Volume 14 Number 11 1986 Nucleic Acids Research

Cloning a selected fragment from a human DNA 'fingerprint': isolation of an extremely polymorphic minisatellite

Zilla Wong, Victoria Wilson, Alec J.Jeffreys and Swee Lay Thein+

Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH and +MRC Molecular Haematology Unit, Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK

Received 8 April 1986; Accepted 7 May 1986

ABSTRACT

A large hypervariable DNA fragment from a human DNA fingerprint was purified by preparative gel electrophoresis and molecular cloning. The cloned fragment contained a 6.3 kb long minisatellite consisting of multiple copies of a 37 bp repeat unit. Each repeat contained an 11 bp copy of the "core" sequences, a putative recombination signal in human DNA. The cloned minisatellite hybridized to a single locus in the human genome. This locus is extremely polymorphic, with at least 77 different alleles containing 14 to 525 repeat units per allele being resolved in a sample of 79 individuals. All alleles except the shortest are rare and the resulting heterozygosity is very high (∿ 97%). Cloned minisatellites should therefore provide a panel of extremely informative locus-specific probes ideal for linkage analysis in man.

INTRODUCTION

Hypervariable tandem-specific regions of the human genome provide highly informative multi-allelic genetic markers ideal for linkage analysis in human pedigrees [1-8]. The tandem repeat units in a subset of these human minisatellites share a common 10-15 bp "core" sequence which may be a recombination signal implicated both in the generation of minisatellites and in the maintenance of polymorphism by allelic alteration of the number of repeats at a minisatellite locus [9]. DNA probes comprised of tandem repeats of the core sequence hybridize to a large number of dispersed autosomal hypervariable minisatellites [9-11]. The resulting DNA "fingerprint" is individual-specific [10] and can be used to study the segregation of multiple informative loci in large human pedigrees, and to search for linkage between hypervariable DNA fragments and disease loci [11].

To determine whether the largest and most polymorphic of the minisatellite fragments in a human DNA fingerprint can be cloned to provide locus-specific probes, we describe the isolation and properties of a

specified minisatellite fragment which has previously been shown to cosegregate apparently with the heterocellular form of hereditary persistence of foetal haemoglobin (HPFH) [11].

MATERIALS AND METHODS

General methods

DNA was isolated from white blood cells [10] and from an Epstein-Barr virus transformed lymphoblastoid cell line [12] derived from individual III 9 of the HPFH pedigree described by Jeffreys et al. [11]. Restriction digestion and Southern blot hybridizations were performed as described elsewhere [9-11]. Double-stranded DNA probe fragments were isolated by electrophoresis onto DE81 paper [13] and labelled with ^{32}P by random oligonucleotide priming [14]; single-stranded minisatellite probe 33.15 was prepared as described elsewhere [9].

Cloning hypervariable fragment g

600 µg individual III 9 lymphoblastoid cell line DNA digested with Sau3A was fractionated by electrophoresis through a 0.5% agarose gel, and relevant size fractions collected by electroelution onto dialysis membrane [15]. After two cycles of preparative gel electrophoresis fragment g was ∿ 1000 fold purified (yield 150ng DNA). 20ng of this partially purified fraction was ligated to 60ng λL47.1 arms isolated after cleavage with BamHI [16], packaged in vitro and plated onto E. coli WL95 (803, supE, supF, $hsdR_K^-$, $hsdM_K^+$, tonA, $trpR^-$, metB, lysogenic for P2) [16,17] to select for recombinants. The resulting library of ∿1500 recombinant phage was screened by plaque hybridization with minisatellite probe 33.15. Four positive plaques were replated and rescreened on E. coli ED8910 (803, supE, supF, recB21, recC22, hsdS) [16], and recombinant phage DNA prepared by the method of Blattner et al. [18]. The Sau3A insert of recombinant phage λg3 was subcloned into the BamHI site of pUC13 [19] and propagated in a recA derivative of E. coli JM83 (JM83, Δ(recA-srlR)$_{306::Tn10}$) [20].

DNA sequence analysis

pλg3 DNA was sonicated, size-selected and shotgun cloned into the SmaI site of M13mp19 [21]. To determine the DNA sequence of the tandem-repetitive region of pλg3, 12 random clones were sequenced by the dideoxynucleotide chain-termination method [22,23]. 8 of these clones were derived from the tandem-repetitive minisatellite region of pλg3 and were used to define the consensus repeat sequence. M13 clones containing the 5'

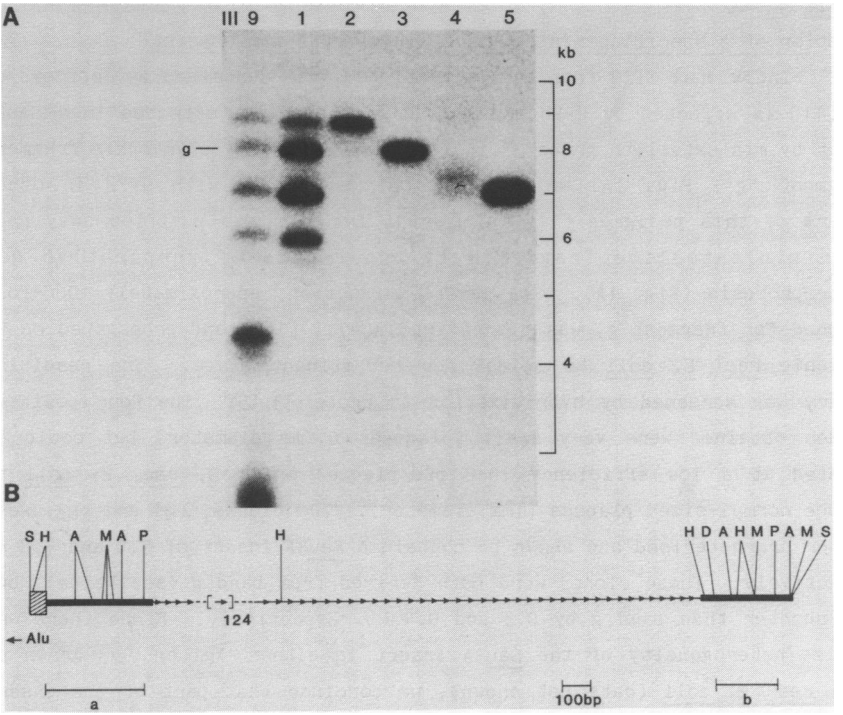

Figure 1. Isolation and characterization of a specified hypervariable DNA fragment from a DNA fingerprint.
A. Gel electrophoretic purification. DNA from individual III9 in the pedigree described by Jeffreys et al. [11] was digested with Sau3A and 6-9 kb fragments collected by preparative gel electrophoresis (fraction 1). These fragments were re-electrophoresed to give fractions 2-5. Aliquots of III9 DNA digested with Sau3A and of each fraction were electrophoresed through a 0.8% agarose gel and Southern blot hybridized with the minisatellite probe 33.15. Fragment g (8.2 kb), which tends to cosegregate with HPFH [11], is approximately 1000-fold purified in fraction 3. This fraction was cloned into λL47.1 [16] and fragment g subcloned into pUC13 [19] to give plasmid pλg3.
B. Organization of DNA fragment g. The Sau3A insert in pλg3 was mapped with restriction endonucleases AluI (A), DdeI (D), HaeIII (H), MboII (M), PstI (P) and Sau3A (S). There are no cleavage sites for HinfI or RsaI. The 7.14 kb insert contains ∿171 tandem repeats of a 37 bp sequence (see Fig. 2) plus 747 bp flanking DNA. The 5' flanking region contains the beginning of an inverted Alu element (hatched). The origins of unique sequence flanking probes a and b are shown.

and 3' flanking regions were detected by hybridization with ^{32}P-labelled flanking probes a and b (Fig. 1) and sequenced to establish the flanking region sequence and the beginning and end points of the minisatellite.

RESULTS

Isolation of a specified minisatellite from a DNA fingerprint

Individual III9 from the Gujarati pedigree described by Jeffreys et al. [11] is affected by HPFH and her DNA fingerprint, detected in a Sau3A digest by minisatellite probe 33.15, contains a 8.2 kb polymorphic fragment (fragment 'g'; Fig. 1A) which tends to cosegregate with HPFH in other members of this pedigree [11]. This DNA fragment was purified away from other minisatellite fragments by two rounds of preparative gel electrophoresis (Fig. 1A). The purified fraction, approximately 1000-fold enriched for fragment g, was cloned into λL47.1 [16], and propagated on P2 lysogenic rec$^+$ E. coli to select for recombinant phage. The resulting library was screened by hybridization to probe 33.15. The four positive plaques obtained were very small (plaques <0.1mm diameter) but could be replated at a low efficiency (0-8 pfu/plaque) on recB, recC E. coli to produce normal-sized plaques (1mm diameter). Two clones, λg1 and λg3, were further characterised and shown to contain a Sau3A insert of 7.7 and 7.8 kb respectively. These clones were both derived from band g (see below), but were shorter than band g by 0.5 and 0.4 kb respectively. Since there was no size heterogeneity of the Sau3A insert in either λg1 or λg3 grown on recB, recC E. coli (data not shown), we conclude that part of the insert has been lost from each clone during their initial propagation in rec$^+$ E. coli.

Yields of λg1 and λg3 DNA prepared by the Blattner method [18] were very low (∿1% of the normal yields of λL47.1 recombinant DNA), again pointing to abnormal growth properties of these minisatellite clones. The Sau3A insert was therefore subcloned into pUC13 [19] and propagated in a recA derivative of E. coli JM83 [20] to minimise rearrangement of the insert. The resulting subclone, pλg3 (Fig. 1B), contained a 7.1 kb Sau3A insert, 0.7 kb shorter than the insert in λg3.

Organisation of minisatellite fragment g

The structure of the minisatellite fragment in pλg3 was determined by restriction mapping (Fig. 1B) and DNA sequencing (Fig. 2). The clone contains a 6.3 kb minisatellite devoid of restriction sites, except for a single internal HaeIII site. The minisatellite is comprised of ∿171 repeats of a 37 bp unit which contains the sequence GTGGGCAGG; this sequence corresponds precisely to the most invariant part of the 11-16bp core sequence previously identified as being shared by several different human minisatellites [9] (Fig. 2). The repeat units are not completely

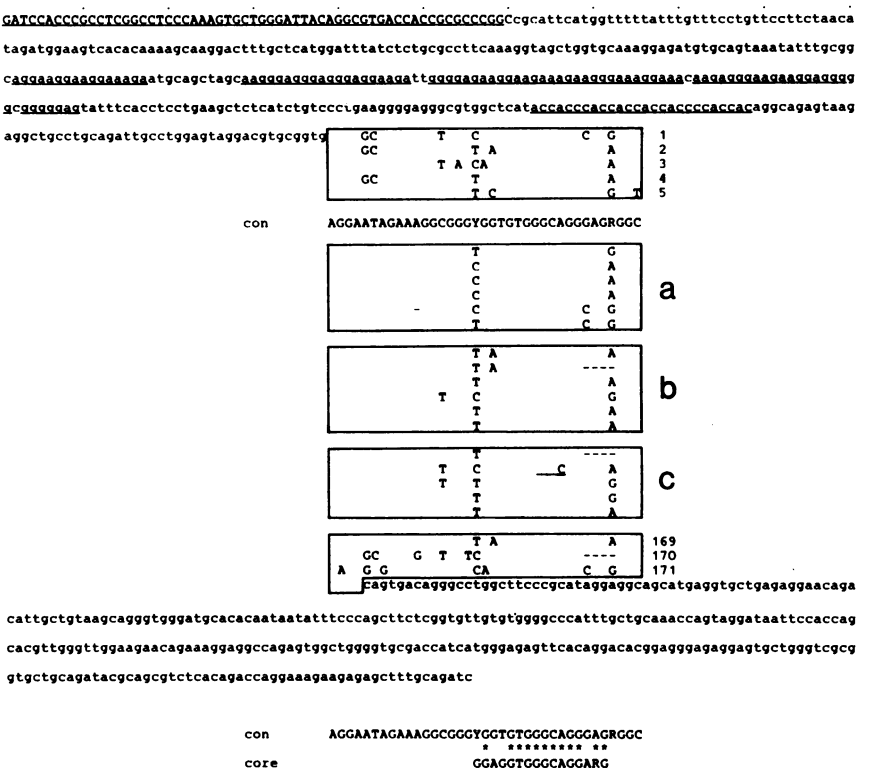

Figure 2. DNA sequence of hypervariable fragment g. The sequence of the 5' and 3' flanking regions and the beginning and end of the minisatellite is shown together with the repeat sequences of three random regions (a-c) from the minisatellite. The beginning of the inverted Alu element is shown in underlined uppercase, and several simple sequence regions in the 5' flanking region are underlined. The consensus sequence (con) of the 37 bp minisatellite repeat unit is shown, and differences from this consensus are given for individual repeat units. The second repeat unit in region c contains a HaeIII cleavage site (underlined), and this region therefore spans the internal HaeIII site in the minisatellite (Fig. 1B). The consensus repeat sequence is also shown aligned with the "core" sequence [9].

homogeneous; sequence analysis of several randomly-selected regions from within the minisatellite revealed a limited amount of repeat sequence variation. Most variants, including a 4 bp deletion which produces a subset of 33 bp repeat units, are diffused over more than one repeat unit. One variant (an A → C transversion in region c) creates the unique internal HaeIII site and is therefore probably only found in one repeat unit. The beginning and end repeat units of the minisatellite are noticeably more

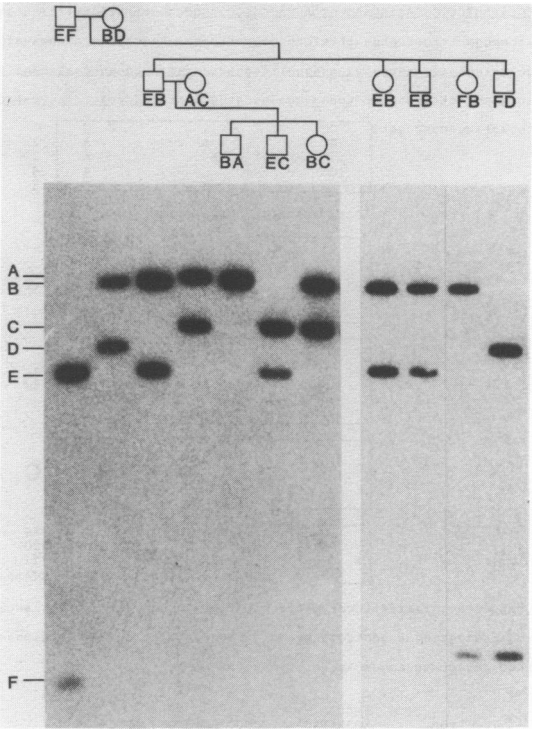

Figure 3. Mendelian inheritance of polymorphic DNA fragments detected by pλg3. 8 μg samples of human DNA were digested with HinfI and electrophoresed through a 0.8% agarose gel. Digests were Southern blot hybridized with the Sau3A insert of pλg3, in 1 x SSC at 65° in the presence of 6% polyethylene glycol and 50 μg/ml alkali-sheared human placental competitor DNA, and washed after hybridization in 0.2 x SSC at 65°. pλg3 detects a single locus which is heterozygous in all individuals shown. Inheritance of alleles (indicated by letters) is Mendelian.

diverged in sequence than are the internal repeats.

The minisatellite in pλg3 is flanked by non-repeated DNA containing the normal density of restriction sites (Fig. 1B). The beginning of the 5' flanking region is comprised of the head of an inverted Alu element. The remaining 5' and 3' flanking regions, defined by hybridization probes a and b (Fig. 1B), are unique sequence DNA and hybridize only to this locus in total human DNA (data not shown). The 5' flanking region contains a considerable amount of simple sequence DNA [polypurine and $(ACC)_n$] (Fig. 2).

Minisatellite fragment g detects a single polymorphic locus

To determine whether the entire cloned minisatellite fragment can

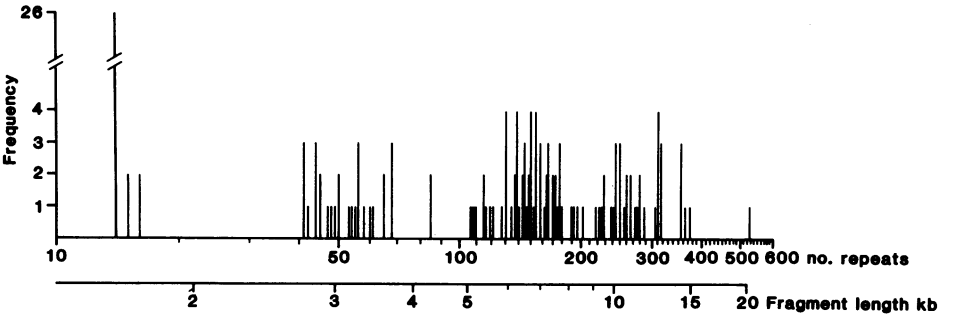

Figure 4. Distribution of minisatellite allele sizes. DNA from 79 randomly-selected British caucasians was digested with HinfI and minisatellite alleles were detected by hybridization to pλg3 (Fig. 3), both in individual samples and in pools of 14 individuals (1 µg DNA per individual) which were electrophoresed through a 35 cm long 0.7% agarose gel to maximise allele resolution. The length of the minisatellite in the shortest common allele was determined by genomic mapping of this fragment in a homozygote, using flanking single-copy probes a and b (Fig. 1B) (data not shown). The number of repeats per allele is approximate and depends on the ratio of 37 to 33 bp repeat units in the minisatellite. Excluding the short common allele, the remaining alleles were sampled either once (42 alleles), twice (12 alleles), three times (12 alleles) or four times (5 alleles). This distribution does not differ significantly from that predicted if all of these alleles are uniformly rare ($\hat{q}$=0.0081, χ^2[4 d.f.] = 4.0), and is therefore consistent with a simple model of one common short allele (q=0.165) and 103 equally rare alleles (q=0.0081 per allele) being present in the population.

be used as a hybridization probe to detect specifically the corresponding locus in human DNA, the Sau3A insert from pλg3 was hybridized in the presence of human competitor DNA to human DNA digested with HinfI, the restriction enzyme routinely used for DNA fingerprinting [9-11] (Fig. 3). At high stringencies (0.2 x SSC, 65°), either one or two hybridizing fragments were detected in all individuals examined. pλg3 detected the 8.2 kb fragment g in previously-tested relatives of III 9 [11], confirming that band g had been cloned (data not shown). At lower stringencies (1 x SSC, 65°), additional faintly hybridizing polymorphic DNA fragments were detected (data not shown).

DNA fragments detected by pλg3 at high stringencies segregate in a Mendelian fashion as alleles of a single locus (Fig. 3). This locus is not sex-linked and also showed no significant but incomplete sex-linkage in two large pedigrees studied (61 progeny tested, $\hat{z}$ = 0.18 at $\hat{\theta}$ = 0.43; data not shown); it therefore behaves as an autosomal locus.

Extreme polymorphic variation at this minisatellite locus

HinfI digests of DNA from 79 randomly-selected British caucasians were screened with the insert from pλg3, first singly and then in pools of 14 people to maximise the resolution of different alleles. The lengths of the different alleles detected in this population sample are shown in Fig. 4, together with the estimated number of repeat units in each allele and the allele frequencies.

At least 77 different alleles could be resolved in this population sample, with repeat numbers ranging from 14 to ∿525 per allele. The shortest allele is relatively common (q=0.165) and the only three homozygotes in the population sample were all homozygous for this allele. All other alleles resolved were rare, with a mean population frequency estimated at q=0.008 (Fig. 4 legend). These estimates of numbers of alleles and the mean frequency of rare alleles are limited by gel resolution, and the true number of different length alleles in this sample may be greater.

The distribution of allele lengths does not appear to be completely random, but shows evidence of trimodality (Fig. 4), with short (14-16 repeats), medium (41-68 repeats) and long (107-525 repeats) classes of alleles. A bimodal distribution of allele lengths has been previously noted in caucasians for the hypervariable region located 5' to the human insulin gene [3], though this bimodality is not evident in negroes [6].

DISCUSSION

The isolation of fragment g demonstrates that the large and highly polymorphic DNA fragments in a DNA fingerprint are in principle amenable to molecular cloning to provide locus-specific probes suitable for studying individual hypervariable regions. While fragment g could be cloned in bacteriophage λ, the resulting clones showed abnormal growth properties on *rec*$^+$ *E. coli*. A similar phenomenon has previously been noted for regions of the human genome containing fold-back DNA [24], and will lead to a depletion of minisatellite clones from conventional amplified human libraries in λ phage. The cloned minisatellite is unstable both in λ and on subcloning into pUC13 in *recA E. coli*; the final recombinant pλg3 had lost about 30 repeat units compared with fragment g. The physical map of pλg3 (Fig. 1B) does not therefore completely reflect the organization of fragment g.

The cloned DNA fragment has the structure expected for a

minisatellite, establishing that fragment g is not derived from a longer satellite DNA. Despite the presence both of a "core" sequence in each repeat unit and part of an Alu sequence in the 5' flanking region, the cloned fragment g acts, in the presence of competitor human DNA, as a locus-specific probe. The failure of pλg3 to detect efficiently other core-containing minisatellites is due to the additional non-core DNA present in each repeat unit which probably interferes with cross-hybridization to other minisatellites [9].

The hypervariable locus defined by pλg3 shows no significant sex-linkage and is therefore not located on the X or Y chromosomes nor, probably, on the pseudoautosomal region of these chromosomes which appears to be richly-endowed with highly polymorphic DNA [see 25]. The availability of a locus-specific probe will permit the autosomal localization of this hypervariable locus.

The cloned minisatellite shows extreme length polymorphism, presumably due to allelic variation both in the number of repeat units and in the ratio of 37 and 33 bp repeat types in each allele. In a random population sample of 158 chromosomes, one common and at least 76 rare alleles could be resolved. This locus in much more polymorphic than most human hypervariable regions so far characterized [3-6,8,26], including the selection of relatively short cloned minisatellites initially used to define the "core" sequence [9]. The heterozygosity at this locus is at least 96.6%, with most homozygotes arising from the short common allele.

New alleles at this locus are generated by alteration in the repeat number, either by slippage during DNA replication or by unequal exchange driven by the "core" sequence, a putative recombination signal [9]. Under the neutral mutation - random drift hypothesis, the parameter $4N_e v$ (θ), where N_e is the effective population size and v is the mutation rate per gamete to a new length allele, can be most accurately estimated from the number of different alleles scored in a population sample, using the infinite allele model [27]. We estimate θ at 60-90 for this locus, depending on whether the common short allele is included or not. Since N_e for man is $\sim 10^4$ [28], the mutation rate v to resolvable new length alleles at this locus is approximately 0.002 per gamete. The average length of minisatellite DNA at this locus is 5kb, and thus the mutation (recombination) rate per kb minisatellite is $\sim 4 \times 10^{-4}$, compared with $\sim 10^{-4}$ per kb estimated for other shorter core-containing minisatellites [9] and a mean meiotic recombination rate of 10^{-5} per kb for human DNA [9,29]. Thus

the rate of generation of new alleles at this minisatellite locus is remarkably high, consistent with previous suggestions that these core-rich regions may be recombination hotspots [9]. Presumably, the mutation rate is relatively low for short alleles, which might explain how the shortest allele has drifted to achieve a significant frequency in the population without being disrupted by unequal exchange during this process.

DNA fingerprints have proved a powerful method for individual identification [30] and for establishing family relationships in for example paternity and immigration disputes [10,31]. The accuracy of this method is determined by the low mean probability x that a band is apparently shared by two randomly-selected individuals. For British caucasians, x has been estimated empirically at 0.2 for DNA fingerprint HinfI fragments larger than 4 kb [10,31]. The allele distribution in Fig. 4 enables us to calculate x for alleles >4kb; for this specific minisatellite locus, x = 0.016, an order of magnitude lower than the estimate from multi-locus DNA fingerprints. If this cloned minisatellite is typical of loci represented in DNA fingerprints, then most bands shared between unrelated DNA fingerprints are due to fortuitous comigration of different minisatellite fragments. The important practical consequence of this is that, in cases of non-exclusion in for example paternity disputes where all paternal bands in a child are precisely present in the DNA fingerprint of the putative father, the very low probability of false inclusion previously calculated for x = 0.2 is a gross overestimate of the true probability (for x = 0.016).

Using DNA fingerprint probes 33.6 and 33.15, it is possible to score the segregation of up to 34 dispersed autosomal loci in large human sibships [11]. For most loci examined, only one of the two alleles is scorable in the set of larger (>4 kb) resolved DNA fingerprint fragments, suggesting that large size differences exist between minisatellite alleles, with many alleles being located in the poorly resolved (<4 kb) region of the DNA fingerprint. This is also seen for the cloned minisatellite; HinfI alleles at this locus vary from 1.7 to 20.4 kb in length, and the allele frequency distribution in Fig. 4 shows that both alleles of this locus would be resolved in the DNA fingerprints of only 40% of individuals, while neither allele would be scorable in 14% of people.

Minisatellite probes 33.6 and 33.15 detect together about 60 hypervariable loci [11], many of which should now be amenable to cloning to provide a bank of highly informative single locus probes ideal for linkage

studies in man. Linkage analysis of inherited disorders is also possible in single large families using the entire DNA fingerprint [11]. If a polymorphic fragment is detected which appears to cosegregate with the disease, it is essential that this fragment is cloned to provide a locus-specific probe for extending the linkage analysis to additional affected families. Fragment g was selected for cloning due to its possible linkage to HPFH; we are now using pλg3 to study other affected families to obtain further evidence concerning possible linkage to this haemoglobin disorder.

ACKNOWLEDGEMENTS

We are grateful to Stephen Harris for advise on cloning and to Raymond Dalgleish for providing human DNA samples. A.J.J. is a Lister Institute Research Fellow, and this work was supported by a grant to A.J.J. and a Training Fellowship to S.L.T. from the Medical Research Council. The DNA probes are the subject of Patent Applications. Commercial enquiries should be addressed to the Lister Institute of Preventive Medicine, Brockley Hill, Stanmore, Middlesex, U.K.

REFERENCES

1. Wyman, A. and White, R. (1980) Proc. Nat. Acad. Sci. USA. 77, 6754-6758.
2. Higgs, D.R., Goodbourn, S.E.Y., Wainscoat, J.S., Clegg, J.B. and Weatherall, D.J. (1981) Nucleic Acids Res. 9, 4213-4224.
3. Bell, G.I., Selby, M.J. and Rutter, W.J. (1982) Nature 295, 31-35.
4. Capon, D.J., Chen, E.Y., Levinson, A.D., Seeburg, P.H. and Goeddel, D.V. (1983) Nature 302, 33-37.
5. Goodbourn, S.E.Y., Higgs, D.R., Clegg, J.B. and Weatherall, D.J. (1983) Proc. Nat. Acad. Sci. USA. 80, 5022-5026.
6. Lebo, R.V., Chakravarti, A., Buetow, K.H., Cheung, M.-C., Cann, H., Cordell, B. and Goodman, H. (1983) Proc. Nat. Acad. Sci. USA. 80, 4808-4812.
7. Reeders, S.T., Breuning, M.H., Davies, K.E., Nicholls, R.D., Jarman, A.P., Higgs, D.R., Pearson, P.L. and Weatherall, D.J. (1985) Nature 317, 542-544.
8. Stoker, N.G., Cheah, K.S.E., Griffin, J.R. and Solomon, E. (1985) Nucleic Acids Res. 13, 4613-4622.
9. Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985) Nature 314, 67-73.
10. Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985) Nature 316, 76-79.
11. Jeffreys, A.J., Wilson, V., Thein, S.L., Weatherall, D.J. and Ponder, B.A.J. (1986) Am. J. Hum. Genet., in press.
12. Nilsson, K., Klein, G., Henle, W. and Henle, G. (1971) Int. J. Cancer 8, 443-450.
13. Dretzen, G., Bellard, M., Sassone-Corri, P. and Chambon, P. (1981) Anal. Biochem. 112, 295-298.

14. Feinberg, A.P. and Vogelstein, B. (1984) Anal. Biochem. 137, 266-267.
15. Yang, R.C.-A., Lis, J. and Wu, R. (1979) Meth. Enzymol. 68, 176-182.
16. Loenen, W.A.M. and Brammar, W.J. (1980) Gene 20, 249-259.
17. Jeffreys, A.J., Barrie, P.A., Harris, S., Fawcett, D.H., Nugent, Z.J. and Boyd, A.C. (1982) J. Mol. Biol. 156, 487-503.
18. Blattner, F.R., Williams, B.G., Blechl, A.E., Denniston-Thompson, K., Faber, H.E., Furlong, L., Grunwald, D.J., Kiefer, D.O., Moore, D.D., Schumm, J.W., Sheldon, E.L. and Smithies, O. (1977) Science 194, 161-169.
19. Vieira, J. and Messing, J. (1982) Gene 19, 259-268.
20. Matfield, M. (1983) Ph.D. thesis, University of Leicester.
21. Yanis-Perron, C., Vieira, J. and Messing, J. (1985) Gene 33, 103-119.
22. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. USA. 74, 5463-5467.
23. Biggin, M.D., Gibson, T.J. and Hong, H.F. (1983) Proc. Nat. Acad. Sci. USA. 80, 3963-3965.
24. Wyman, A.R., Wolfe, L.B. and Botstein, D. (1985) Proc. Nat. Acad. Sci. USA. 82, 2880-2884.
25. Rouyer, F., Simmler, M.-C., Johnsson, C., Vergnaud, G., Cooke, H.J. and Weissenbach, J. (1986) Nature 319, 291-295.
26. Krontiris, T.G., DiMartino, N.A., Colb, M. and Parkinson, D.R. (1985) Nature 313, 369-374.
27. Ewens, W.J. (1972) Theor. Popul. Biol. 3, 87-112.
28. Morton, N.E. (1982) Outline of genetic epidemiology. Basel, Karger.
29. Botstein, D., White, R.L., Skolnick, M. and Davis, R. (1980) Am. J. Hum. Gen. 32, 314-331.
30. Gill, P., Jeffreys, A.J. and Werrett, D.J. (1985) Nature 318, 577-579.
31. Jeffreys, A.J., Brookfield, J.F.Y. and Semeonoff, R. (1985) Nature 317, 818-819.