Int J Legal Med (2003) 117 : 271–275
DOI 10.1007/s00414-003-0382-7

ORIGINAL ARTICLE

Thore Egeland · Ingvild Dalen · Petter F. Mostad

Estimating the number of contributors to a DNA profile

Received: 20 December 2002 / Accepted: 25 April 2003 / Published online: 14 August 2003

Abstract The broad topic of this paper is the evaluation of DNA evidence in criminal cases. More specifically, we deal with mixture evidence which refers to cases where there are, or could be, several contributors to a biological stain based on, e.g., blood or semen. The present paper adresses DNA mixtures based on single nucleotide polymorphism (SNP) markers, i.e., diallelic markers. Based on STR analysis, it is in most cases easy to identify the presence of a mixture since three or four bands will show up with a high probability for at least one locus. Obviously, this will not be the case for diallelic markers and interpreting mixtures will be a great challenge. We address this problem by first approaching the more general problem of estimating the number of contributors to a stain. In addition we discuss how the markers should be selected and how many are required.

Keywords SNP · Mixture evidence · Likelihood · Estimates · Bayes

Introduction

The statistical interpretation of forensic DNA mixtures is well understood for many practical purposes and considerable progress has been made since Evett et al. (1991). The general formula for likelihood calculations in Weir et al. (1997) has been discussed and generalized, for instance by Fukshansky and Bär (1998), Curran et al. (1999) and Fung and Hu (2000). Several computer programs are available, see e.g. Mortera et al. (2002), Fung and Hu (2000, http://www.hku.hk/statistics/staff/wingfung/), and Curran et al. (1999, http://statgen.ncsu.edu/storey/). Some recent references are Hu and Fung (2003) and Fung and Hu (2002). There are however remaining challenges and some of these are addressed in the present paper. In particular, the numbers of contributors to a stain cannot normally be known with certainty. This problem has been handled in various ways, Weir (1995) presents alternative calculations assuming different number of contributors while Brenner et al. (1996), Buckleton et al. (1998), Lauritzen and Mortera (2002) and others give bounds on likelihood ratios that can be used when the number of donors to a stain cannot be agreed upon. The above approaches do not use data to estimate the number of contributors in a formal manner beyond observing that a stain indicates the minimum number of contributors, for instance at least three persons must have contributed if five different alleles are seen in a profile. Stockmarr (2000) estimated the number of contributors in a specific example by maximizing the likelihood. This paper continues this effort in a setting where it appears to be particularly relevant, namely for SNP (single nucleotide polymorphism) markers. For these diallelic markers, each locus will display one or two alleles. Consequently, it is more difficult to assess whether more than one person have contributed. In fact, Gill (2001) pointed out that "...the greatest challenge will be to identify and interpret mixtures". We address the question ("Is it a mixture?") by first approaching the more general problem of estimating the number of contributors to a stain. In addition we discuss how the markers should be selected and how many are required.

The next section discusses the methods. In particular the likelihood for SNP markers is written in a form that makes it easy to estimate the number of contributors and determine whether a stain is a mixture or not. The result section presents three examples based on simulated data. Based on these examples and the general methods, we draw some conclusions in the last section regarding the number of markers required and how these should be chosen.

T. Egeland (✉)
Biostatistics, Rikshospitalet University Hospital,
0027 Oslo, Norway
Tel.: +47-228-51148, Fax: +47-228-51313,
e-mail: thore.egeland@basalmed.uio.no

I. Dalen
University of Oslo, Oslo, Norway

P. F. Mostad
Chalmers Technical University, Göteborg, Sweden

Methods

We start this section by fixing some notation and reformulating the main research problems in more precise terms. For a specific marker we denote the less frequent variant B and the more frequent B^c. Let p_i denote the frequency of B at locus i. An individual profile may be summarized by a vector of length N. Element i is 0, 1 or 2 depending on whether only B, only B^c or both alleles are seen. A stain from $x \geq 1$ persons may be summarized similarly by a vector of length N. A main problem may now be phrased and exemplified: "Have more than one persons contributed to a specific stain, say (0, 1, 1, 2, 0, 1)?"

Likelihood and estimation

Using the notation explained above, the probabilities of observing 0, 1, or 2 for marker i may be written:

$$\begin{aligned} p_{0i} &= p_i^{2x} \\ p_{1i} &= (1-p_i)^{2x} \\ p_{2i} &= 1-p_i^{2x}-(1-p_i)^{2x}. \end{aligned} \tag{1}$$

This follows by a direct argument and agrees with the more general formula in Weir et al. (1997). The above equation assumes independence between the two alleles from a person and independence between persons contributing. These independence assumptions may be relaxed, leading to modified versions of Eq. 1. If the markers are independent, the probability of observing

$$z = (0,1,1,2,0,1)$$

equals

$$p_{01}p_{12}p_{13}p_{24}p_{05}p_{16} \tag{2}$$

Generally, the likelihood for a profile $(z_1,...,z_N)$ may be written using indicator functions $I(.)$

$$\begin{aligned} L(x) &= P(data|x) = \prod_{i=1}^{N} p_{0i}^{I(z_i=0)} p_{1i}^{I(z_i=1)} p_{2i}^{I(z_i=2)} \\ &= \prod_{i=1}^{N} p_i^{2xI(z_i=0)} (1-p_i)^{2xI(z_i=1)} \left(1-p_i^{2x}-(1-p_i)^{2x}\right)^{I(z_i=2)} \end{aligned} \tag{3}$$

Consider next the case where all $p_i=p$ and let n_0 and n_1 count the number of occurrences of 0's and 1's. In this particular case, all relevant statistical information is contained in the sufficient statistic (n_0,n_1) and the probabilities are given by the multinomial formula

$$P(n_0,n_1|x) = a(n_0,n_1)p^{2xn_0}(1-p)^{2xn_1}\left(1-p^{2x}-(1-p)^{2x}\right)^{N-n_0-n_1} \tag{4}$$

where $a(n_0,n_1)=N!/(n_0!n_1!(N-n_0-n_1)!)$. In the general case, the unknown number of contributors may be estimated by maximizing Eq. 3 with respect to x. If all $p_i=p$, one may choose to use Eq. 4. A large number of computer programs will handle the maximization. Apparently, there is no simple formula for the maximum likelihood estimator except for the trivial case when all $p_i=0.5$. Then:

$$x^* = \frac{1}{2\log 2}\log\frac{n_0+n_1}{2N}.$$

The above estimator is finite if and only if $n_0+n_1 \geq 1$. In the Appendix it is shown that the general likelihood Eq. 3 also has a unique and finite maximum if and only if $n_0+n_1 \geq 1$.

Note that:

$$P(n_0+n_1 \geq 1) = 1-\prod_{i=1}^{N}\left(1-p_i^{2x}-(1-p_i)^{2x}\right) = 1-\left(1-p^{2x}-(1-p)^{2x}\right)^N \tag{5}$$

where the last equality assumes $p_i=p$, $i=1,...,N$. The right hand side of Eq. 5 is minimized for $p=0.5$ for fixed x.

Is it a mixture?

We next consider the question of determining whether the stain is a mixture or not. Two approaches are outlined, a frequentist and a bayesian.

Frequentist approach. The parameter x can be considered fixed but unknown and two hypotheses formulated in the usual way:

H_0 : One person contributed, i.e., $x=1$
H_1 : More than one person contributed, i.e., $x \geq 2$.

A reasonable approach is to reject H_0 when

$$K = \frac{\max_{j=1,2,3,...} P(data|x=j)}{P(data|x=1)} > c. \tag{6}$$

The specific value of c can be determined by simulating K under H_0. Since K is discrete, it is not possible to achieve a precise level of significance. Example 2 in the next section indicates that a reasonable and simple solution is to reject H_0 and claim that a stain is a mixture when $K>1$.

Bayesian approach. Sometimes there is information available in addition to the SNP markers. Different sources of data may be combined as explained below. Bayes theorem gives:

$$P(x=i|data) = \frac{P(data|x=i)\alpha(i)}{\sum_{j=1}^{\infty}P(data|x=j)\alpha(j)},$$

where $P(x=j)=\alpha(j)$ is the prior distribution. The posterior odds for the stain to be a mixture can be written

$$\begin{aligned} \frac{P(x>1|data)}{P(x=1|data)} &= \sum_{j=2}^{\infty}\frac{P(x=j|data)}{(x=1|data)} \\ &= \sum_{j=2}^{\infty}\frac{P(data|x=j)\alpha(j)}{P(data|x=1)\alpha(1)} \end{aligned}$$

To continue, some prior assumptions are required and a formulation in terms of the prior odds for being a mixture, $R=\sum_{j=2}^{\infty}\alpha(j)/\alpha(1)$, seems reasonable. The posterior odds will depend on not only R but the entire x distribution. However, we can find an upper bound for the posterior odds:

$$\begin{aligned} &\sum_{j=2}^{\infty}\frac{P(data|x=j)\alpha(j)}{P(data|x=1)\alpha(1)} \\ \leq\ &\frac{M\sum_{j=2}^{\infty}\alpha(j)}{P(data|x=1)\alpha(1)} \\ =\ &\frac{M}{P(data|x=1)}R, \end{aligned} \tag{7}$$

where

$$M = \max_{j=2,3,4...} P(data|x=j).$$

Observe that the above approach may only be used to statistically show that there is only one contributor. The previous frequentist approach applies more generally. However, if one is willing to assume more a priori data, the restriction on the Bayesian approach disappears. For instance, specifying the alternative hypothesis "$x=2$" corresponds to assuming $\alpha(j)=0$ for $j>2$ and the posterior odds

$$\frac{P(data|x=2)\alpha(2)}{P(data|x=1)\alpha(1)}$$

can be used to distinguish between the alternatives for a specified prior on $\alpha(2)/\alpha(1)$.

Results

The previous section has presented results regarding (1) estimation of the number of contributors to a stain, (2) testing if a stain is a mixture or not and (3) verification of a non-mixture allowing for inclusion of prior information or data. Three examples follow to demonstrate the practical implementation of the methods. The examples are based on 1000 simulated datasets in S-PLUS 6.0.

Example 1

This example discusses the number of loci required to accurately estimate the number of contributors. We provide detailed explanation of the first line of Table 1. Column 1 shows that the data is simulated with x equal to 1, followed by a column indicating the number of loci, N=50 in this case. The two next columns list the fraction correctly identified for p=0.1 and p=0.5. In the former case 0.965 or 96.5% were correctly classified whereas there were no errors for p=0.5. As expected, the precision increases in N and decreases in x. If the number of contributors is 3 or less, the correct classification rate is always above 87% for N=200. Observe that the case with 5 contributors may not be resolved satisfactorily even with 1,000 markers for p=0.5. Figure 1 shows the standard deviation of the estimator of x as a function of p<0.5. Two intuitive results are confirmed, the uncertainty *increases* in x and *decreases* in p. Figure 2 displays the number of loci required to secure a finite estimate of the number of contributors with probability b. The plot is based on inequality Eq. 5 and explains to some extent why it is difficult to estimate cases with many contributors for p close to 0.5.

Table 1 The fraction of correctly identified number of contributors is shown in the two rightmost columns for p=0.1 and p=0.5 for various values of x (the number of contributors) and number of markers (N)

x	N	Correct (p=0.1)	Correct (p=0.5)
1	50	0.965	1.000
2	50	0.717	0.876
3	50	0.590	0.421
4	50	0.424	0.000
5	50	0.417	0.000
1	100	0.990	1.000
2	100	0.854	0.969
3	100	0.753	0.802
4	100	0.623	0.000
5	100	0.545	0.000
1	200	1.000	1.000
2	200	0.960	0.999
3	200	0.902	0.870
4	200	0.797	0.396
5	200	0.756	0.000
1	500	1.000	1.000
2	500	0.999	1.000
3	500	0.989	0.990
4	500	0.964	0.851
5	500	0.934	0.000
1	1000	1.000	1.000
2	1000	1.000	1.000
3	1000	1.000	0.998
4	1000	0.997	0.929
5	1000	0.986	0.458

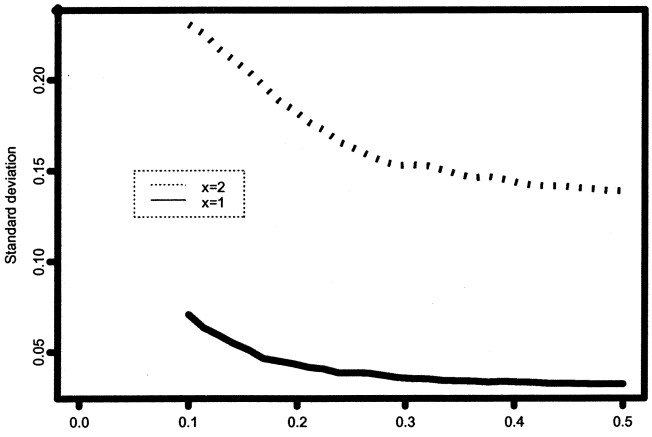

Fig. 1 The standard deviation of the estimate of the number of contributors is plotted as a function of p for one contributor, i.e., x=1, (solid line) and x=2 based on a simulation exercise with 200 markers. The uncertainty *increases* in x and *decreases* in p

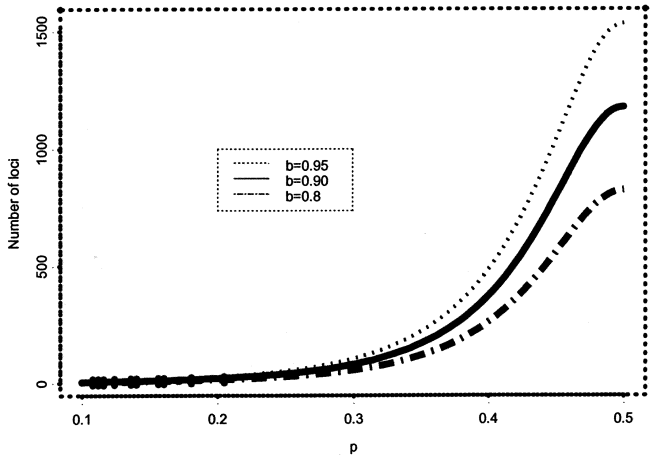

Fig. 2 The number of loci required to secure a finite maximum likelihood (ML) estimate of the number of contributors with probability b is plotted as a function of p based on Eq. 5

Example 2

Data was simulated first assuming x=2. The test statistic K defined in Eq. 6 was calculated and the null hypothesis was rejected when K>1. In other words, we conclude that two or more persons contributed if the maximal likelihood

Table 2 The properties of the *K*-statistic for mixtures defined in Eq. 6 are shown for various values of *N* and for *p*=0.1 and *p*=0.5

N	p=0.1; x=2	p=0.5; x=2	p=0.1; x=1	p=0.5; x=1
50	0.912	0.998	0.074	0.004
100	0.982	1.000	0.023	0.000
200	0.998	1.000	0.002	0.000
500	1.000	1.000	0.000	0.000

assuming $x \geq 2$ exceeds the likelihood assuming $x=1$. Table 2 summarizes the results for varying N (50, 100, 200, 500) and p (0.1 and 0.5). The power is high, or equivalently, the probability of a type II error is small. For N≥100 the probability of reaching the correct conclusion is 0.982 ($N=100$, $p=0.1$) or higher. It remains to check the significance level of the test and we simulated data with $x=1$ for this purpose. The two rightmost columns of Table 2 show that the test also performs well with respect to type I errors, i.e., the probability of falsely claiming a mixture is low.

Example 3

Recall that Eq. 7 could be useful to prove that a stain is not a mixture, when that is indeed the case. We simulated data for $x=1$, $N=100$, and $p=0.1$, and computed the ratio $M/P(data|x=1)$. In 95% of the cases, the ratio was smaller than 0.0008, reducing any prior odds for a mixture substantially towards zero. For $p=0.5$, the ratios are even smaller.

Discussion and concluding remarks

The examples of the paper have been based on simulated data and so we are able to see how the methods perform in cases where the truth is known. Another reason for simulating is that relevant case data using SNP markers do not appear to be available. Gill (2001) considered 50–150 markers. For practical forensic case work, confusing a mixed profile and a profile from a single person, could have serious consequences as a match between a stain and reference person could be missed. Based on our results, it seems fair to conclude that a decision regarding mixture or not can be reached for the number of markers in the indicated range. Based on Table 2, we recommend 100 markers. In this case the type II error, i.e., the probability of missing a mixture stain, ranges from 0 to 0.018 while the type I error lies between 0 and 0.023. It is harder to estimate the precise number of contributors, particularly if a large number, say five or more, cannot a priori be excluded. Table 1 shows that with 1, 2 or 3 contributors, the correct classification rate is 75% or higher. This accuracy may be acceptable for investigating purposes, but insufficient for a court. It is possible to obtain a posterior distribution on the number of contributors. The evidence may then be weighed according to this distribution.

It remains to be seen what numbers will be available and if the problems of interpretation of data based on conventional markers (see Evett and Weir 1998; Evett et al. 1998) will be reduced for SNPs. Different contributors to a stain could have donated varying amounts and this information could be used to improve estimates.

The calculations are simplified by the diallelic structures of SNPs. For conventional markers similar calculations are obviously more complicated. However, numerical or simulation-based results are always obtainable. Moreover, the formulation of hypotheses would typically differ for conventional markers; the question of a mixture or not is typically not relevant. For instance, if one marker displays 5 alleles and the other fewer, one might want to test the null hypothesis $x \leq 3$ against the alternative $x>3$. The test procedure we have suggested extends easily to this case.

Acknowledgements This work was supported by the Leverhulme Trust.

Appendix

Consider the likelihood $L(x)$ given in Eq. 3. Observe that the likelihood increases in x if all $z_i=2$ and decreases in x if all $z_i<2$. Assume that not all z_i equals 2. Then we show below that $L(x)$ always has a single maximum for some $x \geq 1$.

Maximizing $L(x)$ is equivalent to maximizing

$$\begin{aligned} f(x) &= \log L[x] \\ &= \sum_{i=1}^{N} I(z_i = 0) 2x \log p_i + I(z_i = 1) 2x \log(1 - p_i) \\ &+ I(z_i = 2) \log\left(1 - p_i^{2x} - (1 - p_i)^{2x}\right) \\ &= Cx + \sum_{i=1}^{N} I(z_i = 2) \log\left(g_i(x)\right), \end{aligned}$$

where $C<0$ and

$$g_i(x) = 1 - p_i^{2x} - (1 - p_i)^{2x}$$

We see by direct computation that $g_i''(x) < 0$ for all $x>0$. Defining $h_i(x)=log(g_i(x))$, it follows that $h_i''(x) < 0$ for all $x>0$, and thus that $f(x)<0$ for all $x>0$. Further, we get that $lim_{x\to\infty} g_i(x)=1$, that $lim_{x\to\infty} h_i(x)=0$, and that $lim_{x\to\infty} f(x)=-\infty$. These two facts about f show together that f, and thus L, has a unique maximum for some x≥0. For discrete x, one or two consecutive positive integers maximize L.

References

Brenner C, Fimmers R, Baur M (1996) Likelihood ratios for mixed stains when the number of donors cannot be agreed. Int J Legal Med 109:218–219

Buckleton J, Evett IW, Weir BS (1998) Setting bounds for the likelihood ratio when multiple hypotheses are postulated. Sci Justice 38:23–26

Curran JM, Triggs CM, Buckleton J, Weir BS (1999) Interpreting DNA mixtures in structured populations. J Forensic Sci 44: 987–995

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

Evett IW, Buffery C, Willott G, Stoney D (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 J (1998) Taking account of peak areas when interpreting mixed DNA profiles. J Forensic Sci 43:62–69

Fukshansky N, Bär W (1998) Interpreting forensic DNA evidence on the basis of hypothesis testing. Int J Legal Med 111:62–66

Fung WK, Hu YQ (2000) Interpreting forensic DNA mixtures: allowing for uncertainty in population substructure and dependence. J R Statist Soc A 163:241–254

Fung WK, Hu YQ (2002) The statistical evaluation of DNA mixtures with contributors from different ethnic groups. Int J Legal Med 116:79–86

Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. Int J Legal Med 114:204–210

Hu YQ, Fung WK (2003) Interpreting DNA mixtures with the presence of relatives. Int J Legal Med 117:39–45

Lauritzen SL, Mortera J (2002) Bounding the number of contributors to mixed DNA stains. Forensic Sci Int 130:125–126

Mortera J, Dawid AP, Lauritzen SL (2003) Probabilistic expert systems for DNA mixture profiling. Theor Popul Biol 63:191–205

Stockmarr A (2000) The choice of hypotheses in the evaluation of DNA profile evidence. In: Gastwirth JL (ed) Statistical science in the courtroom. Springer, Berlin Heidelberg New York, pp 143–160

Weir BS (1995) DNA statistics in the Simpson matter. Nat Genet 11:366–368

Weir BS, Triggs C, Starling L, Stowell L, Walsh K, Buckleton J (1997) Interpreting DNA mixtures. J Forensic Sci 47:213–222