Long range genome structure around the human α-globin complex analysed by PFGE

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

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

Received May 13, 1987; Revised and Accepted June 26, 1987

ABSTRACT

A map encompassing 300 kilobases (kb) in and around the human α-globin gene complex shows features with important implications for understanding the structure and function of the human genome. In contrast to other segments of the mammalian genome that have been analysed by pulsed field gradient electrophoresis (PFGE), this region contains an unusually high density of sites for infrequently cutting restriction enzymes that recognise GC rich motifs including the under-represented CpG doublet. This suggests that the 26 kilobase (kb) stretch of DNA containing the α-globin gene family, which is known from sequence analysis to be 60% GC rich, is itself embedded within a region of high GC content. This long-range structure, identified by PFGE, corresponds to a class of GC rich isochores that are thought to represent early replicating DNA present in Giemsa negative chromosomal bands. The identification of such regions by PFGE will be of value in understanding the organisation of human chromosomes and will influence the strategies used to construct a physical map of the genome.

INTRODUCTION

Pulsed field gradient gel electrophoresis (PFGE) and Southern blotting (1,2) can be used to separate and identify very large fragments of DNA in the order of 20-2000 kb. These techniques have been used to establish long range restriction maps of genomic DNA spanning several megabases around the human major histocompatability complex (3,4) and the Duchenne muscular dystrophy locus (5-7). The large DNA fragments necessary for this type of analysis are generated by using restriction enzymes that recognise eight-base pair cleavage sites and/or contain CpG in their recognition sequence. The dinucleotide CpG is frequently methylated (60-90%) and under-represented (20-25% of the expected frequency) in vertebrates (8). Enzymes containing CpG in their recognition sequence are often sensitive to the presence of methylation and hence may cut only a subset of the relatively small number of potential sites. The specificity of these restriction enzymes confers on them a predilection to cleave unusual regions of the genome that contain a high frequency of the dinucleotide CpG and that are unmethylated in the tissues studied.

We have used PFGE and Southern blotting to construct a long-range map around the human α-globin gene complex. The map is radically different from that described for other loci in that there is a high density of cleavable sites for CpG-containing enzymes throughout a region of at least 300 kb in and around the functional α globin genes precluding the construction of a 'megabase map'. This suggests that the α-globin complex which is known from sequence analysis to be C+G rich (60% over 24 kb) may be imbedded within a much longer segment of C+G rich DNA. This unusual structure identified by PFGE may correspond to a C+G rich isochore (9,10) and thus represents a new feature of the genome that can be recognised by this type of analysis.

MATERIALS AND METHODS

A complete restriction map for 5 CpG containing enzymes (*Not* I, *Sal* I, *Mlu* I, *Xho* I, *Sst* II) as well as *Sfi* I and *Asp* 718 was determined for 150 kb of cloned DNA spanning the human α-globin gene complex. A corresponding map from genomic DNA was also obtained allowing the accessibility of certain sites in genomic DNA to be assessed and additional sites, flanking the cloned DNA, to be identified.

Analysis of cloned DNA around the α complex

Cosmids cα3'Bg, cSG1, cRN24, CNFG2, cRN2103, cRA36, and the plasmid pHT-HVR1 have been previously described (see figure 1) (11,12). The region of cloned DNA was extended by isolating the cosmid cNFG9 (figure 1) from a human genomic library (CV007K) prepared in the cosmid vector pCV007. This library was kindly made available by K.H. Choo of the Murdoch Birth Defects Research Institute, Royal Children's Hospital, Parkville, Australia. The construction of the vector and library has been described (13). Plating and screening techniques were as described (11) except that 500,000 colonies were analysed for the primary screen which represents 3-4 human genome equivalents.

A restriction map of the cloned DNA was constructed by analysis of sequence data (14) (from the beginning of the ζ2 gene until just beyond the 3' HVR with gaps of 4.6 kb between the ζ genes and 6.6 kb between the α1 and 3' HVR) together with single, double and triple digests of cloned DNA and purified fragments.

Analysis of genomic DNA around the α-globin complex

Restriction sites in genomic DNA were analysed by standard DNA isolation and Southern blot hybridisation (15,16) as well as by PFGE. Single and double digests were analysed. DNA was obtained from semen, blood, Epstein-Barr virus immortalized lymphocytes and K562 cells. High molecular weight DNA (>400 kb) was prepared by incubation of ~10^8 cells with buffer (10mM NaCl, 10mM Tris-HCl, pH = 8.0), proteinase K (40 µg/ml) and 10% SDS for 4 hours (for semen extractions 400 µl 1M DTT was added), followed by phenol-chloroform extraction and ethanol precipitation. The integrity of the DNA was maintained by mixing the phases very gently, centrifuging at below 1200 rpm and using wide bore pipette tips throughout. The yield of DNA was approximately 250 µg per 10^8 cells. DNA was digested for 4 hours with 10 µg of DNA per sample in a volume of 80 µl using a 5-10 fold excess of restriction enzyme. Samples were loaded on 1.2% agarose gels in 0.5 TBE(0.05 M Tris-HCl, 0.05 M Boric acid, 1 mM EDTA) and run for 18 hours at 11 v/cm in an orthogonal field electrophoresis apparatus. The apparatus was as described by Brown and Bird (17). Depending on the required size range the switching time was varied between 5 and 50 seconds (at 5-10 seconds, maximum resolution was between 20 and 150 kb and with increasing switching time larger fragments could be resolved). Cooled buffer (4-14^{o}C) was continuously circulated. After electrophoresis, gels were stained with ethidium bromide and exposed to a 254 nm transilluminator for 120 seconds. The DNA was then denatured in 1 M NaOH for 30 minutes, neutralized (3 M NaCl, 1 M Tris-HCl, pH 7.5) for 60 minutes, transferred to a nylon membrane (Hybond, Amersham) for 12-24 hours and crosslinked to the membrane by exposure for 120 seconds to U.V. irradiation (254 nm). DNA probes (see legend to figure 1) were hexanucleotide labelled with [α-^{32}P] dCTP (18) to a specific activity of at least 4 x 10^6 cpm/ng. Nylon filters were hybridised at 42^{o}C in 50% formamide, 1 x SSC and washed at high stringency (0.1 x SSC, 0.1% SDS, 65^{o}C) before autoradiography. Filters were rehybridised on several occasions, after washing for 30 minutes at 45^{o}C with 0.4 M NaOH and then 0.1 x SSC/0.1% SDS/0.2 M Tris-HCl, pH 8.0 according to manufacturer's instructions (Hybond, Amersham) to remove previously hybridised probe.

RESULTS

The long range restriction map around the α-globin complex

Restriction sites for seven enzymes are shown in figure 1 and examples of genomic mapping, using PFGE, are shown in figures 2 and 3. Within the cloned DNA the map is complete except for Sfi I for which additional sites may exist in the immediate vicinity of the sites at -27 kb, -58 kb and -61 kb. Beyond the cloned DNA, additional sites may exist which are only accessible in selected tissues, or at specific

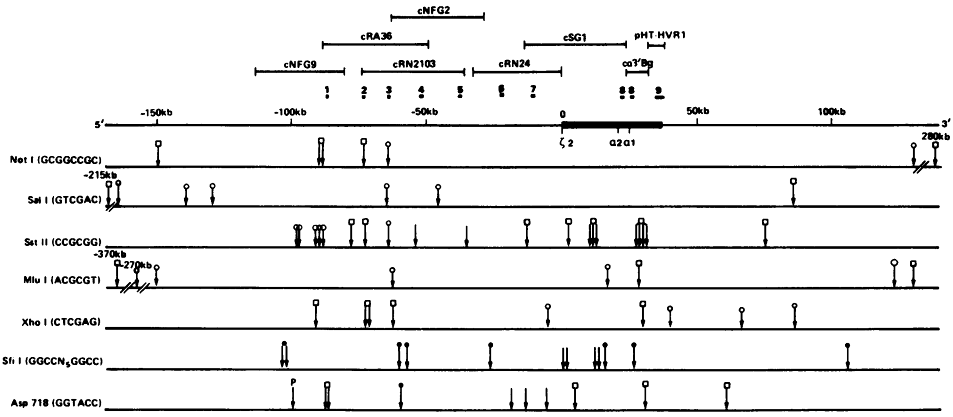

Figure 1 Restriction Enzyme Map in and around the α-Globin Locus. Restriction sites for 7 enzymes are shown by arrows. Sites denoted by (□) cut to completion in all the samples examined. Sites denoted by (○) have been shown to cut genomic DNA partially in at least one tissue, while non-methylation sensitive sites for which partials have been obtained are shown by (●). For sites shown thus (↓) no information is available in genomic DNA. An Asp 718 site which exists only in some of the individuals and has not been seen in the cloned DNA is denoted by (P). Coordinate 0 represents the mRNA cap site of the ζ2 gene and the mRNA cap sites of the α genes are shown within the black rectangle that represents the α globin locus. The probes used are denoted 1 = RA 2.2, 2 = RA 0.3, 3 = RA 1.0, 4 = RA 330, 5 = RA 1.4, 6 = L2, 7 = L1, 8 = α(Pst I), 9 = 3' HVR (Hinf I) (11). Cloned DNA from -113 kb to +37 kb is shown above.

developmental stages. The frequency of sites for CpG containing enzymes throughout this 300 kb stretch of DNA is far greater than in the previously published maps (3-7) or than observed in random cosmids containing segments of human DNA (19). In the latter more than half of the cosmids assessed had no sites for these rare-cutting enzymes.

Accessibility of restriction sites

The fact that we had mapped sites for CpG containing enzymes in cloned DNA provided the opportunity to investigate the causes of variation in their accessibility in genomic DNA. Using DNA from sperm or EBV transformed lymphocytes, some sites (eg. Mlu I at coordinate +28 and Xho I at coordinate +29) are always accessible but many are only partially cut by these enzymes. Three recalcitrant sites (Not I at -64, Mlu I at +16 and Xho I at -6) were examined in detail.

The accessibility of these sites was independent of DNA size, conditions of digestion and protein/DNA interaction (see figure 4). Furthermore digestion was independent of the method used for DNA preparation (data not shown). Filters with DNA prepared in agarose blocks (kindly

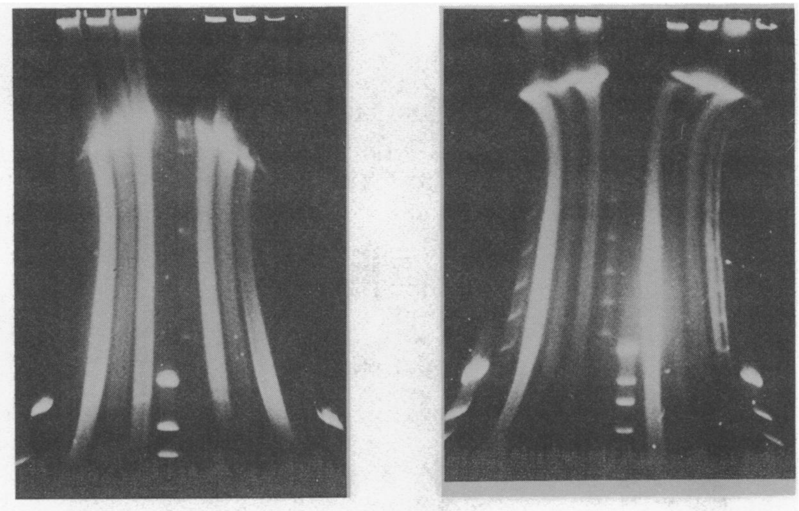

Figure 2 Representative Digests of Genomic DNA Stained with Ethidium Bromide. Gels were run for 18 hours with 5 seconds (left) and 20 seconds (right) switching time. These conditions give good resolution up to 150 kb and from 50 to 400 kb respectively. λ Hind III and λ multimers were run in both central and peripheral tracks. DNA fragments ranging from ~10 to 400 kb were seen following single digests with a variety of restriction enzymes. The identification of large genomic fragments demonstrates the integrity of DNA prepared in solution.

provided by Susan Kenwrick and Mark Patterson; for method of preparations see reference 7) gave identical digestion patterns to those using DNA prepared in solution. In addition, one semen sample was divided into 4 aliquots and DNA extracted from each of them. Hybridisation, using probe 9, to Mlu I or Xho I digests of these aliquots gave reproducible partial digestion patterns with respect to the Mlu I site at +16 and Xho I site at -6.

The ability of Not I to cut the site at -64 was assessed in twelve different tissues by double digests (data not shown). In ten of them this site was not cut, but partial cleavage was seen in fetal skin (30%) and fibroblasts (50%) indicating a tissue-specific modification of the DNA was responsible for the incomplete cleavage of this site. Since all three sites (Not I, -64, Mlu I +16 and Xho I at -6) could be fully digested in cloned DNA in which CpG is unmethylated it appears that the major determinant of accessibility is the degree to which these and other (17) CpG containing recognition sequences are methylated.

The high frequency of restriction sites for CpG containing enzymes

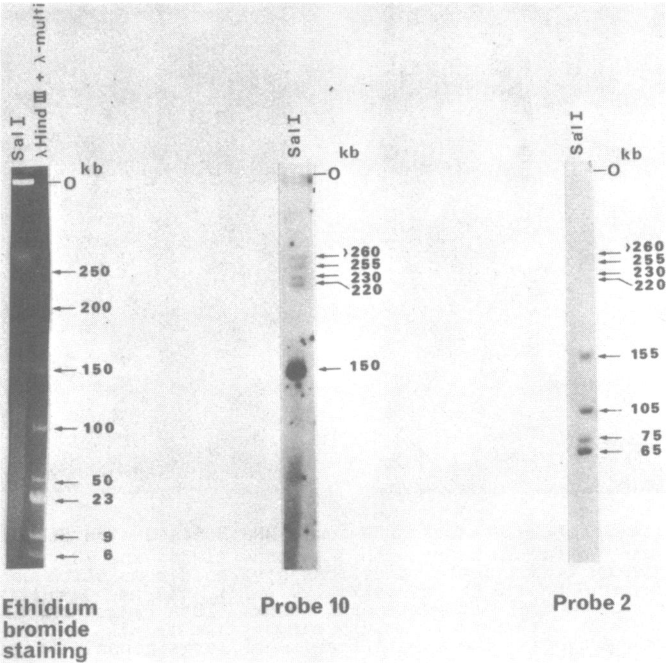

Figure 3 Normal Sal I Restriction Pattern.

Ethidium bromide stain and two autoradiographs of a Sal I digest hybridised with probes 2 and 10. Sites at +85 kb and -215 kb have cut to completion whereas partial digestion is seen at -65 kb, -130 kb, -140 kb and -170 kb, and no digestion at -45 kb (see figure 1).

detected in cloned DNA is reflected in the relatively small sizes of the genomic fragments observed. This suggests that a large proportion of these sites are cut in the tissues studied. In fact, thirteen of the thirty four methylation-sensitive sites that map within the cloned segment always cut to completion in genomic DNA. A further nine cut partially in at least one tissue. Some sites could not be easily assessed because they clustered such that any resulting partials would not have been detected in the gel systems used here (eg. the cluster of Sst II sites at +29 kb). In addition, no site studied in detail was found to be methylated in all tissues.

DISCUSSION

The most striking feature of the map established here is the density of sites for what are normally considered to be infrequently cutting enzymes. The map is markedly different to those around the human major

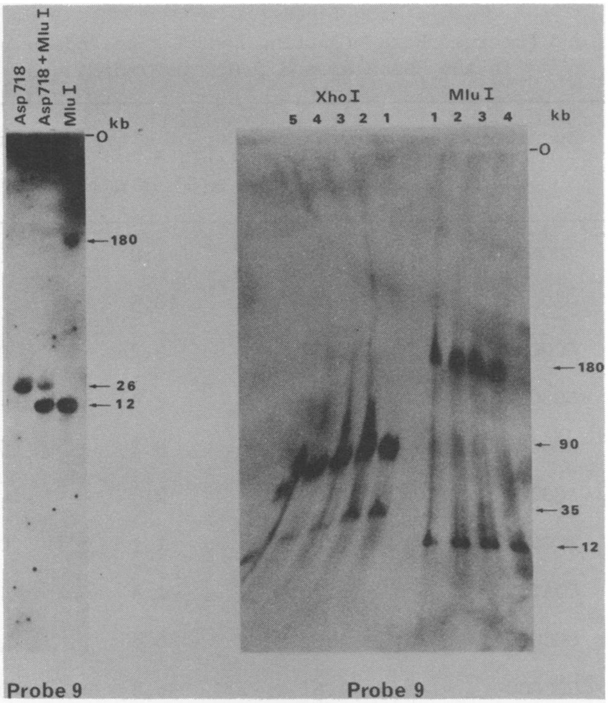

Figure 4 Site Accessibility is Independant of DNA Size, Conditions of Digestion and Protein/DNA Interaction.

The right hand panel depicts the partially digested fragments of 180 kb (Mlu I) and 90 kb (Xho I) hybridised to probe 9. Sample 1: Digestion for 4 hours with 10-fold excess of the appropriate enzyme. Sample 2: Digestion for 18 hours with 20-fold excess of enzyme. Sample 3: After digestion for 4 hours with 10-fold excess of enzyme, incubation for 4 hours with proteinase K, inhibition of proteinase K activity by adding PMSF and subsequent redigestion for 4 hours. Sample 4: As sample 3 but phenol/chloroform extraction and ethanol precipitation after proteinase K and before redigestion. Samples 1 to 4 are all from the same semen sample. Sample 5 was an unrelated Xho I digest.

The left hand panel shows Asp 718, Mlu I and Asp 718, and Mlu I digests hybridized with probe 9 to evaluate the accessibility of the Mlu I site at +16 kb. Asp 718 cuts to completion at +4 kb and +30 kb and the Mlu I site cuts to completion at +28 kb. The relative intensities of the bands at 180 kb and 12 kb with Mlu I alone appear identical to the relative intensities of the 26 kb and 12 kb bands obtained with Asp 718 and Mlu I. Therefore, the Mlu I site at +16 kb is no more accessible in the double digest compared with the single digest.

histocompatibility (3,4) or Duchenne Muscular Dystrophy loci (5-7), for example. To some extent the frequency of sites within the GC rich α globin cluster could have been predicted. On theoretical grounds (see legend to table 1) such enzymes should cut much more frequently in GC rich segments of the genome than in normal or GC poor regions especially when the effect of CpG suppression (8) is taken into account. A comparison of the frequency of

TABLE 1.
Predicted and Observed Rare - Cutting Restriction Endonuclease Sites in the Human α- and β-Globin Locus.

Restriction enzyme	Specificity	Number of sites /100 kb			
		Predicted		Observed	
		β(39.5% GC)	α(60.6% GC)	β	α
Not I	GCGGCCGC	0.013	1.0	0	0
Sfi I	$GGCC(N)_5GGCC$	0.29	18.5	1.4	20.8
Sst II	CCGCGG	0.37	9.2	0	12.5
Bssh II	GCGCGC	0.30	7.4	1.4	62.5
Xma III	CGGCCG	0.37	9.2	0	25.0
Mlu I	ACGCGT	0.80	2.0	1.4	4.2
Nru I	TCGCGA	0.67	1.1	0	0
Pvu I	CGATCG	0.70	1.3	0	0
Nae I	GCCGGC	1.5	36.8	2.8	20.8
Nar I	GGCGCC	1.5	36.8	0	41.7
Sma I	CCCGGG	1.9	46.5	5.5	174.9
Aat II	GACGTC	3.2	2.9	4.1	12.5
Xho I	CTCGAG	3.6	4.5	2.8	0
Sal I	GTCGAC	3.2	2.9	1.4	0
Fsp I	TGCGCA	5.4	30.3	1.4	8.3
Asu II	TTCGAA	6.1	1.3	2.8	8.3
Cla I	ATCGAT	6.8	2.4	6.8	0
Snab I	TACGTA	6.4	0.9	4.1	0
Total		42.8	215.0	36.8	391.5

18 infrequently cutting restriction enzymes as listed by Van Ommen and Verkerk (33) were analysed. Predicted number of sites per 100 kb was calculated independently for a GC content of 60.6% (α-globin complex) and 39.5% (β-globin complex), with the additional assumption that CpG occurs only at 20% of its expected frequency in the human genome (8) and is accompanied by a commensurate increase of TG and CA doublets (26). The theoretical frequencies for each dinucleotide were then calculated. The probability for the sequence of one restriction enzyme was obtained by the product of the conditional probabilities for each nucleotide taking into account the

immediate preceding one. CpG suppression has been reported to be less in GC rich parts of the genome (34) (CpG's occur at 51.1% of their expected frequency in the α-globin complex while only at 17.5% of their expected frequency in the β-globin complex). Since the predicted number of sites is very sensitive to the amount of CpG suppression the difference between GC rich and GC poor DNA would be even more prominent, e.g. in 51.1% CpG suppressed DNA the total predicted number of sites in the α globin complex would be 421.9 instead of 215.0. It should be noted that in this calculation the clustering of CpG's (as occurs in HTF islands) (25,26) is ignored. Furthermore, it is not yet clear to what extent the reduced CpG suppression in GC rich segments of the genome (34) might reflect the clustering of HTF islands within such regions.

The observed number of sites was obtained by analysing the sequence of 24010 bp of the α-globin locus and 73304 bp of the β-globin locus (14). The number of sites was then expressed per 100 kb. It should be noted that not all sites identified in this way will be cut in genomic DNA due to methylation.

such sites in the sequenced DNA around the human α (24010 bp 60.6% G+C) and β (73104 bp 39.5% G+C) globin gene clusters for example reveals 10.6 times more sites in the α complex per unit length (see table 1 and legend). In this study we have shown that the high frequency of sites extends for at least 300 kb around the locus, presumably reflecting a similarly high GC content in this flanking DNA. It has previously been shown that the α-globin genes, and other expressed sequences including several oncogenes, are contained within a GC rich fraction (H3) of the genome that can be separated from other fractions by virtue of its buoyant density (9,10). It was suggested that these fractions constituted long stretches (>200 kb) of the genome with relatively homogeneous GC content (isochores) and that the nucleotide sequence of genes within such segments reflect this GC content. It is of interest that these GC rich isochores are thought to correspond to the early replicating DNA present in Giemsa negative (light) chromosomal bands obtained at high resolution by Giemsa and Reverse staining (20-24). The map established here is thus consistent with the concept that the α-globin locus is embedded within a larger segment of GC rich DNA.

A second feature of this map is the clustering of restriction sites (e.g. at +28 kb and -90 kb) within the 300 kb segment. It has been suggested that the sites for these CpG sensitive enzymes should not be distributed randomly throughout the genome but should be concentrated into islands of undermethlated, CG rich DNA (so-called Hpa II tiny fragment (HTF) islands) and that these islands are found at the 5' ends of expressed sequences (25,26). It has been calculated that most cleavable sites, and particularly clusters of sites, will lie within such islands (26). Although the α globin genes are known to be associated with HTF islands (27) we did not observe any clustering of sites for the six enzymes used in this study around these genes. Nevertheless, analysis of the sequence of the α globin cluster for all 18 enzymes listed in table 1 showed prominent clustering of sites at the 5' ends of the α-genes, demonstrating that HTF islands may be missed when only a subset of these enzymes are analysed. Another cluster of hypomethylated sites was seen in the region of the newly described θ1 gene (28) (+27 kb to +29 kb), whose function is as yet undetermined. It will be interesting to see whether this and similar regions (eg -73 kb and -90 kb) fulfil the criteria for HTF islands (26) and if so whether they conform to the general hypothesis that these islands correspond to the 5' ends of expressed sequences.

The readily cleavable sites at the 3' end of the α complex between coordinates +27 kb and +29 kb, presumably indicate a hypomethylated GC rich segment of the genome. We have previously identified this area as a region wherein the 3' breakpoints of many large, naturally occurring deletions of the α-complex are found (11). In view of previous observations in prokaryotes implicating hypomethylated GC rich areas in recombination (29) it will be interesting to see if these features are found at other breakpoint cluster regions.

The α-globin complex may not be a particularly unusual locus with respect to the high frequency of sites for 'rare cutters', especially in view of the similar frequency of such sites in a subgroup of a random selection of cosmids from human chromosome 3 (19). Such findings have important implications for strategies to establish a long range map of the human genome or to 'chromosome hop' to regions of interest (30,31). Attempts to map or 'hop' through a segment of DNA such as this would be forestalled by the frequency of sites for the enzymes used in such strategies. It will therefore be necessary to adopt alternative ways of releasing large genomic fragments, such as partial digestion or chemical cleavage (32), to overcome this problem for the α globin locus and structurally similar regions of the genome.

ACKNOWLEDGEMENTS

We thank D.J. Weatherall, S. Kenwrick, M. Patterson, S.T. Reeders, T.E.A. Peto, W.G. Wood and J.B. Clegg for helpful discussions and comments on the manuscript. C. Hatton for hybrid cell lines and L. Rose for typing the manuscript. N.F.-G. was supported by a Fellowship from the Hematology/Oncology Department of the Children's Hospital, Boston, USA.

*Present address: Howard Hughes Medical Institute, Children's Hospital, 300 Longwood Avenue, Boston, MA 02115, USA

REFERENCES

1. Carle, G.F. and Olson, M.V. (1984) Nucleic Acids Res. 12, 5647-5664.
2. Schwartz, D.C. and Cantor, C.R. (1984) Cell 37, 67-75.
3. Hardy, D.A., Bell, J.I., Long, E.O., Lindsten, T. and McDevitt, H.O. (1986) Nature 323, 453-455.
4. Lawrance, S.K., Smith, C.L., Srivastava, R., Cantor, C.R. and Weissman, S.M. (1987) Science, 235, 1387-1390.
5. Burmeister, M. and Lehrach, H. (1986) Nature 324, 582-585.
6. Van Ommen, G.J.B. et al. (1986) Cell 47, 499-504.
7. Kenwrick, S., Patterson, M., Speer, A., Fischbeck, K. and Davies, K. (1987) Cell 48, 351-357.
8. Russel, G.J., Walker, P.M.B., Elton, R.A. and Subak-Sharpe, J.H. (1976) J. Molec. Biol. 108, 1-23.
9. Bernardi, G. et al. (1985) Science 228, 953-958.
10. Zerial, M., Salinas, J., Filipski, J. and Bernardi, G. (1986) Eur. J. Biochem. 160, 479-485.
11. Nicholls, R.D., Fischel-Ghodsian, N. and Higgs, D.R. Cell, 49, 369-378.
12. Nicholls, R.D. et al. (1985) Nucleic Acids Res. 13, 7569-7578.
13. Choo, K.H., Filby G., Greco, S., Lau, Y.-F. and Kan, Y.W. (1986) Gene 46, 277-286.
14. Sequence available on request.
15. Old, J.M. and Higgs, D.R. (1983) In: The Thalassemias. Methods in Hematology 6, 74-102.
16. Southern, E.H. (1987) J. Molec. Biol. 98, 503-517.
17. Brown, W.R.A. and Bird, A.P. (1986) Nature 322, 477-481.
18. Feinberg, A.P. and Vogelstein, B. (1983) Anal. Biochem. 132, 6-13.
19. Smith, D.I., Golembieski, W., Gilbert, J.D., Kizyma, L. and Miller, O.J. (1987) Nucleic Acids Res., 15, 1173-1184.
20. Goldman, M.A., Holmquist, G.P., Gray, M.C., Caston, L.A. and Nag, A. (1984) Science 224, 686-692.
21. Holmquist, G., Gray, M., Porter, T. and Jordan, J. (1982) Cell 31, 121-129.
22. Furst, A., Brown, E.H., Braunstein, J.D. and Schildkraut, C.L. (1981) Proc. Natl. Acad. Sci. USA 78, 1023-1027.
23. Yunis, J.J. (1981) Hum. Genet. 56, 293-298.
24. Aota, S. and Ikemura, LT. (1986) Nucleic Acids Res., 14, 6345-6355.
25. Bird, A.P., Taggart, M., Frommer, M., Miller, O.J. and Macleod, D. (1985) Cell 40, 91-99.
26. Bird, A.P. (1986) Nature 321, 209-213.
27. Bird, A.P., Taggart, M.H., Nicholls, R.D. and Higgs, D.R. (1987) EMBO Journal 6, 999-1004.
28. Marks, J., Shaw, J. and Shan, C-K. (1986) Nature 321, 789-793.
29. Korba, B.E. and Hays, J.B. (1982) Cell 28, 531-541.
30. Collins, F.S. and Weissman, S.M. (1984) Proc.Natl.Acad.Sci.USA 81, 6812-6816.
31. Poustka, A., Pohl, T.M., Barlow, D.P., Frischauf, A.M. and Lehrach, H., (1987) Nature 325, 353-355.
32. Schultz, P.G. and Dervan, B.D. (1983) Proc. Natl. Acad. Sci. USA 80, 6834-6837.
33. Van Ommen, G.J.B. and Verkerk, J.M.H. (1986) In: Human genetic diseases. A practical approach, 113-133, (IRL Press, Oxford).
34. Adams, R.L.P. and Eason, R. (1984) Nucleic Acids Res. 12, 5869-5877.