment upstream of ζ2 (p Bam 6.8); an EcoRI-NcoI fragment, IZ (E-N), from the recombinant pSac 2.7; an Alu 1 fragment containing the interzeta hypervariable region IZHVR (15); a 3.1 kb Bam Hl/EcoRI fragment containing the ζ gene (15); a 0.3 kb Bam HI fragment from pCSG1 that contains part of the ψα2 gene; and a 0.8 kb Bam HI fragment (θ1) containing sequences from the human θ1 globin gene (8). The positions of these probes are given in Figure 1.

RESULTS

HTF island within the ζ2 gene

The DNA sequence from the end of the first intron of the ζ2 gene up to 200 bp beyond the translation termination signal (TAA) has an average G+C content of 76.1%. Within this segment of the ζ gene the ratio of observed to expected frequencies of CpG is 0.84 and hence there is very little suppression of this dinucleotide. Furthermore, this region contains many restriction sites for Hpa II/Msp I, Hha I and other enzymes whose recognition sequences contain CpG (figure 2). At the 5' end of the ζ2 gene none of these features is present. Since current sequence data extend only 350 bp beyond the ζ2 gene the 3' border of this potential HTF island is less certain. However, the last 150 bp of the sequenced DNA have a G+C content of only 43.7%, no CpG dinucleotides and thus no Hpa II/Msp I, Hha I sites or sites for rare cutting restriction enzymes. Hence sequence analysis

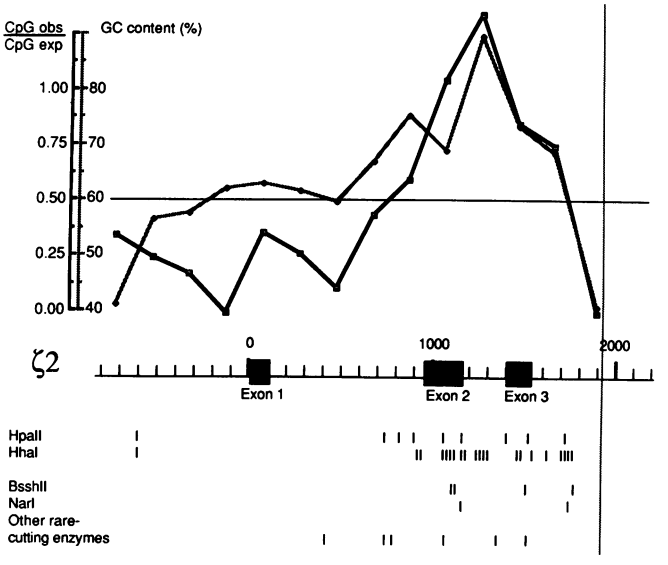

Figure 2: Base composition (▨), ratio of observed to expected CpG dinucleotide frequencies (■) and methylation sensitive enzymes across the ζ2 gene. Numbers are in basepairs, 0 being the cap site of the ζ2 m-RNA. Above the gene are the G+C content and the ratio of observed to expected (from the G+C content) CpG dinucleotides. The horizontal line represents the mean value for those parameters over the whole of the sequenced α-globin complex (see table 1). The vertical line at 1920 bp represents the end of the sequenced DNA (see figure 1). Sites for CpG containing enzymes are shown below the gene.

indicated that an HTF island exists between coordinates +800 and +1800 of the ζ2 globin gene.

Most of this putative HTF island is contained within a 1250 bp Pvu II fragment that can be identified and distinguished from other ζ-specific fragments using the 3.1 Bam HI/EcoRI ζ probe (Figure 1). This 1250 bp band disappears completely in double digests with Pvu II and Hpa II or Msp I (Figure 3) indicating that most of the Hpa II sites within this stretch of DNA are unmethylated in DNA obtained from semen or bone marrow. In addition, at least one Nar I site and one Bssh II site within this region are cut and hence presumably unmethylated in semen and K562 cells, as judged from double digests (Hind III/Nar I and Hind III/Bssh II) hybridised with the IZ (N-E) and the LO probe (Figure 3).

On the 5' side of this region, the 848 and 904 bp bands were only trimmed by Hpa II due to a small overlap of the putative HTF island at the

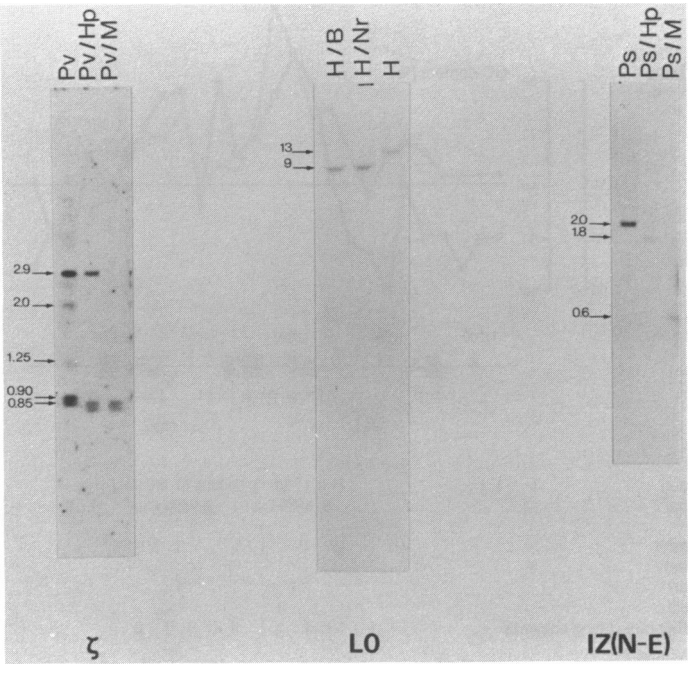

Figure 3: Methylation of CpG containing enzyme sites at the ζ2 gene in sperm DNA. Probes used are shown below the autoradiographs. Sizes are in kilobases. Enzymes used were Pv, PvuII; Hp, HpaII; M, MspI; H,Hind III; B, BssHII; Nr, NarI; Ps, PstI.

3' ends of these fragments. The 2.9 kb ζ-2 specific band was uncut by Hpa II although it disappeared when cut with Msp I. Similarly, a 2 kb Pst I fragment 3' of the HTF island detected with the IZ (E-N) probe was either uncut or only partially cut by Hpa II but cut to completion by Msp I (Figure 3). However a Nar I site and an Sst II site beyond the putative island appear to be ummethylated. Hence the region within the 3' portion of the ζ2 gene conforms to the criteria laid down for the definition of an HTF island, in that it is a segment of ummethylated CpG rich sequence surrounded by methylated or partially methylated DNA.

HTF island at the ψζ1-ψα2 gene

The DNA sequence from the end of the first intron of the ψζ gene to the 3rd exon of the ψα2 gene (2000 bp) has an average G+C content of 72.2%. The ratio of observed to expected frequencies of the dinucleotide CpG is again high (0.79) and this segment also contains many sites for CpG containing,

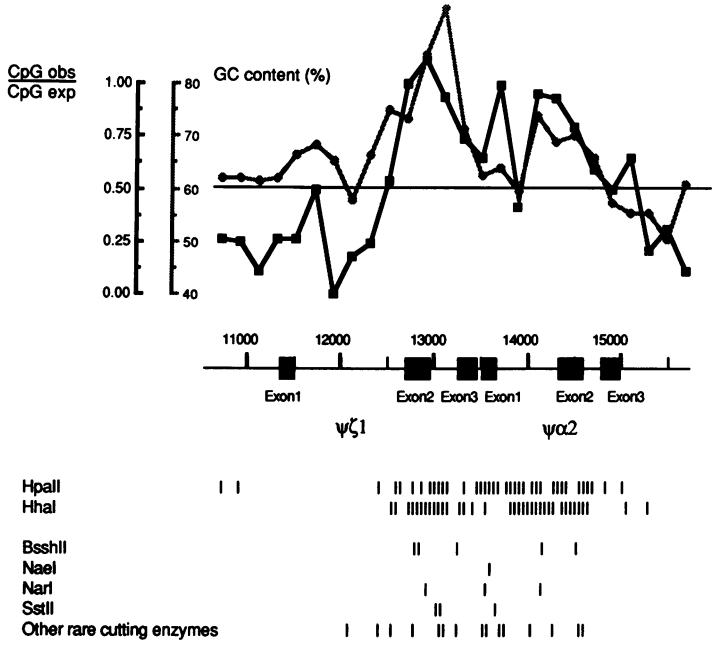

Figure 4: Base composition (▰), ratio of observed to expected CpG dinucleotide frequencies (▬), and methylation sensitive enzymes across the ψζ1/ψα2 genes. Numbers are in basepairs, 0 being the cap site of the ζ2 m-RNA (see figure 1). Above the gene are the G+C content and the ratio of observed to expected (from the G+C content) CpG dinucleotides. The horizontal line represents the mean value for those parameters over the whole of the sequenced α-globin complex (see table 1). Sites for CpG containing enzymes are shown below the genes.

rare-cutting restriction enzymes indicating that this region is a putative HTF island (Figure 4).

A 2.0 kb Pvu II fragment, identified by the 3.1 kb EcoRl/Bam HI ζ probe, contains most of this putative island. Digestion of this region with Hpa II/Msp I indicates that the majority of sites within this region are unmethylated (Figure 3). Furthermore, this region also contains unmethylated sites for the enzymes Sst II, Bssh II, Nae I, and Nar I as judged from double digests hybridised with the IZHVR and ψα2 probes (Figure 5). Since some of these sites are very close to each other it was not always possible to determine exactly which site had been cut. We have previously shown that regions surrounding this segment are fully or partially methylated in DNA from a variety of tissues by comparison of Hpa II and Msp I digests (2).

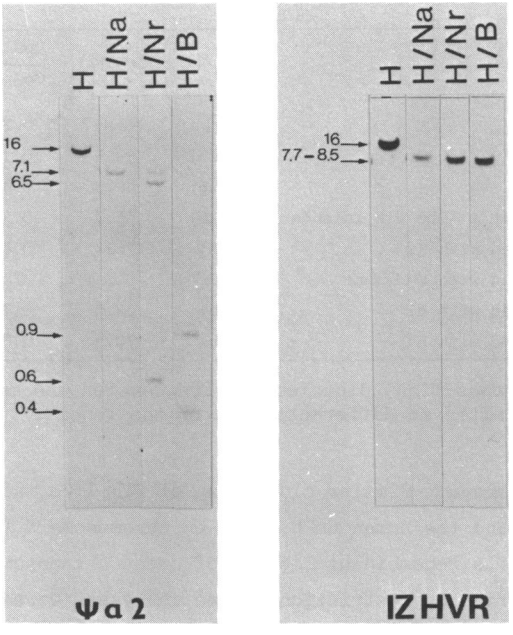

Figure 5: Methylation of CpG containing enzyme sites at the ψζ1/ψα2 genes in sperm DNA. Probes used are shown below the autoradiographs. Sizes are in kilobases. Enzymes used were H, HindIII; B, BssHII; Nr, NarI; Na, NaeI.

HTF island at the θ1 gene

Although the sequence of the human θ1 gene has not yet been determined the homologous gene from Orang-Utan shows a G+C content of 70.8%, a ratio of observed to expected frequencies of CpG dinucleotides of 0.78 and many potential sites for Hpa II/Msp I, Hha I and other CpG containing restriction enzymes. Previous mapping of the homologous human α-globin complex using rare cutting restriction enzymes demonstrated that the segment of DNA containing the human θ1 gene also contains unmethylated sites for Sst II, Mlu I and Xho I (16). Using this segment as a probe (θ1), Pst I digested DNA cut with Hpa II or Msp I indicated that this region of DNA is also unmethylated and hence corresponds to another HTF island (data not shown).

DISCUSSION

It has been calculated that, on average, there will be one HTF island for every 100 kb of the genome although variation around this estimate is likely to be quite large (17). Within the 35 kb of the human α-globin gene complex there are at least 5 HTF islands (see Figure 1) which is far above

TABLE 1

		C+ G(%)	$\frac{\text{CpG obs}}{\text{CpG exp}}$	Hpall sites/kb
Whole genome	($3x10^6$kb)	40.0	0.25	0.6
β-globin complex	(73.3kb)	39.5	0.17	0.3
α-globin complex	(24.3kb)	60.8	0.45	5.6
α-globin complex without the HTF islands	(17.7kb)	56.7	0.33	2.9
HTF island associated with $\zeta 2$	(1.1kb)	76.1	0.84	7.2
HTF island associated with $\psi\zeta 1/\psi\alpha 2$	(2.5kb)	72.2	0.79	13.6
HTF island associated with $\alpha 2$	(1.7kb)	69.0	0.74	12.9
HTF island associated with $\alpha 1$	(1.3kb)	70.9	0.82	15.4

Table 1: Base composition, dinucleotide frequencies and methylation sensitive Hpa II sites in different parts of the genome.

this estimated average. Similar clustering of HTF-like islands has also been reported around the human G6PD locus on chromosome X (21, 22) and in a recent analysis of a recombinant library of human chromosome 3 many HTF-like sequences, identified by restriction enzyme analysis, were shown to be less than 10 kb apart (18). Such clustering may reflect HTF islands within multigene families, such as the α globin complex. However the organisation of the genome may be such that genes may also cluster in certain regions, particularly within the extended stretches (>>200kb) of homogeneous G+C rich DNA that constitute heavy isochores (19, 20). It is of interest that the α globin gene family lies within such a G+C rich isochore (16,19,20) and that recently we have been able to identify other HTF-like islands closely flanking the α-globin complex (unpublished observations).

Hitherto nearly all HTF islands have been found at the 5' end of functional genes (1). The only exceptions to this are the two HTF islands at the 3' end of the G6PD gene (21,22). However, these islands lie approximately 40 kb 3' of the G6PD gene and are associated with the 5' ends of two previously unidentified genes that lie beyond the G6PD gene (Toniolo, D. et al., in press). The consistent association of HTF islands with the 5' ends of functional genes, particularly housekeeping genes, has led to the proposal that they might signal a particular class of protein-DNA interaction necessary for efficient transcription of the associated gene (1). The HTF island at the 5' end of the $\theta 1$ gene appears to conform to this pattern but as yet the functional status of this gene has not been determined. The embryonic $\zeta 2$ globin gene appears to have a conventional polymerase II eukaryotic promoter (23) but the HTF-island associated with

TABLE 2

	HTF-like islands associated with		
	α-globin	ζ-globin	β-globin
Human	+	+	-
Horse	+	+	NA
Goat	+	+	-
Mouse	-	-	-
Rabbit	+	NA	-
Chicken	+/-	-	+/-
Xenopus	-	-	-

Table 2: Association of HTF like islands with globin genes in different species. This table was constructed by analysing sequencing data from the EMBL and Genbank data libraries for G+C content (%), ratio of observed to expected CpG dinucleotides and HpaII sites. No methylation data was taken into account for species other than human. For Xenopus only the sequence of exons (cDNA) was available for analysis. All the genes considered to be associated with HTF like islands (+) had a GC content of over 69.6% for at least 500 bp, a CpG_{obs}/CpG_{exp} ratio of over 0.68 and over 7 HpaII sites per kb. Genes considered not to be associated with HTF like islands (-) fulfilled at least 2 of the following 3 criteria; GC content under 50%; CpG_{obs}/CpG_{exp} ratio under 0.35; HpaII sites/kb under 2; In chicken the α^D gene was not associated with an HTF like island but the α^A gene and the β-globin gene fulfilled neither the above criteria for presence nor absence of an HTF like island.

this gene is atypical in that it spans the 3' end of the gene. Similar HTF-like sequences are also seen at the 3' portion of the ζ genes in at least two other species (See Table 2). It is possible that, as demonstrated for other globin genes (24) efficient transcription of ζ2 is dependent on sequences at both the 5' and 3' end of the body of the gene. In the light of this unexpected finding it is of interest that the DNase I hypersensitive site, associated with expression of the ζ2 gene in the embryonic cell line K562 also maps to the same 3' region as the HTF island (25). It is possible that both types of analyses are pointing to critical DNA protein interactions at the 3' end of the ζ2 gene that are necessary for its appropriate expression. Interestingly, both HTF islands and DNase I hypersensitive sites associated with the α globin genes are located at the 5' ends of the genes.

The HTF island that spans the 3' end of ψζ1 and the 5' end of ψα2 is the first example of such an island associated with pseudogenes; the previously analysed ψα1 gene appears to have lost its HTF-island following

gene inactivation (2). Inevitably, insufficient time will have elapsed for recently inactivated pseudogenes to lose HTF-like characteristics associated with their previously functional status. This argument could apply to the $\psi\zeta 1$ gene but not to the $\psi\alpha 2$ gene which is estimated to have been inactive for over 60 MY. It is also possible that an HTF island could be maintained in a pseudogene by gene conversion against a corresponding HTF island within a functional homologue. Such a mechanism is unlikely to explain the $\psi\zeta 1$-$\psi\alpha 2$ HTF island because it extends beyond the previously defined $\zeta 2/\psi\zeta 1$ conversion unit into a region where the restriction maps indicate that there is little homology between $\zeta 2$ and $\psi\zeta 1$. Thus at present it is not clear why this HTF island should be kept in association with the non-functional, $\psi\zeta 1$ and $\psi\alpha 2$ pseudogenes.

These unexpected observations on HTF islands within the human α-globin gene family may be important in understanding how HTF islands have arisen and been maintained throughout evolution. Comparison of our results with analysis of sequence data from several other species provides a basis for preliminary speculation about the evolutionary history and importance of these HTF-islands. Although the α and β globin genes are derived from the same ancestral gene (Reviewed in 26) and are coordinately expressed in erythroid precursors, the α like genes are frequently associated with HTF-like islands whereas the β globin genes appear not to be associated with such sequences (See Table 1). Furthermore, the appearance of HTF islands in the α globin cluster may have occurred around the time of the mammalian radiation (85 MY ago) since they are not present in Xenopus or chicken but are present in all mammals that have been studied except mouse in which neither α nor ζ like genes are associated with islands (See table 2). It is possible that these findings reflect the evolution of a significant difference in the mechanisms by which the expression of α-and β-like genes are controlled. The behaviour of human α and β genes in various expression systems has been shown to be quite different (24, 27) and in the light of these observations further studies comparing the function of α like genes with (eg. human, rabbit) or without (eg. mouse) HTF islands will be instructive.

Despite the uncertainties in our understanding of the evolution and function of HTF islands, they are clearly of value in the search for unidentified genes. However, two new principles emerge from this study; first, that a proportion of HTF islands will be associated with pseudogenes and second, that such islands may be located within the 3' portions of functional genes.

REFERENCES

1. Bird, A.P. (1986). Nature, 321, 209-213.
2. Bird, A.P., Taggart, M.H., Nicholls, R.D. and Higgs, D.R. (1987). EMBO J., 6, 999-1004.
3. Bird, A.P., Taggart, M.H., Frommer, M., Miller, O.J. and Macleod, D. (1985). Cell, 40, 91-99.
4. Lindsay, S. and Bird, A.P. (1987). Nature, 327, 336-338.
5. Estivill, X., Farrall, M., Scambler, P.J., Bell, G.M. Hawley, K.M.F., Lench, N.J., Bates, G.P., Kruyer, H.C., Frederick, P.A., Stanier, P., Watson, E.K., Williamson, R. and Wainwright, B.J. (1987). Nature, 326, 840-845.
6. Lauer, J., Shen, C.-K. and Maniatis, T. (1980). Cell, 20, 119-130.
7. Hardison, R.C., Sawada, I., Cheng, J.-F., Shen, C.J. and Schmid, C.W. (1986). Nucleic Acids Res., 14, 1903-1911.
8. Marks, J., Shaw, J.-P. and Shen, C.J. (1986). Nature, 321, 785-788.
9. Van Ommen, G.J.B. and Verkerk, J.M.H. (1986). In: Human genetic diseases. A practical approach, 113-133, (IRL Press, Oxford).
10. Thompson, S., Stern, P.L., Webb, M., Walsh, F.S., Engstrom, W., Evans, E.P., Shi, W.-K., Hopkins, B. and Graham, C.F. (1984). J. Cell. Sci., 72, 37-64.
11. Rutherford, T.R., Clegg, J.B. and Weatherall, D.J. (1979). Nature, 280, 164-165.
12. Old, J.M. and Higgs, D.R. (1983). In: The Thalassemias. Methods in Hematology, 6, 74-102.
13. Feinberg, A. and Vogelstein, B. (1984). Anal. Biochem., 137, 266-267.
14. Hill, A.V.S., Nicholls, R.D., Thein, S.L. and Higgs, D.R. (1985) Cell, 42, 809-819.
15. Higgs, D.R., Wainscoat, J.M., Flint, J., Hill, A.V.S., Thein, S.L., Nicholls, R.D., Teal, H., Ayyub, H., Peto, T.E., Falusi, A.C., Jarman, A.P., Clegg, J.B. and Weatherall, D.J. (1986). Proc. Natl. Acad. Sci., USA, 83, 5165-5169.
16. Fischel-Ghodsian, N., Nicholls, R.D. and Higgs, D.R. (1987) Nucleic Acids Res., 15, 6197-6207.
17. Brown, W.R.A. and Bird, A.P. (1986). Nature, 322, 477-481.
18. Smith, D.I., Golembieski, W., Gilbert, J.D., Kizyma, L. and Miller, O.J. (1987). Nucleic Acids Res., 15, 1173-1184.
19. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985). Science, 228, 953-958.
20. Zerial, M., Salinas, J., Filipski, J. and Bernardi, G. (1986). Eur. J. Biochem., 160, 479-485.
21. Toniolo, D., D'Urso, M., Martini, G., Persico, M., Tufano, V., Battistuzzi, G. and Luzzatto, L. (1984). EMBO J, 3, 1987-1995.
22. Persico, M.G., Viglietto, G., Martini, G., Toniolo, D., Paonessa, G., Moscatelli, C., Dono, R., Vulliamy, T., Luzzatto, L. and D'Urso, M. (1986). Nucleic Acids Res., 14, 2511-2522.
23. Proudfoot, N.J., Gil, A. and Maniatis, T. (1982). Cell, 31, 553-563.
24. Charney, P., Treisman, R., Mellon, P., Chao, M., Axel, R. and Maniatis, T. (1984). Cell, 38, 251-263.
25. Yagi, M., Gelinas, R., Elder, J.T., Peretz, M., Papayannopoulou, T., Stamatoyannopoulos, G. and Groudine, M. (1986). Molec. Cell. Biol., 6, 1108-1116.
26. Bunn, H.F. and Forget, B.G. (1986). In: Hemoglobin: Molecular, Genetic and Clinical Aspects, 126-169.
27. Mellon, P., Parker, V., Gluzman, Y. and Maniatis, T. (1981). Cell, 27, 279-288.