Int J Legal Med (2001) 114 : 204–210

ORIGINAL ARTICLE

Peter Gill

An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes

Received: 11 May 1999 / Accepted: 27 September 1999

Abstract This paper assesses the use of single nucleotide polymorphisms (SNPs) for forensic analysis. It demonstrates that relatively small arrays of approx. 50 loci are comparable to existing short tandem repeat (STR) multiplexes. A quantitative test, however, is a prerequisite for mixture interpretation. In addition, as the mixture proportion becomes low, it will be necessary to distinguish between the allele and background. Relatively small biallelic arrays are also suitable to distinguish between closely related individuals such as brothers.

Keywords SNP · Polymorphisms · Arrays · Multiplexes · Simulation · Biochips

Introduction

There is increasing interest in the use of biallelic markers or single nucleotide polymorphisms (SNPs) for forensic purposes (Syvanen et al. 1993). Several formats have been used for PCR-based biallelic assays: the reverse dot blot (Saiki et al. 1988) applied to HLA DQ-alpha and Polymarker systems, microtitre-based formats (Kostyu et al. 1993) and finally microfabricated arrays on glass (Southern et al. 1992, 1994; Guo et al. 1994). The latter are of special interest since the potential exists to build arrays consisting of hundreds of loci. This paper specifically explores the potential of biallelic arrays, particularly with respect to the analysis of mixtures. All of the platforms are non-electrophoretic.

A crucial aspect of forensic DNA typing is the interpretation of mixtures (Evett et al. 1991; Weir et al. 1997). Until recently, statistical interpretation of mixtures has proceeded without considering differences in signal strength of heterozygotes at a locus. Evett et al. (1998), Clayton et al. (1998) and Gill et al. (1998) reported methods to interpret mixed STR profiles based on identification of the allele peak areas. Although intended for STR (electrophoretic) analysis, the principles can be extended to encompass biallelic loci on non-electrophoretic media.

P. Gill
The Forensic Science Service, Trident Court,
2960 Solihull Parkway, Birmingham Business Park,
Birmingham B37 7YN, UK

How large does an array need to be?

Typically an array of biallelics could comprise several hundred loci that are typed from a single individual. In this paper I consider relatively small arrays of 50–150 loci. Excluding the possibility of genetic 'nulls', a consideration of each locus in turn must fall into one of two categories – either one allele will be visible or two alleles will be seen. The notation *A, B* is used to denote the two alleles where *a* and *b* denote their respective frequencies – only *AA*, *AB* or *BB* are possible (because $a + b = 1$, all formulae could be expressed solely in terms of *a*). If a microfabricated array consists of *n* loci, the match probability can be approximated by making the simplified assumption that the frequency of *A (a)* is constant for every locus in the array. For *n* different loci in an array, the number of *AA* genotypes is a^2n, the number of *AB* genotypes is $2abn$ and the number of *BB* genotypes is b^2n. For example if $n = 100$ and $a = 0.5$, then 50 loci will be heterozygote and 50 will be homozygote (*AA* and *BB* in equal proportion). Therefore, the match probability across the entire array is $(a^2)^{25} \times (2ab)^{50} \times (b^2)^{25}$. If estimated as a likelihood ratio (LR_n):

$$LR_n = \left(\frac{1}{a^2}\right)^{a^2n} \times \left(\frac{1}{2ab}\right)^{2abn} \times \left(\frac{1}{b^2}\right)^{b^2n}$$

Figure 1 shows simulations for arrays ranging from 50–150 loci. A relatively small array of 50 gives likelihood ratios equivalent to approximately 12 STRs over a wide range of $a > 0.2 < 0.8$. Note that the plots in Fig. 1 are symmetrical, so that the $LR_{(a = 0.8)}$ is the same as $LR_{(a = 0.2)}$.

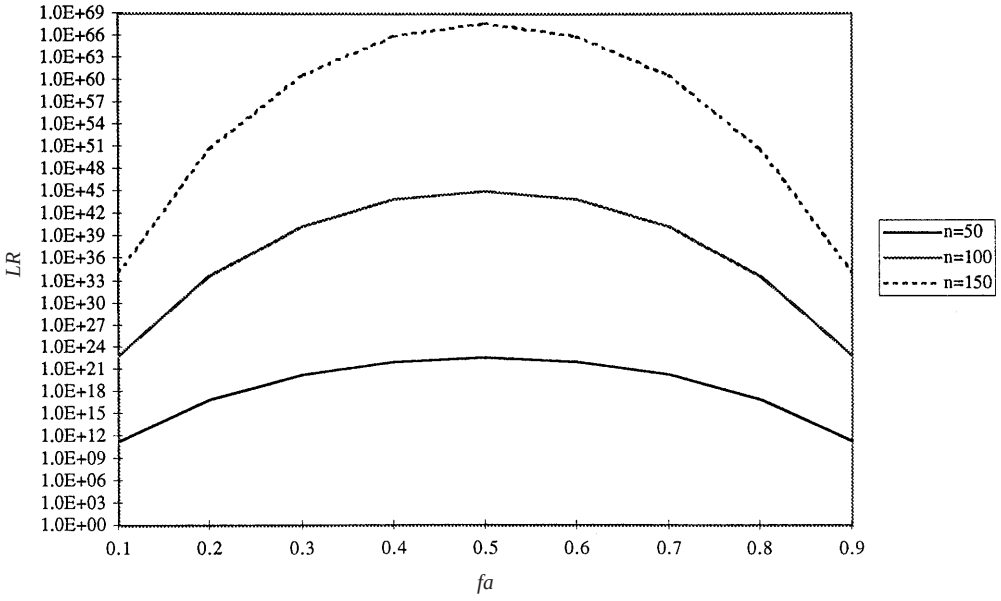

Fig. 1 Estimates of LR_n from arrays of *n* loci, assuming *fa* is constant across the set

Analysis of mixtures

Assuming two contributors to the mixture, if one allele shows then both must be homozygous for the same allele (*AA,AA* or *BB,BB*).

If there are two alleles visible, and assuming that there are two contributors to a mixture, (suspect S and victim V, respectively) then the following genotype combinations are possible: *AA,AB*; *AA,BB*; *AB,BB*; *AB,AB* and all of the reverse possibilities (Weir et al. 1997). This makes a total of nine possible genotype combinations (m = 1...9), all of which may be represented in a mixture. Given a normal outbreeding population, the proportion of observations of all of the above mixture types can be estimated given *a*.

Contributors to the mixture are the suspect and an unknown individual

For example, suppose that a blood stain is retrieved from a crime scene and the genotypes are consistent with a combination of the suspect (S) with an unknown individual (U).

We consider the following conditions in the likelihood ratio:

C: Contributors were the suspect and unknown

$\overline{C}$: Contributors were two unknown individuals

For each locus, calculation of the likelihood ratio depends upon the genotype of the suspect and the alleles observed in the mixture and there are three broad categories to consider.

Category 1

The suspect is homozygous *(AA)* and the mixture is *AB*

(U) must be either *AB* or *BB*

$$C = 2ab + b^2$$

$$\overline{C} = 6a^2b^2 + 4a^3b + 4ab^3$$

$$LR = (2ab + b^2)/(6a^2b^2 + 4a^3b + 4ab^3)$$

Category 2

The suspect is heterozygous *(AB)* and the profile is *AB*.

(U) must be *AA*, *AB* or *BB* and:

$$C = (a + b)^2$$

$\overline{C}$ is the same as in category 1 above.

Category 3

The suspect is homozygous *(AA)* and the profile shows just one allele. (U) is *AA* and the LR is $1/a^2$.

A complete list of numerators and denominators is given in Table 1. The proportion of an array of *n* loci having a particular mixture type *(m)* is *fm*: Each locus has *mp* = 9 possible mixture genotype combinations each (listed in Table 1).

The total $\overline{LR_n}$ of a mixture in an array of *n* loci is:

$$\overline{LR_n} = \prod_{m=1}^{mp} LR^{(fm \times n)}$$

Simulation of typical (average) mixture statistics on the combined $\overline{LR_n}$ for any number of biallelic loci was carried out under the simplified assumption that the allele proportion *(a)* for each locus is the same across loci (Fig. 2). $\overline{LR_n}$ maximises when *a* is high (0.8) or low (0.2). A battery of 50 loci with frequencies of alleles ranging between 0.1–0.9 will give a minimum LR of 10^4.

Table 1 When the mixture comes from an unknown individual and suspect, the LR numerators and denominators for each of the nine possible genotype combinations are calculated from the formulae listed. The proportion of mixture genotypes expected (fm) are also listed

	Mixture type (U, S)								
	AA,AA	*AA,AB*	*AB,AA*	*AA,BB*	*BB,AA*	*AB,AB*	*AB,BB*	*BB,AB*	*BB,BB*
Frequency of observation (fm)	a^4	$2a^3b$	$2a^3b$	a^2b^2	a^2b^2	$4a^2b^2$	$2ab^3$	$2ab^3$	b^4
LR (numerator)	1	$(a+b)^2$	b^2+2ab	a^2+2ab	b^2+2ab	$(a+b)^2$	a^2+2ab	$(a+b)^2$	1
LR (denominator)	a^2	$6a^2b^2+4a^3b+4ab^3$	$6a^2b^2+4a^3b+4ab^3$	$6a^2b^2+4a^3b+4ab^3$	$6a^2b^2+4a^3b+4ab^3$	$6a^2b^2+4a^3b+4ab^3$	$6a^2b^2+4a^3b+4ab^3$	$6a^2b^2+4a^3b+4ab^3$	b^2

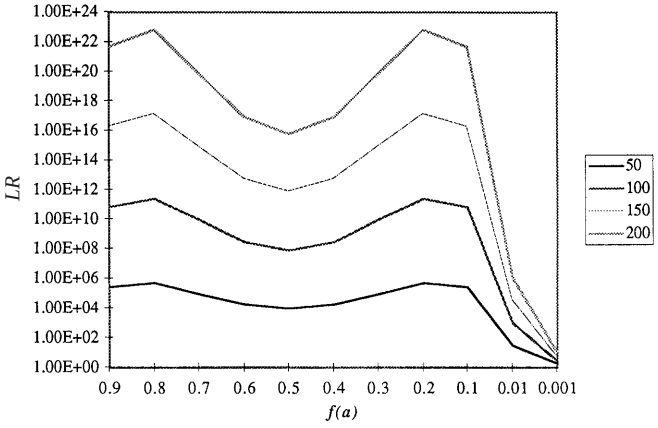

Fig. 2 Likelihood ratio plots for arrays ranging from 50–200 loci when there is a mixture of a suspect with an unknown individual using genotypes listed in Table 1

Mixtures conditioned on victim's profile

An example is a typical rape case where the mixture comprises contributions from the victim and the suspect:

C Contributors are the suspect and victim

$\overline{C}$ Contributors are the victim and unknown

Mixture profile with two alleles:

Category 1

If the profile comprises two alleles *(AB)*; the victim is *AB* and the suspect is either *AA* or *AB* or *BB*.

$C = 1$

$C = (a + b)^2$

true for all suspect genotypes.

$LR = 1/(a + b)^2$

Therefore LR = 1 (regardless of the value of *a*), i.e. the evidence is always neutral!

Category 2

If the profile comprises two alleles *(AB)*; the victim is homozygous *(AA)* and the suspect is *AB* or *BB* therefore:

$C = 1$

$\overline{C} = 2ab + b^2$

(since the perpetrator cannot be *AA*)

$LR = 1/(2ab + b^2)$

Mixture profile with one allele

Both victim and suspect are homozygote *(AA,AA)*:

$LR = 1/a^2$

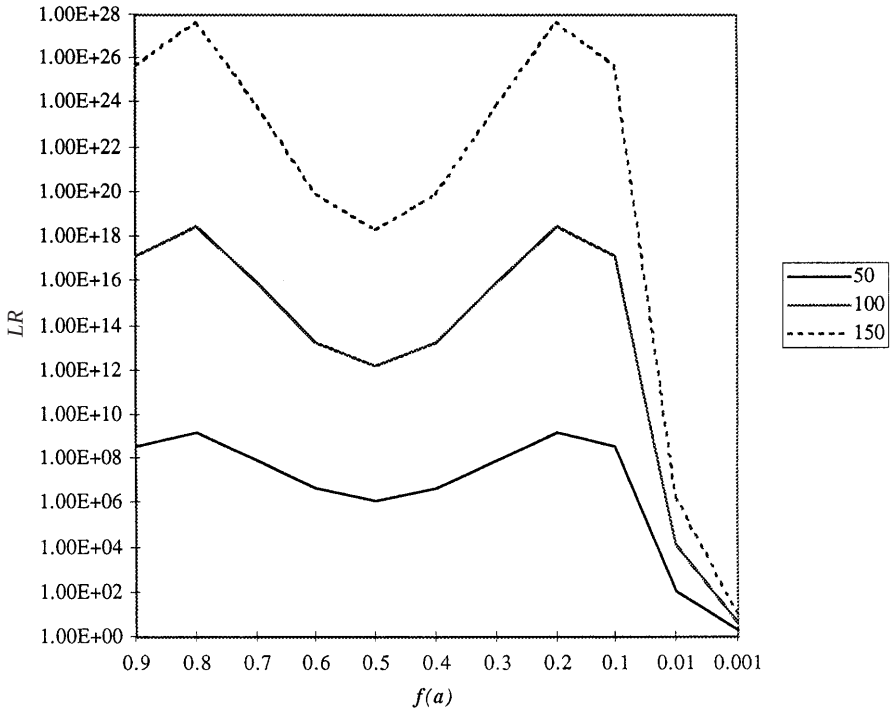

Fig. 3 Likelihood ratio plots for arrays ranging from 50–150 when there is a mixture of a victim with an unknown individual using genotypes listed in Table 2

Table 2 Mixture conditioned on a victim. The LR denominators for each of the nine possible genotype combinations are calculated from the formulae listed

	Mixture type (V, S)								
	AA,AA	*AA,AB*	*AB,AA*	*AA,BB*	*BB,AA*	*AB,AB*	*AB,BB*	*BB,AB*	*BB,BB*
Frequency (fm)	a^4	$2a^3b$	$2a^3b$	a^2b^2	a^2b^2	$4a^2b^2$	$2ab^3$	$2ab^3$	b^4
LR (denominator)	a^2	$2ab+b^2$	$(a+b)^2$	b^2+2ab	a^2+2ab	$(a+b)^2$	$(a+b)^2$	a^2+2ab	b^2

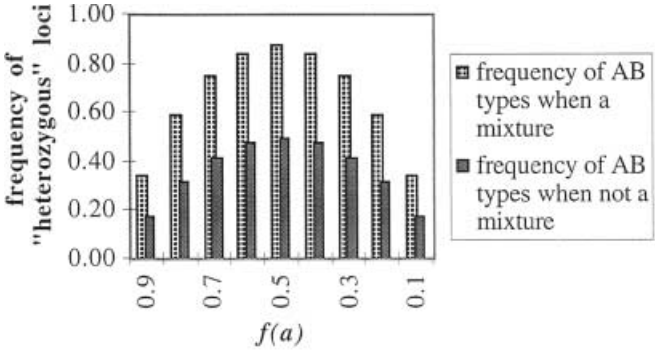

Fig. 4 Under the assumption that A and B alleles can be definitively identified and no allele drop-out occurs, the number of apparent heterozygotes increases in the array when there is a mixture. A comparison of proportions of 'heterozygous' loci in an array between unmixed samples and simple mixtures is given

Simulations were carried out as previously described (Fig. 3) – denominators to calculate likelihood ratios are given in Table 2. Likelihood ratios maximise when a is 0.2 or 0.8; a small array of 50 loci will give a LR>10^6.

Recognising a mixture

With STR analysis, it is relatively easy to identify the presence of mixtures by the presence of 3 or 4 allelic bands at a locus (Clayton et al. 1998). However, with biallelic assays, the lack of allelic variation prevents this method, and is certainly a disadvantage. Simple mixtures will always comprise the five different gene combinations:

a *AAAA*

b *AAAB*

c *AABB*

d *ABBB*

e *BBBB*

Furthermore, in the absence of a quantitative assay it is only possible to distinguish between three mixture types – the two homozygote forms (a) and (e), and those containing both A and B alleles (b, c, d). The key to mixture recognition will first of all depend on observing an increased apparent heterozygosity across the array (Fig. 4). Secondly, the signals of "heterozygotes" will appear imbalanced, i.e. at a given locus, one allele will produce a

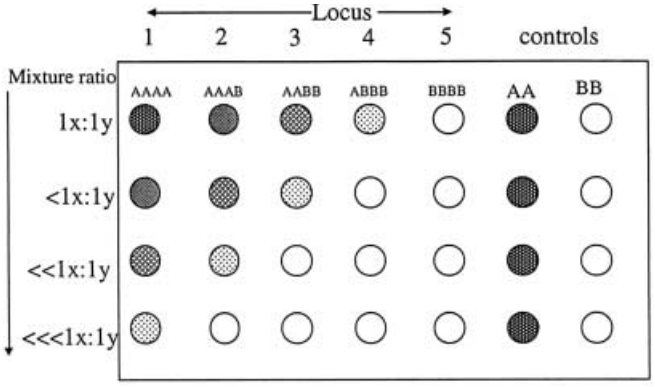

Fig. 5 Consider a mixture of two individuals x and y. They are analysed at five different loci which give genotypes *AA,AA*; *AA,AB*; *AA,BB*; *BB,BB* respectively. When the mixture ratio is 1:1, the presence of A and B alleles can be identified at all but the 5th locus which only has allele B. As the proportion of individual x in the mixture progressively decreases, allele drop-out occurs, beginning with the *AB,BB* genotype

stronger signal than the other. Development of accurate quantitative methods will be important to enable an interpretation strategy equivalent to that outlined by Evett et al. (1998) or Gill et al. (1998) for STRs.

When mixtures containing a minor contribution are analysed, it is inevitable that allele drop-out occurs and this will happen when the signal is indistinguishable from background noise. If the mixture ratio is progressively reduced from 1:1, the first loci to drop out will have just one dose of an allele from the minor contributor i.e. mixture type *ABBB* or *AAAB*, whereas a signal will still be obtained from *AABB* types. In addition, one of the difficulties in processing large numbers of biallelic loci will be to develop efficient multiplex assays i.e. some loci may amplify more efficiently than others such that different thresholds of detection may exist across the array (Fig. 5).

Implications of allele drop-out in bi-allelic arrays

Defining *Mx* as the proportion of the mixture contribution of individual A relative to individual B:

$Mx = A/(A + B)$

Gill et al. (1998) also show that for STRs *Mx* is similar across all loci within the mixture. When contributions from individuals are similar ($Mx \cong 0.5$) then identification of alleles is relatively easy. However, when extreme ($Mx \cong 0.9$ or 0.1) then allele or locus drop-out is likely to occur. This means that an allele found in the suspect will not be observed in the mixture. The limitations of detection are entirely dependent upon levels of background inherent in the assay; a population of negative controls will be needed to calculate cumulative probability density functions for background noise at every locus.

Assuming that the suspect's contribution to the mixture is minor and that the allelic signal is close to the background, $p(B = \text{null})$ is the probability that allele *B* is completely absent in the mixture whereas $p(B \neq \text{null})$ is the probability that allele *B* is present in the mixture (albeit at levels which may not be distinguishable from the background noise).

If S:*BB*; V:*AA* and the profile is *AA*, then the following alternatives are considered:

1. numerator: The profile is *AB* with the probability $p(B \neq \text{null})$ and V and S are the contributors of the profile i.e. allele drop-out of B has occurred.

2. denominator: The profile is *AB* or *BB* with the probability $p(B \neq \text{null})$ i.e. allele drop out has occurred; or *AA* with the probability $p(B = \text{null})$; V and one unknown person are the contributors of the profile.

$$LR = \frac{p(B \neq null)}{[2ab + b^2]p(B \neq null) + a^2 p(B = null)}$$

Continuing with the illustration where the suspect contribution is at low level, it follows that with a borderline profile where S = *AA*; V = *AA* and profile = *AA*, the possibility that the perpetrator is *AB* or *BB* must still be evaluated in the denominator:

1. numerator: The profile is *AA* with the probability $p(B = null)$ and V and S are contributors of the profile.

2. denominator: The profile is *AB* or *BB* with the probability $p(B \neq null)$ or *AA* (with the probability $p(B = null)$; V and one unknown person are the contributors of the profile.

$$LR = \frac{p(B = null)}{[2ab + b^2]p(B \neq null) + a^2 p(B = null)}$$

Here $p(B = null)$ can be directly estimated from the cumulative probability density of the background signal observed in negative controls for each locus – the lower the signal, the lower *p(null)* must be. As the minor contribution of the mixture approaches background, the LR is greatly diminished (Fig. 6).

Paternity testing

In a disputed paternity where the putative father (Fp), mother (M) and child (C) are DNA-tested, under the condition that the relationship of the mother to the child is undisputed, a generalised paternity index (Pi) will take one of three forms dependent upon the genotypes of parents:

a Class 1: Under the condition where all three profiles are the same and homozygous (e.g. M = *AA*; C = *AA*; Fp = *AA*) or any other case where the father must be homozygous (e.g. M = *AA*; C = *AB*; Fp = *BB*), then $Pi = 1/a$.

b Class 2: Alternatively, if the child's profile can be explained by inheritance of either A or B from the father because both mother and child are heterozygous, then the paternity index is $1/(a + b)$ (e.g. where M = *AB*; C = *AB*; Fp = *AA*). The paternity index is always 1.

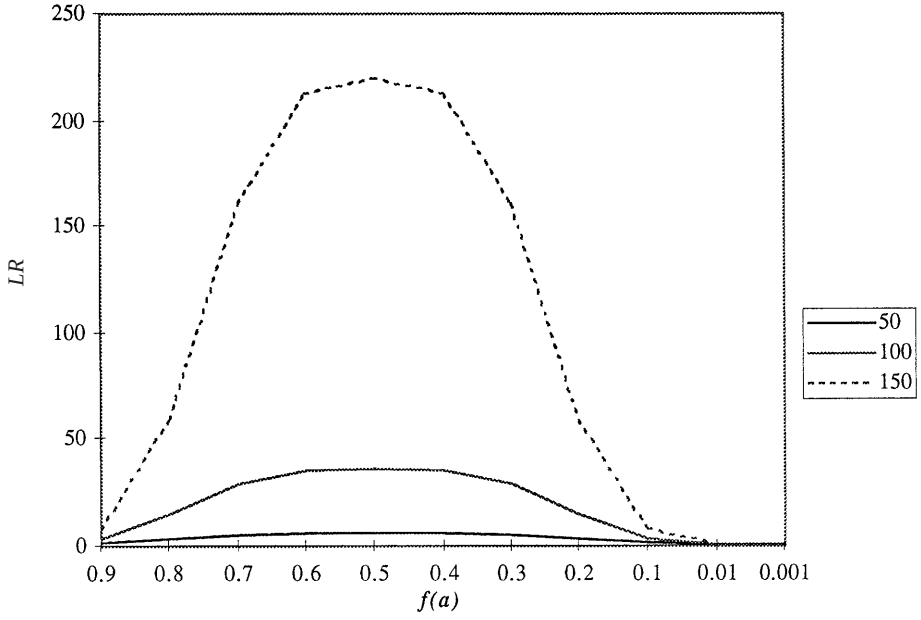

Fig. 6 Simulation of the effect of null alleles generated in arrays where there is a minor contributor to the mixture. The simulation was the same as described in Fig. 2 except that it was assumed that *AA*:*AB* and *BB*:*AB* genotypes may not be distinguished from *AA*:*AA*; and *BB*:*BB* genotypes – *p(null)* = 0.5 for the purposes of this illustration

c Class 3: If the mother is homozygous and the father is heterozygous (e.g. where M = *AA*; C = *AB*; Fp = *AB*) then the paternity index is 1/2*b*.

In a large array, the proportion of mother/father/child combinations that would be expected to occur is dependant upon the proportion of alleles at each locus in the population, the expected proportion of each trio in the array was calculated. Whenever, the mother and child are heterozygote (class 2 trio), then this yields neutral information and the LR is always 1.

Following Evett and Weir (1998), an exclusion probability was calculated for loci for different population allele proportions indicated in Table 3. For example, a locus where $a = 0.5$ gives an exclusion probability of 0.1875. Assuming independence, the exclusion probability was calculated for n loci (where n ranged from 10–50) and demonstrated that 50 loci gave a maximum probability of exclusion of 0.99997. Provided that all loci in the array have $a > 0.2$ and < 0.8 then the probability of exclusion is always >0.99. It would be necessary to ensure that closely linked loci are not used otherwise haplotypes may be inherited which compromise independence assumptions.

Table 3 Probability of paternity exclusion, given the genotype of the mother

Frequency allele A	No. of loci			
	1	10	20	50
0.1	0.08190	0.57450	0.81895	0.98605
0.2	0.13440	0.76386	0.94424	0.99927
0.3	0.16590	0.83700	0.97343	0.99988
0.4	0.18240	0.86652	0.98218	0.99996
0.5	0.18750	0.87462	0.98428	0.99997
0.6	0.18240	0.86652	0.98218	0.99996
0.7	0.16590	0.83700	0.97343	0.99988
0.8	0.13440	0.76386	0.94424	0.99927
0.9	0.08190	0.57450	0.81895	0.98605

Testing categories of relatedness

Several categories of relatedness were tested ranging from full brothers to unrelated individuals using formulae of Weir (1997) (Fig. 7), testing the alternatives:

C The DNA profile originated from the suspect.

$\overline{C}$ The DNA profile originated from a relative (e.g. brother) of the suspect.

Small biallelic arrays were shown to be powerful tests to distinguish individuals at any level of relationship (except for identical twins).

Discussion

Relatively small arrays (ca. 50 loci) are very efficient tools for human identity testing purposes, forensic stains or for distinguishing between close relatives, e.g. brothers, provided that loci are chosen so that alleles range in proportion between 0.2–0.8. The greatest challenge will be to identify and to interpret mixtures. There will be a marked increase in apparent heterozygosity within the array; furthermore there will be marked imbalance of alleles within heterozygotes. For interpretation, the test must show a high degree of quantitative accuracy and be essentially free of background noise. When the minor contributor of a mixture is close to the background threshold level then allele drop-out will be encountered, but interpreta-

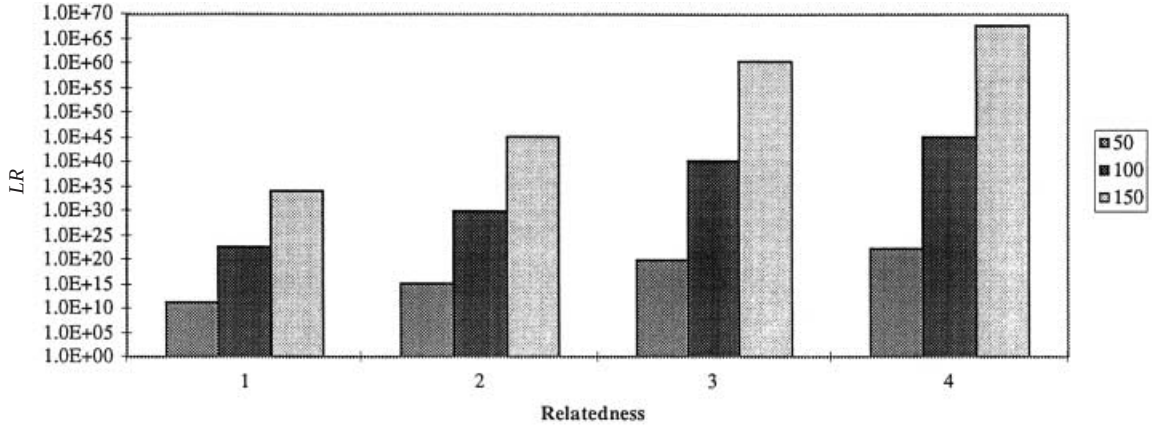

Fig. 7 A comparison of likelihood ratios from arrays of 50, 100 and 150 loci, respectively, under the assumption that the suspect and perpetrator are related. The allele proportion *(fa)* is 0.5 . On the *X-axis*: *1)* full brothers, *2)* father and son, *3)* first cousins, *4)* unrelated

tion can still proceed provided that cumulative probability functions can be used to estimate *p(null)*. Interpretation of more than two individuals contributing to a mixture will present a major challenge. Independence assumptions have not been assessed in this paper; however, it is inevitable that due consideration will be needed with large arrays.

Currently, the greatest problem in developing useful SNP arrays for forensic use is not related to statistical issues, rather, the problems are biochemical. Making a large balanced multiplex of ca. 50 loci from less than 1 ng of genomic template is indeed a daunting prospect.

References

Clayton TM, Whitaker JP, Sparkes R, Gill P (1998) Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Sci Int 91:55–70

Evett IW, Weir BS (1998) Interpreting DNA evidence. Sinauer Associates, Sunderland, Mass.

Evett IW, Buffery G, Willot G, Stoney DA (1991) A guide to interpreting single locus profiles of DNA mixtures in forensic cases. J Forensic Sci Soc 31:41–47

Evett IW, Gill P, Lambert JA (1998) Taking account of peak areas when interpreting mixed DNA profiles. J Forensic Sci 43: 62–69

Gill P, Sparkes R, Pinchin R, Clayton T, Whitaker J, Buckleton J (1998) Interpreting simple STR mixtures using allele peak areas. Forensic Sci Int 91:41–53

Guo Z, Guilfoyle RA, Thiel AJ, Wang R, Smith LM (1994) Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. Nucleic Acids Res 22:5456–5465

Kostyu DD, Pfhol J, Ward FE, Lee J, Murray A, Amos DB (1993) Rapid HLA-DR oligotyping by an enzyme-linked immunosorbent assay performed in microtiter trays. Hum Immunol 38: 148–158

Saiki RK, Bugawan TL, Horn GT, Mullis KB, Erlich HA (1986) Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. Nature 324:163–166

Southern EM, Maskos U, Elder JK (1992) Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. Genomics 13:1008–1017

Southern EM, Case-Green SC, Elder JK, Johnson M, Mir KU, Wang L, Williams JC (1994) Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. Nucleic Acids Res 22:1368–1373

Syvanen AC, Sajantilla A, Lukka M (1993) Identification of individuals by analysis of biallelic DNA markers, using PCR and solid phase minisequencing. Am J Hum Genet 52:46–59

Weir BS, Triggs CM, Starling L, Stowell LI, Walsh KAJ, Buckleton J (1997) Interpreting DNA mixtures. J Forensic Sci 42: 213–222