Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 5022–5026, August 1983
Genetics

Molecular basis of length polymorphism in the human ζ-globin gene complex

(tandem repeats/antenatal diagnosis)

S. E. Y. GOODBOURN, D. R. HIGGS, J. B. CLEGG, AND D. J. WEATHERALL

Medical Research Council Molecular Haematology Unit, Nuffield Department of Clinical Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, England

Communicated by F. Sanger, May 3, 1983

ABSTRACT **The length polymorphism between the human ζ-globin gene and its pseudogene is caused by an allele-specific variation in the copy number of a tandemly repeating 36-base-pair sequence. This sequence is related to a tandemly repeated 14-base-pair sequence in the 5′ flanking region of the human insulin gene, which is known to cause length polymorphism, and to a repetitive sequence in intervening sequence (IVS) 1 of the pseudo-ζ-globin gene. Evidence is presented that the latter is also of variable length, probably because of differences in the copy number of the tandem repeat. The homology between the three length polymorphisms may be an indication of the presence of a more widespread group of related sequences in the human genome, which might be useful for generalized linkage studies.**

The observation (1) of a highly polymorphic locus in human DNA that appeared to be a result of rearrangements suggested the possibility of examining genetic instability at the molecular level in man. Subsequently, other human loci with similar properties have been described in the human α-globin locus (2), the 5′ flanking region of the human insulin gene (3), and in the human λ light chain immunoglobulin locus (4).

To determine the molecular basis of one such length polymorphism, the region between the two human ζ-globin genes has been analyzed in detail. It previously was shown that at least three different lengths of DNA can separate these genes (2). Evidence presented here shows that this variation is caused by different copy numbers of a tandemly repeating 36-base-pair (bp) sequence, part of which is strikingly similar to the 14-bp tandemly repeating sequence of the insulin gene 5′ flanking region, which also causes length polymorphism (5, 6). In addition, there is homology to a sequence in the large introns of the ζ-globin genes that is also tandemly repetitive (7), and evidence is presented that this sequence too can cause length polymorphism.

The existence of common structural features between these loci may be coincidental, but a more likely possibility is that these sequences represent examples of a class of flexible repetitive elements spread throughout the human genome. If this is correct, and other related sequences can be identified and isolated, they may serve as the basis for a comprehensive bank of highly polymorphic loci that could be used to establish the chromosomal location of genes linked to them, including those that are determinants for hereditary disorders.

MATERIALS AND METHODS

The preparation of genomic DNA from peripheral blood and the blot hybridization studies were as described (2). Probe A in Fig. 1 is complementary to the middle exon of the pseudo-ζ-globin gene and is a 350-bp *Hin*cII/*Pvu* II fragment of pBRζ (8). Probe B is complementary to intervening sequence (IVS) 1 of the pseudo-ζ-globin gene and 45 bp of coding sequence; this probe is the large *Bgl* II/*Sac* I fragment from pBRζ. Probe C is also from pBRζ and is prepared as an *Eco*RI/*Sac* I fragment; this probe hybridizes to the *Sac* I fragment that contains the variable length region. Probe D is a *Sac* I/*Pst* I fragment from the 2.7-kb *Sac* I fragment (fragment E in Fig. 1).

Construction of Clones and Subclones. DNA from an individual who was heterozygous for the 4.2-kb and 4.8-kb *Sac* I length alleles was digested with *Sac* I, and the 4- to 6-kb fraction was selected on a 5–20% linear NaCl gradient. The fractionated DNA was ligated to *Sac* I-cleaved phage λ Charon 16A DNA, which had been treated with calf intestinal phosphatase to prevent self-ligation and then was packaged *in vitro* (9) into infective λ particles. The phage were screened by hybridization (10) on *Escherichia coli* 803 (11) without amplification (to minimize potential rearrangements). DNA from positively hybridizing clones was subcloned directly from a small-scale preparation into a *Sac* I-cleaved modified pBR322 vector and was used to transform *E. coli* 1046 (a strongly *rec*A$^-$ strain).

DNA Sequence Analysis. The frequent occurrence of *Sma* I sites in the variable-length region was used to generate a series of overlapping deletions whose endpoints are the *Sma* I sites themselves. This method is based upon that of Frischauf *et al.* (12) except that, instead of DNase I digestion, random partial *Sma* I cleavage was used to introduce blunt-end breaks. *Eco*RI linkers were attached to these breakpoints, and the subclones then were digested with *Eco*RI to cleave the linkers and vector. The fragments were religated, and individual recombinants were isolated after transformation. The proximity of the *Eco*RI site (which now lies at the edge of the region of interest) to the *Hin*dIII site in the plasmid vector sequence enables molecules suitable for DNA sequence assay (13) to be produced by *Hin*dIII/*Eco*RI digestion without any gel purification. DNA prepared from plasmid "mini-preps" (14) proved to be sufficiently pure to obtain sequence data without significant interference from the host DNA. Correlation of the *Sma* I sites derived from the DNA sequence with the original fine structure map demonstrated that no rearrangements occurred during the propagation of these deletion clones.

DNA fragments were end-labeled at their 3′ ends by filling in, then chemically degraded (13), fractionated on 40-cm 6% denaturing polyacrylamide gels of either 0.4 mm or 0.1 mm thickness, prepared, and treated as described by Garoff and Ansorge (15).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: bp, base pair(s); IVS, intervening sequence; HVR, hypervariable region.

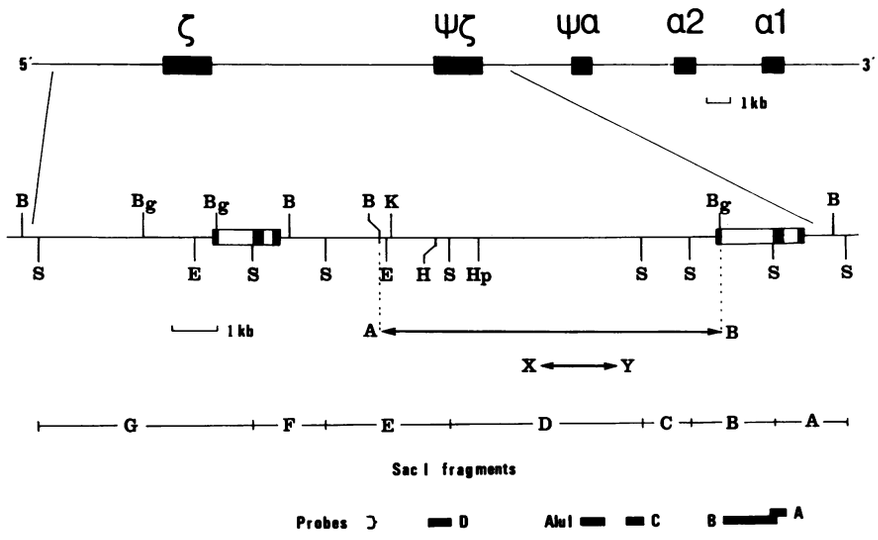

FIG. 1. Restriction enzyme map of the human α-globin gene complex. The top line represents the chromosomal order of the known genes. Below is an expanded map of the embryonic ζ-globin gene region based upon the data of Lauer *et al.* (8) and on analysis of cloned DNA (unpublished data; ref. 7). ■, Coding sequences; □, IVSs. Restriction sites: B, *Bam*HI; S, *Sac* I; Bg, *Bgl* II; E, *Eco*RI; K, *Kpn* I; H, *Hin*dIII; Hp, *Hpa* I. The horizontal arrow between points A and B defines the outer limits of the length variation (2). The region between X and Y corresponds to the sequence determined from the 4.2-kb *Sac* I D fragment allele. *Sac* I fragments and probes referred to in the text are indicated on the lower lines.

RESULTS

Cloning and Characterization of Inter-ζ-Globin Gene Hypervariable Regions (HVR). A restriction enzyme map of the human α-globin gene complex on chromosome 16 is shown in Fig. 1. As previously reported (2), there are differences between individual haplotypes in the length of DNA between the *Bam*HI and *Bgl* II sites marked A and B in Fig. 1 (referred to as inter-ζ-globin gene HVRs). The three length classes for this region will be referred to as "small," "medium," or "large" alleles. The site of variation was localized more precisely to fragment D by analyzing the products of *Sac* I-digested DNA with a variety of probes from cloned DNA (Fig. 1); even at this improved level of resolution there were only a limited number of allele sizes detectable. This region was cloned from an individual heterozygous for the medium- and large-sized alleles, which give 4.2-kb and 4.8-kb *Sac* I D fragments, respectively.

A library of *Sac* I fragments enriched for this size range from a limit digest was cloned in phage λ Charon 16A (16) and then screened with probe C (Fig. 1). Six positive clones were isolated from $\approx 10^6$ screened; restriction enzyme mapping showed that each allele was represented three times. To minimize the number of generations of phage growth in a *recA*$^+$ background and hence reduce the potential for rearrangements, the *Sac* I inserts were subcloned directly from small-scale phage minilysate DNA preparations. For comparison of the two alleles, a subclone for each was selected in which the inserts were in the same orientation with respect to the plasmid vector sequence; these were designated pSac4.2 and pSac4.8 corresponding to

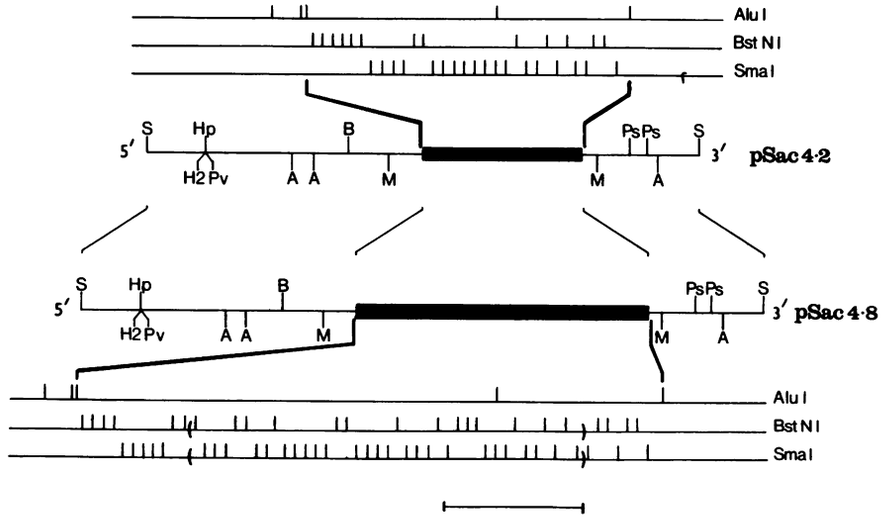

FIG. 2. Fine restriction enzyme maps of the cloned *Sac* I D fragments from an individual heterozygous for the 4.2-kb and 4.8-kb alleles. Data were obtained by using the method of Smith and Birnstiel (17) and a variety of double-digest experiments. The two center maps show some of the sites used to compare the clones pSac4.2 and pSac4.8. Other sites are not shown. These clones are identical outside of the region marked by a heavy box. This box represents the *Sma* I/*Bst*NI cluster, which is shown in an expanded form for each clone above and below the 4.2-kb and 4.8-kb alleles, respectively. The precise order of sites between the parentheses in the pSac4.8 map has not been confirmed by sequence data. The bar at the foot of the diagram represents 1 kb for the center and 500 bp for the outside maps. Restriction enzymes: S, *Sac* I; Hp, *Hpa* I; H2, *Hinc*II; Pv, *Pvu* II; A, *Ava* I; B, *Bal* I; M, *Mbo* II; Ps, *Pst* I.

medium- and large-sized alleles, respectively. Fine-structure restriction enzyme mapping of these subclones demonstrated that the only difference between them resides in a region of DNA that is rich in regularly spaced *Sma* I sites (Fig. 2). These sites occurred with a periodicity of 34–40 bp, except in some cases where the gap was 72, 108, or 144 bp. Wherever a *Sma* I site was missing from the repeating pattern, it was replaced with a *Bst*NI site, and because this latter enzyme recognizes a variety of mutated *Sma* I sites (C-C-C-G-G-G- > C-C-$\frac{A}{T}$-G-G-G or C-C-C-$\frac{A}{T}$-G-G), it suggested that these sites were part of a simple tandemly repeated sequence.

Several studies have shown that recombinant clones containing direct repeats are unstable in *Escherichia coli* (8, 18–20), although less so in *rec*A$^-$ strains (21). Indeed, four human insulin gene clones have been isolated with different lengths of 5′-flanking tandem repeat from the same genomic library (5, 6), despite the fact that there is only one insulin gene per haploid genome (22). Given the apparent tandemly reiterative nature of the inter-ζ-globin gene HVR, it seemed probable that these might also be unstable in *E. coli*. Two lines of evidence show that the difference in *Sma* I/*Bst*NI cluster length observed between pSac4.2 and pSac4.8 is not a cloning artefact and is the real cause of the allele-specific length difference in humans.

First, in four additional clones from this individual (two from each allele) and four 4.2-kb allele clones from other individuals (three of which originate from a library that was constructed from an individual lacking the 4.8-kb allele), there were no restriction site differences outside of the cluster. All of the 4.8-kb allele clones had an identical *Sma* I/*Bst*NI cluster length, and only minor differences were observed in some of the 4.2-kb alleles. Because these latter clones originate from three separate individuals, some of these differences may reflect natural genetic variation (the full extent of which is unknown). Nevertheless, small differences have been detected between some clones from the same individual. In this respect the behavior of such regions in *E. coli* may mimic that in humans and, as previously suggested (18), may be a useful tool for anticipating these events. Second, blot hybridization with a cluster-specific probe (see below) on DNA from individuals with known inter-ζ-globin HVR genotypes showed that all of the interallelic length variation resides within restriction fragments that just span the *Sma* I/*Bst*NI cluster.

DNA Sequence of the Variable-Length Region. Sequences of the entire variable region from the shorter allele (pSac4.2) and part of the larger allele (pSac4.8) were obtained, and they show (Fig. 3) that these are composed of a simple tandemly repeating sequence as expected. The basic repeat is 36 bp long, and there are 32 copies of it in the shorter allele. Although the sequence of the larger allele was not determined in full, an estimate of the copy number of repeats as 58 ± 1 can be made from the mapping data. There are no major differences in the basic 36-bp repeat sequence between the two alleles.

Fig. 3 shows all of the 4.2-kb allele repeats in a 5′-to-3′ order. Interestingly, the pattern of base changes between repeats is distinctly nonrandom; in fact, there appear to be subclasses of the 36-bp sequence in which a given repeat shows the same or similar changes in several places to a second repeat (e.g., repeat 4 is identical to repeat 30 despite showing six changes from repeat 1). This might be explained in terms of multiple unequal crossing-over events that have generated the observed mosaic pattern, but the relative lack of the expected intermediate sizes of *Sac* I D fragment makes this unlikely. An alternative explanation is that multiple rounds of localized gene conversion between misaligned repeats have occurred; for example, if each repeat mutates independently to give a pattern

```
                                    5'...ACACCCATCAATGGGAG
      CACCAGGACACAGATGGAGGCTAATGTCATGTTGTAG
      ACAGGATGGGTGCTGAGCTGACCACACCCACATTATT
      AGAAAATAACAGCACAGGCTTGGGGTGGAGGCGGGAC
      ACAAGACTAGCCAGAAGGAGAAAGAAAGGTGAAAAGC
      TGTTGGTGCAAGGAAGCTCTTGGTATTTTCAACGGCT
 1    TGGGCACAGGCTGTGAGG-GTGCCTGGGACGGCTTGT
 2    G          CA  AG    CA
 3    G        T     AG    CT
 4    G          CA  AG    CA
 5    G        T     AG    CT
 6    G              AG    CA
 7    G              -     C
 8    G        T     AG    C
 9    G        T -   AG    C
10    G        T     AG    C
11    G  A     T     AG
12    G              -           T
13    G        T     AG    C     T
14    G        T     AG    C
15    G        T     AC    C
16    G              -     C     T
17    G          CA  AG    C
18    G              -     C
19    G              -     C        A     C
20    G        T     AG    C
21    G              -
22    G        T     AG    C
23    G        T     -A    C     T
24    G        T     AG
25    G              -     C
26    G              AG
27    G              -A    C     T
28    G        T     G     C
29    G              -     CA    T
30    G          CA  AG    CA
31    G        T     AA    C
32    G                -A  C  C           C
      AGCTACAGGGAGAAAAGACTTGGTGCTGTGGGCCTGC
      CTTGGGGCTGGTGGTACAGCCCTTATCTGCTGCCCTC
      AGGATCTCCCGGCCCCTCTCGTCCAGGCCCCTGCAAC
      CCCATGCCCCAGCCTCTGAGGACCAAAGGCGCCCCTG
      CTTGGGAAGAGGGGGCTCAGGGGAGTCGCCTGACCCG
      GTTCCAAGCCAGGCTGATTTACCGTTGTTAACATCCT
      AGTGCACGCATCCCTCTGCCTCATGCACCCAACTCCA
      AGGCCTGGTACAC...3'
```

FIG. 3. Nucleotide sequence of the inter-ζ-globin gene HVR. The 1,645-bp segment displayed is from the 4.2-kb *Sac* I D allele and corresponds to the region marked in Fig. 1. To emphasize the tandemly reiterative nature of this region, the full sequence is not illustrated. The sequence on line 1 corresponds to the 5′-most copy of the tandemly reiterated element. Only the differences from copy number 1 in each of the subsequent consecutive 31 elements are indicated. Nucleotides that are absent at a given position are indicated by a dash. The DNA sequence from a position equivalent to elements 27–32 inclusive has been determined from the 4.8-kb *Sac* I D allele and shows 10 base changes (see text).

A-B-C-D-E-F, a misaligned conversion event can correct this to A-B-C-A-B-C without altering the total number of repeats. Because the sequence data on the 4.8-kb alleles also revealed the same pattern of changes within repeats, this seems an attractive proposition.

To test if the 36-bp tandem repeat sequence is repeated at other points in the human genome, the *Alu* I fragment containing the 3′ portion of the repeat (see Fig. 1) was used as a probe in genomic blotting experiments. Under stringent conditions this probe only hybridized to one major fragment, and even after prolonged autoradiographic exposure, there were only faint signals from other fragments. These additional signals

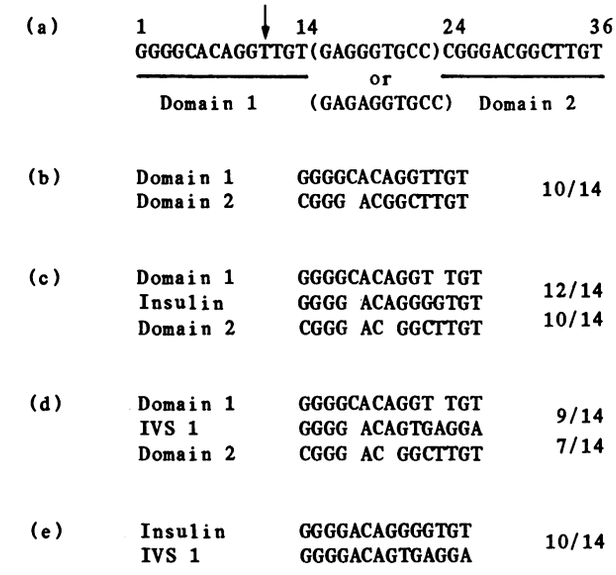

FIG. 4. Comparison of the 36-bp repeat monomer with other sequences (as described in the text). The arrow over nucleotide 11 of the 36-bp repeat monomer indicates that this position is frequently C (see Fig. 3). The nucleotide sequence between 14 and 24 (in parentheses) can be replaced by the sequence below it, which makes the monomer 37 bp long. The two sequences, which are referred to as "domains," are underlined. Gaps have been introduced to maximize homology. Fractions on the right of the diagram indicate the degree of homology.

may come from single copies of the 36-bp unit, but there do not appear to be other regions in the genome where this sequence is tandemly repeated. DNA from individuals who are homozygous for the shortest allele hybridized strongly to this probe, showing that this allele also contains some copies of the tandem repeat.

The structure of the 36-bp repeat monomer has two domains that are closely related to each other (Fig. 4). These could have arisen by a duplication of a 14-bp precursor sequence, which, as a result of the duplication or a subsequent event, became separated by 9–10 bp of unrelated sequence. The two domains appear to have mutated independently before becoming expanded as one intact unit into the variable-length region observed. Despite the fact that the 36-bp repeat was not found elsewhere in the human genome, it appears to be related to other sequences. Comparison of the DNA sequence of this repeat monomer with that of the repeat monomer of the human insulin 5′ flanking length polymorphism (5) revealed a striking homology between these sequences (Fig. 4). Proudfoot *et al.* (7) have already noted a similarity between the insulin flanking sequence and a tandemly repeating sequence in IVS 1 of the ζ- and pseudo-ζ-globin genes, and there is homology between this latter sequence and the two domains of the 36-bp repeat (Fig. 4). As will be shown below, there is length variation in (or close to) IVS 1 of the pseudo-ζ-globin gene, and it is likely that this is caused by a difference in copy number of the 14-bp tandemly repeating sequence in this region.

By assuming that this is the reason for the observed length variation in IVS 1 of the pseudo-ζ-globin gene, it appears that three length polymorphisms, effectively selected at random from the human genome, are related. This strongly suggests that there may be something unusual about the basic sequence that allows tandem expansion; thus, there may be other similar sequences in the human genome, which might also give rise to length variation. However, these may be difficult to detect. If there had been a duplication of the ancestral sequence so that the nonrelated spacer DNA is different in length from the 9- to 10-bp observed for the inter-ζ-globin gene 36-bp sequence, the hypothetical unit may not hybridize to any of these probes. Similarly, a change in the monomer sequence may have a significant effect on hybridization once this is expanded into an oligomeric structure. It is perhaps not surprising that the 36-bp repeat probe did not hybridize to the insulin-linked or ζ-globin IVS 1 sequences because the lengths of the monomers are so different. Bell *et al.* (5) noted that the insulin-linked sequence is also unique, and the probe used here for IVS 1 of the ζ- and pseudo-ζ-globin genes only showed one additional band (which did not appear to be polymorphic). Preliminary experiments with the 36-bp repeat probe under conditions of reduced stringency suggest that there may be other related sequences in the human genome. Furthermore, the weakly hybridizing fragments appeared to be polymorphic. Thus, if the correct conditions can be found, additional sequences may be identified.

An Additional Length Polymorphism. In addition to *Sac* I D, only fragment B of the remaining *Sac* I fragments in the ζ-globin locus (Fig. 1) appeared to show any variation (fragment C cannot be studied because it contains highly repetitive sequences). The commonest size observed for this fragment was 1.85 kb, but about 23% of haplotypes examined produced a band of 1.65 kb, which can be interpreted as a polymorphism of this fragment. Furthermore, three additional sizes for fragment B were observed (1.78 kb, 1.95 kb, and 2.15 kb) at lower frequencies, making site changes an unacceptable explanation for this variation. All of the variation within this fragment lies to the right of the *Bgl* II site (Fig. 1) as shown by *Sac* I/*Bgl* II double digests, and this is reflected by a slight length variation in α-globin gene specific *Bgl* II fragments. Since 1,232 bp out of 1,278 bp of this *Bgl* II/*Sac* I variable region comes from IVS 1 of the pseudo-ζ-globin gene (7), it is probable that the cause of the variation lies in this region. The possibility that this is due to the copy number of a tandemly repeated 14-bp sequence has been discussed, and indeed the copy number does differ between the pseudo-ζ- and functional-ζ-globin gene itself (7).

Although only a limited study on this variable region was performed, there was no evidence of variation within IVS 1 of the ζ-globin gene itself; it is possible that functional restraints may prevent this.

DISCUSSION

The length polymorphism between the human ζ-globin gene and its pseudogene is caused by differing copy numbers of a simple tandemly repeating sequence. A similar length polymorphism occurs in the 5′ flanking region of the human insulin gene (5), and it seems likely that length variations in IVS 1 of the human pseudo-ζ-globin gene also result from such a mechanism. Length variation presumably results from unequal crossing-over between misaligned repeats, but because only a limited repertoire of haplotypes is found, there may be some constraints on this process. The similarity between elements of these sequences causing length variation suggests that they may have a common ancestry, although it is not obvious how such sequences came to reside on more than one chromosome because the sequences flanking the repeat elements are not related and show no unusual properties (except for small direct repeats, which are different for each locus). The relatedness of the inter-ζ-globin, insulin 5′ flanking, and pseudo-ζ-globin IVS 1 variable regions suggests that they may be part of a much larger family of genomic sequences exhibiting length polymorphism, which could be of some significance in the search for a more general approach to antenatal diagnosis of hereditary disorders. As Botstein *et al.* (23) point out, the linkage of detectable poly-

morphisms in DNA to loci responsible for serious genetic illnesses offers a huge potential for their antenatal diagnosis. The diagnosis of some thalassemias through the direct association of polymorphisms with mutant genes is already a routine practice, but a more powerful application would be the linkage of polymorphisms to unknown loci where there is no knowledge of the gene or gene-product causing the illness. This approach first requires the isolation of a bank of probes for known polymorphic sequences. Simple restriction site polymorphisms may be of limited value for these purposes, compared with HVRs of the type described here. The discovery of other similar features in the genome would greatly strengthen their potential in this respect. In addition, the polymorphisms identified in this locus can be used as more specific chromosomal markers. The inter-ζ-globin gene HVRs show population specificity, and some alleles exhibit marked linkage disequilibrium with other markers in this locus (unpublished data).

We thank Drs. Joyce Lauer, T. Maniatis, and N. J. Proudfoot for providing us with clones and Dr. G.-J. van Ommen for helpful suggestions. This work was supported by the Medical Research Council of Great Britain, including a research studentship to S.E.Y.G.

1. Wyman, A. R. & White, R. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 6754–6758.
2. Higgs, D. R., Goodbourn, S. E. Y., Wainscoat, J. S., Clegg, J. B. & Weatherall, D. J. (1981) *Nucleic Acids Res.* **9**, 4213–4224.
3. Bell, G. I., Karam, J. H. & Rutter, W. J. (1981) *Proc. Natl. Acad. Sci. USA* **79**, 5759–5763.
4. Hieter, P. A., Hollis, G. F., Korsmeyer, S. J., Waldmann, T. A. & Leder, P. (1981) *Nature (London)* **294**, 536–540.
5. Bell, G. I., Selby, M. J. & Rutter, W. J. (1982) *Nature (London)* **295**, 31–35.
6. Ullrich, A., Dull, T. J., Gray, A., Phillips, J. A. & Peter, S. (1982) *Nucleic Acids Res.* **10**, 2225–2240.
7. Proudfoot, N. J., Gil, A. & Maniatis, T. M. (1982) *Cell* **31**, 553–563.
8. Lauer, J., Shen, C.-K. J. & Maniatis, T. (1980) *Cell* **20**, 119–130.
9. Scherer, G., Telford, J., Baldari, C. & Pirrotta, V. (1981) *Dev. Biol.* **86**, 438–447.
10. Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180–182.
11. Wood, W. B. (1966) *J. Mol. Biol.* **16**, 118–133.
12. Frischauf, A. M., Garoff, H. & Lehrach, H. (1980) *Nucleic Acids Res.* **8**, 5541–5549.
13. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
14. Ish-Horowicz, D. & Burke, J. F. (1981) *Nucleic Acids Res.* **9**, 2989–2998.
15. Garoff, H. & Ansorge, W. (1981) *Anal. Biochem.* **115**, 450–457.
16. Blattner, F. R., Williams, B. G., Blechl, A. E., Denniston-Thompson, K., Faber, H. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Sheldon, E. L. & Smithies, O. (1977) *Science* **196**, 161–169.
17. Smith, H. O. & Birnstiel, M. L. (1976) *Nucleic Acids Res.* **3**, 2387–2398.
18. Arnheim, M. & Kuehn, M. (1979) *J. Mol. Biol.* **134**, 743–765.
19. Brutlag, D. L., Carlson, M., Fry, K. & Hsieh, T. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 1137–1146.
20. Fritsch, E. F., Lawn, R. M. & Maniatis, T. (1980) *Cell* **19**, 959–972.
21. Albertini, A. M., Hofer, M., Calos, M. P. & Miller, J. H. (1982) *Cell* **29**, 319–328.
22. Bell, G. I., Pictet, R. L., Rutter, W. J., Cordell, B., Tischer, E. & Goodman, H. M. (1980) *Nature (London)* **284**, 26–32.
23. Botstein, D., White, R. L., Skolnick, M. & Davis, R. (1980) *Am. J. Hum. Genet.* **32**, 314–331.