Int J Leg Med (1991) 104:267–272

International Journal of
Legal Medicine

DNA single locus profiles: tests for the robustness of statistical procedures within the context of forensic science

I.W. Evett and R. Pinchin

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

Received May 16, 1991 / Received in revised form July 8, 1991

Summary. We describe a series of experiments, carried out on data from DNA profiles, which have been designed specifically to test the validity of the statistical procedures currently used in the Home Office Forensic Science Service. The tests address issues which have been the subject of topical debate, in particular those of representativeness and band independence. The results confirm the confidence which had already been placed in the established procedures. We recommend that all practitioners in the forensic field should carry out similar testing on their own data collections.

Key words: DNA – RFLP – Statistics

Zusammenfassung. Wir betrieben eine Serie von Experimenten, die an Daten von DNA-Profilen durchgeführt wurden. Diese Experimente sind spezifisch mit dem Ziel erstellt worden, die Validität der statistischen Verfahren zu testen, welche derzeit vom „Home Office Forensic Science Service" benutzt werden. Die Tests konzentrieren sich auf solche Ziele, welche Gegenstand der aktuellen Diskussion sind, insbesondere Fragen der Repräsentativität und der Unabhängigkeit der Banden. Die Resultate bestätigen die Zuverlässigkeit, welche bereits in den etablierten Methoden vorhanden war. Wir empfehlen, daß alle Praktiker im forensischen Feld ähnliche Testmethoden an ihren eigenen Datenkollektiven durchführen sollten.

Schlüsselwörter: DNA – RFLP – Statistik

Introduction

The statistical methods for interpreting data from hypervariable loci have attracted considerable interest in forensic science; issues such as sample size, representativeness and population stratification have been debated, for example by Lander [1], and by others. Such debates tend to be from the perspective of population genetics and the arguments relating to the wisdom of conventional assumptions – such as independence between alleles and loci – have their basis in sound theoretical considerations. For the forensic scientist and the criminal justice system it is necessary to establish whether theoretically valid effects have any practical impact on operational procedures. To do this effectively, it is necessary to devise and apply tests which are appropriate to that particular context.

Evett and Gill [2] have described investigations of the consequences of population stratification and of the use of small databases. Gill et al. [3] have described the databases and procedures in current use in the Home Office Forensic Science Service (FSS); they have also presented estimates of the discriminating power for four probes in three different racial groups. This paper describes investigations into representativeness and into the validity of the assumption of statistical independence between the sizes of the two bands in the profile from one locus.

Materials and methods

The data collections are those described by Gill et al. [3]. To reflect the demands of casework in England and Wales, data were collected from three racial groups: Caucasian, Afro-Caribbeans and Asians from the Indian sub-continent (here referred to as Asians for brevity). Four probes were used: YNH24, pMLJ14, MS31 and MS43A. The numbers of individuals in the twelve databases are shown in Table 1.

For casework, the FSS uses a guideline of 2.8% for matching and relative frequencies are estimated from 5.6% sliding windows,

Table 1. Numbers of individuals probed by the four probes for each of the three ethnic groups

	YNH24	pMLJ14	MS31	MS43A
Caucasian	272	239	214	213
Afro-Caribbean	224	200	196	222
Asian	238	220	224	214

or bins, on the band weight distributions. Because of sampling effects, a default minimum frequency of 1% is used. This methodology formed the basis of the experimentation.

Results and discussion

The first two experiments were designed to simulate certain casework conditions and computer programs were used to carry out many thousands of match/binning comparisons using our standard procedures. In the event of a match, the binning stage leads to an estimate of the relative frequency of the relevant combination of band weights. In the context of a case where a stain at the crime scene is matched to a sample from a suspect, the likelihood ratio which measures the strength of the evidence is the inverse of the relative frequency of the observed phenotype. To illustrate the results of experiments such as those described in this paper, it is useful to assign verbal predicates to logarithmic ranges of the likelihood ratio. The scale, which is the same as that used previously by Evett and Gill [2], is shown in Table 2.

Such a scale is inevitably arbitrary and it is not claimed that Table 2 is a prescriptive formula for casework. However, experience with presentations and training exercises suggests that the convention is a valuable aid to comprehension.

Most of the profiles in the databases consist of 2 bands. The minority which contain only one band arise either because an individual is a true homozygote or has 2 bands which are too close in weight to be resolved or because there is a second allele which is too low in molecular weight to be detected by the system in use. In the first 2 experiments the one banded profiles are excluded for the sake of simplicity. The third experiment employs all of the data.

Experiment 1: regional variation

Gill et al. [3] have explained that the Caucasian databases were created from blood samples collected during routine casework at the operational laboratories of the FSS. There are 6 such laboratories, each serving a different region of the country. It is desirable to establish the validity of a laboratory in Lancashire, say, using a data collection which includes people from, for example, the South East of England. A conventional approach to addressing such issues may be to carry out comparisons between the band weight distributions for the 6 laboratories using chi-squared or Kolmogorov-Smirnoff tests of fit. However, significance testing on such a scale – there are fifteen different inter-laboratory comparisons – is unimaginative and not likely to prove helpful for 2 reasons. First, goodness-of-fit tests on databases of the sizes under consideration are not powerful. Second, even if such a test led to a result which might be argued to be statistically significant, it would not necessarily mean that it was *practically* significant.

Table 2. An illustrative verbal convention for evidence strengths in cases where there is a match

Likelihood ratio in the range		Evidence strength
1	30	Weak
30	100	Fair
100	300	Good
300	1000	Strong
Greater than	1000	Very strong

For these reasons the following experiment, which is of a kind which we call a "crossed-database" experiment, was designed to test questions which are directly relevant to the operational context. The experiment involves the visualisation of an artificial case in which a scientist in one laboratory carries out a comparison between a crime sample and a sample from an innocent suspect, having access to data, not from his *own* laboratory but from the other laboratories. For simplicity, the experiment was restricted to 2-banded profiles.

For each laboratory in turn two databases were set up: the test database consisting of the profiles from samples collected at that laboratory, and a reference database consisting of the profiles from samples collected at all of the other laboratories. Typically the test database was in the size range 40–60 profiles and the reference database in the range 150–200 profiles.

For a test database of size n, all of the $n(n-1)/2$ between person comparisons were carried out using 2.8% as a match criterion. In the event of both bands matching the evidential significance was assessed from the reference database, using 5.6% windows and a default minimum band frequency of 1%. This was repeated for all 4 probes, for all of the laboratories which contributed sufficient data. The total number of matches from each run was divided by the total number of comparisons to give an estimate of the probability of a match (PM) between 2 different individuals. In a similar manner, estimates were made of the probabilities of concluding each of the various degrees of evidence strengths. The results are presented in Table 3a–d.

The estimates for PM are, of course, comparable to those reported by Gill et al. [3], though not precisely so because that analysis included one-banded profiles. Variation in PM between laboratories can be expected because of sampling effects and simple chi-squared tests based on the numbers of matches were not significant for any of the 4 probes. With regard to evidence strength, bearing in mind that "good evidence" or stronger corresponds to quoting a match and a frequency of 1 in 100 or smaller, the probability of this mistakenly happening was on no occasion estimated as greater than 1 in 240 – on most occasions it was far less. This finding reflects the conservative nature of our procedures, even with regard to regional variation.

Experiment 2: racial variation

The availability of data from 3 different racial groups enables the forensic scientist to choose the most appropriate data given the circumstances of each particular crime investigation. There are occasions, however, when the

Table 3. Experiment 1: crossed-database experiment for regional variation

	YNH24					MS31				
Laboratory	1	2	4	5	6	1	2	4	5	6
Test file size	41	38	39	54	86	40	46	13	48	48
Number of comparisons	820	703	741	1431	3655	780	1035	78	1128	1035
Probability of match	0.0098	0.0085	0.012	0.0091	0.010	0.01538	0.00966	0.01282	0.01507	0.00870
Probabilities of:										
Weak evidence	0	0	0	0	0	0	0.00097	0	0.00089	0
Fair	0.0061	0.0085	0.0081	0.0080	0.0063	0.0141	0.00676	0.01282	0.01064	0.0058
Good	0.0037	0	0.0041	0.0007	0.0038	0	0.00097	0	0.00355	0.00193
Strong	0	0	0	0	0	0.00128	0.00097	0	0	0
Very strong	0	0	0	0.0007	0	0	0	0	0	0.00097
	pMLJ14					MS43A				
Laboratory	1	2	3	4	5	1	2	4	6	
Test file size	38	45	53	30	49	40	48	26	81	
Number of comparisons	703	990	1378	435	1176	780	1128	325	3240	
Probability of match	0.0128	0.00606	0.00798	0.0023	0.00765	0.01282	0.0133	0.01231	0.01265	
Probabilities of:										
Weak evidence	0.00142	0	0	0	0	0.00128	0	0.00308	0.00309	
Fair	0.00996	0.00101	0.00218	0	0.00255	0.00769	0.00798	0.00308	0.00494	
Good	0.00142	0.00404	0.00581	0	0.00255	0.00256	0.00532	0.00308	0.00494	
Strong	0	0.00101	0	0.0023	0.00255	0.00128	0	0	0.00031	
Very strong	0	0	0	0	0	0	0	0	0	

circumstances do not permit a clear choice and it may happen, for one reason or another, that the wrong database is used. It is clear from the work of Gill et al. [3] that there are marked differences in distributions between the racial groups. Standard significance tests are quite superfluous in this context: on the other hand, crossed-database experiments enable the consequences of an incorrect choice of database in the operational environment to be investigated. Attention was concentrated on Caucasians and Afro-Caribbeans, because it can be seen from Gill et al. [3] that the Asian distributions tended to fall between these extremes. As in Experiment 1, data on 2-banded profiles alone was used, for the sake of simplicity.

For each probe in turn, all of the between-person comparisons in the Caucasian database were carried out; on each occasion that a match occurred, the Afro-Caribbean database was used for reference. In a similar way, all Afro-Caribbean comparisons were carried out using Caucasian files for reference.

The verbal predicates were again used for describing evidence strength and the results are presented graphically. Figure 1a, for example, shows, as a dotted line plot, the outcome of the experiment using Caucasian YNH24 profiles as the test database. This is a plot of aggregated probabilities viz: weak evidence *or better,* fair evidence *or better,* and so on. For comparison, the between-person comparisons were also run using the correct (i.e. Caucasian) database used for reference and the results are shown on the same graph as an unbroken line. Figure 1b–d show the results for the other probes and Fig. 1e–h show the comparable runs for Afro-Caribbeans using Caucasian databases.

As would be expected, the use of the wrong database can change substantially the magnitudes of the probabilities of incorrect evidence assessments. In extreme cases, for example, the probability of incorrectly concluding "good" evidence or better increases almost by a factor of 10. However, to keep the effects in perspective it is necessary to bear in mind the overall magnitude of these probabilities. Thus, in one of the worst cases, that of pMLJ14 Caucasians, the probability of incorrectly concluding good evidence or better is about 1 in 75. For this to occur the case circumstances would have to be such that: only one probe is used; the suspect is Caucasian; the true criminal is another Caucasian; and the scientist uses an Afro-Caribbean database. It is difficult to imagine how such a set of circumstances would prevail but it is useful to gain a view of the effects because real casework effects can be expected to be less severe.

Experiment 3: independence

Lander [1] pointed out the weaknesses of conventional tests for Hardy-Weinberg equilibrium in the forensic use of hypervariable loci. Tests based on Wahlund's principle are problematic because one-banded profiles are not necessarily from homozygous individuals. Once again it is necessary to establish a test which is sensitive to weaknesses in the procedures which are used in the relevant operational context.

A file of data from single locus profiles can be organised as two columns of numbers; each line corresponding to one profile, with one-banded profiles being represented by the same number entered twice. If the data

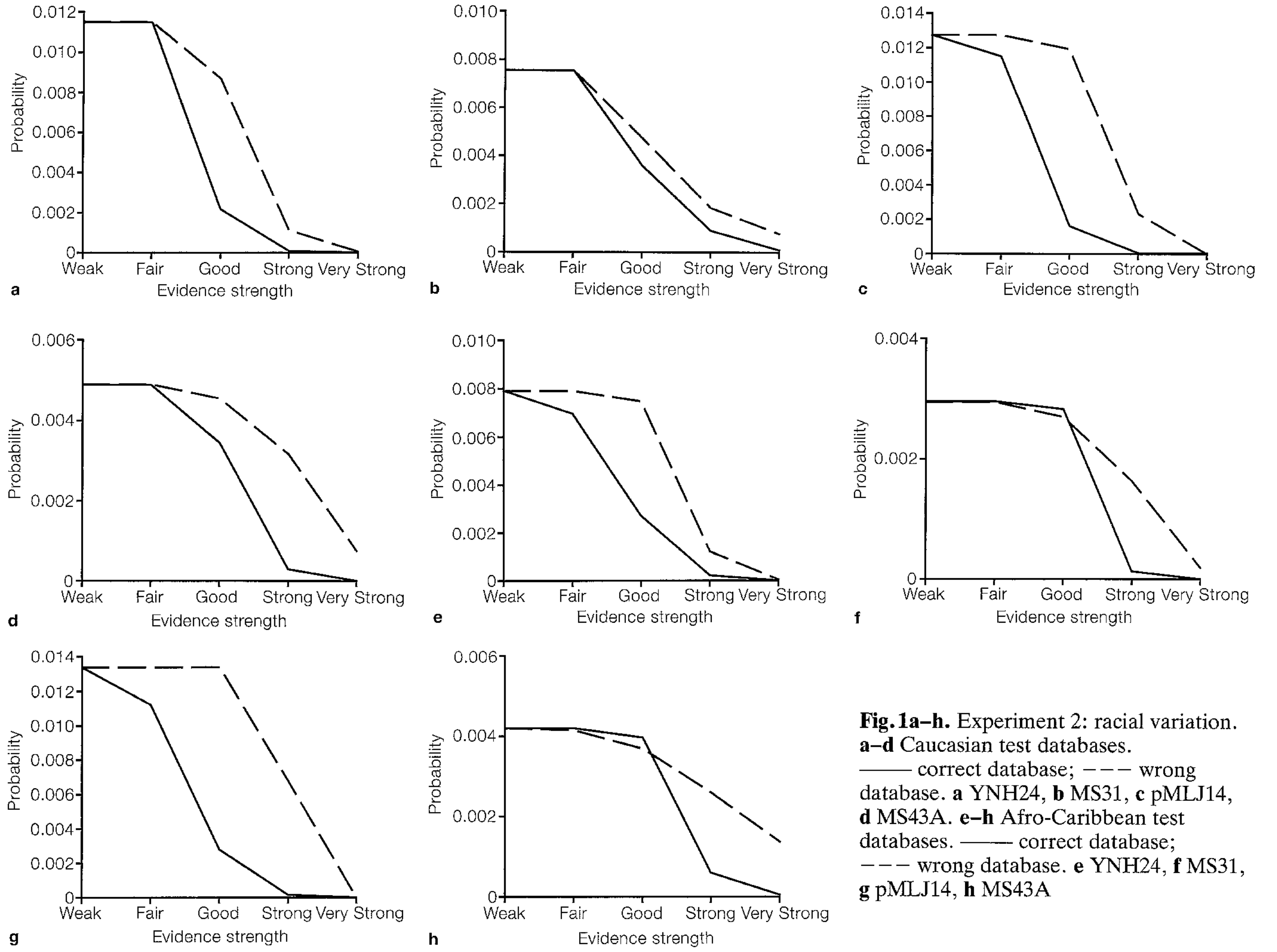

Fig. 1a–h. Experiment 2: racial variation. **a–d** Caucasian test databases. —— correct database; – – – wrong database. **a** YNH24, **b** MS31, **c** pMLJ14, **d** MS43A. **e–h** Afro-Caribbean test databases. —— correct database; – – – wrong database. **e** YNH24, **f** MS31, **g** pMLJ14, **h** MS43A

Table 4. Experiment 3: contingency table for the independence analysis for Caucasian YNH24 profiles

		Outcome of band 1 comparisons		Totals
		Match	Non-match	
Outcome of band 2 comparisons	Match	203	2590	2793
	Non-match	2399	31936	34335
	Totals	2602	34526	37128

Number of profiles: 273

are organised so that, for each entry, the smaller weight band precedes the larger weight band then a plot of the file would produce a graph similar to Fig. 1 in the paper by Devlin et al. [4], with the one-banded profiles falling on the 45° line through the origin. Starting with the file configured thus, in 50% of the rows selected at random the larger weight bands were placed first and the smaller weight bands second; this effectively represented our state of ignorance about which band was paternal and which band was maternal.

A program was then used to carry out the following: all intercomparisons for the first column of bands; all intercomparisons for the second column of bands; and all intercomparisons for pairs of bands taken together. The numbers of band 1 matches, the number of band 2 matches and the number of occasions when both bands matched were all counted. The results of such a run can be presented as a two by two contingency table. That for YNH24 Caucasian profiles is shown, for illustration, as Table 4.

For 273 profiles there were: 37128 intercomparisons; 2602 band 1 matches; 2793 band 2 matches; and 203 occasions when both bands matched. From these figures, all cells of the table can be completed. If there is independence between the interitance of the two bands in a profile then the expected number of occasions on which both bands would match is simply 2602 × 2793/37128 = 196. A simple chi-squared test (using Yates' correction) was used to look for possible departures from independence. The results for all four probes and the 3 racial groups are shown in Table 5.

For 10 out of 12 of the analyses the test statistic was very small. The remaining 2 tests were, however, significant at the 5% level and so were made the subject of further analysis.

The explanation for the pMLJ14 Caucasian result was quite straightforward. The unexpectedly large number

Table 5. Experiment 4

Probe	Ethnic group	Number of profiles	Number of comparisons	Observed number of matches	Expected number of matches	Chi^2 statistic
YNH24	Caucasian	272	36,856	203	197	0.06
	Afro-Caribbean	224	24,976	93	73	5.25
	Asian	238	28,203	176	171	0.05
MS31	Caucasian	214	22,791	159	153	0.07
	Afro-Caribbean	196	19,110	49	45	0.20
	Asian	244	24,976	118	128	0.99
pMLJ14	Caucasian	239	28,441	166	141	4.22
	Afro-Caribbean	200	19,900	45	39	0.73
	Asian	220	24,090	43	46	0.11
MS43A	Caucasian	213	22,578	152	148	0.01
	Afro-Caribbean	222	24,531	54	55	0.01
	Asian	214	22,791	94	99	0.34

of matches on both bands is caused by a disproportionally large number of one-banded profiles for this probe. Many of these are clearly the result of the presence of low molecular weight alleles: we estimate that at least half of the one banded profiles arise from this effect. Accordingly, using random numbers one half of the one-banded profiles were removed from the file and the experiment re-run: on this occasion the test statistic fell close to zero so giving no cause to doubt the assumption of independence. It is interesting to note that the test statistic was very low for Afro-Caribbean pMLJ14 data in spite of a comparable proportion of one-banded profiles. When 50% of those were removed the number of occasions of both bands matching was considerably *less* than the number expected from the independence assumption.

The investigation of the YNH24 Afro-Caribbean data was by no means straightforward there being only a small number of one-banded profiles. Detailed inspection of the output from the between person comparisons suggested that there might be some clustering of data points in the range 2.2–3.2 kb. It is conceivable that this may indicate some degree of population substructuring but our main concern was to gain an appraisal of its practical impact. First, we found that for each of the 43 profiles in this region which matched at least one other profile the independence assumption resulted, on average, in a relative frequency estimate which was 70% of the estimate gained without making the assumption. On average, this would mean quoting a frequency of 1 in 84 instead of 1 in 50. When this is viewed in the context of Fig. 1e, we consider that it has negligible practical significance. Next, each of these 43 samples had been profiled using at least one other probe and after using only one additional probe *none* of the intercomparisons resulted in a match. Bearing in mind that operational procedures currently are based on 4 probes this effect, such as it is, has no more than academic interest.

One further point is worth making. When all 903 intercomparisons were made on this file of 43 (using the main database for reference) there were 66 matches. However, the Bayesian likelihood ratio calculation as described by Berry et al. [5], which embodies no indepence assumption, resulted in only 24 occasions when the likelihood ratio exceeded one.

Conclusion

The results of the experiments described here provide further confirmation of the robustness of our operational procedures.

The crossed-database experiments provide useful insights into the issue of representativeness. Even in the extreme case of using an Afro-Caribbean instead of a Caucasian database the consequences are not serious: particularly in the light of the results of Gill et al. [3] which show that use of a second probe is almost certain to lead to an exclusion of an innocent man. It is now clear that the precise shapes of the band weight frequency distributions are not particularly important in the context of forensic science. What is important, of course, is the degree of heterogeneity in the population. Concern would be justified in a case where there was good reason to believe that the suspect and the true offender were both members of some highly inbred population sub-group. Such cases are extremely unusual in the United Kingdom but we realise that they might be more frequent in other countries. The situation which might occur more frequently is one where the true offender is a close relative of the suspects: this has been the subject of a paper by Evett [6].

Whereas the results here confirm our confidence in our procedures we are aware of the dangers of complacency and experimentation of a similar nature to that described here will be a continuing feature in the future evolution of DNA technology in the FSS. Other workers in the field will, no doubt, find our results useful but we stress that our tests for robustness have been set within the context of our own procedures and databases: it is our recommendation that all workers in this field should satisfy themselves similarly by carrying out experiments similar to those described here.

References

1. Lander ES (1989) 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
2. 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
3. Gill P, Woodroffe S, Lygo JE, Millican ES (1991) Population genetics of four hypervariable loci. Int J Leg Med 104 (in press)
4. Devlin P, Risch N, Roeder K (1990) No excess of homozygosity at loci used for DNA fingerprinting. Science 249:1416–1420
5. Berry DA, Evett IW, Pinchin R (1991) Statistical inference in crime investigations using DNA profiling: single locus probes. J R Statist Soc (Series C – Applied Statistics) (in press)
6. Evett IW (1991) Evaluating DNA profiles in a case where the defence is "It was my brother". J Forensic Sci Soc (in press)