Biochem. Soc. Trans. (1987), 15, 309-317

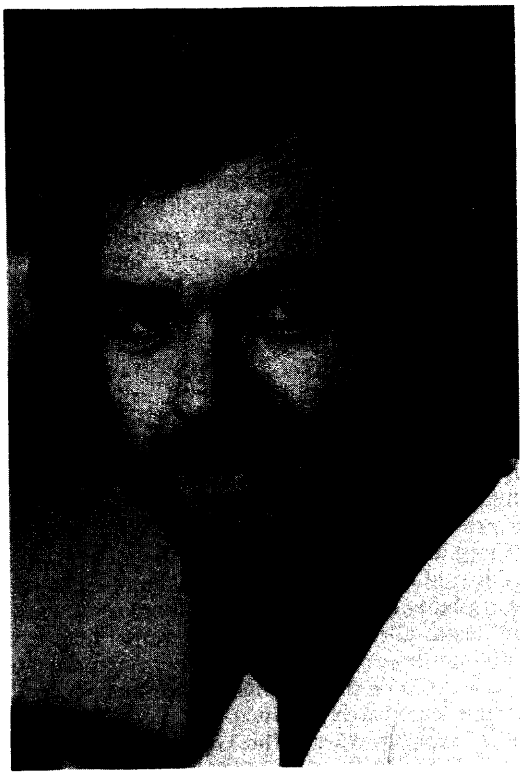
Highly variable minisatellites and DNA fingerprints

ALEC J. JEFFREYS

Department of Genetics, University of Leicester, Leicester LE1 7RH, U.K.

Twenty-Third Colworth Medal Lecture

Delivered on 25 September 1986 at Trinity College, Dublin



DR. A. J. JEFFREYS, F.R.S.

Genetic markers form the basis of any genetic analysis. Until the late 1970s, almost all human biochemical markers were confined to protein polymorphisms detected serologically or by gel electrophoresis. With the advent of DNA cloning and Southern blot detection of single copy genes in total human DNA, it became possible to analyse variability directly at the level of genomic DNA. An early analysis of restriction fragment variation in the human β-globin gene cluster revealed that DNA polymorphisms do exist at a reasonable frequency and provide useful co-dominant markers (Jeffreys, 1979). The restriction fragment length polymorphisms (RFLPs) detected all resulted from the gain/loss of a single restriction endonuclease cleavage site to produce dimorphisms whose usefulness as genetic markers is limited by their low heterozygosity: for a given diallelic marker, the maximum frequency of heterozygotes obtainable in a population in the absence of selection is 50%.

Since this early survey, many examples of RFLPs detected by human gene probes or random cloned DNA segments have been reported (see Cooper & Schmidtke, 1984). In almost every instance, variability results from restriction site gain/loss by base substitution or microdeletion/insertion. The overall variability in human DNA is low, with a mean heterozygosity per base pair (bp) of about 0.002 (Jeffreys, 1979; Cooper & Schmidtke, 1984). Variable sites are not uniformly dispersed: some regions such as the highly polymorphic HLA gene cluster are rich in RFLPs (Wake *et al.*, 1982), whereas other genes, for example thyroglobulin (Baas *et al.*, 1984), are markedly deficient in DNA variants. Thus, given a cloned DNA segment, the discovery of associated RFLPs can be tedious, and can involve screening with numerous restriction endonucleases before a suitable polymorphism is found. If more than one RFLP is discovered within or adjacent to a gene of interest, the gain in heterozygosity over one marker is often poor, as the result of linkage disequilibrium arising presumably through neutral drift leading to only a limited number of haplotypes of closely linked markers in a given population (see Collins & Weissman, 1984).

RFLPs can be more efficiently detected using long probes to screen multiple restriction sites at a given locus (Litt & White, 1985), and by using restriction endonucleases such as *Taq*I which cleave at sequences containing the mutable CpG doublet (Barker *et al.*, 1984). Recent developments in denaturing gradient gels make it possible to detect additional DNA polymorphisms which do not create or destroy restriction endonuclease cleavage sites (Myers *et al.* 1985). While the latter approach provides an exciting new method for studying DNA variation at defined and clinically important loci, it remains to be seen whether the technology can be simplified sufficiently for widescale linkage analysis in man.

Despite the limitations of RFLPs, they have revolutionized human genetics over the last few years, and have provided for the first time an almost unlimited source of genetic markers in man. The construction of detailed linkage maps of human chromosomes has become a real possibility (Botstein *et al.*, 1980; White *et al.*, 1985), and already considerable progress has been made in mapping the human X-chromosome (Drayna & White, 1985) and several autosomes (see White *et al.*, 1986). RFLPs have also been extensively used to search for markers linked to disease loci whose gene product is unknown. Recent spectacular advances have been the discovery of markers linked to Huntington's chorea (Gusella *et al.*, 1983), adult polycystic disease of the kidney (Reeders *et al.*, 1985) and cystic fibrosis (Tsui *et al.*, 1985). Despite this progress, it should be stressed that the logistics of detecting linkage with randomly selected markers are formidable: given that the human genome is about 3300 map units (cM) long, at least 115 randomly dispersed markers would have to be screened before there was even a 50/50 chance that one marker would be linked within 10 cM of a defined disease locus. Given that most RFLPs are diallelic and would be uninformative in most pedigrees, the prior odds of detecting linkage between a disease locus and a random marker in a given pedigree are even lower. This second problem could be circumvented by using more highly polymorphic markers.

RFLPs closely linked to, or within, disease loci can be used for antenatal diagnosis, carrier detection and counselling, provided that sufficient pedigree data exist for a given family to establish linkage phase between marker alleles and the disease (see Weatherall, 1985). RFLPs have also been applied to the analysis of chromosome rearrangements in cancer, most notably in retinoblastoma and Wilms' tumour where RFLP marker loss, resulting from partial or complete loss of a specific chromosome via mitotic recombination or non-disjunction, is associated with the unmasking of a functionally recessive oncogene as an integral part of tumour development (Cavanee *et al.*, 1983; Koufos *et al.*, 1984; Orkin *et al.*, 1984; Reeve *et al.*, 1984; Friend *et al.*, 1986). Other applications of RFLPs include studies on the population genetics of linked haplotypes of RFLPs, which have led to the discovery of meiotic recombination hotspots in gene clusters (Chakravarti *et al.*, 1984) and have illuminated our understanding of the emergence and diversification of human races (Wainscoat *et al.*, 1986).

Hypervariable loci in human DNA

As already mentioned, most RFLPs result from restriction site gain/loss usually as the result of base substitution. However, other mutational mechanisms, such as transposition, unequal and illegitimate recombination and replication slippage, all act on DNA and might lead to the generation of localized regions of high variability. The chance discovery by Wyman & White (1980) of a random human DNA segment which defined a multiallelic locus was the first direct demonstration that hypervariable regions (HVRs) exist in human DNA, although the variable DNA region itself has only recently been cloned (Wyman *et al.*, 1985).

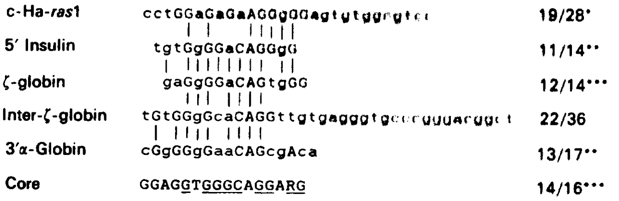
More recently, a number of other HVRs have been discovered by chance in human DNA, including a region 5′ to the human insulin gene (Bell *et al.*, 1982), another 3′ to the c-Ha-*ras1* oncogene (Capon *et al.*, 1983) and no less than three HVRs in and around the α-globin gene cluster (Higgs *et al.*, 1981; Proudfoot *et al.*, 1982; Goodbourn *et al.*, 1983; Jarman *et al.*, 1986). In each case, the HVR consists of tandem repeats of a short sequence. Hypervariability at these 'minisatellites' results from changes in the number of repeats, presumably driven either by unequal recombination between misaligned minisatellites or by slippage at replication forks leading to the gain or loss of repeat units. The resulting length variability can be high, with, in some cases, scores of different length alleles, and the frequency of heterozygotes can sometimes approach 100%. Furthermore, detection of RFLPs at these HVRs is no longer dependent on the restriction endonuclease used (provided that it does not cleave the minisatellite repeat unit), and these loci provide ideal markers for human genetics (Reeders *et al.*, 1985).

The total number of hypervariable loci in human DNA is unknown but is likely to be large. Knowlton *et al.* (1986) have recently screened 1680 different recombinants from a human genomic library, and have discovered at least 12 clones which contain highly polymorphic regions. At face value, this suggests that the human genome might contain at least 1500 HVRs, more than enough to saturate the human linkage map with highly informative markers provided that some general method is developed for their isolation.

The minisatellite 'core' sequence

Fig. 1 shows the repeat units from several different tandem-repetitive HVRs in human DNA. Most show an unusual base composition, with a marked purine/pyrimidine asymmetry between strands. In addition, there is some similarity in repeat unit sequence between the insulin HVR and the three α-globin clusters HVRs (Proudfoot *et al.*, 1982; Goodbourn *et al.*, 1983; Jarman *et al.*, 1986), suggesting that certain classes of sequence might be predisposed towards forming minisatellites. As predicted, some of these HVRs do weakly hybridize at low stringency to other human DNA fragments, some of which are polymorphic (Goodbourn *et al.*, 1983; Jarman *et al.*, 1986).

During DNA sequence analysis of the human myoglobin gene we discovered a small minisatellite comprised of four repeats of a 33 bp sequence within one of the introns (Weller *et al.*, 1984). Although monomorphic, this minisatellite again showed some sequence similarity to other HVRs and furthermore cross-hybridized weakly to multiple loci. A random selection of eight of these loci were cloned and characterized (Jeffreys *et al.*, 1985*a*). All were minisatellites with repeat units between 16 and 64 bp long repeated from 3 to 29 times depending on the clone. Four of these minisatellites showed allelic variation in repeat unit number, although the level of variability was limited; the most highly polymorphic locus obtained had eight alleles containing between 12 and 25 repeat units and with a heterozygosity of about 80%.



```
c-Ha-ras1        cctGGaGaGaAGGgGGagtgtggrgtcc             19/28*
                     | |   |||||
5′ Insulin        tgtGgGGaCAGGgG                          11/14**
                   | |||||||| ||
ζ-globin           gaGgGGaCAGtgGG                         12/14***
                     ||| ||||
Inter-ζ-globin   tGtGGgGcaCAGGttgtgagggtgcccgggacggct     22/36
                  | |||| ||||
3′α-Globin       cGgGGgGaaCAGcgAca                        13/17**
Core             GGAGGTGGGCAGGARG                         14/16***
```

Fig. 1. *Repeat unit sequences of hypervariable loci in human DNA*

Consensus repeat sequences are shown for the insulin 5′ HVR (Bell *et al.*, 1982), ζ-globin intron HVR (Proudfoot *et al.*, 1982), inter-ζ-globin HVR (Goodbourn *et al.*, 1983), the HVR 3′ to the α-globin gene (Jarman *et al.*, 1986) and 3′ to the c-Ha-*ras*1 oncogene (Capon *et al.*, 1983). Possible alignments of these sequences with the minisatellite core sequence (Jeffreys *et al.*, 1985*a*) are shown in uppercase, but are of weak statistical significance. Most sequences show a marked purine/pyrimidine strand asymmetry (number of purines/number of bases; *, **, ***, bias significant at the 10%, 5% and 1% level respectively).

Comparison of the repeat units of the myoglobin minisatellite and the eight derived minisatellites showed that they shared, with minor variations, a common 11–16 bp 'core' sequence GGAGGTGGGCAGGARG, present once per repeat unit. Since the remainder of the repeat units showed no similarity between different minisatellites, it is most unlikely that this set of core-containing minisatellites constitutes a family of evolutionarily related sequences. Instead, it appears probable that the core sequence assists minisatellite production, by promoting the initial duplication of a DNA segment containing a core sequence and/or by aiding the subsequent changes in repeat number driven by unequal exchange or DNA slippage at replication to produce long and highly variable minisatellites. Since we seen no reason to invoke directionality in this process, this model predicts that the majority of core-containing loci will contain core monomers or oligomers, and that only a minority of minisatellites will have fortuitously amplified until they have attained a high repeat copy number and extensive polymorphism. If this model is correct, different core-containing minisatellites are the products of convergent evolution and do not constitute a DNA sequence family in the accepted sense.

The function of the core sequence

While its role is unknown, there are several lines of evidence which suggest that the core sequence might serve as a recombination signal in human DNA:

(1) The repeat units of even the most polymorphic minisatellites are never completely homogeneous. Instead, several closely related repeat motifs can exist in a single allele, which furthermore are seldom confined to a single repeat unit but tend to be diffused across several non-adjacent repeats (Fig. 2). This evidence of large-scale duplication/deletion points towards the recombinational processes of unequal exchange and gene conversion as being the pre-

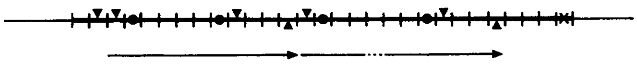


Fig. 2. *DNA sequence variation at a minisatellite*

The cloned minisatellite 33.15 (Jeffreys *et al.*, 1985*a*) is comprised of 29 repeats of a 16 bp sequence present in the five variant forms shown. Note the large duplication, together with the diffusion of most variants across the minisatellite as expected for incomplete homogenization by cross-over fixation (Smith, 1976; Dover, 1982).

dominant mechanisms of allelic change, rather than replication slippage which would tend to cause the gain/loss of one or a few repeat units and result in the lateral diffusion of new repeat unit variants into adjacent repeats (Smith, 1976; Dover, 1982).

(2) We have discovered several instances of new 'mutations' (DNA fragment length changes) arising in minisatellites (Jeffreys *et al.*, 1985*a*; Thein *et al.*, 1987; I. Patel & A. J. Jeffreys, unpublished work). In most cases, the allelic length change is considerable, involving up to kilobases of minisatellite, again consistent with substantially staggered unequal exchanges.

(3) Assuming no selection on minisatellites, it is possible to use the neutral mutation-random drift model to estimate the rate of unequal exchange required to maintain the observed level of minisatellite repeat number variability seen in human populations. Despite the uncertainties implicit in such calculations, the estimated rate is high (about 10^{-4} per kb of minisatellite; Jeffreys *et al.*, 1985*a*; Wong *et al.*, 1986) and is consistent with the number of new mutant alleles that we have so far observed in human pedigrees. In comparison, the mean rate of meiotic recombination along human DNA is much lower ($\sim 10^{-5}$ per kb; Botstein *et al.*, 1980). Core-containing minisatellites therefore appear to be recombination hotspots, as predicted if the core sequence drives recombination. However, it is not yet known whether meiotic recombination or germ-line mitotic recombination (sister-chromatid exchange) are the predominant forces acting on minisatellites to generate variability. This can only be directly assessed by studying the exchange or otherwise of markers flanking new mutant minisatellite alleles (see below).

(4) The E_β gene in the mouse major histocompatibility complex (MHC) contains a known meiotic recombination hotspot. Sequence analysis of this region has revealed a short tandem-repetitive region closely related to the core sequence (Steinmetz *et al.*, 1986). While it is tempting to speculate that this region is responsible for driving meiotic recombination, other meiotic recombination hotspots, between the $A_{\beta3}$ and $A_{\beta2}$ genes in the murine MHC (Uematsu *et al.*, 1986) and in the human β-globin gene cluster (Chakravarti *et al.*, 1984), do not contain core minisatellites. The core sequence is therefore most unlikely to be the sole mediator of mammalian recombination.

(5) The core sequence is similar in length and G-richness to chi, the cross-over hotspot initiator sequence GCTGGTGG of *Escherichia coli*, a sequence believed to initiate generalized recombination by serving as the recognition sequence for the binding of the *rec*BC gene product, endonuclease V (Smith *et al.*, 1981; Smith, 1983). It remains to be seen whether the core sequence is the functional mammalian homologue or analogue of chi, though if it is, one can readily envisage how the core sequence could drive an initial localized duplication by strand displacement and repair synthesis at an incipient Holliday junction, as well as by promoting the subsequent unequal exchange required for the formation of long and variable minisatellites (Jeffreys *et al.*, 1985*a*).

There is no clear evidence yet for other minisatellite-associated core sequences in human DNA. Jarman *et al.* (1986) have proposed a second core, GNGGGG(N)ACAG, based on a comparison of the insulin and three α-globin HVRs, but this sequence might represent a variant of the myoglobin core. We should stress that the core sequence defined by hybridization to the myoglobin minisatellite is also likely to be biased towards that version present in the myoglobin gene. A fuller assessment of the numbers and classes of core sequences can only be made by sequence analysis of a much larger spectrum of minisatellites.

Polycore probes and DNA fingerprints

The myoglobin minisatellite used to define the core sequence contains 17 bp of non-core sequence per repeat unit in addition to the 16 bp core. This non-core sequence actively impedes cross-hybridization to other core-containing minisatellites (Jeffreys *et al.*, 1985*a*), and it therefore follows that hybridization probes consisting of tandem repeats of only the core sequence should detect minisatellites more effectively. Two such probes, termed 33.6 and 33.15 and containing different versions of the core sequence, were isolated as minisatellites from human DNA (Jeffreys *et al.*, 1985*a*). Other polycore probes have since been chemically synthesized. Such probes cross-hybridize efficiently to a large number of fragments in human DNA, many of which show substantial variability. The complex Southern blot profiles detected can be clarified by digesting human DNA with a restriction endonuclease such as *Hin*fI which cleaves at a 4 bp recognition sequence, to remove the bulk of flanking DNA from a minisatellite and maximize the resolution of allelic length variation. The resulting Southern blot profile consisting of a complex set of large and highly variable DNA fragments is termed a DNA 'fingerprint' (Fig. 3) (Jeffreys *et al.*, 1985*a,b*). There is, in addition, an irresolvably complex mixture of shorter hybridizing DNA fragments presumably derived from short and relatively monomorphic minisatellites.

Properties of human DNA fingerprints

The DNA fingerprint pattern is sensitive to the core sequence present in the polycore probe, and the two probes, 33.6 and 33.15, detect almost completely different sets of hypervariable loci to produce independent DNA fingerprints (Figs. 3 and 4). The patterns almost always show somatic stability, are identical in monozygous twins and are maintained in cultured cell lines (Jeffreys *et al.*, 1985*b*). Only two instances of somatic variation have been found. First, minor variation between tissues has been found with restriction endonucleases such as *Hin*fI and *Sau*3A whose sites can be blocked by methylation (for example, a *Sau*3A site terminating in mCG,GATmCG, is resistant to *Sau*3A cleavage; see McClelland & Nelson, 1985). This DNA fingerprint variability probably arises through tissue differences in methylation, and disappears when endonucleases such as *Alu*I and *Hae*III, which cannot be blocked by CpG methylation, are used. The second exception to the somatic stability rule can occur in tumours (see below).

DNA fingerprints also show substantial germ-line stability, with all hypervariable DNA fragments in a child being traceable back to the parents and in turn to the grandparents. In a large survey of many families where parentage is beyond dispute (see below), we estimate that roughly one offspring fragment in 300 cannot be detected in either parent (Jeffreys *et al.*, 1985*a*; I. Patel & A. J. Jeffreys, unpublished work). This rough estimate of the mutation rate to new length alleles, which is consistent with previous population genetic estimates of mutation rate (Jeffreys *et al.*, 1985*a*), is of course a mean over all hypervariable loci scored, and is likely to vary from locus to locus (see below).

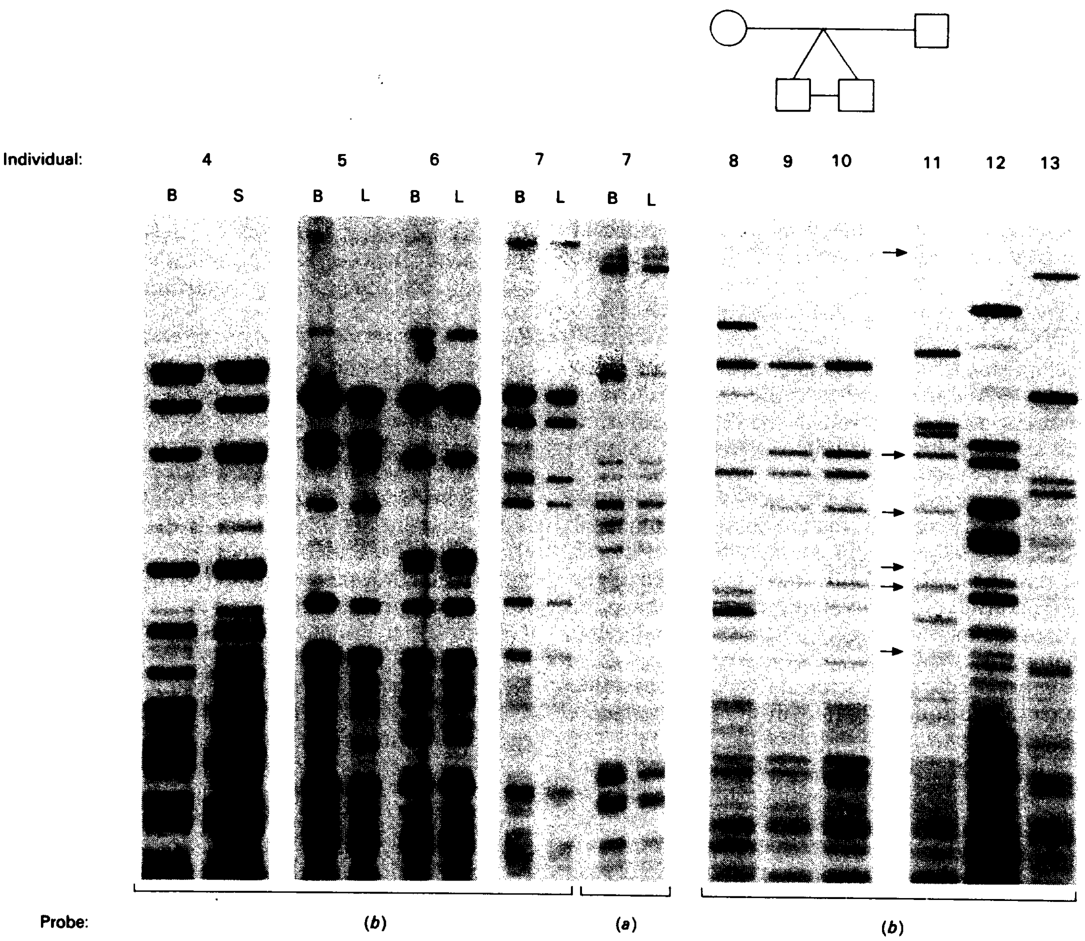


Fig. 3. *Human DNA fingerprints detected by digestion of human DNA with HinfI followed by Southern blot hybridization with minisatellite probes 33.6 (a) or 33.15 (b)*

The hybridizing fragments shown range from 1.5 kb to >20 kb in length; many additional small fragments have been electrophoresed off the gel. Somatic stability is shown by indistinguishable patterns in blood (B) and sperm (S), in identical twins 9 and 10, and in comparison of blood (B) versus Epstein–Barr virus transformed lymphoblastoid cell lines (L). Germ-line stability is indicated by the family group 8–11, with paternal bands in offspring 9, 10 being arrowed. Three unrelated males (11–13) illustrate the individual specificity of DNA fingerprints. From Jeffreys *et al.* (1985*b*).

Genetic complexity of DNA fingerprints

The number and distribution of hypervariable loci can be studied by scoring DNA fingerprint fragment segregation in large human sibships (Fig. 4) (Jeffreys *et al.*, 1986). As expected, most of the resolved parental fragments behave as single heterozygous Mendelian characters and are transmitted on average to half of the offspring. Most parental fragments do not have resolvable allelic partners, suggesting that large size differences can exist between alleles at a given locus, and that alleles are frequently present in the irresolvably complex set of small DNA fragments. Instances of linkage between two or three heterozygous parental fragments have been found; these presumably arise from long minisatellites containing the occasional internal restriction endonuclease cleavage site to produce a 'haplotype' of linked fragments (Jeffreys *et al.*, 1986).

Table 1 summarizes the genetic properties of DNA fingerprints detected by probes 33.6 and 33.15, deduced from pedigree analysis. The large resolvable fragments are derived from a pool of, very approximately, 60 hypervariable loci. In addition, there are likely to be many other variable loci whose alleles are confined to the short (<3 kb) and irresolvable region of the gel. In a given parent, segregation information can be obtained in his or her children on about 30 of these hypervariable loci simultaneously, using both polycore probes. Different combinations of loci will therefore be represented in the set of resolved fragments in different

Table 1. *Summary of hypervariable loci which can be scored in a DNA fingerprint*

These data are taken from the DNA fingerprints detected by probes 33.6 and 33.15 in the family shown in Fig. 4.

	Father	Mother
Total no. of fragments scored	41	32
No. of allelic pairs	6	6
No. of linked pairs	1	1
No. of distinct loci scored	34	25
Total no. of hypervariable loci in DNA fingerprint*	~66	~43

*The approximate total number of HVRs detected which can have large resolved and scorable alleles can be estimated from the low proportion of variable fragments which can be paired into alleles (Jeffreys *et al.*, 1986).

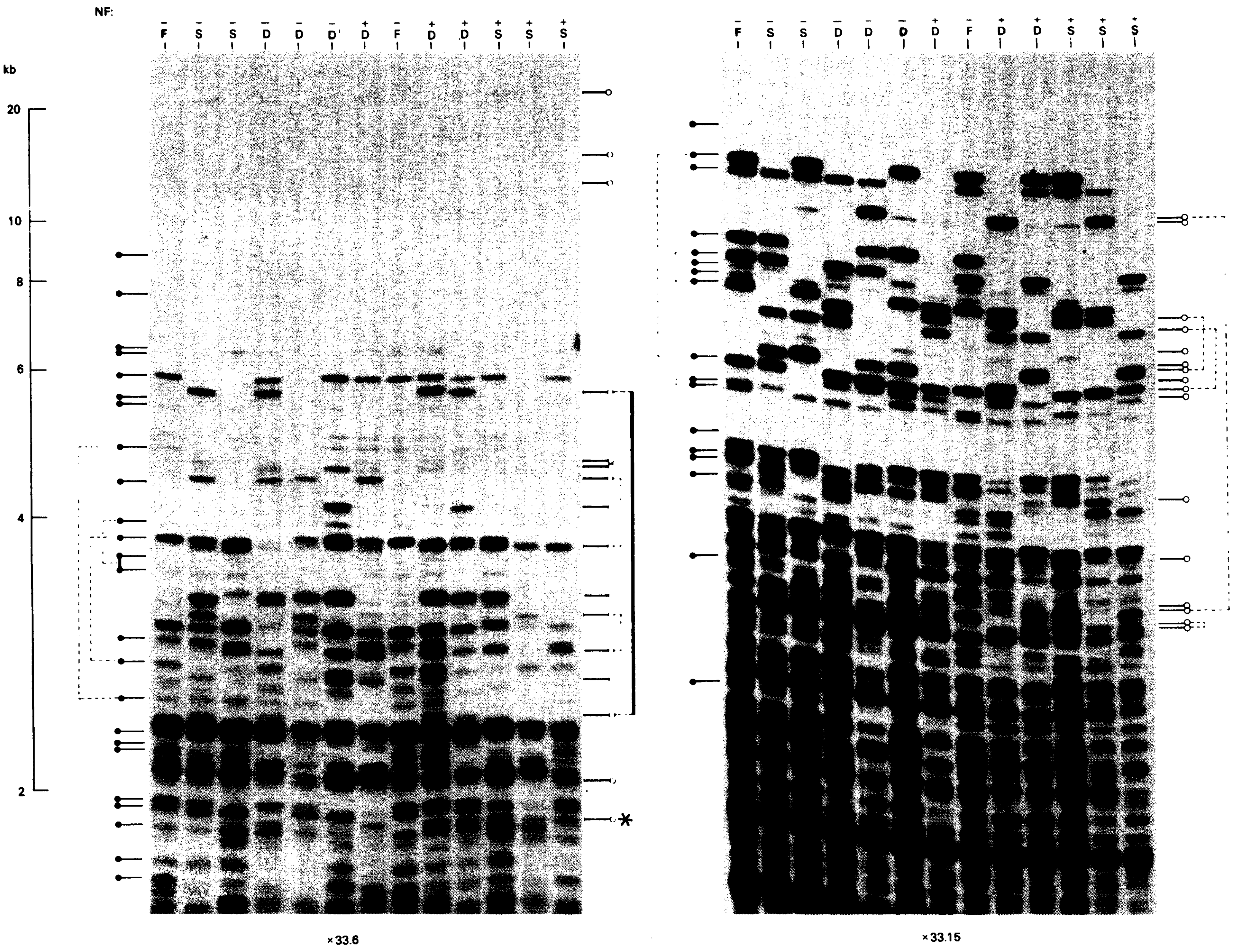


Fig. 4. *DNA fingerprints of a single large sibship*

F, Father; S, son; D, daughter. The mother was unavailable but maternal DNA fragments can be readily identified as fragments present in offspring but not in the father. Parental DNA fragments whose segregation could be scored in offspring are indicated (●, paternal; ○, maternal). Pairs of fragments connected by dotted lines are allelic, and pairs connected by solid lines are linked in coupling. This sibship is also segregating for the autosomal dominant disease, neurofibromatosis (NF), inherited from the affected mother; +, child affected; –, child unaffected. The DNA fragment marked with an asterisk shows weak evidence for co-segregation with neurofibromatosis. From Jeffreys *et al.* (1986).

individuals. The loci scored are all autosomal and appear to assort independently, suggesting widespread dispersal around the human genome (Jeffreys *et al.*, 1986).

Individual specificity of human DNA fingerprints

DNA fingerprints vary substantially between unrelated people (Fig. 3) and even between closely related individuals such as siblings (Fig. 4). On average, 36 bands > 3 kb long can be scored per individual, using two DNA fingerprint probes. The mean level of band sharing, s, between unrelated individuals is 25%, both for North Europeans (Jeffreys *et al.*, 1985*b*) and for individuals from the Indian subcontinent (I. Patel & A. J. Jeffreys, unpublished work). This band sharing probability s is heterogeneous, and falls to < 10% for the largest DNA fragments. In addition, many instances of apparent band sharing between two unrelated individuals may well arise through fortuitous co-migration of different minisatellites, rather than representing allelic identity.

Since most DNA fingerprint bands assort independently, the probability of each band's presence is not conditioned significantly by the presence of other bands. We can therefore conservatively estimate the probability that all 36 bands in one individual A are present in a second unrelated individual B as $s^{36} = 2 \times 10^{-22}$. Heterogeneity in s will reduce this probability. Similarly, the chance that A and B have identical DNA fingerprints can be estimated at $(1 - 2s + 2s^2)^{36/s} = 4 \times 10^{-30}$. Finally, the probability that two first degree relatives, for example siblings, are identical is about 3×10^{-14} (Jeffreys & Morton, 1987). Since all of these probabilities are rather less than the reciprocal of the world population, it seems reasonable to conclude that these DNA fingerprints are completely individual-specific, with the exception of monozygous twins.

DNA fingerprint applications based on individual-specificity

DNA fingerprints provide a rich source of genetic marker information and can be applied to a wide range of problems in human genetics. They provide a reliable approach to zygosity testing which obviates the need for screening large numbers of conventional markers in the search for sibling discordancies which would indicate dizygosity, and provides a far more certain estimate of the probability of monozygosity than is achievable using normal markers (Hill & Jeffreys, 1985). DNA fingerprints also provide a ready source of markers for monitoring bone-marrow transplants (except in the rare instances of donor and recipient being monozygous twins) (Thein *et al.*, 1986).

DNA fingerprints can also be used to study chromosome and DNA changes in cancer: in a survey of 35 different tumours, clear alterations in DNA fingerprints were detectable in 10 cases (Thein *et al.*, 1987). Most changes resulted in band loss or shifts in hybridization intensity, presumably as the result of processes such as chromosome loss and DNA amplification known to occur in tumour cells. Interestingly, three cases of adenocarcinoma of the gastrointestinal tract showed novel DNA fragments not present in normal DNA. These 'mutant' bands presumably arise by somatic changes in minisatellite allele length, perhaps by sister-chromatid exchange, and establish that minisatellite length changes are not confined to the germ-line. It is not yet known whether these minisatellite length changes are associated with tumour development, or whether they reflect the product of normal processes in somatic tissue which only become apparent on clonal expansion of a tumor cell. In any event, DNA fingerprint changes in tumours provide interesting new markers for studying tumour clonality and progression.

DNA fingerprints can be obtained from as little as 50 μl of blood, 5 μl of semen or 15 hair roots. In addition, DNA frequently survives sufficiently intact for DNA fingerprint analysis in dried blood and semen stains up to 5 years old, and sperm DNA fingerprints can be obtained from vaginal swabs 7 h or more after intercourse (Gill *et al.*, 1985). The forensic implications, particularly for the conclusive identification of rapists, are evident, and represent a dramatic improvement on the currently available semen markers which can provide at the very best only a 99% certainty of correct identification (Sensabaugh, 1982).

DNA fingerprints in establishing family relationships

Since DNA fingerprints give information on a large number of informative Mendelian markers, most of which are rare in a population, they provide a powerful new method for establishing family relationships in, for example, paternity disputes (Jeffreys *et al.*, 1985*b*) (Fig. 5). By comparing the DNA fingerprints of a mother and her child, it is possible to identify DNA fragments in the child which are absent from the mother and must therefore have been inherited from the biological father. Typically, the two polycore probes routinely used identify 10–15 such bands. If the claimed father is not the biological father, most of these paternal bands will be absent from his DNA fingerprints, producing multiple exclusions (Fig. 5). If he is the father, then all of the child's paternal bands should be present in his DNA fingerprints (Fig. 3). The chance that a randomly picked man would accidentally contain all of these paternal bands can be conservatively estimated at s^{10} to $s^{15} = 10^{-6}$ to 10^{-9}. Thus paternity can be established with a level of certainty which far exceeds that obtainable using a substantial battery of conventional genetic markers (Dodd, 1985). The DNA fingerprint test has also been applied to a number of immigration disputes, usually where a sponsor in the U.K. is attempting to bring what he claims are his wife and children to this country. In the absence of adequate documentary evidence, DNA fingerprints can generally conclusively establish one way or the other the claimed relationship by testing whether all of the children's DNA fragments are present in the claimed mother and/or father. Depending on the relatives available, it is even possible sometimes to conduct the test where one parent is deceased or otherwise unavailable, by making use of undisputed relatives to reconstruct partially or wholly the DNA fingerprint of the missing parent (see Fig. 4) (Jeffreys *et al.*, 1985*c*; Hill, 1986).

DNA fingerprint evidence demonstrating paternity has already been presented in an English court of law, and has also been accepted by the U.K. Home Office as satisfactory evidence for resolving immigration disputes (Jeffreys *et al.*, 1985*c*). I would, however, like to point out that, contrary to statements in the popular press, this test is not foolproof. It cannot necessarily detect blood sample substitutions, whether accidental or deliberate. More significantly, the use of highly polymorphic markers means that family groups

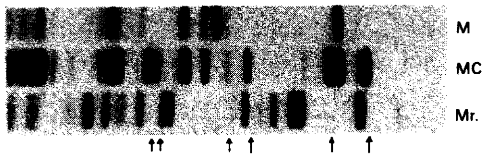


Fig. 5. *Resolution of a paternity dispute by DNA fingerprint analysis*

M, Mother; MC, mother's child; Mr. Y, the man accused of being the child's father. Note that many of the child's paternal DNA fragments (arrowed) are not present in Mr. Y, establishing through multiple exclusions that he cannot be the father of this child.

containing a child with a new mutant minisatellite fragment will inevitably arise, albeit fairly infrequently. The presence of a single unascribable band could indicate new mutation in an otherwise biologically correct family group. Alternatively, it might be possible, depending on the case under investigation, that one of the claimed parents is a very close blood-relative of the true parent, thereby accounting for the very close but not perfect degree of band matching. The relative likelihood of these two possibilities can be calculated provided that the new mutation rate is known. At the moment, the mutation rate estimate is approximate (see above).

The above discussion demonstrates that DNA fingerprint analysis in resolving for example paternity disputes is markedly different from tests using conventional genetic markers. In the latter, the paradigm is exclusion: markers are tested until an exclusion is found, and, if no exclusion emerges, the probability of false inclusion of a non-father is then calculated. With DNA fingerprints and their associated statistical power, the paradigm is inclusion: if all offspring bands fit the parents, then the relationship is established with legal, if not strictly mathematical, certainty (Hill, 1986). If there are parental mismatches, then a variety of hypotheses need to be considered, including switching of samples, new mutation and a genuinely incorrect relationship. Finally, we appreciate the ethical dilemmas which will sometimes arise when family relationships are defined by strict biological criteria rather than by the everyday definition of the family group.

DNA fingerprints and inherited disease

As mentioned above, the two polycore probes that we routinely use detect at least 60, and probably many more, hypervariable loci. Furthermore, the set of detectable HVRs should be expandable by using new modified polycore probes and by the discovery, if they exist, of new core sequences. The potentially rich variety of HVRs detectable by these approaches can be applied to the search for markers linked to disease loci in two ways: first, by cloning individual HVRs from DNA fingerprints and producing locus-specific HVR probes for linkage analysis, and second, by screening multilocus DNA fingerprints for a hypervariable fragment which co-segregates with a disease locus in a large pedigree (Jeffreys *et al.*, 1986). Fig. 4 shows an example of the second approach with a large sibship segregating for neurofibromatosis, an autosomal dominant cancer. One fragment (marked with an asterisk) is transmitted from the affected parent to her affected offspring, though one unaffected child also inherits the band. The lod score (logarithm of the odds in favour of linkage) is 1.5, well short of the value of 3 considered to be very strong evidence for linkage.

While DNA fingerprints provide a large amount of genetic marker information suitable for linkage analysis, there are several drawbacks to the approach. First, disease diagnosis has to be accurate in the limited number of members of the pedigree. Second and more important, linkage between a DNA fingerprint fragment and a disease locus cannot be further tested in other affected families, since DNA fragments allelic to that found in the first pedigree tested cannot be clearly identified. Having identified such a putatively linked fragment in one pedigree, it is necessary to clone this fragment to produce a locus-specific HVR probe for studying linkage to this locus in additional affected families, and to localize the locus within the genome. Thus DNA fingerprints cannot be used to establish linkage, but, given a dominant disorder and a suitably large horizontal pedigree, they can be used to screen for HVRs with a high probability of linkage. Assuming random dispersal of minisatellites and disease loci around the human linkage map, it is possible to calculate the probability that a given co-segregating band, if cloned, would prove to be linked to the disease locus (A. J. Jeffreys, unpublished work). For the neurofibromatosis DNA fragment (Fig. 4), the probability of genuine linkage within 20 cM of the disease locus, rather than chance co-segregation, is about 16%, compared with 1.2% for a randomly picked marker.

Isolation of locus-specific HVR probes

Individual bands in a DNA fingerprint can be readily isolated, first by preparative gel electrophoresis to separate the require band from all other hybridizing fragments, and second by cloning into bacteriophage λ or plasmids (Wong *et al.*, 1986; Z. Wong, V. Wilson & A. J. Jeffreys, unpublished work). Under high-stringency hybridization conditions, the cloned band consisting of minisatellite and a small amount of flanking DNA almost always acts as a locus-specific HVR probe.

One such HVR probe, initially identified as a large fragment in a DNA fingerprint which appeared to co-segregate with a form of hereditary persistence of foetal haemoglobin apparently controlled by an autosomal dominant gene unlinked to the globin gene clusters (Jeffreys *et al.*, 1986), has been studied in detail (Wong *et al.*, 1986). The minisatellite is comprised of 37 bp repeat units containing the core sequence, and is extremely variable (Fig. 6). Only the shortest allele containing 14 repeat units is common (population frequency = 0.165). All remaining alleles containing between 15 and more than 500 repeat units are rare (mean frequency = 0.008), and 97% of individuals are heterozygous at this locus. This illustrates the extreme variability of the largest fragments in a DNA fingerprint, and establishes that cloned core-containing minisatellites will provide a rich new source of highly informative genetic markers ideal for human linkage analysis. Indeed, given present estimates of the number of variable minisatellites, it is quite conceivable that sufficient markers could be isolated to provide a low-to-medium resolution linkage map of highly informative loci spanning the entire human genome.

DNA fingerprints of animals

If the human core sequence is a recombination signal, then it is likely to be conserved in evolution and it therefore follows that human polycore probes should cross-hybridize to minisatellites in non-human DNA. Indeed, in a preliminary survey both probes 33.6 and 33.15 cross-hybridized to multiple variable DNA fragments in every vertebrate tested, ranging from mammals to birds, reptiles, amphibians and fish. Furthermore, the intensity and complexity of the hybridization pattern showed no obvious systematic decrease with phylogenetic distance from man (A. J. Jeffreys, J. Hillel & D. B. Morton, unpublished work).

In dogs, the complexity and genetic properties of the DNA fingerprints are similar to those of man, with highly variable fragments being derived from multiple dispersed autosomal canine loci (Jeffreys & Morton, 1987). The DNA fingerprints of cats are rather less complex, though still highly informative. In contrast, farm animals (sheep, goats, pigs, cows) produce faint and relatively uninformative DNA fingerprints, and it seems possible that the core sequence may have shifted in a stem ancestor of these artiodactyls (J. Hillel, A. J. Jeffreys, V. Wilson & D. B. Morton, unpublished work). In addition, probe 33.15 cross-hybridizes strongly to the core-containing bovine 1.720 satellite DNA (Pöschl & Streek, 1980), obscuring hybridization to conventional minisatellites.

There are many possible applications of DNA fingerprints in animal breeding, for example, identification of stolen animals, verification of semen samples for artificial insemination, determination of pedigree, linkage analysis and the search for quantitative trait loci of economic inportance (Beckmann

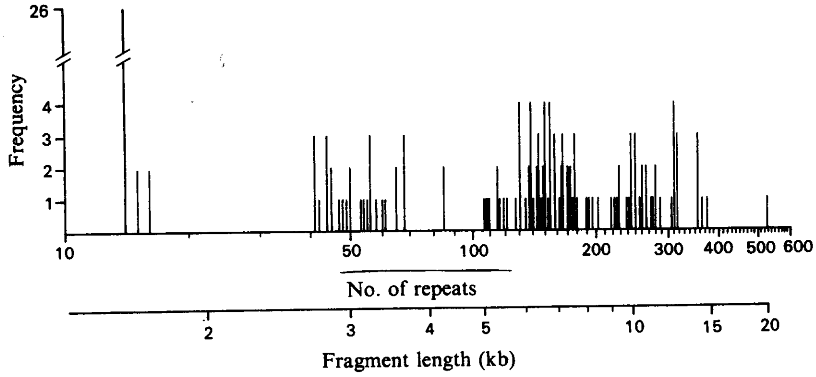


Fig. 6. *Allelic variation at a cloned hypervariable minisatellite.*

An 8 kb *Sau*3A fragment was cloned from a human DNA fingerprint to produce a locus-specific HVR probe termed pλg3. The histogram shows the distribution of allele lengths and allelic repeat unit numbers at this locus in a survey of 79 unrelated English individuals. At least 76 different alleles could be resolved, and all are rare apart from the shortest allele containing 14 repeat units. From Wong *et al.* (1986).

& Soller, 1983; Soller & Beckmann, 1983). Parent–offspring testing in the wild could lead to a more detailed understanding of the genetic structure of natural populations, and could also be used in endangered species to maximize outbreeding in captive colonies.

The cross-hybridization of human polycore probes to animal DNA also opens up the investigation of minisatellites in genetically well-defined animals. Recent analysis of DNA fingerprints in recombinant inbred strains of mice has permitted the localization of individual DNA fingerprint fragments on mouse chromosomes (A. J. Jeffreys, V. Wilson, B. A. Taylor & G. Bulfield, unpublished work). As in man, the variable loci are autosomal and dispersed, and, more important, are not preferentially associated with regions of repetitive DNA such as centromeres and telomeres. Furthermore, using mouse strains of known genealogy it is possible to estimate the level of germ-line instability of different minisatellites. Remarkably, this rate varies dramatically from locus to locus. One mouse minisatellite in particular appears to be extraordinarily unstable, with a mutation rate to new length alleles indirectly estimated at > 1% per gamete.

Such highly unstable loci will provide an ideal system for studying more directly the rates and mechanisms of allelic length change at minisatellites, and for investigating further the hypothesis that the core sequence is a recombination signal in vertebrate DNA.

I extend my profound thanks to all my colleagues who have contributed to this story, and in particular to Vicky Wilson, Zilla Wong, Ila Patel, Steve Harris, Nicky Hartley, Jossi Hillel, David Morton, John Brookfield, Robert Semeonoff, Bruce Ponder and Swee Lay Thein for their interest, help and encouragement. This work was supported by a grant from the M.R.C. A.J.J. is a Lister Institute Research Fellow. The minisatellite probes are the subject of Patent Applications, and commercial enquiries should be addressed to The Lister Institute of Preventive Medicine, Brockley Hill, Stanmore, Middx. HA7 4JD, U.K.

Baas, F., Bikker, H., van Ommen, G.- J. & de Vijlder, J. J. M. (1984) *Hum. Genet.* **67**, 301–305

Barker, D., Schafer, M. & White, R. (1984) *Cell* **36**, 131–138

Beckmann, J. S. & Soller, M. (1983) *Theor. Appl. Genet.* **67**, 35–43

Bell, G. I., Selby, M. J. & Rutter, W. J. (1982) *Nature (London)* **295**, 31–35

Botstein, D., White, R. L., Skolnick, M. & Davis, R. (1980) *Am. J. Hum. Genet.* **32**, 314–331

Capon, D. J., Chen, E. Y., Levinson, A. D., Seeburg, P. H. & Goeddel, D. V. (1983) *Nature (London)* **302**, 33–37

Cavanee, W. K., Dryja, T. P., Phillips, R. A., Benedict, W. F., Godbout, R., Gallie, B. L., Murphree, A. L., Strong, L. C. & White, R. L. (1983) *Nature (London)* **305**, 779–784

Chakravarti, A., Buetow, K. H., Antonarakis, S. E., Waber, P. G., Boehm, C. D. & Kazazian, H. H. (1984) *Am. J. Hum. Genet.* **36**, 1239–1258

Collins, F. S. & Weissman, S. M. (1984) *Progr. Nucleic Acid Res. Mol. Biol.* **31**, 315–462

Cooper, D. N. & Schmidtke, J. (1984) *Hum. Genet.* **66**, 1–16

Dodd, B. E. (1985) *Nature (London)* **318**, 506–507

Dover, G. (1982) *Nature (London)* **299**, 111–117

Drayna, D. & White, R. (1985) *Science* **230**, 753–758

Friend, S. H., Bernards, R., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M. & Dryja, T. P. (1986) *Nature (London)* **323**, 643–646

Gill, P., Jeffreys, A. J. & Werrett, D. J. (1985) *Nature (London)* **318**, 577–579

Goodbourn, S. E. Y., Higgs, D. R., Clegg, J. B. & Weatherall, D. J. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 5022–5026

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E. & Martin, J. B. (1983) *Nature (London)* **306**, 234–238

Higgs, D. R., Goodbourn, S. E. Y., Wainscoat, J. S., Clegg, J. B. & Weatherall, D. J. (1981) *Nucleic Acids Res.* **9**, 4213–4224

Hill, W. G. (1986) *Nature (London)* **322**, 290–291

Hill, A. V. S. & Jeffreys, A. J. (1985) *Lancet* **i**, 1394–1395

Jarman, A., Nicholls, R. D., Weatherall, D. J., Clegg, J. B. & Higgs, D. R. (1986) *EMBO J.* **5**, 1857–1863

Jeffreys, A. J. (1979) *Cell* **18**, 1–10

Jeffreys, A. J. & Morton, D. B. (1987) *Anim. Genet.* in the press

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985*a*) *Nature (London)* **314**, 67–73

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985*b*) *Nature (London)* **316**, 76–79

Jeffreys, A. J., Brookfield, J. F. Y. & Semeonoff, R. (1985*c*) *Nature (London)* **317**, 818–819

Jeffreys, A. J., Wilson, V., Thein, S. L., Weatherall, D. J. & Ponder, B. A. J. (1986) *Am. J. Hum. Genet.* **39**, 11–24

Knowlton, R. G., Brown, V. A., Braman, J. C., Barker, D., Schumm, J. W., Murray, C., Takvorian, T., Ritz, J. & Donnis-Keller, H. (1986) *Blood* **68**, 378–385

Koufos, A., Hansen, M. E., Lampkin, B. C., Workman, M. L., Copeland, N. G., Jenkins, N. A. & Cavanee, W. K. (1984) *Nature (London)* **309**, 170–172

Litt, M. & White, R. L. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 6206–6210

McClelland, M. & Nelson, M. (1985) *Nucleic Acids Res.* **13**, r201–r207

Myers, R. M., Lumelsky, N., Lerman, L. S. & Maniatis, T. (1985) *Nature (London)* **313**, 495–498

Orkin, S. H., Goldman, D. S. & Sallan, S. E. (1984) *Nature (London)* **309**, 172–174

Pöschl, E. & Streek, R. E. (1980) *J. Mol. Biol.* **143**, 147–153

Proudfoot, N. J., Gil, A. & Maniatis, T. (1982) *Cell* **31**, 553–563

Reeders, S. T., Breuning, M. H., Davies, K. E., Nicholls, R. D., Jarman, A. P., Higgs, D. R., Pearson, P. L. & Weatherall, D. J. (1985) *Nature* (*London*) **317**, 542–544

Reeve, A. E., Housiaux, P. J., Gardner, R. J. M., Chewings, W. E., Grindley, R. M. & Millow, L. J. (1984) *Nature* (*London*) **309**, 174–176

Sensabaugh, G. F. (1982) *Curr. Top. Biol. Med. Res.* **6**, 247–282

Smith, G. P. (1976) *Science* **191**, 528–535

Smith, G. R. (1983) *Cell* **34**, 709–710

Smith, G. R., Kunes, S. M., Schultz, D. W., Taylor, A. & Triman, K. L. (1981) *Cell* **24**, 429–436

Soller, M. & Beckmann, J. S. (1983) *Theor. Appl. Genet.* **67**, 25–33

Steinmetz, M., Stephan, D. & Lindahl, K. F. (1986) *Cell* **44**, 895–904

Thein, S. L., Jeffreys, A. J. & Blacklock, H. A. (1986) *Lancet* **ii**, 37

Thein, S. L., Jeffreys, A. J., Gooi, H. C., Lotter, F., Flint, J., O'Conner, N. T. J. & Wainscoat, J. S. (1987) *Br. J. Cancer* in the press

Tsui, L. C., Buchwald, M., Barker, D., Braman, J. C., Knowlton, R., Schumm, J. W., Eiberg, H., Mohr, J., Kennedy, D., Plavsic, N., Zsiga, M., Markiewicz, D., Akots, G., Brown, V., Helms, C., Gravius, T., Parker, C., Rediker, K. & Donis-Keller, H. (1985) *Science* **230**, 1054–1057

Uematsu, Y., Kiefer, H., Schulze, R., Fischer-Lindahl, K. & Steinmetz, M. (1986) *EMBO J.* **5**, 2123–2129

Wainscoat, J. S., Hill, A. V. S., Boyce, A. L., Flint, J., Hernandez, M., Thein, S. L., Old, J. M., Lynch, J. R., Falusi, A. G., Weatherall, D. J. & Clegg, J. B. (1986) *Nature* (*London*) **319**, 491–493

Wake, C. T., Long, E. O. & Mach, B. (1982) *Nature* **300**, 372–374

Weatherall, D. J. (1985) *The New Genetics and Clinical Practice*, 2nd edn., Oxford University Press, Oxford

Weller, P., Jeffreys, A. J., Wilson, V. & Blanchetot, A. (1984) *EMBO J.* **3**, 439–446

White, R., Leppert, M., Bishop, D. T., Barker, D., Berkowitz, J., Brown, C., Callahan, P., Holm, T. & Jerominski, L. (1985) *Nature* (*London*) **313**, 101–105

White, R., Nakamura, Y., Julier, C., Silva, A., O'Connell, P., Leppert, M., Lathrop, M. & Lalouel, J.- M. (1986) in *DNA Probes: Applications in Genetic and Infectious Disease and Cancer* (Lerman, L. S., ed.), pp. 43–47, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Wong, Z., Wilson, V., Jeffreys, A. J. & Thein, S. L. (1986) *Nucleic Acids Res.* **14**, 4605–4616

Wyman, A. & White, R. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 6754–6758

Wyman, A. R., Wolfe, L. B. & Botstein, D. (1985) *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2880–2884

Received 16 December 1986