ORIGINAL ARTICLE

I. W. Evett · P. D. Gill · J. A. Lambert · N. Oldroyd
R. Frazier · S. Watson · S. Panchal · A. Connolly
C. Kimpton

Statistical analysis of data for three British ethnic groups from a new STR multiplex

Received: 19 June 1996 / Received in revised form: 9 September 1996

Abstract Data have been collected from 602 Caucasians, 190 Afro-Caribbeans and 257 Asians of Indo/Pakistani descent who have been profiled using a new six locus short tandem repeat (STR) multiplex. The data have been analysed by conventional significance testing methods: the exact test, homozygosity, and conventional goodness of fit to Hardy-Weinberg proportions. Frequency tables are given and the expected performance in British forensic casework is discussed.

Key words DNA · STR · Statistics · Exact test · Wahlund · Hardy-Weinberg · British · Population genetics · Caucasian · Afro-Caribbean · Asian

Introduction

The analysis of polymorphic short tandem repeat systems (STR) by automated fluorescence is becoming commonplace for human identification purposes. Previously, we reported the characteristics of a quadruplex system comprising four loci (Kimpton et al. 1994; Gill and Evett 1995; Evett et al. 1996a). However, the discriminating power of this system is relatively low compared to restriction fragment length polymorhism (RFLP) analysis. To augment the power of the existing multiplex, a new system of six STR loci and a sex test was devised consisting of the following loci:

D8S1179 (CHLC, accession nr. 374), D18S51 (Straub et al. 1993), HUMVWFA31/A (Kimpton et al. 1992), HUMTH01 (Polymeropoulos et al. 1991), HUMFIBRA(FGA) (Mills et al. 1992), D21S11 (Sharma and Litt 1992), and the amelogenin sex test described by Sullivan et al. (1993).

This system is the basis of the UK DNA national criminal intelligence database. There are two loci which are in common with the quadruplex (HUMTH01) and HUMVWFA31/A). HUMFIBRA, D21S11 and D18S51 are classified as highly polymorphic, complex loci (Urquhart et al. 1994). This has the advantage of significantly increasing the discriminating power of the multiplex system.

The characteristics of four STR loci (HUMTH01, HMVWFA31/A, HUMFES/FPS and HUMF13A1) were previously reported by Gill and Evett (1995) and Evett et al. (1996a). There were two conclusions reached – within ethnic group population substructure was low ($F_{st} < 0.01$); secondly, any apparent departures from independence which were recorded were trivial in their effect. These findings reinforced the recent recommendations of the National Research Council (NRC) (1996).

Materials and methods

The primer specifications and amplification conditions are described by Kimpton et al. (1996) and Oldroyd et al. (1996). DNA samples were analysed using ABD 377 sequencers. Alleles were identified by reference to control allelic ladder standard markers as described by Gill et al. (1996).

Nomenclature is described by Gill et al. (1996). The scheme used follows the recommendations of the DNA commission (1994). Designations for the following loci are HUMTH01 (Puers et al. 1993), HUMVWFA31/A and D21S11 (Urquhart et al. 1994), HUMFIBRA (Barber et al. 1996), D18S51 and D8S1179 (Barber and Parkin 1996).

The databases described in this paper comprise the predominant racial groups within the UK and they were from samples collected as follows:

1. 602 Caucasians from police staff, Forensic Science Service (FSS) staff, and casework.
2. 190 Afro-Caribbeans from police and FSS staff, Metropolitan Police Forensic Science Laboratory (MPFSL) staff, and casework.
3. 257 Asians originating from the Indian sub-continent. From patients at hospitals in Birmingham and Oxford, from police and FSS staff, and from immigration paternity testing (Dr Paul Debenham).

As far as possible, close relatives were excluded from the databases.

I. W. Evett (✉) · P. D. Gill · J. A. Lambert · N. Oldroyd
R. Frazier · S. Watson · S. Panchal · A. Connolly · C. Kimpton
Forensic Science Service,
Priory House, Gooch St North,
Birmingham B5 6QQ, UK
FAX: +44 (121) 622 2051

Statistical testing

We have written elsewhere about the limitations of statistical independence testing – see, in particular Evett (1996). Nevertheless, we recognise that tradition calls for such testing so, while not subcribing to that tradition, we recognise that it is widespread and we describe the results of the following tests:

1. Conventional comparison between the observed numbers of genotypes with the numbers expected from Hardy-Weinberg equilibrium (HWE) using the usual $(Obs–Exp)^2/Exp$ test statistic. There are six tests per database, one for each locus.
2. Comparison of the total number of homozygotes with the total expected assuming HWE using the $(Obs–Exp)^2/Exp$ test statistic. Again, six tests per database.
3. The exact test for within- and between-locus testing follows the method of Zaykin et al. (1995). For each database there are 6 within-locus tests, 15 two-locus, 20 three-locus, 15 four-locus, 6 five-locus, and one six-locus test, a total of 63 tests. The within-locus tests are of HWE; the between locus tests are composite tests of HWE and linkage equilibrium (LE).

Significance levels were determined for each test by comparing the test statistic with a distribution created by randomly shuffling the alleles independently at each locus 1000 times. This testing regime is described in more detail in Evett et al. (1996a) and Lee et al. (1996).

The results of the testing were as follows:

Caucasian (602 profiles): the p-values for all tests were in excess of 0.05.

Afro-Caribbean (190 profiles): all p-values exceeded 0.05 except for the homozygosity and exact tests for VWA (0.02 and 0.03 respectively), and the following composite exact tests

Combination	p-value
D21/VWA	0.03
FGA/VWA	0.04
D18/FGA/VWA	0.02
D21/FGA/VWA	0.02
D18/D21/FGA/VWA	0.03

We consider that these results do not have practical significance for the following reasons:

1. The excess of homozygosity in VWA has not been significant in other Afro-Caribbean databases that we have reported on – see Evett et al. (1996 a). We note that Drozd et al. (1994) reported a tendency for more VWA homozygotes than expected in a Caucasian database, though this has not been found in other studies on Caucasians (Evett et al. 1996 b). We consider that the effect, if that is what it is, has negligible practical significance, particularly in view of the corrections that we will apply to frequencies in casework as described later, but we will continue to keep it under review.

2. We have noted,in previous studies that an excess of homozygosity at one locus tends to manifest itself as apparent interactions in combinations which include that locus, so it is likely that all of the above higher order effects stem from the observation on VWA. Nevertheless, we repeated the exact tests but keeping genotypes fixed at each locus as described by Evett et al. (1996 a) the p-values for the first four of these combinations of loci were all comfortably in excess of 0.05. The four locus combination cannot be tested in this way because of the size of the database.

Asian (257 profiles): all p-values exceeded 0.05 with the exception of the following composite tests:

Combination	p-value
D18/D21	0.05
D18/D8	0.04
D21/D8	0.01
THO1/FGA	0.04

When the tests were rerun with genotypes fixed at each locus the first two of these gave p-values in excess of 0.05. For the last two the p-values fell to less than 0.01; however, there were no instances of low p-values for three locus combinations involving these pairs and the combination of all four together was not significant. The allele frequencies for all three databases and six loci are shown in Table 1.

Expected performance in casework

The probability of a match (PM) between two unrelated individuals was estimated for each ethnic group and each locus by counting the number of matches from all between person comparisons. These results are shown in Table 2. Multiplying PM's between loci suggest the overall six locus match probability to be of the order 10^{-8} for each of the groups.

In casework it is our intention to estimate single locus match probabilities using the formulae recommended in the second NRC report (1996), from the paper by Balding and Nichols (1994):

For homozygous genotype A_iA_i:

$$\frac{(2\theta + (1-\theta)p_i)(3\theta + (1-\theta)p_i)}{(1+\theta)(1+2\theta)}$$

For heterozygous genotype A_iA_j:

$$\frac{2(\theta + (1-\theta)p_i)(\theta + (1-\theta)p_j)}{(1+\theta)(1+2\theta)}$$

Where θ is a measure of population subdivision and p_i and p_j are estimated allele frequencies. Following the tenor of the NRC recommendations, the value of θ used for any given case will depend upon the circumstances. Estimates of θ from the databases described here will be the subject of a separate paper, but, as the NRC report (1996) implies, there is now ample published evidence that values in the

Table 1 Allele frequencies estimated from the three databases: (A) Caucasian (602); (B) Afro-Caribbean (190); (C) Asian from the Indian subcontinent (257)

	Ethnic group		
Allele	A	B	C
D18			
8	0.000	0.003	0.000
9.2	0.001	0.000	0.000
10	0.008	0.000	0.006
11	0.012	0.008	0.019
12	0.139	0.079	0.080
13	0.125	0.074	0.134
14	0.164	0.063	0.247
14.2	0.000	0.003	0.000
15	0.145	0.147	0.167
16	0.137	0.171	0.154
17	0.115	0.174	0.076
18	0.080	0.103	0.029
19	0.041	0.095	0.035
19.2	0.000	0.005	0.000
20	0.017	0.053	0.039
21	0.010	0.013	0.010
22	0.005	0.011	0.002
23	0.001	0.000	0.002
24	0.002	0.000	0.000
D21			
53	0.000	0.003	0.000
54	0.001	0.000	0.000
57	0.001	0.003	0.002
59	0.031	0.074	0.016
61	0.160	0.258	0.177
63	0.226	0.184	0.185
64.1	0.000	0.000	0.002
64	0.000	0.003	0.000
65	0.258	0.147	0.171
66	0.027	0.029	0.029
67	0.069	0.066	0.051
68	0.093	0.068	0.109
69	0.018	0.016	0.004
70	0.090	0.071	0.185
71	0.001	0.011	0.000
72	0.022	0.034	0.066
73	0.000	0.008	0.000
74	0.002	0.000	0.002
75	0.000	0.018	0.002
77	0.000	0.008	0.000
TH01			
5	0.002	0.005	0.000
6	0.241	0.142	0.292
7	0.194	0.384	0.169
8	0.108	0.203	0.101
8.3	0.001	0.000	0.000
9	0.140	0.126	0.267
9.3	0.304	0.129	0.158
10	0.012	0.011	0.012
10.3	0.000	0.000	0.002

Table 1 (continued)

	Ethnic group		
Allele	A	B	C
D8			
8	0.018	0.008	0.010
9	0.013	0.008	0.000
10	0.094	0.034	0.167
11	0.066	0.032	0.068
12	0.143	0.132	0.111
13	0.333	0.211	0.198
14	0.209	0.311	0.200
15	0.088	0.213	0.161
16	0.031	0.039	0.072
17	0.004	0.011	0.012
18	0.000	0.003	0.000
FGA			
18	0.025	0.011	0.006
18.2	0.000	0.013	0.000
19	0.056	0.053	0.045
19.2	0.000	0.005	0.000
20	0.143	0.066	0.105
20.2	0.002	0.005	0.000
21	0.187	0.134	0.154
21.2	0.002	0.000	0.008
22	0.165	0.132	0.154
22.2	0.011	0.000	0.004
23	0.139	0.234	0.169
23.2	0.004	0.003	0.002
24	0.146	0.124	0.187
24.2	0.002	0.000	0.006
25	0.075	0.079	0.099
25.2	0.000	0.000	0.004
26	0.035	0.074	0.037
27	0.007	0.029	0.018
28	0.000	0.013	0.004
29	0.000	0.013	0.000
30	0.001	0.000	0.000
30.2	0.000	0.005	0.000
31	0.000	0.003	0.000
45.2	0.000	0.003	0.000
46.2	0.000	0.003	0.000
VWA			
11	0.000	0.005	0.000
13	0.001	0.016	0.002
14	0.105	0.079	0.117
15	0.080	0.218	0.082
15.2	0.000	0.000	0.002
16	0.216	0.208	0.241
17	0.270	0.211	0.284
18	0.219	0.161	0.183
19	0.093	0.068	0.084
20	0.014	0.029	0.004
21	0.002	0.005	0.002

Table 2 Probability of a match between two unrelated people (PM) estimated for each locus from between-person comparisons. Combined six-locus PM estimated by multiplying across loci

	D18	D21	TH01	D8	FGA	VWA	Combined
Caucasians	0.028	0.049	0.079	0.061	0.032	0.063	1.36E-08
Afro-Caribbeans	0.024	0.041	0.103	0.075	0.025	0.049	9.48E-09
Asians	0.035	0.041	0.085	0.043	0.031	0.065	1.08E-08

range 0.01 to 0.03 will be conservative in typical forensic casework. For the estimates of allele frequencies we shall use the sampling corrections derived by Balding and Nichols (1994):

For homozygotes: $p_i = \dfrac{x_i + 4}{n + 4}$.

For heterozygotes: $p_i = \dfrac{x_i + 2}{n + 4}$, $p_j = \dfrac{x_j + 2}{n + 4}$.

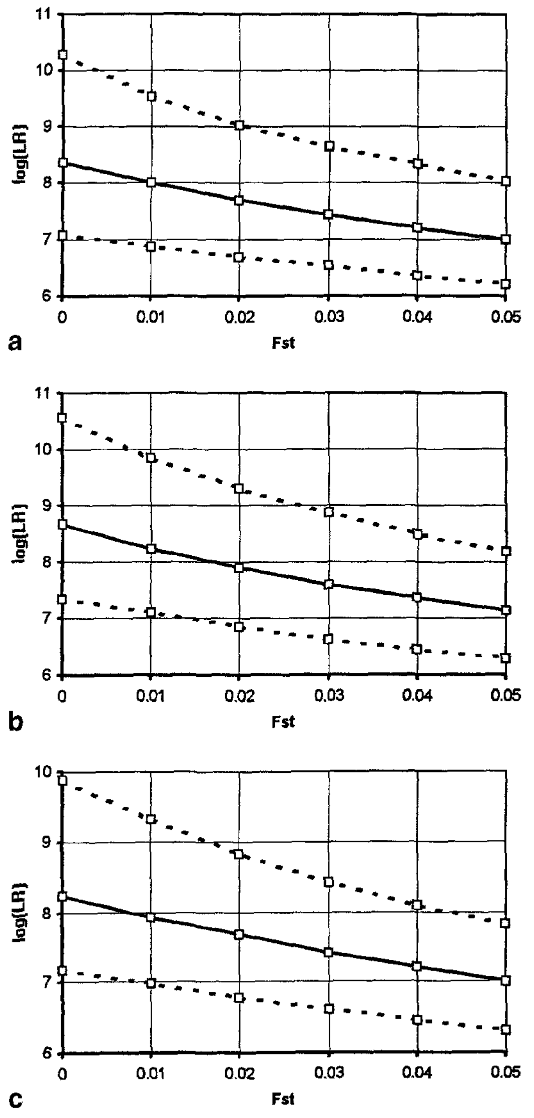

Fig. 1a–c Distribution of $\log_{10}$(LR) for various values of F_{ST}. The graphs are of the median, 5 and 95 percentiles. **a** Caucasian, **b** Afro-Caribbean, **c** Asian

Where x_i, x_j are the frequencies of alleles *ij* in a database of *n* alleles. This is equivalent to adding the genotype (*ii* in the homozygote and *ij* heterozygote cases) twice to the database of genotypes.

The single locus match probabilities calculated in this way are multiplied across loci, a procedure supported by the between locus studies we report above.

The performance of the technique in the case in which the suspect is truly the offender can be assessed for each ethnic group by calculating the six-locus genotype frequency for each member of the relevant database. Figure 1 summarises the distributions of likelihood ratios calculated in this way (likelihood ratio taken to be the inverse of the match probability in this context) for each of the three databases and for various values of θ (= F_{ST}).

Acknowledgements We wish to express our thanks to the staff of the DNA database unit for their help in generating some of the data used in this study, and Dr Paul Debenham for providing a number of the Asian samples.

References

Balding DJ, Nichols RA (1994) DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 64: 125–140

Barber MD, Parkin BH (1996) Sequence analysis and allelic designation fo the two short tandem repeat loci D18S51 and D8S1179. Int J Legal Med (in press)

Barber MD, McKeown BJ, Parkin BH (1996) Structural variation in the alleles of a short tandem repeat system at the human alpha fibrinogen locus. Int J Legal Med 108: 180–185

DNA recommendations – 1994 report concerning further recommendations of the DNA commission of the ISFH regarding PCR-based polymorphisms in STR (short tandem repeat) systems. Int J Legal Med 107: 159–160

Drozd MA, Archard L, Lincoln PJ, Morling N, Nelleman LJ, Phillips C, Soteriou B, Syndercombe Court D (1994) An investigation of the HUMVWA31A locus in British Caucasians. Forensic Sci Int 69: 161–170

Evett IW (1996) Expert evidence and forensic misconceptions of the nature of exact science. Sci Justice 36: 118–122

Evett IW, Gill PD, Scranage JK, Weir BS (1996a) Establishing the robustness of short tandem repeat statistics for forensic applications. Am J Hum Genet 58: 398–407

Evett IW, Lambert JA, Buckleton JS, BS Weir (1996b) Statistical analysis of a large file of data from STR profiles of British Caucasians to support forensic casework. Int J Legal Med (in press)

Gill P, Evett IW (1995) Population genetics of short tandem repeat (STR) loci. Genetica 96: 69–87

Gill P, Urquhart A, Millican E, Oldroyd N, Watson S, Sparkes R, Kimpton CP (1996) A new method of STR interpretation using inferential logic – development of a criminal intelligence database. Int J Legal Med (in press)

Kimpton CP, Walton A, Gill P (1992) A further tetranucleotide repeat polymorphism in the VWF gene. Hum Mol Genet 1: 287

Kimpton C, Fisher D, Watson S, Adams M, Urquhart A, Lygo JE, Gill P (1994) Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. Int J Legal Med 106:302–311

Kimpton C, Oldroyd NC, Watson SK, Frazier RRE, Johnson PE, Millican ES, Urquhart A, Sparkes RL, Gill P (1996) Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification. Electrophoresis. Submitted

Lee LD, Fairley M, Lambert JA and Evett IW (1996) Validation of a frequency database for four STR loci for use in casework in the Strathclyde Police Forensic Science Laboratory. Forensic Sci Int 79:43–48

Mills KA, Even D, Murray JC (1992) Tetranucleotide repeat polymorphism at the human alpha fibrinogen locus (FGA). Hum Mol Genet 1:779

National Research Council (1996) The evaluation of forensic DNA evidence. National Academy Press. Washington DC

Oldroyd NJ, Urquhart AJ, Kimpton CP, Millican ES, Watson SK, Frazier RRE, Gill P (1996) Development and optimisation of a highly discriminating multiplex PCR system suitble for forensic identification. In: Carracedo A, Brinkmann B, Bär W (eds) Advances in forensic haemogenetics 6. Springer, Berlin, Heidelberg New York, pp 198–200

Polymeropoulos MH, Xiao Hi Rath DS, Merril CR (1991) Tetranucleotide repeat polymorphism at the human tyrosine hydrolase gene (TH). Nucleic Acids Res 19:3753

Puers C, Hammond HA, Jin HL, Caskey CT, Schumm JW (1993) Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01 [AATG]n and reassignment of alleles on population analysis by using a locus specific ladder. Am J Hum Genet 53:953–958

Sharma V, Litt M (1992) Tetranucleotide repeat polymorphism at the D21S11 locus. Hum Mol Genet 1:67

Straub RE, Speer MC, Luo Y, Rojas K, Overhauser J, Ott J, Gilliam TC (1993) A microsatellite genetic linkage map of human chromosome 18. Genomics 15:48–56

Sullivan KM, Mannucci A, Kimpton CP, Gill P (1993) A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. Biotechniques 15:636–641

Urquhart A, Kimpton CP, Downes TJ, Gill P (1994) Variation in short tandem repeat sequences – a survey of twelve microsatellite loci for use as forensic identification markers. Int J Legal Med 107:13–20

Zaykin D, Zhivotovsky L, Weir BS (1995). Exact tests for association between alleles at arbitrary numbers of loci. Genetica 96:169–178