Int J Leg Med (1991) 104:221–227

International Journal of
Legal Medicine

Population genetics of four hypervariable loci

Peter Gill, Susan Woodroffe, Joan E. Lygo, and Emma S. Millican

Central Research and Support Establishment, Home Office Forensic Science Service, Aldermaston, Reading, Berkshire RG7 4PN, UK

Received February 12, 1991 / Received in revised form April 16, 1991

Summary. Populations of white Caucasians, Afro-Caribbeans and Asians residing within the UK have been analysed at 4 different hypervariable loci. A computerised system was used to store and to analyse the data. Simulation experiments were carried out in order to determine whether there was any evidence for population stratification, which would lead to non-independence of allelic distributions.

Key words: DNA-Hypervariable loci – Matching window – Races – Population study

Zusammenfassung. Populationen weißer Europäer von "Afro-Caribbeans" und von Asiaten, welche in Großbritanien leben, wurden an 4 unterschiedlichen hypervariablen Loci untersucht. Ein computerisiertes System wurde benutzt, um die Daten zu lagern und zu analysieren. Simulationsexperimente wurden durchgeführt, um zu bestimmen, ob irgendein Beweis für Populations-Schichtung besteht; ein Befund, welcher auf fehlende Unabhängigkeit der Allelverteilung schließen lassen würde.

Schlüsselwörter: DNA – Hypervariable Loci – Matching window – Rassen – Populations-Studie

Introduction

Problems associated with the determination of molecular weight of continuous distributions of alleles from hypervariable loci have been discussed by Gill et al. (1990). Evett and Gill (1990) introduced the use of a 2.8% kB match guideline. This figure was based on a survey of 437 samples analysed in duplicate, giving 95266 comparisons, using YNH24. The guideline is not rigid; this series of experiments demonstrated that 2.1% of duplicates actually fell outside the 2.8% limit (i.e. bands were > 2.8% kB apart), whereas 0.78% of profiles from randomly chosen individuals were included. However, use of the Bayesian model proposed by Evett et al. (1990) circumvents the need for using a rigid criterion for deciding whether 2 bands match (where $P = 0$ or $P = 1$). The Bayesian model calculates a likelihood ratio which can be either positive or negative (i.e. P of a match is not binary but is somewhere between 0 and 1).

Recently, Lander (1989a, b) and Cohen (1990) have raised issues relating to the possible effect of linkage disequilibrium resulting from population stratification, where a population consists of mixed racial sub-groups which do not interbreed. The effect of population stratification has been examined by Evett and Gill (1990) using a model which consisted of an artificial population comprised of a mixture of Afro-Caribbeans and white Caucasians (the latter group had 2 kB added to each band in the database in order to accentuate the stratification). They showed that using this extreme example of a stratified population, the probability of observing chance associations of YNH24 was not changed.

Population stratification could result in associations between alleles from different loci, i.e. linkage disequilibrium which is not dependent upon physical linkage of loci (Lander 1989a, b). Surveys of the population structure of hypervariable loci are not yet extensive. Baird et al. (1986), Balazs et al. (1989) and Odelberg et al. (1989) found different allelic frequencies among different ethnic groups; Flint et al. (1989) have examined allelic distributions in Polynesian islanders. This paper includes a detailed analysis of the population variation in 3 different ethnic groups (white Caucasian, Afro-Caribbean and Asian). Between 200–300 people were analysed in each ethnic group using up to 4 different probes. Using the match guideline of 2.8%, a simple computer program was used to search for chance matches between randomly chosen individuals, where each had been analysed at 2 or more loci. This work was carried out in order to determine whether there was evidence for non-independence between loci examined in each of the different ethnic groups tested.

Offprint requests to: P. Gill

Table 1. Results of comparisons between different races

	YNH24	MS31	pMLJ14	MS43a	Total Pm
White Caucasian					
No. of HETS	381	236	182	253	
No. of HOMS	5	33	57	22	
No. of observations	272	214	239	213	
No. of comparisons	36856	22791	28441	22578	
Pm	0.011	0.012	0.0084	0.012	1.27^{-8}
Heterozygosity (%)	94.9	90.2	90.0	91.6	
Afro-Caribbean					
No. of HETS	177	85	47	92	
No. of HOMS	1	1	18	3	
No. of observations	224	196	200	222	
No. of comparisons	24976	19110	19900	24531	
Pm	0.006	0.004	0.002	0.003	4.06^{-10}
Heterozygosity (%)	96.9	95.4	89.5	94.6	
Asian					
No. of HETS	289	218	73	142	
No. of HOMS	17	7	5	32	
No. of observations	238	224	220	214	
No. of comparisons	28203	24976	24090	22791	
Pm	0.011	0.009	0.003	0.008	2.42^{-9}
Heterozygosity (%)	93.7	93.8	94.1	87.4	

Between individual comparisons of each sample in the database were made; the number of heterozygote (HETS) and homozygote (HOMS) chance associations (within a 2.8% window) were recorded. The total number of comparisons made is (n*n-1)/2 where n is the sample size of the population. The expected combined chance association of 4 probes is obtained by multiplication of Pm (match probability) for each probe

Materials and methods

The populations. The population results of white Caucasians were based on blood samples collected during the course of casework from forensic science laboratories in England and Wales; the Afro-Caribbean population originated from Manchester and the Asian population originated from Oxford and Edgbaston, England.

Electrophoretic system and probes used. Aliquots of DNA (2–3 µg) were extracted following the procedure described by Gill et al. (1987). The DNA was digested with approximately 30X excess HinfI (Boehringer) and run overnight on 20 × 25 cm agarose gels (0.04 *M* Tris; 0.02 *M* Na Acetate; 0.2 m*M* EDTA). Gels were depurinated, denatured and Southern blotted onto nylon membranes (Amersham Hybond) following the method of Gill et al. (1987). Membranes were hybridised with oligolabelled probes YNH24, pCRE1.2 (a sub-clone of pMLJ14 described by Nakamura et al. 1987), MS31 and MS43a (Wong et al. 1987). The protocol has been described by Gill et al. (1990).

Analysis of band positions and frequency determination. The method of band analysis used was as described by Gill et al. (1990) except that frequencies were calculated using a ± 2.8% sliding window fit. Sizes of band fragments were determined using the method of Elder and Southern (1987) and profiles were sized by reference to 3 Lambda ladder markers (Amersham Catalogue No. NK8668). In addition, each plate contained at least 1 genomic control.

Analysis of probabilities of chance association. To determine the effect of window size on the probability of chance association, data were analysed as follows:

(i) Only samples which had been analysed using at least 2 probes were included.

(ii) Each race code was taken separately. Taking each probe in turn, each sample was compared with every remaining sample in the database so that the total number of comparisons = n*(n-1)/2 where n = number of individuals analysed for each probe. The numbers of homoyzgote and heterozygote matches were recorded and are shown in Table 1. The experiment was repeated several times, increasing the window in steps from 2.8% to 11.2%.

(iii) For each sample and taking each race code sequentially, band positions for every pair of probes (i.e. a total of 6 different combinations; Table 2) were compared with each remaining sample in the database. This experiment was carried out using a 2.8% window only.

(iv) Finally, each profile was compared with all other profiles in the database in order to determine whether matches using 3 and 4 probes could be observed.

Results and discussion

Population databases

Population frequency histograms for white Caucasians, Afro-Caribbeans and Asians are illustrated in Figs. 1–4. All ethnic groups showed marked differences from each other.

Characteristics of YNH24

A major peak was found in white Caucasians centred at 2750 bp. Afro-Caribbeans at 3000 bp and Asians at 2850 bp. The range of alleles observed was between 1700 bp (Asian) and 8400 bp (white Caucasian).

Table 2. Comparison of pairs of probes for each database

Probe 1	Probe 2	Number of comparisons		Expected	Observed
		$(n*n\text{-}1)/2$	n		
White Caucasian database					
YNH24	MS31	15225	175	1.88	2
YNH24	pMLJ14	13695	166	1.21	1
YNH24	MS43a	16110	180	2.06	1
MS31	pMLJ14	7626	124	0.76	0
MS31	MS43a	6328	113	0.91	0
pMLJ14	MS43a	6903	118	0.71	1
Total		65887		7.51	5
Proportion				1.141^{-4}	7.589^{-5}
Afro-Caribbean database					
YNH24	MS31	17578	188	0.56	0
YNH24	pMLJ14	16290	181	0.38	1
YNH24	MS43a	16836	184	0.46	1
MS31	pMLJ14	13530	165	0.20	1
MS31	MS43a	12561	159	0.22	0
pMLJ14	MS43a	14706	172	0.19	0
Total		91501		2.01	3
Proportion				2.198^{-5}	3.279^{-5}
Asian					
YNH24	MS31	21945	210	2.14	2
YNH24	pMLJ14	22791	214	0.80	2
YNH24	MS43a	19701	199	1.63	0
MS31	pMLJ14	19701	199	0.57	0
MS31	MS43a	19110	196	1.31	2
pMLJ14	MS43a	17578	188	0.43	0
Total		120826		6.90	6
Proportion				5.712^{-5}	4.966^{-5}

The legend is the same as for Table 1 except that between individual comparisons of pairs of probes were made. Expected frequencies were calculated: $(\mathrm{Pm}(1) \times \mathrm{Pm}(2) \times (n*n\text{-}1)/2)$ where Pm was taken from Table 1 and $(n*n\text{-}1)/2$ from Table 2

Characteristics of pMLJ14

A major peak existed in both white Caucasians and Asians at 1560 bp, the frequencies of the remaining alleles were below 0.04. In Afro-Caribbeans, alleles were more common (up to 0.08) between the 2–3.7 kB range. The overall distribution of alleles in pMLJ14 ranged between 1–21 kB, all alleles above 7 kB were rare (< 0.01).

Characteristic of MS31

All 3 databases peaked at similar positions (6700 bp, white Caucasian; 6760 bp, Afro-Caribbean; 6900 bp, Asian). Alleles ranged between 3–14 kB; Afro-Caribbeans had additional low molecular weight alleles down to 1.6 kB.

Characteristics of MS43a

Generally, alleles ranged between 3.4–13.6 kB in white Caucasians, with rare alleles being found up to 20 kB in Asians and Afro-Caribbeans.

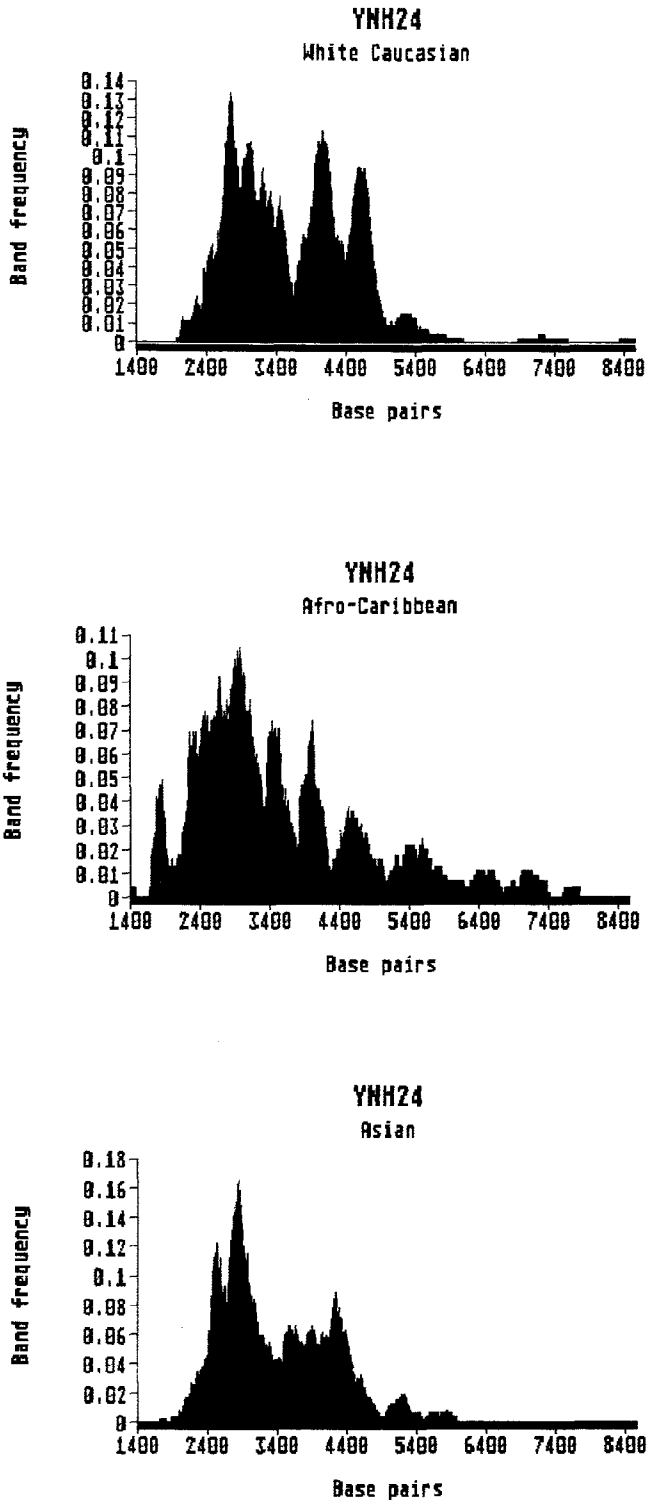


Fig. 1. Frequency distribution of probe YNH24 using a ±2.8% sliding window fit. The sliding window fit is described by Gill et al. (1990). Essentially, it consists of a "bin" which moves at 5 bp intervals. The histogram is a compilation of all possible "bin" frequencies

Biological significance of observed allelic distributions

The mutation rate of hypervariable loci is significantly greater than for other loci in the human genome (Jeffreys et al. 1988). Frequency distributions described in

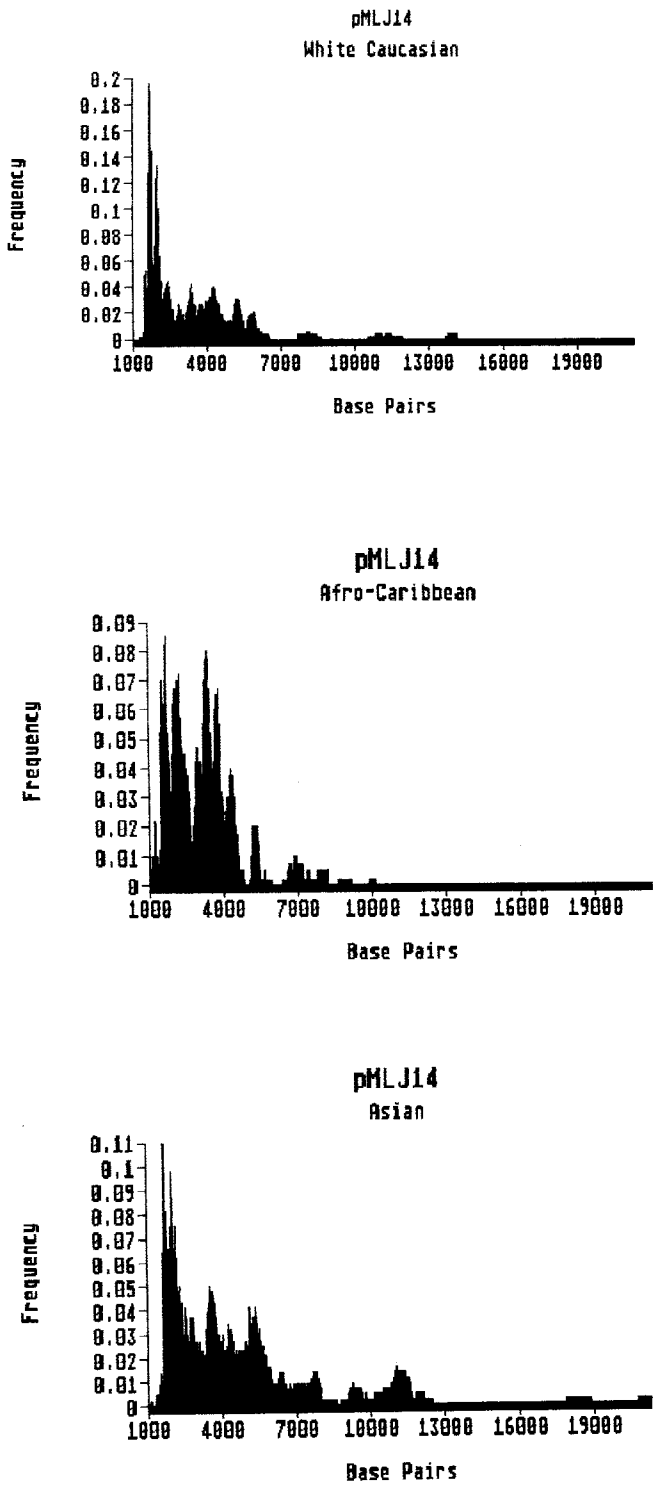


Fig. 2. Frequency distribution of probe pMLJ14 using a ± 2.8% sliding window fit

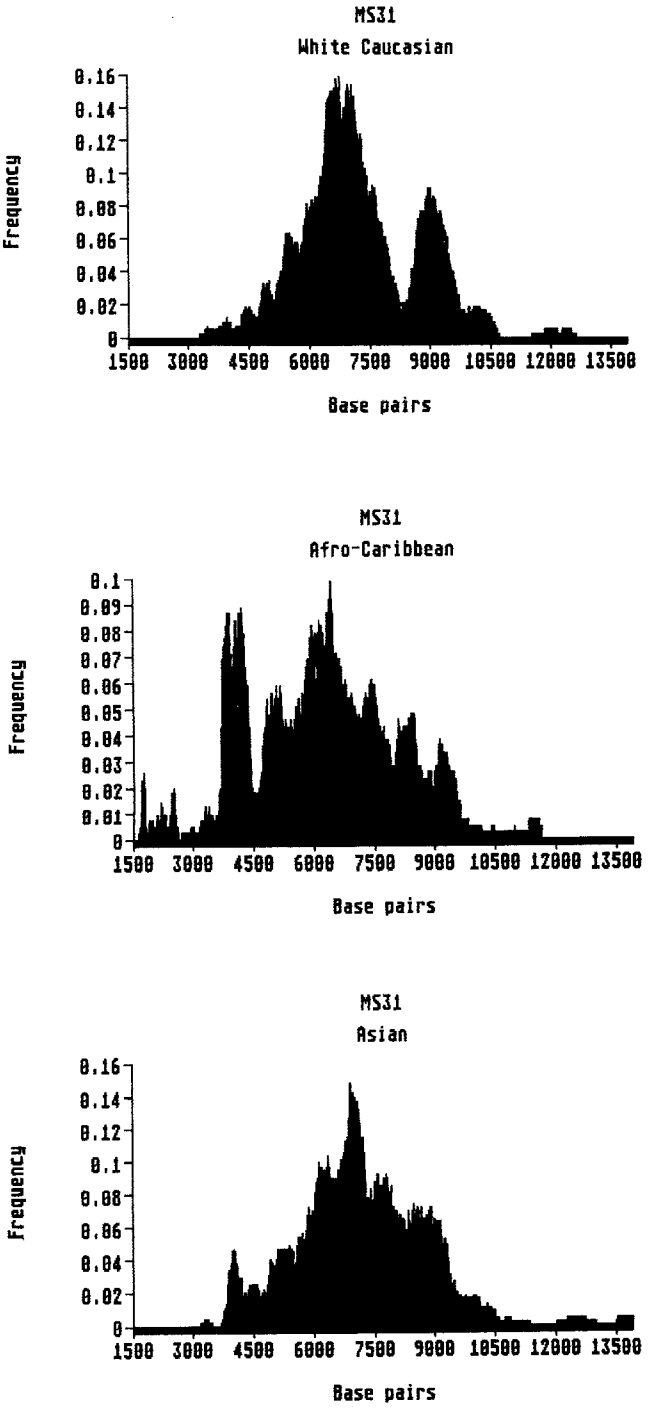
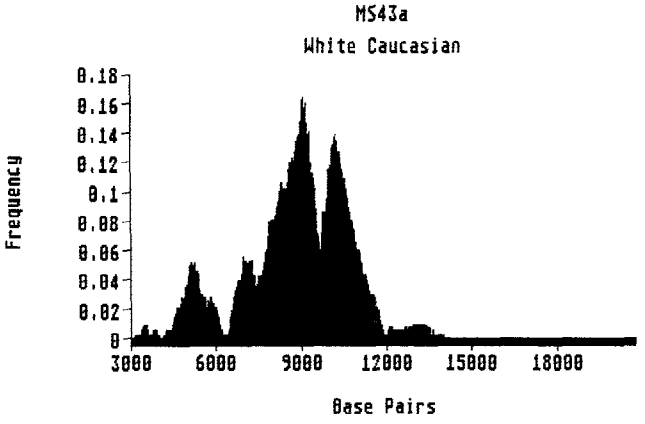
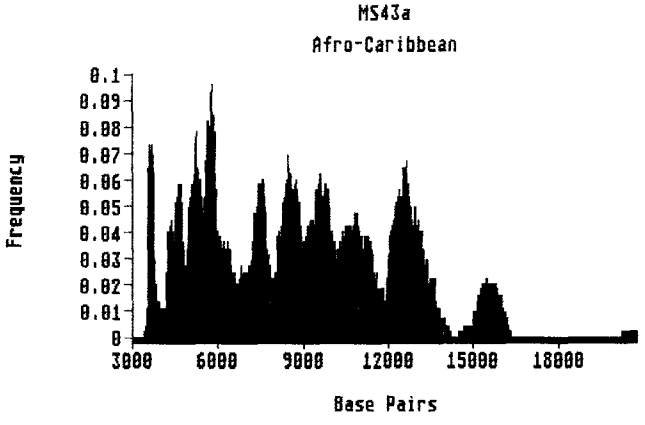


Fig. 3. Frequency distribution of probe MS31 using a ± 2.8% sliding window fit

this paper all become rare towards the high molecular weight end of the distribution. This trend is well illustrated by YNH24 Afro-Caribbeans where the frequency progressively declines from 2.6 kB down to 7.4 kB. Alleles detected by pMLJ14 also become progressively rare above 5 kB in all races. Distributions of MS31 and MS43a are both shifted towards the higher molecular weights, with the latter locus showing a high molecular weight peak at 9 kB for white Caucasians and Asians. Comparison with the Afro-Caribbean population shows a progressive decline in frequency of bands greater than 5.4 kB. If the ancestral population of the human race has originated from Africa (Cann et al. 1987) then it would be expected that greater heterogeneity would be found in African populations since the two effects of recurrent mutation and genetic drift would have operated over a longer period of time. It is probable that high molecular





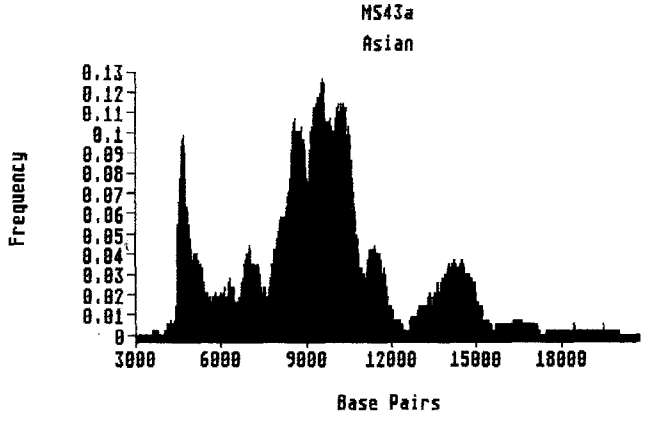


Fig. 4. Frequency distribution of probe MS43a using a ± 2.8% sliding window fit

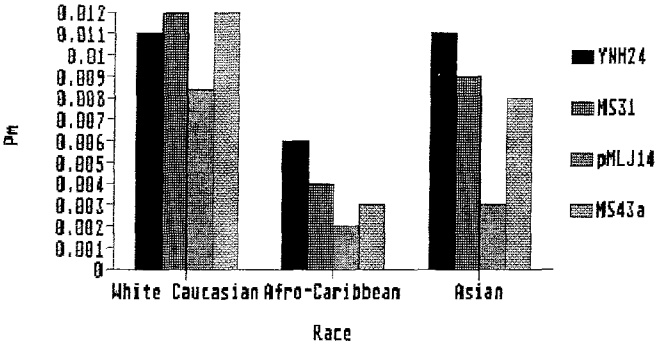


Fig. 5. Match probabilities of 4 probes for 3 different ethnic groups using a 2.8% window guideline

Table 3. Probability of chance association for 4 probes according to window size

Window size (%)	Observed Pm using 4 probes		
	White	Afro-Caribbean	Asian
2.8	1.27^{-8}	4.06^{-10}	2.42^{-9}
5.6	1.27^{-6}	4.52^{-8}	3.92^{-7}
8.4	1.81^{-5}	9.71^{-7}	6.66^{-6}
11.2	1.83^{-4}	7.69^{-6}	4.42^{-5}

The effect of window size on match probability using 4 probes

weight alleles are more likely to mutate than low molecular weight alleles. This effect would account for their rarity in the high molecular weight regions and is supported by the observation that the frequency of bands using multi-locus probes decreases as the molecular weight increases (Jeffreys et al. 1985; Gill et al. 1987).

Independence of alleles between hypervariable loci

For each population database, the number of matches between heterozygotes and homozygotes was recorded using the 2.8% window guideline for each ethnic group (Table 1, Fig. 5). Probe pMLJ14 gave the lowest probability of chance association for each race. Afro-caribbeans were more heterogeneous than both white Caucasian and Asian populations for 3 of the 4 probes tested (Table 1).

Table 2 shows comparisons between pairs of probes using a 2.8% window guideline. The probability of chance association using 2 probes was approximately 10^{-5} and was no different to that expected. This is a strong indication that the populations analyses are independent at the 4 loci examined.

No matches were observed when 3 or 4 probes were compared for each sample. The probability of chance association using 4 probes was calculated as 10^{-8} and 10^{-10} (Table 3) using a 2.8% window.

Population stratification

Jeffreys et al. (1991) pointed out that the effect of recurrent mutation will counteract the effect of genetic drift, and will therefore prevent the occurrence or tendency for alleles to become fixed, even in highly inbred populations. Furthermore, observed mutation rates recorded by Jeffreys et al. (1985, 1988, 1991) and Armour et al. (1989) are almost certainly underestimates, since many mutations could occur which would never be detected because of the limited resolving power of the electropho-

retic system utilised or because changes can occur within the tandem repeat which does not produce a change in the size of the fragment. Jeffreys et al. (1990) showed that even alleles of identical length can be different and have arisen by convergent evolution. It follows that if alleles within a bin are different, then this requires a very high mutation rate to new length alleles. It is the high mutation rate alone which prevents any given allele from reaching a significant population frequency.

Relationship between probability of chance association and heterozygosity

Given a continuous allelic distribution in hypervariable loci it follows that observed heterozygosity estimates (Table 1) are underestimated because they are partly dependent upon the resolving power of the electrophoretic system utilised. An apparent homozygote could be 2 bands differing in size by only a few base pairs. Alternatively, observations of homozygotes are likely to be greater than expected because of the loss of low molecular weight alleles which run off the end of the gel and the possibility of the occurrence of null alleles. It is not possible to count the number of alleles in a hypervariable system; clear peaks in the allele frequency distribution usually represent collections of alleles which on internal mapping are clearly related in origin (Alec Jeffreys, Pers. Comm.), i.e. the 2 kb peak for YNH24 Afro-caribbeans might be a collection of very few different alleles. Hence measurements of Hardy Weinberg equilibrium are problematical.

Measurements of observed heterozygosity (Ht) at the loci examined do not give a good indication of population heterogeneity. This is best achieved by determination of chance matches by simulation. The lowest probability of chance association (Pm) was found in pMLJ14 Afro-Caribbean (Pm = 0.003; Ht = 89.5%); the greatest was in MS43a (Pm = 0.012; Ht = 91.6%).

Searching databases

It may be a future requirement in forensic laboratories to search databases for profiles which have been generated using different electrophoretic systems in different laboratories. Different electrophoretic systems and protocols result in difficulties in obtaining results which are directly comparable between laboratories (Rose and Keith 1989). This could result in an interlaboratory error of greater than 2.8% (where protocols are different). Simulations were carried out in order to determine probabilities of chance association using 4 probes and these are shown in Table 3. Even the use of a window of 11.2% resulted in a probability of chance association of 10^{-5}. This indicated that a large window could be utilised for searching purposes provided that at least 4 different single locus probes are used. To carry out a meaningful statistical comparison on results from different laboratories using different protocols would require a demonstration of compatibility of results, coupled with suitable quality controls to ensure that fragment sizes were within designated limits.

If it was suspected that 2 samples match, the alternative approach would be to re-run samples from different laboratories on the same electrophoretic plate; direct statistical comparisons could then be made.

Conclusion

Data have been analysed for 4 different probes using 3 different ethnic groups. Each race has a different distribution of alleles for each probe. Examination of data for band matches using computer simulation demonstrated independence of bands (and therefore lack of linkage disequilibrium) in all populations analysed. The greatest heterogeneity was found in Afro-Caribbean populations and this could be associated with the possible ancestry of the human race.

Acknowledgement. The authors are grateful for Professor Alec Jeffreys for making valuable comments on the manuscript.

References

Armour JAL, Patel I, Thein SL, Fey MF, Jeffreys AJ (1989) Analysis of somatic mutations at human minisatellite loci in tumours and cell lines. Genomics 4:328–334

Baird M, Balazs I, Giusti A, Miyazaki L, Nicholas L, Wexler K, Kanter E, Glassberg J, Allen F, Rubinstein P, Sussman L (1986) Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. Am J Hum Genet 39:489–501

Balazs I, Baird M, Clyne M, Meade E (1989) Human population genetic studies of five hypervariable DNA loci. Am J Hum Genet 44:182–190

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cohen JE (1990) DNA fingerprinting for forensic identification: potential effects on data interpreatation of sub-population heterogeneity and band number variability. Am J Hum Genet 46:358–368

Elder JK, Southern EM (1987) Computer aided analysis of one-dimensional restriction fragment gels. In: Bishop MJ, Rawlings CJ (eds) Nucleic acid and protein sequence analysis. IRL Press, Oxford, pp 165–172

Evett IW, Gill P (1991) A discussion of the robustness of methods for assessing the evidential value of DNA single locus profiles in crime investigations. Electrophoresis 12:226–230

Evett IW, Werrett DJ, Pinchin R, Gill P (1990) Bayesian analysis of single locus DNA profiles. In: The International Symposium on Human Identification 1989. Promega Corporation, pp 77–101

Flint J, Boyce AJ, Martinson JJ, Clegg JB (1989) Population bottlenecks in Polynesia revealed by minisatellites. Hum Genet 83:257–263

Gill P, Lygo JE, Fowler SJ, Werrett DJ (1987) An evaluation of DNA fingerprinting for forensic purposes. Electrophoresis 8: 35–38

Gill P,Sullivan K, Werrett DJ (1990) The analysis of hypervariable DNA profiles: problems associated with the objective determination of the probability of a match. Hum Genet 85:75–79

Jeffreys AJ, Wilson V, Thein SL (1985) Individual specific "fingerprints" of human DNA. Nature 316:76–79

Jeffreys AJ, Royle NJ, Wilson V, Wong Z (1988) Spontaneous mutation rates to new length alleles at tandem repetitive hypervariable loci in human DNA. Nature 332:278–281

Jeffreys AJ, Turner M, Debenham P (1991) The efficiency of multi-locus DNA fingerprint probes for individualisation and

establishment of family relationships, determined from extensive casework. Am J Hum Genet 48:824–840

Jeffreys AJ, Neumann R, Wilson V (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell 60:473–485

Lander ES (1989a) DNA fingerprinting on trial. Nature 339:501–505

Lander ES (1989b) Population genetic considerations in the forensic use of DNA typing. In: Ballantyne J, Sensabaugh G, Witkowski J (eds) Banbury Report 32: DNA Technology and Forensic Science. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp 143–156

Odelberg SJ, Plaetke R, Eldridge JR, Ballard L, O'Connell P, Nakamura Y, Leppert M, Lalouel M, White R (1989) Characterisation of eight VNTR loci by agarose gel electrophoresis. Genomics 5:915–924

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimot E, Hoff M, Kumlin E, White R (1987) Variable number of tandem repeats (markers) for human gene mapping. Science 235:1616–1622

Rose SD, Keith TP (1989) Standardization of systems: Essential or desirable. In: Ballantyne J, Sensabaugh G, Witkowski J (eds) DNA technology and Forensic Science. Banbury Report, 32. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp 319–326

Wong Z, Wilson V, Patel I, Povey S, Jeffreys AJ (1987) Characterisation of a panel of highly variable minisatellites clone from human DNA. Ann Hum Genet 51:269–288